UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: UNIVERSITÀ DEGLI STUDI DI PADOVA
Dipartimento di INGEGNERIA DELL'INFORMAZIONE

Corso di Dottorato di Ricerca in INGEGNERIA DELL'INFORMAZIONE
Curriculum: Scienza e Tecnologia dell'Informazione (ICT)
Ciclo: XXXIII

EXPLOITING THE TEMPORAL DIMENSION IN CLINICAL DATA MINING

Coordinatore: Ch.mo Prof. Andrea Neviani
Supervisore: Ch.mo Prof. Barbara Di Camillo

Dottoranda: Erica Tavazzi

Abstract

Based on the incremental amount of data being collected in the healthcare sector, healthcare analytics is creating a paradigm shift in many research areas from patient care to public health management. A number of different techniques, combined with an efficient use of data, allow to perform analyses for descriptive, predictive, and prescriptive purposes, providing means for making informed and efficient healthcare decisions.

In the clinical context, data collected in the medical practice are exploited for gaining evidence-based insights on the patients' condition with the aim of both improving care and expanding medical knowledge. In this framework, longitudinally-collected clinical data constitute an invaluable resource. Their dynamic nature provides a number of advantages: for instance, by continuously characterizing the evolution of the clinical condition, they allow to detect how the relationships among the observed features change as time passes. However, their employment also requires some precautions, mainly due to the need to properly manage their time dimension, as well as their possibly heterogeneous nature.

This thesis focuses on how to treat longitudinal clinical datasets for gaining valuable knowledge out of them. Ranging from data preprocessing to the development of descriptive and predictive models, issues, challenges, and potential of this kind of data are identified and discussed. To fill some of the gaps identified in literature, *ad hoc* developed methodologies are proposed. For each technique, relevant application examples in different clinical contexts are provided as well.

First, the issue of missing values in the data is addressed, by proposing two imputation approaches based on the similarity assessed among visits or patients over time. By managing the longitudinal and heterogeneous nature of data while making the best use of the available information, the developed methodologies are designed to meet different cases of use commonly present in clinical databases.

Then, descriptive and predictive modeling techniques are explored. The temporally-evolving information is exploited to study how features interact and influence the prognosis with the passing of time. Moreover, patient patterns are investigated in terms of succession and timing of consecutive events, providing a characterization of the population's pathways. By learning on the whole evolution dynamics, disease mechanisms and treatment effects can be investigated, obtaining models able to accurately describe and effectively simulate the patients' behaviour. Such tools can constitute a valuable mean to support physicians in clinical decision making and accompany patients in disease management.

Beside addressing the needs related to data complexity, in all the presented methodologies a special focus was given to interpretability, explainability, and communication effectiveness of the results.

Ringraziamenti

Questo Dottorato è stato occasione di crescita da un punto di vista sia professionale che personale. Ci tengo molto a ringraziare chi è stato al mio fianco durante questi tre anni, supportando i miei passi e prendendo parte a questa esperienza.

Un primo ringraziamento, di cuore, va al mio supervisore Prof.ssa Barbara Di Camillo per avermi accolta e guidata in questi miei primi anni nel mondo della ricerca, e per avermi trasmesso col suo esempio come energia e passione siano elementi fondamentali e imprescindibili da affiancare alle competenze.

Un sentito ringraziamento va anche al CORIS, specialmente nella persona della Dott.ssa Teresa Gasparetto, per avermi coinvolta nelle progettualità del Consorzio, aprendomi alla realtà sanitaria regionale e supportando economicamente questo mio Dottorato.

Ringrazio inoltre il Prof. Olivier Michielin, il Prof. Michel Cuendet ed il Dott. Roberto Gatta del Precision Oncology Center del CHUV di Losanna, per avermi accolta ed inclusa nel loro gruppo, supportandomi nella mia esperienza internazionale di questo percorso. Poche cose aprono gli orizzonti come potersi mettere in gioco in un ambiente nuovo.

Pur essendo il Dottorato un percorso personale, nel mio caso non sarebbe sicuramente stato lo stesso senza la preziosa collaborazione dei colleghi con cui ho sviluppato i miei lavori scientifici. Con alcuni l'affiatamento è stato tale che considerarvi solo colleghi sarebbe riduttivo: grazie Sebastian, Rosanna, Roberto G, non riesco ad immaginare nessuno di più adatto a voi nei progetti svolti insieme.

Vorrei naturalmente ringraziare anche tutti gli altri colleghi dei gruppi di ricerca con di cui ho fatto parte in questi anni, primi tra tutti i colleghi del gruppo di *Sysbiobig* del DEI: Giacomo, Ilaria, Marco C, Marco M, Alessandro, Alessandra, Mehdi. Le pause pranzo e caffè, le riunioni di gruppo, le parole scambiate insieme in corridoio hanno contribuito a creare un ambiente di lavoro di condivisione e scambio costruttivo, base di tutto.

Grazie anche a tutto il personale del CORIS, specialmente ai miei cari colleghi del *team VB-HC* Marianna e Giacomo, per aver messo a disposizione le proprie competenze e professionalità, contribuendo a costituire un gruppo multidisciplinare e dinamico.

Grazie inoltre ai colleghi del CHUV per la compagnia e le condivisioni durante il periodo trascorso insieme. Un doppio ringraziamento, personale e professionale, a Cami, per il sorriso con cui mi hai accolta fin dal mio primo giorno e per la sempre grande disponibilità nel condividere la tua passione e competenza professionale. *Ad maiora*!

Una tesi in questo ambito non può fare a meno di un quasi costante supporto clinico: un doveroso ringraziamento va a tutto il personale medico con cui ho avuto modo di confrontarmi in questi anni, per la loro pazienza nell'introdurmi al contesto, alle dinamiche ed alle necessità del

settore. Ho avuto la fortuna di incontrare grandi professionisti, motivati e capaci di trasmettere la loro passione per la materia e la cura per i pazienti.

Un ringraziamento speciale vorrei indirizzarlo anche a tutti i pazienti sui cui dati ho avuto modo di lavorare: senza la loro scelta di condivisione, la ricerca non potrebbe proseguire. In alcuni casi questa scelta è stata fatta pur sapendo che, quasi certamente, i frutti sperati non sarebbero arrivati in tempo per loro. C'è un immenso amore per il prossimo, in questo, che ammiro e rispetto, e spero di onorare col piccolo della mia ricerca.

Qualsiasi progresso in campo scientifico richiede sguardi esperti e punti di vista talvolta totalmente esterni: grazie a tutti coloro, noti o anonimi, che hanno supportato la mia ricerca con le loro attente azioni di supervisione e revisione. Tra questi, un sentito ringraziamento al Dott. Roberto Gatta e alla Dott.ssa Lucia Sacchi per aver pazientemente e competentemente revisionato questa tesi.

Anche al di fuori della sfera professionale, sono molte le persone che ci tengo a ringraziare per il loro contributo in questi anni, durante i quali mi sono state accanto rispettando, assecondando, ed anche perdonando i miei ritmi e i miei umori nelle varie fasi di questo percorso.

Primi tra tutti, un enorme ringraziamento va alla mia famiglia, da sempre supporto costante qualsiasi percorso io decida di intraprendere.

Grazie mamma Paola e papà Fernando per avermi educata al rispetto, al pensiero critico, alla curiosità, ed alla creatività: mi avete dato le basi e i mezzi per adoperarmi quotidianamente in quello che faccio, attività di ricerca compresa. E quando provo qualche smarrimento, siete lì a ricordarmi chi sono e a consigliarmi, vicini al di là di ogni distanza.

Grazie Fabio per il tuo esempio professionale di intraprendenza e personale di tenacia: guardandoti mi sento invitata a guardare al "nuovo" e al "difficile" con la fiducia di avere strumenti da impiegare, certa che una soluzione la si trova o la si crea.

Un pensiero affettuoso va anche ai miei cari nonni Beatrice e Lorenzo, per me emblema di cultura e grandi insegnanti di come la mente vada mantenuta sempre allenata, informata e curiosa di sapere: il vostro esempio risuona nelle mie giornate.

Grazie a Simone: sei stato il primo a nominarmi la prospettiva di fare un Dottorato di ricerca e, dopo anni ed un intreccio di cammini, sei stato anche il primo a leggere questa mia tesi. Il tuo supporto si declina sotto molteplici aspetti, alcuni espliciti ed altri più silenziosi e riservati, forse semplicemente come ho bisogno che sia. Grazie.

Un pensiero di immensa gratitudine va ai miei amici storici, compagni dei miei principali momenti di crescita personale e professionale, sempre pronti a supportarmi nelle difficoltà e a brindare insieme ai traguardi. Silvia, Cecilia, Chiara B, Chiara E, Alessandro K, Annamaria, Alice: ciascuno di voi riesce a cogliere ed accogliere parti di me che non sempre escono a parole. Vi sono profondamente grata. Grazie a Roberto C, per avermi trasmesso tramite il tuo esempio come la curiosità sia il potente motore alla base della ricerca, e per avermi supportata nella scelta di intraprendere questo percorso, tranquillizzandomi su come provare voglia dire scoprirsi e non ci sia da temere alcuna irreversibilità. Un intimo ringraziamento anche alla Dott.ssa CS, supporto fondamentale e guida nel mio più personale percorso di ricerca. Infine, un nostalgico omaggio al mio coro Caterina Ensemble, che ha accompagnato in musica i primi anni di questo Dottorato.

Nel raggiungere questo mio primo traguardo professionale, vorrei infine rivolgere un omaggio a tutte quelle persone (tra cui molte donne) che hanno reso possibile che mai in questi anni percepissi lungo il mio percorso una qualsiasi disparità di genere. Nel piccolo della mia esperienza di vita ed accademica, mi sono sentita libera: libera di studiare, libera di pensare e di proporre, libera di riuscire nei miei obiettivi e libera di sentirmi fiera dei miei successi, libera di avere i miei limiti e libera di poter scegliere come lavorarci su. Grazie.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Similarly to what is happening in many fields, the increased availability of structured data represents an invaluable resource in healthcare.

Historically, the healthcare sector has progressively been generating large amounts of data, driven by record keeping, compliance and regulatory requirements, and patient care [161]. During the last decades, the rise of biomedical sciences (*e.g.* the omics), health-related technologies (*e.g.* medical monitoring devices), the implementation of administrative healthcare information systems, as well as the digital transition from paper medical records to electronic health record (EHR) systems have led to an exponential growth of produced and available structured data, constituting the so-called *big data in healthcare* [106, 120, 181]. Recent technology advancements in hardware and software are making it easier to collect, transfer, store, aggregate, and analyze this information, even when derived from multiple sources [99].

Healthcare Analytics: from Data to Knowledge

By exploiting the information being collected, *healthcare analytics* is transforming the healthcare industry both in terms of cost optimisation and ever improving quality of care [55], by applying quantitative and qualitative techniques to extract knowledge from the available data [182]. In general, analytics methods include the use of mathematical and algorithmic processing of data resources, to gain insight from data for making informed and efficient healthcare decisions [39, 107].

Based on the collection, organization, manipulation, and mining of health and medical data, healthcare analytics mainly develops into three branches (and thus possible operational steps), namely *descriptive*, *predictive*, and *prescriptive* analytics [41, 107].

Descriptive analytics gives an overview on the data, by categorizing and converting them into useful information for better understanding and characterizing healthcare decisions, implications, outcomes and quality [162]. The extracted information is often summarized in the form of tables or graphical representations, making it more easier for the user to answer specific questions or identify patterns, thus providing a broader view for evidence-based practice [101].

Predictive analytics is a slightly more advanced type of analytics, that emphasizes the use of information in an effort to infer the future. By examining historical or summarized health data

through supervised or unsupervised techniques, it allows to detect patterns of relationships and extrapolate behaviours to forecast [96, 155].

Finally, prescriptive analytics exploits the knowledge obtained from the other forms of analysis to determine the best course of action with respect to the desired outcomes, allowing to make proactive decisions [162, 228].

Further areas are sometimes included in the definition of healthcare analytics, such as *diagnostic* analytics, that involves the investigation of historical data to delineate why something happened and to detect the variables causally linked to the outcomes being investigated [228], and *discovering* analytics (also called exploratory analytics), which supports users to reveal new scientific evidence starting from large volumes of data with plenty of detail [176].

Figure 1.1 reports a brief summary of this classification.

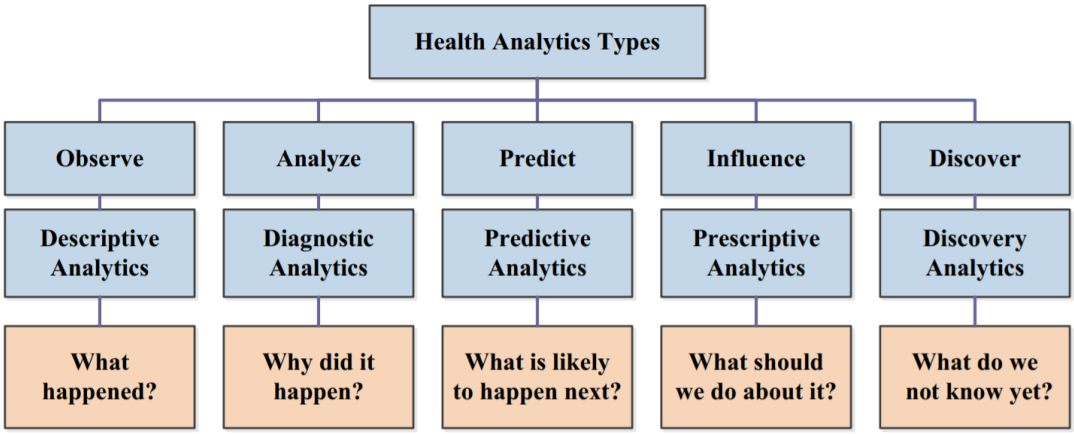


Figure 1.1: *Types and functions of healthcare analytics. Adapted from [100].*

Healthcare analytics is virtually creating a paradigm shift in the whole healthcare sector, from basic research to clinical and management applications [95, 1]. The possible advantages of such analyses could vastly improve patients' lives and benefit society as a whole. From an economic perspective, the use of these techniques to improve practice efficiency results in a more affordable, high-quality healthcare [48]. Besides, from a clinical point of view, the possible improvements in medical knowledge, as well in diagnosis and prognosis capabilities, allow higher health standards. Studies like survival analyses can evidence risk factors and detect the effect of specific treatments both in disease progression and quality of life [77], moving towards a personalised care system. Moreover, an enhanced knowledge of pathologies can be translated into computer-aided tools, offering clinicians a valid support in decision making.

Big Data in Healthcare: Characteristics and Types

With these aims, many different kinds of data are nowadays collected in the practice routine, constituting the so-called *big data in healthcare*.

Big data in general are characterized by some peculiar features, commonly referred to as the classical *5-V*s: first of all, they are are constituted by a high *volume* of information, that continuously increases over time [113]. Their rapid generation rate and potential analysis speed, especially when data are automatically streamed, characterizes them with the *velocity* feature, that is self-evidently linked to their volume increase [115]. Then, big data can present a wide *variety*, by potentially including heterogeneous information, differently structured, and gathered from different sources. As defined by IBM [180], big data exhibit *veracity* as a further dimension, which refers to their level of reliability (in terms of truthfulness, accuracy, or correctness) and, consequently, to the quality of the resulting analyses. Finally, big data undoubtedly constitute a huge potential in terms of *value*, that requires on the other hand accurate methodologies and plausible time frames to reap benefits out of them.

Specific to healthcare big data, Dinov [52] introduced two more important characteristics, that are, energy and life-span. The *energy* corresponds to the amount of information content included in the data: the higher the amount of data, the more precise the description of the phenomenon of interest, and then the more beneficial the analysis. The *life-span* is a concept strongly associated to the data value: as time passes, more data are being collected, but their usability can also be reduced for reasons of obsolescence. This phenomenon, also known as information devaluation, materializes as a decay of the lifespan and value of healthcare data at an exponential rate.

Some further characteristics should be considered when practically working with big data, specifically in healthcare. Among others, we find *accessibility* (can we really reach the collected information?) and, connected to it, *timing* (can we access them at the right time to make appropriate decisions?), *security and privacy preservation* (can we maintain their content safe and confidential?), *interoperability* (can we use them in combination?), *completeness* (are they exhaustive?), and *manageability* (can we handle their complexity?).

All these characteristics evidence how big data in healthcare constitute a rich informative basis to healthcare analytics applications, but also require special care in their use due to their potentially overwhelming complexity [67].

With respect to the possible types of big data employed in healthcare analytics, they can mainly be categorized into three categories, namely administrative, clinical, and behavioural/ environmental.

Administrative data include the information collected for business or organizational purposes and related to the contacts of the patient to the healthcare system/structure, such as hospital admissions or drug prescriptions. Typical analytics applications based on these data are studies of patient management, investigations of the employed resources, or estimations of the care costs for reimbursement purposes [94].

The second category comprehends all the *clinical information* generated in the medical practice and related to the patients' health status. This information can be recorded in a variety of sources, such as EHRs, patient databases, clinical trial registers, medical imaging repositories, and laboratory information systems [10, 141], and be employed in evidence-based medicine, for developing decision support systems, or for investigating condition risk factors.

Finally, *behavioural and environmental data* are recently emerging for their important contri-

bution as health-related information. Life-style data such as physical activity frequency, drinking or smoking habits, and occupational information are being collected through self-report questionnaires, wearable sensors/apps, or social media. Together, environmental monitoring systems allow to record the information related to possible substances or factors the subjects have been exposed to, such as air pollutants or electromagnetic fields. The combined use of this information together with that collected in the other two data categories allows to detect the influence of habits and exposition factors on disease spread, short- and long-term health conditions, or possible related genetic mutations [59, 66].

Mining Clinical Data with a Temporal Dimension

Among the different kinds of data just introduced, this thesis will deal with the employment of analytics techniques on clinical data. As introduced above, this category includes all the variables related to the medical status of a patient. This information can be captured as part of the standard care delivery process, or during activities such as clinical research trials or studies. In the clinical practice, real-time analysis of these data allows the delivery of timely, appropriate care to the patient. The aggregation and analysis of multiple patient information is the basis for a system that continually learns from the patients' conditions, therefore both improving care and advancing the frontiers of knowledge in medicine [139].

Clinical data represent a multidimensional description of the patient status in a specific observation moment, recording for instance symptoms, administered treatments, or lab test results. We could say that such recording constitutes a sort of photograph of the situation, potentially capturing detailed information of what is happening in that specific moment, possibly from a number of different perspectives.

Often, this acquisition is repeated over a follow-up period, in order to monitor the evolution of the patient's clinical history under natural or treated conditions. Such may be the case of a sequence of screening visits, or consecutive clinical evaluations over multiple days of hospitalization. In this case, what strongly characterizes the collected information is one remarkable additional dimension: *time*. In our metaphor, the patient's story is now described by a sequence of snapshots, with its evolution constituting a sort of movie. By properly observing and employing all this collected information, clinicians can detect the progression of the patient's health status, evaluate the path of care, and plan the next health treatments. Such sequence of observations of clinical parameters taken at different time moments is known as longitudinal clinical data collection. It represents a typical structure of many medical data collections, one example among all being clinical registers.

The availability of data longitudinally collected over time provides a number of advantages. First of all, the patient's clinical condition is widely characterized: specific variables can be tracked over time, by delineating their evolution; moreover, the effect or relation of one or more variables on the other at different time instants can be inspected. Noticeably, the time dimension not only allows to focus on the observed values and how they change as long as time passes, but also permits to inspect how much time elapses among distinct events. For instance, the time occurring between the administration of a certain therapy and the appearance of a certain clinical phenomenon can be investigated [22].

When extending the study to more subjects, further possible analysis targets can be added. For instance, for each patient the pattern of events characterizing his/her evolution can be determined, and compared with the others: in this way, similarities or deviations among subjects can emerge, possibly highlighting interesting behaviours or providing stratification of the population based on similar clusters of occurrences. Moreover, clinical conditions can be studied at a population level, by assessing outcome occurrence, risk factors, or disease evolution patterns.

However, managing the temporal dimension when analyzing clinical data is not trivial. Let's introduce some of the possible features – and related criticalities – of this kind of data with an example.

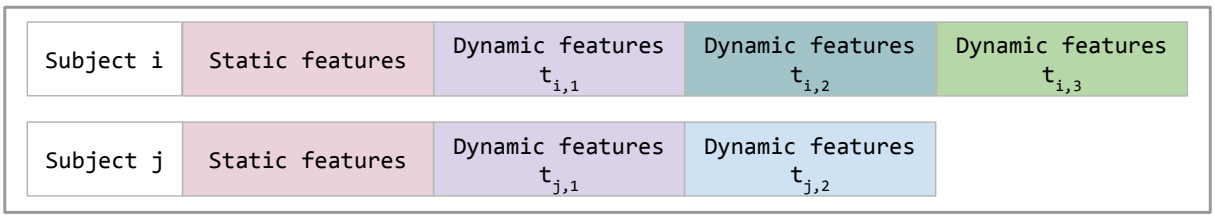


Figure 1.2: *General structure of longitudinal clinical records collected for two distinct subjects.*

Figure 1.2 schematically reports an example of longitudinal clinical records collected for two distinct patients i and j. First, according to their temporal dimension, the variables collected for each subject can be classified in two categories, that are, static and dynamic variables. *Static* variables correspond to all information which is constant throughout the patient's clinical history, such as sex or age at disease onset. On the other hand, *dynamic* variables are those that vary in time, such as blood pressure or sugar levels. Generally, static variables are collected at one given time point, like for instance when providing the anamnesis to the doctor; alternatively, when the clinical database results from the aggregation of multiple sources, they can correspond to the information collected in static databases, such as administrative records containing demographic information of patients. On the other hand, dynamic variables are iteratively collected over subsequent visits, and progressively added to the database.

Therefore, the first aspect to take care of when structuring the data or designing an analysis is how to properly handle this distinction. Additionally, dynamic data can be *temporally sparse*. In general, they can be recorded over a noncontinuous scale, according to the patient monitoring schedule or the specific variable granularity, and be asynchronous with respect to the population. This results in a sampling grid that, in general, may vary both along the patient-specific observation period and from patient to patient [18]. In the previous example of Figure 1.2, the observations for the two subjects differ in terms of both number of acquisitions and timing.

When these kind of data are mined, for instance with the goal of performing descriptive and/or predictive analyses, it is necessary to thoroughly deal with all these aspects. Nevertheless, not all the algorithms are able to correctly manage the temporal aspect together with the other data dimensions, that are those related to the acquisition of a multiple number of variables for a number of distinct subjects. Most of the traditional automatic data exploration/mining techniques usually treat temporal data as unordered collections of events, ignoring their temporal information: these techniques mainly focus on analyzing data occurring at the same time, neglecting the

inspection of the relations at time points [8]. Treating data as unrelated aggregates of individual data elements does not allow to identify trends in the values, to assess the cross-influences between variables as long as the condition progresses, or to detect similar profiles occurring over different time scales or having different baseline values [145].

In other cases, the temporal information is only marginally used for determining the values of static analysis outcomes, such as short- or long-term mortality rates. Moreover, when employing the other collected variables to correlate with/predict these outcomes, their use is often limited to baseline values, thus preventing the detection and assessment of their (possibly varying) influence over time. In addition, only an accurate analysis of the whole clinical evolution can allow to identify possible crossroad events that can act on the condition progression or on the outcome occurrence.

To summarize, when dealing with clinical data with a temporal dimension, the key concept is that "*the temporal dimension of data is a fundamental variable that should be taken into account in the mining process and returned as part of the extracted knowledge*", as accurately formulated by Berlingerio *et al.* [22].

Thesis Goal and Structure

Based on the considerations above, this thesis focuses on how to properly manage and exploit the temporal dimension when performing analyses on longitudinal clinical datasets.

Ranging from data preprocessing to the development of descriptive and predictive models, issues, challenges, and potential related to the temporal dimension of data are identified and discussed. To fill some of the gaps identified in literature, *ad hoc* developed methodologies are proposed. For each introduced technique, an application example or case study is also provided.

In particular, Chapter 2 outlines the possible issues related to the presence of missing values in real-world datasets, describing how an imputation step can be included in the preprocessing in order to obtain complete datasets for the analyses. The limits of state-of-the-art imputation methods when applied to longitudinally-collected clinical data are discussed, and two distinct developed imputation techniques are presented. Based on possibly different data properties, the proposed methodologies are able to exploit the similarity among visits or clinical evolution patterns in order fill in the missing information, while properly handling the temporal dimension of data and the possible feature type heterogeneity.

Chapter 3 focuses on the potential of clinical data with a temporal dimension in capturing the progression of clinical conditions. By effectively employing modelling methodologies able to manage the features' dynamism, it is possible to build descriptive and predictive models that thoroughly catch the progression of the condition, allowing to delve into the relationships among features over time as well as prognosis forecasting. As a case study, in this Chapter a model of progression of Amyotrophic Lateral Sclerosis based on Dynamic Bayesian Networks is proposed.

Chapter 4 presents an alternative approach to longitudinally collected clinical data, namely Process Mining. This family of analytics methods bases on data structured as sets of consecutive events, a representation that lends itself well in the case of data dynamically evolving over time. Through the employment of supervised and unsupervised techniques, it allows to discover the

processes that generated the observed data, to describe how the care patterns evolve for different subjects, and to analyze how much they comply with expected data behaviours. A Process Mining-oriented approach is adopted to perform the steps of a classical statistical analysis, by specifically focusing on the formalisms provided for querying data for inferential analyses and exploring the model outputs provided in terms of easily accessible graphical workflows.

Finally, in Chapter 5 the main innovation aspects, strengths, and limitations of the proposed methodologies are discussed.

Chapter 2

Missing Data Imputation

This Chapter deals with one of the possible issues to address in the preprocessing phase, that is, the presence of missing values in the data. In the specific context of longitudinal clinical data collection, this matter requires in general to simultaneously manage the temporal dimension of data, the possible feature type heterogeneity, and the medical information content.

First, state-of-the-art imputation techniques are presented and their limitations when applied to clinical data with a temporal dimension discussed. Then, two innovative methodologies are introduced and thoroughly described, by additionally providing practical examples of application on real-world clinical datasets. The developed methods are designed to meet different cases of use commonly present in clinical databases, by not only managing the longitudinal data nature, but also exploiting the richness of the information collected increasingly over time.

2.1 The Missing Data Issue

When collecting healthcare information, the type and frequency of acquired data may vary based on the specific application field, patients' clinical conditions, and/or administrative requirements. A typical issue when working with these real-world datasets – in healthcare, but more generally in many other domains – is the presence of missing values. In the specific context of clinical data, medical tests and treatments can be carried out at different times even if patients exhibit the same symptoms, resulting in different information densities. This, together with human factors (poor handwriting, missing charts or pages, measurements being documented in inconsistent locations, etc.), can results in many aspects of a patient's clinical condition being unmeasured or unrecorded at different time points.

Missing values may be clinically important, but cannot be handled by most analytics algorithms [215] and can significantly affect the conclusions drawn from the data [80]. For instance, missing data can introduce bias in the results of randomised controlled trials, negatively affecting the derived clinical decisions and ultimately patient care [172]. When performing survival analysis, missing data can occur in one or more risk factors. The standard response of simply excluding the affected individuals from the analysis could lead to invalid results if the excluded group is selective with respect to the entire sample, and to a waste of costly collected data [198]. In re-

mote health monitoring settings, missing data is a prevalent issue affecting long-term monitoring systems which can lead to failure in decision making [11]. For electronic health records, missing values frequently outnumber observed ones, mainly because they were designed to record and improve patient care and streamline billing rather than collecting data for research purposes [14].

Many kinds of analyses, from simple statistics to advanced data mining and machine learning methods, either fail altogether in dealing with missing data or end up producing biased estimates of the investigated associations when simple curing techniques (such as complete case analysis, overall mean imputation, or the missing-indicator methods) are applied [53]. To utilise all clinical data and achieve optimal performance of the used algorithms, the missing data issue must be addressed, and thus a preliminary step of imputation (*i.e.* "filling in" the gaps with plausible values) is often performed in the preprocessing phase.

2.2 Types of Missing Data

Missing values can be of three general types: *missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR). When missing data are MCAR, the presence and/or absence of data is completely independent of observable variables and parameters of interest. In this case, the set of subjects with no missing data is also a random sample from the source population. This represents the best possible type of missing data as any analysis performed will be unbiased [81], although it is a highly unlikely scenario.

Missing data are MAR when the propensity for a value to be missing depends on some observed patient characteristic. For instance, males are less likely to fill in a depression survey. This kind of missing data can induce bias in the resulting analysis especially when the data is unbalanced because of many missing values in a certain category.

Finally, we are in the MNAR scenario when the missing values are neither MCAR nor MAR. For instance, when asking subjects for their income level it might well be that missing data are more likely to occur when the income level is relatively high. Or, as another example, when asking a subject about his/her habits such as alcohol drinking, the answer is more likely to be missing/non declared in case of excessive consumption. Here, the reason for missingness obviously is not completely at random, but it is related to unobserved patient characteristics, and/or to the specific context the variables belong to.

2.3 Previous Work on Missing Data Imputation

Several methods for handling the presence of missing data are already available to date [17].

The simplest approach when dealing with missing data is applying filtering techniques to exclude from the analyses all unrecorded information. This can consist in completely dropping all cases where at least one variable is missing (listwise deletion), or by only deleting cases having missing values in one of the variables being considered in the specific evaluated model (pairwise deletion). As a limitation, these approaches completely neglect the relationships among variables, possibly causing severe information loss and worsening the statistical power and standard

errors of the analyses [154, 217]. In addition, especially for pairwise deletion, not all the algorithms allow/include the option to only work with the available cases. Finally, filtering could not be a viable option when the sample cardinality is limited and/or the percentage of missing data is too high.

Simple statistical approaches, such as mean/median/mode filling or value propagation (Last Observation Carried Backward or Next Observation Carried Forward), are often applied. Despite being fast and easily interpretable, these methods may lead to low accuracy and biased estimates of the investigated associations [53, 124].

More advanced methods which take into account cross-sectional relationships among the data have been proposed. Regression approaches estimate missing values by regressing them from other related variables [230], especially time [147]. Specifically, deterministic regression imputes the data by using the exact prediction of the regression model. This, however, can produce an overestimation of the correlation among the variables, sometimes even introducing spurious correlations. To overcome this issue stochastic methods can be employed, where a random error term is added to the predicted value in order to recover a part of the data variability [164].

Multivariate imputation by chained equations (MICE) [199] is one of the most prominent methods in the literature [12]. In this imputation procedure, a series of regression models are run whereby each variable with missing data is modelled conditional upon the other variables in the dataset. This means that each variable is modelled according to its distribution, with, for example, predictive mean matching for continuous data, logistic regression for binary data, polytomous logistic regression for categorical data and proportional odds for ordinal data.

3D-MICE, recently introduced in [123], combines MICE with Gaussian process (GP) [163, 90] predictions, thus imputing missing data based on both cross-sectional and longitudinal patient data information. MICE is used to carry out cross-sectional imputation of the missing values, while a single-step GP is used to perform longitudinal imputation. The estimates obtained by the two methods are then combined by computing a variance-informed weighted average. 3D-MICE can adequately impute continuous longitudinal patient data, but is unable to handle categorical and static variables.

A non-parametric method based on a random forest that can cope with different types of variables simultaneously, called missForest, was introduced by Stekhoven *et al.* [186]. This method is based on the idea that a random forest intrinsically constitutes a multiple imputation scheme by averaging over many unpruned classification or regression trees. While not requiring assumptions about the statistical distribution of the data, missForest requires the observations to be pairwise independent, that is, distinct samples must present no cross-relations. This is of course hardly the case when the dataset consists of a longitudinal collection of data belonging to the same subjects recorded over different time points (as in, for instance, clinical registers with several visits for each patient).

Another popular imputation method for cross-sectional time series data is Amelia II [89], which performs multiple imputation by implementing an Expectation-Maximisation with Bootstrapping algorithm. Amelia II is able to impute cross-sectional, time-series, and time-series-cross-section data, also allowing the incorporation of observation and data-matrix-cell level prior information. At the same time, this method requires all variables in the dataset to be multivariate

normally (MVN) distributed. This requirement reduces the applicability of the method especially when dealing with non-normalisable and/or categorical variables.

Recently, a number of deep learning frameworks for estimating missing values in multi-time-series clinical data have been proposed [25, 121, 225]. These methods achieved impressive results on benchmark datasets due to the high-quality representations extracted from large amounts of data, which means that their applicability is limited when less data are available.

Finally, "nearest neighbours" (NN) methods are among the most popular imputation procedures [7, 20]. Missing values of samples with missing data are replaced by values extracted from other similar samples with respect to observed characteristics. NN imputation approaches are donor-based methods where the imputed value is either a value that was actually measured for another record in a database (1-NN) or the average/median/mode of measured values from k records (k-NN). While many imputation methods require the missing data to be MCAR, or at least MAR, imputation based on a k-nearest neighbours approach is applicable in any of the three above-described (MCAR, MAR and MNAR) situations, as long as there is a relationship between the variable with the missing value and the other variables [20]. NN methods were often shown to outperform other imputation techniques [224], even though results depend heavily on the choice of the metric used to measure the similarity between samples.

2.4 Open Issues and Contribution

When approaching the issue of imputing clinical data, some needs related to the specific nature of the collected information emerge.

First of all, as introduced in Chapter 1, one of the characterizing properties of clinical data is often the dependence of some variables on time.

The multiple acquisition of one or more variables over time can significantly contribute in terms of exploitable information for imputing possible missing values. Indeed, by considering a single variable its consecutive monitoring can suggest possible trends to be exploited in inferring the unrecorded values. Let's think for instance to weight monitoring in a patient: if one value is unrecorded and the sampling is dense enough, interpolation or regression techniques could be applied to infer the missing acquisitions, by actually treating the feature as a time series. Besides, observing the parallel evolution of more variables together can bring out possible relationships that can be exploited to impute missing values in present or future time points. It could be the case, for example, of variables with possible contemporary inter-related trends, or features that are predictive for some conditions at subsequent times.

Although potentially useful for addressing the issue of missing collected data, the temporal dimension can also be extremely challenging to manage, being often noncontinuous and asynchronous, as introduced in Chapter 1 (see Figure 1.2). At a single-patient level each subject can have indeed different times for recorded symptoms and findings, performed diagnostic studies, and provided treatments. By considering the study population in its entirety, the time intervals between subsequent observations can also widely vary between patients and/or clinical settings, resulting in data collected over a potentially highly-sparse grid of time points [123].

With reference to the state-of-the-art methods illustrated in the previous section, most of

them are designed for cross-sectional imputation (measurements at the same time point) and thus not able to explicitly handle the temporal nature of longitudinal patient data [91]. For some algorithms (as for instance missForest [186]), moreover, the assumption of independence between samples has to hold, thus making not immediate their extension to datasets consisting of multiple acquisitions per patient. Furthermore, because data collection periods may vary across patients, aggregated samples may not be directly comparable.

Another characteristics of the data collected in the clinical context is the potentially vast heterogeneity in the type of variables. According to the temporal dimension discussed above, variables in this domain can first be classified as either *static* if constant throughout the patient's clinical history, or *dynamic* if varying in time, as introduced in Chapter 1.Besides, they can be *continuous* when representing measurements in a range of continuous values, *ordinal* when the values fall in a discrete ordered set, or *categorical* when describing a qualitative property out of a finite number of categories or distinct groups without any order relations.

An adequate imputation method should therefore be able to handle this data complexity altogether. However, many of the available imputation methods are restricted to only one type of variable (as in 3D-MICE [123], that is able to impute continuous longitudinal patient data only). For mixed-type data, the different variable types are usually handled separately, thus ignoring possible relations among variables of different types. Moreover, most of them make strong assumptions on the characteristics of the missing data, such as locality in Gaussian Process based models [90], low-rankness and temporal regularity in matrix factorisation models [226] and multivariate normality in Expectation-Maximisation methods [89]. For methods based on sample comparisons, like the k-NN ones, the implemented similarity metrics are often not designed to handle data of different nature at the same time, nor they take into account the possibly unbalanced contribution of static and dynamic variables, with the latter recursively adding information over time.

The following sections of this chapter present my contributions to address these open issues, consisting of two imputation methods specifically designed for different types of longitudinally-collected clinical data with missing values. The methods have been designed to meet possible cases of use, that may differ in terms of variable characteristics, frequency of data acquisitions, or clinical similarities among patients.

Section 2.5 outlines a methodology developed in the context of the 2019 ICHI Data Analytics Challenge on Missing data Imputation (DACMI – `https://www.ieee-ichi.org/2019/challenge.html`). The algorithm has been designed to handle missing data in longitudinal data collections where no *a priori* assumption of clinical similarity among patients can be provided. Based on that, an intra-patient approach was performed, that is, the missing values of a patient are inferred directly from his/her previous/following acquisitions. To do that, a further characteristics of the data is required, that is, patients with missing data should have a sufficiently high number of acquisitions. This algorithm can be only applied on datasets consisting of all continuous features.

The proposed methodology bases on the combination of linear interpolation and a weighted k-NN procedure. In the first case, a feature missing value is imputed by using the information collected over the previous and next acquisitions of the same feature. In the k-NN, instead,

some cross-information among variables is included, by defining a similarity metric that employs the Maximal Information Coefficient (MIC) as weights. The choice of combining two distinct imputation approaches was determined after observing that distinct features can exhibit different characteristics in terms of information: some of them could evolve in such a way that any missing value can better be inferred from the close-in-time acquisitions of the same variable; on the contrary, some others can benefit more from the cross-contributions of other features collected at corresponding time points.

The procedure has been tested and validated on an Intensive Care Unit (ICU) database, according to the task proposed for the challenge. A cross-validation scheme has been set up to choose the optimal value of the k-NN parameter and assess the overall method performance, designing a masking procedure where known values were first removed from the data, then imputed and finally compared with the true ones.

This work was presented in the context of the 7th IEEE International Conference on Healthcare Informatics (ICHI 2019) [44], and published as extended paper on the Journal of Healthcare Informatics Research [45]. The algorithm was implemented in R, and has been released as freely available package.

Section 2.6 describes an imputation methodology completely based on a weighted k-NN approach. The implemented technique is able to manage the simultaneous presence of missing information in static and dynamic mixed-type variables, thanks to an *ad hoc* similarity metric that handles both the presence of multiple missing values and the different nature of the features. In this case, the algorithm has been designed for datasets where clinical similarity hypotheses among patients can be formulated (such as registries of patients affected by the same pathology). Implementing an inter-patients approach, the algorithm assesses and exploits the similarity of the patients' clinical evolution over time, based on the assumption that patients with analogous disease course over time can also share similar data values.

Thus, the algorithm proceeds in two steps: first, for each subject with missing data to be imputed, k-NN samples are adaptively created based on the available visits, effectively capturing the temporal evolution of the features over time. Then, a weighted k-NN algorithm is run, identifying the patients who share similar disease progression patterns and using their known values to infer plausible estimates of the missing ones. Also in this case we employed a weight in the similarity metric to include some cross-information between features when comparing the subject variables, hereby computing the Mutual Information metric.

This method was applied on a disease register on Amyotrophic Lateral Sclerosis (ALS): this dataset presents all the above-described characteristics, from the presence of mixed data types and multiple missing values per sample, to the chance to formulate hypotheses of clinical similarity among subjects. Here, the algorithm has been validated both (i) in terms of imputation performance, by employing a cross-validation and masking procedure similar to the previous work, and (ii) by assessing the accuracy improvement of a survival classifier trained on the imputed data.

This work has been presented in the context of the 16th International Conference on Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB 2019) [189], and published as extended paper on BMC Medical Informatics and Decision Making [188]. The implemented algorithm, developed in R, was structured as package and released on CRAN.

2.5 A Combined Interpolation and Weighted k-Nearest Neighbours Approach for the Imputation of Longitudinal Clinical Data

In the framework of the 2019 ICHI Data Analytics Challenge on Missing data Imputation, an imputation task for longitudinal ICU laboratory test data was shared. The challenge centered on the single task of imputing missing data in a clinical dataset of longitudinal multivariable laboratory test results. The provided dataset was an extraction of the large real world ICU database MIMIC-III developed by the MIT Lab for Computational Physiology (`https://mimic.physionet.org/`, [97]), consisting of clinical laboratory test results for 13 commonly measured analytes.

This section describes the methodology developed for the challenge, which consists of the combination of linear interpolation and a k-Nearest Neighbours procedure enriched with a MIC-based weighting scheme [167] to further exploit the relationships among variables. Starting from the characteristics of the dataset provided for the challenge, that does not allow to formulate any *a priori* clinical similarity among patients, and that is constituted by a quite high number (a couple of tens, on average) of available data acquisitions for each patient, the method was built up as intra-patient approach, *i.e.* using the values of the other visits of the same patient to impute his/her missing values in a specific acquisition. Besides, the employment of the MIC computed among pairs of features as weights in the k-NN similarity metric allows to integrate cross-contribution information available at a population level. The MIC was chosen as statistical measure thanks to its capability to capture both linear and nonlinear relationships among features. The algorithm proceeds by first independently testing the linear interpolation and weighted k-NN imputation as distinct approaches for imputing the missing features. Then, the two approaches are combined by selecting for each feature the best performing one, with the imputation performance assessed in a CV setting where some known data are masked and then imputed with the two approaches. In this way, the method optimally meets the nature of each variable, that can either benefit from the cross-contributions of the other features (in the k-NN algorithm) or from the intra-feature information (in linear interpolation).

The implemented methodology was validated on an independent test set against 3D-MICE, *i.e.* the baseline imputation algorithm proposed by the DACMI organisers. The developed method demonstrated statistically significant improvements in 11 out of 13 analytes, with an average performance gain of 8.1%, as well as a considerably reduction of the required computational time.

2.5.1 Material: Continuous Laboratory Test Data

The datasets [122] provided by the DACMI organisers to the challenge participants were derived from MIMIC-III [78, 97], a large real-world database containing de-identified information regarding the clinical care of patients who stayed within the intensive care units (ICU) at Beth Israel Deaconess Medical Centre. Both a training and a test set were provided in order to develop and validate the imputation methodology on two independent sets of data, each one consisting of

inpatient test results for 13 analytes (laboratory tests): Chloride (PCL), Potassium (PK), Bicarbonate (PLCO2), Sodium (PNA), Hematocrit (HCT), Hemoglobin (HGB), Mean Cell Volume (MCV), Platelets (PLT), White Blood Cell count (WBC), Red blood cells Distribution Width (RDW), Blood Urea Nitrogen (PBUN), Creatinine (PCRE), and Glucose (PGLU). Each visit is composed of 13 analyte measurements and is identified by the time in minutes from the first visit (which is identified by timestamp 0). The training set consists of the test results of 8 267 subjects for a total of 199 695 visits, while the test set consists of the test results of 8 267 other subjects for a total of 199 936 visits.

In order to evaluate the performance of the developed imputation algorithms, we employed a randomly masked version of both the training and test sets (see Tables 2.1 and 2.2); one result per analyte per patient-admission was randomly removed, *i.e.* each patient had 13 results masked across the various visits (time points), thus creating cases with known ground truth results. Then, both natively missing and masked data were imputed together; finally, the imputed vs measured values were compared for masked data elements to evaluate the performance of the imputation method.

Table 2.1: *Characteristics of the training set.*

Analyte	Units	Interquartile range	Native missing rate (%)	Missing rate after masking (%)
Chloride	mmol/L	100–108	1.18	5.32
Potassium	mmol/L	3.7–4.4	1.34	5.48
Bicarb.	mmol/L	22–28	1.39	5.53
Sodium	mmol/L	135–142	1.26	5.4
Hematocrit	%	26.8–32.7	12.51	16.65
Hemoglobin	g/dL	8.9–11	15.09	19.23
MCV	fL	86–94	15.23	19.37
Platelets	k/μL	130–330	14.55	18.69
WBC count	k/μL	7.1–14.1	14.8	18.94
RDW	%	14.5–17.4	15.34	19.48
BUN	mg/dL	16–43	0.74	4.88
Creatinine	mg/dL	0.7–1.9	0.7	4.84
Glucose	mg/dL	100–148	2.7	6.84

2.5.2 Methods: Combined Linear Interpolation and MIC-Weighted k-NN Imputation

2.5.2.1 Linear Interpolation Imputation

First, a simple imputation algorithm based on linear interpolation has been implemented as follows. Given an analyte value to be imputed in a certain visit, the other visits from the same

Table 2.2: *Characteristics of the test set.*

Analyte	Units	Interquartile range	Native missing rate (%)	Missing rate after masking (%)
Chloride	mmol/L	100–108	1.20	5.40
Potassium	mmol/L	3.7–4.4	1.28	5.48
Bicarb.	mmol/L	22–28	1.41	5.60
Sodium	mmol/L	136–142	1.26	5.45
Hematocrit	%	26.8–32.6	12.45	16.64
Hemoglobin	g/dL	8.9–11	14.93	19.13
MCV	fL	87–94	15.04	19.24
Platelets	k/μL	133–332	14.42	18.62
WBC count	k/μL	7.1–14.1	14.69	18.89
RDW	%	14.4–17.3	15.16	19.36
BUN	mg/dL	15–42	0.77	4.97
Creatinine	mg/dL	0.7–1.8	0.75	4.94
Glucose	mg/dL	100–147	2.63	6.83

patient are inspected. If the missing data are located between known measurements, they are estimated by linear interpolation in the specific time points. Otherwise, if the missing data correspond to the first or last visits of a given patient, these values are imputed by simply carrying the next observation backward or the last observation forward. When the values of an analyte are missing in all the visits of a given patient, they are imputed with the corresponding average over the population.

2.5.2.2 Weighted k-Nearest Neighbours Imputation

Then, an intra-patient imputation procedure based on a weighted k-NN algorithm has been implemented. Given a missing value in a patient visit, the algorithm uses the other visits from the same patient as neighbours. The k-NN algorithm can be used for imputing missing data by finding the k neighbours closest to the observation with missing data, and then imputing them using the non-missing values from the neighbours [21]. In this way, the algorithm substitutes the missing data with plausible values that are close to the true ones. The similarity among samples is assessed through a distance metric, that compares the values of the available features for each couple of samples. In this work, a MIC-weighted and normalised Euclidean distance metric was employed as a similarity measure.

The following sections detail the MIC and its use in the k-NN distance metric.

Maximal Information Coefficient

In order to exploit the inter-patient analyte dependencies, the cross-information over analytes – computed as the MIC – was integrated in the k-NN procedure as weight in the distance metric.

Unlike correlation metrics, the Mutual Information (MI) and the MIC are measures able to assess the strength of both linear and nonlinear associations among features.

For two discrete variables X and Y whose joint probability distribution is $p_{XY}(x,y) = P(X=x, Y=y)$, and marginal probability distributions are, respectively, $p_X(x) = P(X=x)$ and $p_Y(y) = P(Y=y)$, the MI between them, denoted $\text{MI}(X,Y)$, is computed as:

$$\text{MI}(X,Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y)\ln\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} \quad . \tag{2.1}$$

The marginal and joint probability distributions of X and Y are determined empirically from the data by a frequentist approach, as explained below.

Introduced by Reshef *et al.* [167], the MIC extends the MI based on the idea that if a relationship exists between two variables, then a grid can be drawn on the scatterplot of the two variables that partitions the data to encapsulate that relationship. According to the rationale of the MIC, the grid bins should be chosen in such a way that the MI between the variables is maximal.

Formally, to calculate the MIC of a set of two-variable data X and Y, all grids up to a maximal grid resolution (dependent on the sample size) are explored, computing for every pair of integers (x,y) the largest possible mutual information achievable by an x-by-y grid applied to the data.

Let $D = (X,Y)$ be the set of n ordered pairs of elements of X and Y. The data space is partitioned in x-by-y grids, grouping the X and Y values in x and y bins respectively. The MIC is defined as:

$$\text{MIC}(X,Y) = \max_{xy<B(n)} \frac{MI^*(D,x,y)}{log(min(x,y))} \quad , \tag{2.2}$$

where $B(n) = n^\alpha$ is the search-grid size, with n being the sample size and α usually set to 0.6, and $MI^*(D,x,y)$ is the maximum MI of the distribution induced by D on all the grids having x and y bins (where the probability mass on a cell of the grid is the fraction of points of D falling in that cell) [6].

The normalization in Eq. 2.2 ensures the comparison between grids of different dimensions by correcting for the scale factor, reducing the range of possible values from the $[0,+\infty]$ interval of the MI to $[0,1]$: in this way, high MIC values correspond to strongly associated variables, while low ones correspond to weak associations. When trying to discover associations among pairs of variables, the statistic used to measure the dependence should exhibit two heuristic properties: *generality* and *equitability* [167]. Generality is the ability of a given statistic to capture a wide range of interesting associations, not limited to specific function types (such as linear, exponential, or periodic) or to functional relationships, provided that the sample size is sufficiently large. This property is essential because many important relationships are not well modelled by a specific function. Equitability, on the other hand, is the property of a given statistic to give similar scores to equally noisy relationships of different types. The MIC was shown to outperform several other methods in terms of generality and equitability, including mutual information estimation, distance correlation, Spearman's rank correlation coefficient, principal curve-based methods, and maximal correlation [167].

In this work, the MIC was thus chosen as measure of association and computed among all pairs of analytes on the whole training set using the *minerva* R package v1.5.8 [6]. A heatmap of the cross-sectional MIC among analytes on the training dataset is shown in Figure 2.1. By using the MIC values as weights in the distance metric, we ensure that intra- and inter-patient information are integrated in the imputation procedure.

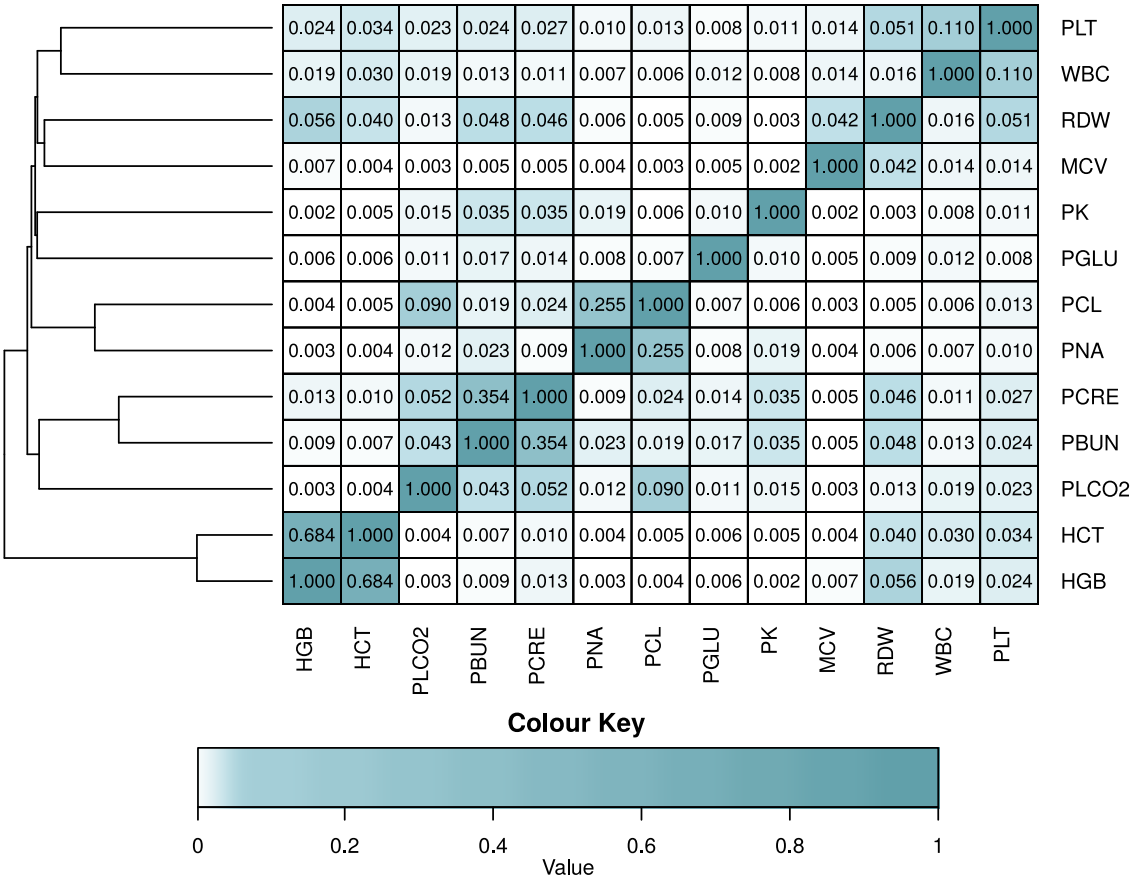


Figure 2.1: *Heatmap and dendrogram of the cross-sectional MIC among analytes computed on the training set.*

k-NN Distance Metric and Imputation Procedure

For a given patient with missing data in at least one variable measurement, the values of his/her 13 analytes (all continuous features) are first normalised to the $[0, 1]$ interval, in order to account for the differences among the analyte ranges. Let $Y_{p,a,i}$ be the measured value for analyte a of patient p at time i, and let $Y_{p,a}$ be the set of all known values for analyte a of patient p. The normalised value is given by:

$$nY_{p,a,i} = \frac{Y_{p,a,i} - \min(Y_{p,a})}{\max(Y_{p,a}) - \min(Y_{p,a})} \quad . \tag{2.3}$$

Then, given the patient visit composed of the measurements of its 13 analytes $\mathbf{v} = (v_1, v_2, \ldots, v_{13})$ with missing data to be imputed in index $i \in \{1, \ldots, 13\}$, the algorithm computes the weighted Euclidean distance with the other visits of the current patient that do not have missing data in position i as:

$$d(\mathbf{v}, \mathbf{u}) = \frac{\sqrt{\sum_{j \in N} \text{MIC}_{i,j} \cdot (v_j - u_j)^2}}{\sum_{j \in N} \text{MIC}_{i,j}} \ , \tag{2.4}$$

where N is the set of indices corresponding to non-missing values in both visits $\mathbf{v}$ and $\mathbf{u}$, and $\text{MIC}_{i,j}$ is the maximal information coefficient between analytes i and j computed on the whole dataset. By dividing the numerator in Eq. 2.4 by the quantity $\sum_{j \in N} \text{MIC}_{i,j}$ we are normalising the distance in order to account for other possible missing values (other than the one being currently imputed) and their importance. This favours candidate neighbouring visits that have many analytes highly associated with the one being currently imputed, and penalises candidate neighbouring visits that have missing values instead (a visit can have several missing values).

Once the distances to all the visits of the current patient have been computed, the nearest k candidates are selected and the missing value is imputed using the average of the corresponding values in the k candidate visits, each weighted by the corresponding distance.

If a visit has multiple missing values to be imputed, the k-NN procedure is repeated for each one of them separately, as the MIC weights (and consequently the distance values, see Eq. 2.4) depend on the specific analyte being imputed. In this implementation, values previously imputed by the k-NN are not used in distance computations and subsequent imputations. Again, if the values of an analyte for a given patient are missing in all his or her visits, the average over the population is used for the imputation of that analyte.

Selection of the Optimal Number of Nearest Neighbours *k*

The k-NN methods require to set a single hyperparameter, that is, the number of neighbours k. To select the optimal value of k in this work, a 10-fold Cross Validation (CV) has been performed at patient level on the training set.

In Machine Learning – and, more in general, in statistical modelling – K-fold Cross Validation[1] is a procedure that allows to thoroughly assess the performance of a model [187]. Given a dataset, it is first subdivided in a fixed number of distinct partitions, named folds: in turn, each fold is selected as internal testing set, while the others are used as internal training set. The internal training set is used to develop the model of interest, and the internal testing set is employed to assess the model predictive performance through a selected evaluation metric (usually corresponding to the one chosen for assessing the performance of the final model outside the CV setting). Then, the single performances obtained in turn over all the folds are combined into a global one (such as the average), that is used to evaluate the global model.

[1]The K of the K-fold Cross Validation indicates the number of folds the dataset is partitioned into and has, in this context, not to be confounded with the k (number of neighbours) parameter of the k-NN methodology, current object of the paragraph.

With respect to a single model developed on all the training set and validated on the test set, the CV allows to flag problems like *overfitting* (model excessively customized on the set of training data points and thus unable to generalize on new independent data) and *selection bias* (model built on a set of data not representative of the population intended to be analyzed) [26].

Since testing a model on various subsets enforces the reliability of experimental design and evaluation, CV can be used not only in performance assessment, but also in all aspects of model learning, such as feature selection, model type selection and, like in this case, hyperparameter tuning [212].

According to the chosen 10-fold CV setting, the 8 267 subjects of the training dataset were randomly split into 10 disjoint folds. In this framework, different values of the hyperparameter k for the k-NN algorithm were set and tested, ranging from 1 to 15. Being the k-NN procedure performed intra-patient, the only dataset-dependent item of the algorithm is the MIC. Thus, in turn, the visits of the subjects in a given fold were imputed using the MIC computed over the remaining 9 folds. The imputation performance was assessed by using the nRMSD metric presented in the next section. The results are shown in Table 2.3: the best average nRMSD values were obtained for $k \in \{3, 4\}$.

Table 2.3: *Results of the 10-fold cross validation procedure on the training set for the weighted k-NN algorithm.*

Analyte	Number of selected neighbours														
	$k=1$	$k=2$	$\mathbf{k=3}$	$\mathbf{k=4}$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$	$k=11$	$k=12$	$k=13$	$k=14$	$k=15$
Chloride	0.2267	0.2028	0.1966	0.1977	0.2004	0.2038	0.2083	0.2124	0.2167	0.2210	0.2244	0.2275	0.2301	0.2323	0.2341
Potassium	0.2967	0.2633	0.2539	0.2507	0.2494	0.2492	0.2498	0.2507	0.2519	0.2530	0.2541	0.2549	0.2556	0.2562	0.2567
Bicarb.	0.2654	0.2376	0.2335	0.2328	0.2358	0.2380	0.2406	0.2434	0.2467	0.2497	0.2520	0.2540	0.2558	0.2576	0.2591
Sodium	0.2450	0.2199	0.2143	0.2132	0.2145	0.2171	0.2199	0.2230	0.2265	0.2292	0.2318	0.2341	0.2360	0.2377	0.2392
Hematocrit	0.1554	0.1438	0.1445	0.1497	0.1560	0.1630	0.1704	0.1782	0.1851	0.1903	0.1944	0.1977	0.2003	0.2025	0.2043
Hemoglobin	0.1540	0.1435	0.1453	0.1500	0.1568	0.1641	0.1716	0.1795	0.1863	0.1917	0.1960	0.1993	0.2021	0.2044	0.2063
MCV	0.3032	0.2710	0.2626	0.2607	0.2613	0.2630	0.2643	0.2662	0.2684	0.2704	0.2720	0.2734	0.2746	0.2757	0.2765
Platelets	0.2608	0.2334	0.2304	0.2328	0.2369	0.2423	0.2474	0.2526	0.2576	0.2617	0.2649	0.2677	0.2699	0.2718	0.2733
WBC counts	0.2772	0.2483	0.2430	0.2440	0.2471	0.2503	0.2536	0.2569	0.2597	0.2617	0.2635	0.2649	0.2661	0.2671	0.2679
RDW	0.2732	0.2435	0.2402	0.2418	0.2453	0.2493	0.2532	0.2575	0.2616	0.2646	0.2676	0.2702	0.2719	0.2734	0.2747
BUN	0.2563	0.2291	0.2231	0.2240	0.2269	0.2305	0.2334	0.2373	0.2409	0.2443	0.2477	0.2503	0.2528	0.2547	0.2563
Creatinine	0.2605	0.2353	0.2307	0.2294	0.2303	0.2330	0.2360	0.2393	0.2424	0.2453	0.2479	0.2500	0.2519	0.2535	0.2550
Glucose	0.3174	0.2820	0.2717	0.2672	0.2661	0.2654	0.2646	0.2648	0.2654	0.2657	0.2662	0.2666	0.2671	0.2675	0.2680
Average	0.2532	0.2272	**0.2223**	**0.2226**	0.2251	0.2284	0.2318	0.2355	0.2392	0.2422	0.2448	0.2470	0.2488	0.2503	0.2516

Best performance is highlighted in bold.

2.5.2.3 Imputation Evaluation Metrics

The DACMI task required the participant teams to employ the normalised root-mean-square deviation (nRMSD) metric to evaluate the performance of the developed imputation methods and compare them with the performance of 3D-MICE. Let $X_{p,a,i}$ be the test result prediction for analyte a of patient p at time i and let $Y_{p,a,i}$ be the true measured value for that analyte. Also, let $I_{p,a,i}$ be 1 if the value of analyte a for patient p at time i is missing, and 0 otherwise. The nRMSD of analyte a is calculated as:

$$\text{nRMSD}(a) = \sqrt{\frac{\sum_{p,i} I_{p,a,i} \left(\frac{|X_{p,a,i} - Y_{p,a,i}|}{\max(Y_{p,a}) - \min(Y_{p,a})} \right)^2}{\sum_{p,i} I_{p,a,i}}} \quad . \tag{2.5}$$

The nRMSD is frequently used to measure the differences between values predicted by a model and the ones observed [123]. The normalisation at the patient level facilitates performance comparisons on analytes with different scales and dynamic ranges. The nRMSD measure has been used in the development of the methodology for both assessing the performance of the algorithm and determining the optimal value of the number of neighbours k in the k-NN approach (see previous Section).

Nevertheless, as a limitation the nRMSD constitutes a metrics obtained over all the imputed values, thus implying that possible outlier values heavily impact on the overall assessed performance.

Therefore, we also computed the normalised absolute error (nAE) of each single imputed value, in order to gain more insight on the quality of the imputation by comparing the distribution of the error. The nAE for analyte a of patient p at time i is given by:

$$\text{nAE}(p, a, i) = \frac{|X_{p,a,i} - Y_{p,a,i}|}{\max(Y_{p,a}) - \min(Y_{p,a})} \quad . \tag{2.6}$$

2.5.2.4 Combined Imputation Method

The performance of the linear interpolation and weighted k-NN imputation methods on the training set is reported in Table 2.4. It emerges how the interpolation-based imputation performs better than the k-NN-based one in 7 out of 13 analytes, namely for Bicarbonate, MCV, Platelets, WBC count, RDW, BUN and Creatinine. For this reason, we decided to impute these analytes using linear interpolation, and the remaining ones with the k-NN-based approach, actually implementing a combined interpolation algorithm.

The interpolation is run on each feature separately, thus its results do not depend on the k-NN step. On the other hand, the k-NN could use the imputed values from the interpolation step during distance computation. For this reason we tested the imputation on the training set by combining the methods in both directions: by running the k-NN first and the interpolation second (KNN+Interp.), and vice versa (Interp.+KNN). In the latter case, we tested a few values for k in cross validation to confirm the optimality of the previously selected values: $k = 3$ was selected as the optimal value (the results are shown in Table 2.4).

2.5.3 Results: Imputation Performance Assessment

2.5.3.1 Performance Comparison on the Training Set

The developed imputation procedures were assessed on the training set using the nRMSD. The results in Table 2.4 show that the combined methods outperform 3D-MICE on 11–12 analytes out of 13. The average nRMSD values are equal to 0.2055 for KNN+Interp. $k = 3$ and 0.2043

Table 2.4: *Imputation performance comparison based on the nRMSE metric for each analyte and imputation method.*

Analyte	Training set						Test set	
	Interp.	KNN $k=3$	selected method	KNN+Interp. $k=3$	Interp.+KNN $k=3$	3D-MICE	Interp.+KNN $k=3$	3D-MICE
Chloride	0.2017	0.1966	KNN	**0.1966 (1.6%)**	**0.1915 (4.1%)**	0.1997	**0.1921 (4.0%)**	0.2000
Potassium	0.2590	0.2539	KNN	**0.2539 (2.9%)**	**0.2505 (4.2%)**	0.2614	**0.2542 (3.4%)**	0.2632
Bicarb.	0.2165	0.2335	Interp.	**0.2165 (6.9%)**	**0.2165 (6.9%)**	0.2326	**0.2118 (8.5%)**	0.2314
Sodium	0.2242	0.2143	KNN	0.2143 (-0.3%)	**0.2113 (1.1%)**	0.2136	**0.2085 (2.8%)**	0.2145
Hematocrit	0.2248	0.1445	KNN	**0.1445 (0.1%)**	**0.1434 (0.9%)**	0.1447	0.1518 (-0.9%)	**0.1505**
Hemoglobin	0.2282	0.1453	KNN	0.1453 (-1.7%)	0.1451 (-1.5%)	**0.1429**	**0.1485 (0.2%)**	0.1488
MCV	0.2582	0.2626	Interp.	**0.2582 (3.6%)**	**0.2582 (3.6%)**	0.2679	**0.2644 (2.5%)**	0.2713
Platelets	0.1778	0.2304	Interp.	**0.1778 (21.3%)**	**0.1778 (21.3%)**	0.2260	**0.1794 (21.8%)**	0.2294
WBC counts	0.2183	0.2430	Interp.	**0.2183 (14.6%)**	**0.2183 (14.6%)**	0.2555	**0.2198 (14.1%)**	0.2560
RDW	0.2100	0.2402	Interp.	**0.2100 (15.8%)**	**0.2100 (15.8%)**	0.2493	**0.2056 (16.4%)**	0.2458
BUN	0.1521	0.2231	Interp.	**0.1521 (17.7%)**	**0.1521 (17.7%)**	0.1848	**0.1546 (16.3%)**	0.1846
Creatinine	0.2130	0.2307	Interp.	**0.2130 (7.0%)**	**0.2130 (7.0%)**	0.2291	**0.2135 (8.7%)**	0.2338
Glucose	0.2817	0.2717	KNN	**0.2717 (1.9%)**	**0.2683 (3.1%)**	0.2769	**0.2677 (3.3%)**	0.2769
Average	0.2204	0.2223	–	**0.2055 (7.4%)**	**0.2043 (7.9%)**	0.2219	**0.2055 (8.1%)**	0.2235

Best performance is highlighted in bold. The percentage of improvement over 3D-MICE is given in parentheses.

for Interp.+KNN $k = 3$, which corresponds to an improvement of 7.4% and 7.9% respectively, compared to the baseline (0.2219).

2.5.3.2 Performance Comparison on the Test Set

The best performing method Interp.+KNN was validated on the independent test set using the selected optimal $k = 3$ value, the MIC, and the population average values computed on the training set. Figure 2.2 schematically depicts the Interp.+KNN imputation procedure for a given subject. Performance is presented in the last two columns of Table 2.4. The average nRMSD value obtained for Interp.+KNN $k = 3$ on the test set, equal to 0.2055, is 8.1% lower than the 3D-MICE baseline (0.2235). Similarly to the training set, the combined method Interp.+KNN outperforms the baseline on average and on 12 out of 13 analytes, although reversing the sign of the improvement for the features Hematocrit and Hemoglobin.

To assess the statistical significance of the improvement, a one-tailed paired Wilcoxon signed-rank test [222] was performed on the nAEs obtained on the test set with 3D-MICE and with Interp.+KNN for each analyte. The Wilcoxon signed-rank test is a non-parametric statistical test used to assess whether the population mean ranks differ in a paired samples setting. This test can be used to determine whether two paired samples were selected from populations having the same distribution.

Since the nRMSD can be directly derived from the nAE values (see Eq. 2.5 and Eq. 2.6), the performed statistical tests can be used to assess the significance of the improvement in terms of both error measures. The test results in p-values < 0.001 for 11 out of 13 analytes, while the features Hematocrit and Hemoglobin, whose p-values are equal to 0.787 and 0.095 respectively,

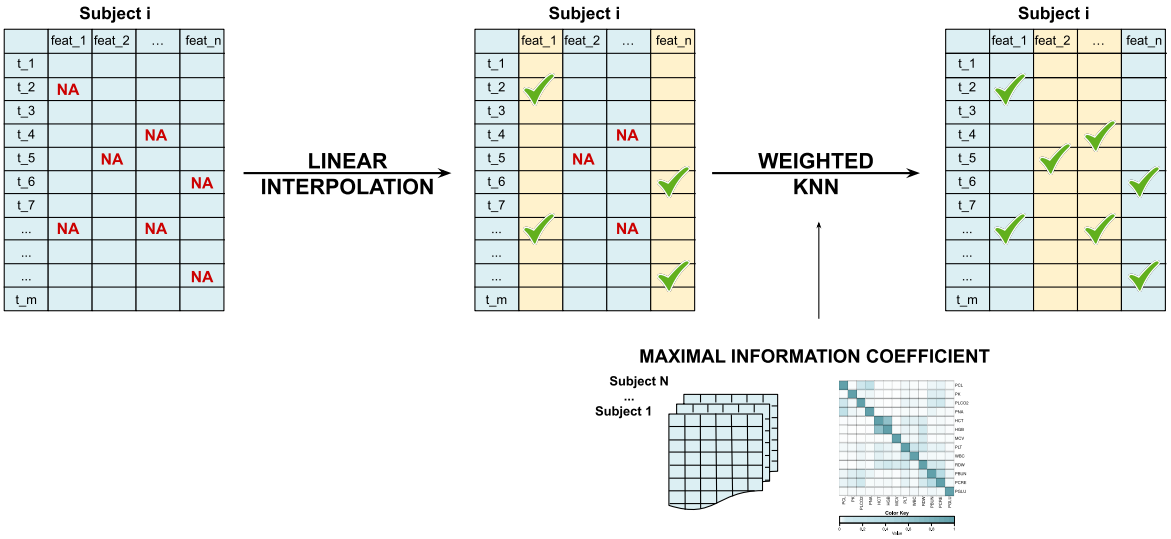


Figure 2.2: *Interp.+KNN imputation procedure. For each subject with missing values, 7 out of 13 analytes are first imputed with linear interpolation. The remaining missing values on the other analytes are then imputed with the k-NN algorithm using the MIC values computed on the training set as weights for the distance metric.*

show no statistically significant improvement in terms of imputation error. This result is also confirmed by both the exiguous difference in the nRMSD values (less than 1% on the test set) obtained by our method compared to those of 3D-MICE for Hematocrit and Hemoglobin, and the reversal of the sign of the improvement on these analytes between training and test set. Figure 2.3 compares the nAE distributions on the test set, showing the shift to lower error values for the Interp.+KNN method with respect to the baseline.

2.5.3.3 Computation Time Comparison

The computation times required for imputing the datasets by using the proposed combined method and the baseline competitor were compared. On a workstation with an Intel® Xeon® W3680 CPU (6 cores/12 threads @ 3.33GHz, 12MB L3 cache) and 24GB of DDR3 RAM, running Ubuntu Linux 16.04 LTS, the combined method can impute a whole dataset of 8 267 subjects with roughly 200,000 visits in less than a minute; the baseline method 3D-MICE requires several hours to impute the same dataset.

2.5.4 Discussion: Applicability and Advantages of a Combined Imputation Approach

Both the interpolation-based and the k-NN-based approaches always yield imputed values in the range of the existing data; more specifically, the intra-patient implementation preserves the analyte dynamic range of each patient.

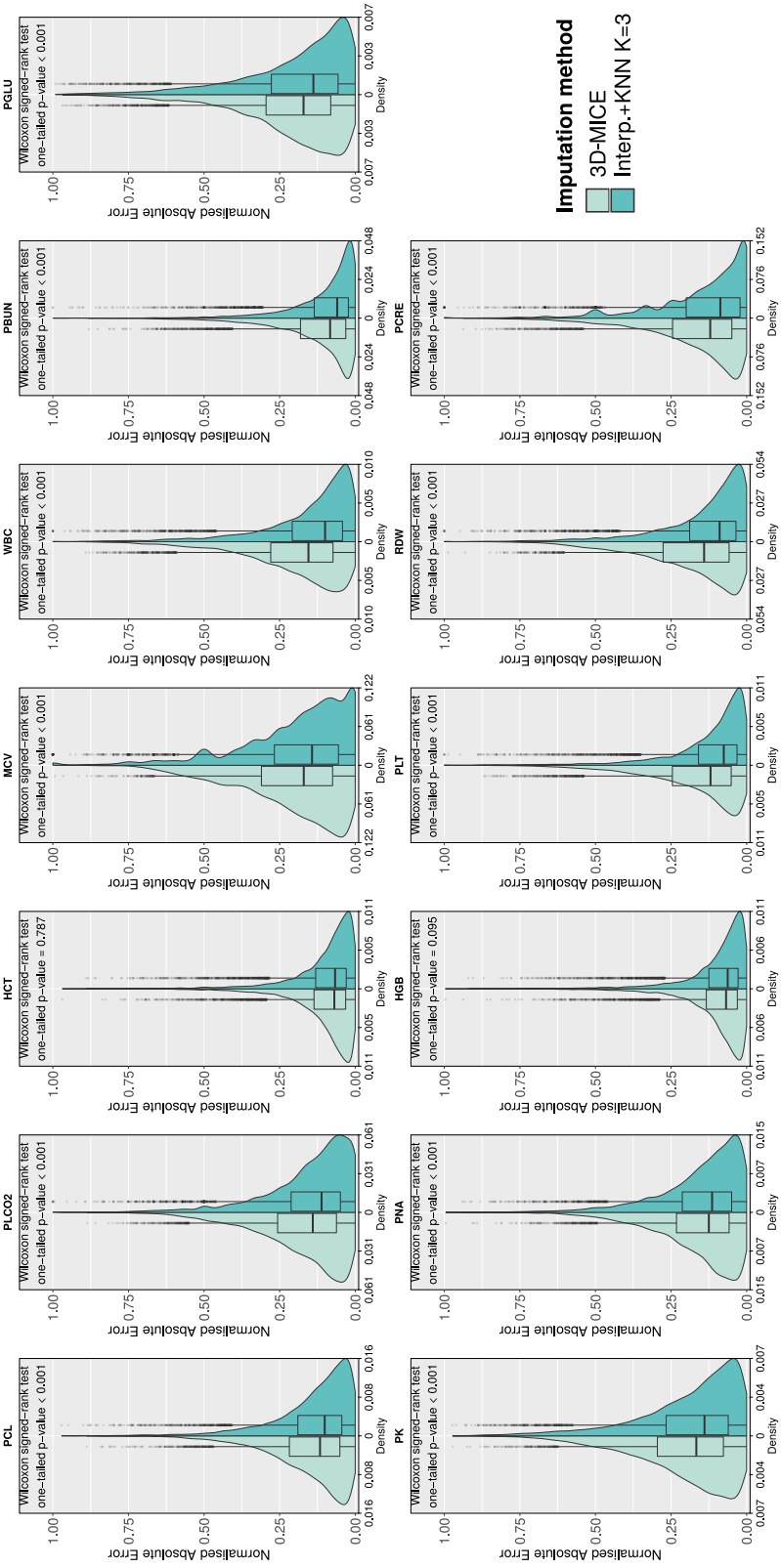


Figure 2.3: *Normalised absolute error distributions obtained with 3D-MICE and Interp.+KNN with $k = 3$ on the test set.*

The integration of the MIC in the weighted k-NN approach adds some data-driven knowledge to the procedure. In the MIC computation (see Figure 2.1), a few relationships that can be expected from the clinical literature emerge. Hematocrit (HCT) and Hemoglobin (HGB), that present a normal ratio of 1:3 in healthy subjects and possibly altered values in the pathological ones [160], have the highest MIC value; similarly, the MIC value for Blood Urea Nitrogen (BUN) and Creatinine (PCRE) is also high, being these analytes both referred to the renal function and with a normal ratio ranging from 10:1 to 20:1 [136]. It is interesting to observe how these pairs of features perform differently when this cross-sectional information is incorporated in the k-NN imputation procedure, though they both have high MIC values. The weighted k-NN outperforms the linear interpolation approach for Hematocrit and Hemoglobin, while falling behind for Blood Urea Nitrogen and Creatinine. This could be due to two possible reasons: 1) some analytes follow a linear trend in the intervals containing the missing values, or 2) the information included in the features themselves, exploited by the interpolation, is stronger than the cross-information. In the specific case of Blood Urea Nitrogen and Creatinine, the intra-feature information could be prevailing due to the low missingness rate (less than 5% after masking) which reinforces the latter hypothesis. The presence of specific patterns in the patients' missing values is another fact that could promote the effectiveness of one method against the other. The absence of many analytes in one visit could decrease the effectiveness of the weighted k-NN procedure while recurring missing measures of one specific analyte could penalise the interpolation-based approach.

In the k-NN approach, the selection of a small k parameter ensures a good compromise between imputation performance and the need to preserve the original distribution of the data – a very important characteristic any imputation method should satisfy. Indeed, as a rule of thumb, it is advisable to limit the number of k neighbours, because of the risk of severely impairing the original variability of the data [21]. This matter requires particular care, since using the imputation accuracy (as measured for instance by the nRMSD) as the sole parameter selection criteria could lead to the choice of a large k value, while completely neglecting the data distortion aspect.

In general, with a k-NN approach the imputation precision is subject to the degree of dependencies the feature with missing data has with other features in the dataset; imputing features with little or no dependencies could lead to a lack of precision, and could introduce spurious associations by considering dependencies where they do not exist [21]. In the presented approach, this risk is realistically mitigated by selecting the best performing method for each feature, the interpolations replaces the k-NN approach on those analytes where the latter performs poorly. Moreover, the choice of combining the two imputation techniques allows each feature to benefit from the imputation technique that better meets its nature in terms of information contribution (intra-feature, in the linear interpolation, or among features, in the k-NN). Finally, it is worth noticing that the proposed algorithm is very time efficient.

The proposed imputation algorithm was implemented in R, and is freely available at: `https://www.github.com/sebastiandaberdaku/PD_Impute`.

2.6 An Adaptive k-Nearest Neighbours Algorithm for the Imputation of Static and Dynamic Mixed-Type Clinical Data

This section presents an adaptive Mutual Information-weighted k-Nearest Neighbours (wk-NN MI) imputation algorithm developed to explicitly handle missing values of continuous/ordinal/categorical and static/dynamic features conjointly. With respect to the method presented in Section 2.5, the current one is applicable on mixed-type longitudinal collections of data referred to a population that allows to formulate clinical similarity hypotheses (for instance, a cohort affected by the same disease).

The implemented methodology bases on the assumption that, when considering a population affected by analogous clinical conditions, the unrecorded information of a specific patient can be derived from the available data of the other subjects with similar clinical status. In this work, the similarity among subjects has been assessed in terms of clinical progression over a fixed time interval, thus thoroughly exploiting the information acquired over the temporal dimension of the data.

Therefore, starting from a dataset to impute, the developed algorithm includes as first step the definition of specific adaptive feature vector samples, constituted by the information collected over the time interval of interest. These samples embody the description of the patient's progression. They are then used in the k-NN procedure to select, among the patients, the ones with the most similar temporal evolution of clinical history over time. An *ad hoc* similarity metric has been implemented for the sample comparison, capable of handling the different nature of the data, as well as the presence of multiple missing values. As in the previous case, the employed similarity metric includes cross-information among features, hereby captured by the MI statistic (and not the MIC, as in Section 2.5, for the sake of limiting computational complexity, as detailed in the following of this Section).

When considering the heterogeneity of the data recorded in the clinical setting, a typical example of mixed-type variables dataset is represented by disease registers. Thus, as case-study, the proposed methodology was applied and validated on a subset of the Piemonte and Valle d'Aosta Amyotrophic Lateral Sclerosis (PARALS) register [36], an epidemiological register.

On this dataset, the methodology was compared to three other state-of-the-art imputation algorithms, namely Amelia II [89], missForest [186] and MICE [199], which are among the main representatives of the methods currently available in the literature (see Section 2.3). The performed experiments show that the implemented method outperforms the competitors in the imputation of most of the features and on average.

Moreover, to further validate the imputation performance compared to the competitors and to assess the possible impact of the proposed method in a concrete scenario, a simple application of the imputed data in a survival classification task is also provided. A naïve Bayes (NB) classifier was used to distinguish between patients with long and short survival times by using only the information in their first months of screening visits. The results show that imputing the training set with the proposed method improves the prediction performance of the NB classifier on a hold-out test set, also achieving better performance than the classifier built on the training set imputed with the top competitor (MICE). By asserting the effectiveness of the proposed imputation

method in enhancing the training data for a very simple classification algorithm with naïve hypotheses, we confirm its applicability in more complex and sophisticated analyses. Noticeably, the proposed methodology could also be of great aid to clinicians since it enables the survival prediction of patients by employing only the information from a few of their visits, regardless of possible missing values.

2.6.1 Material: Longitudinal Heterogeneous Register Data

The dataset used in this work is extracted from PARALS, that is a clinical prospective epidemiological register of patients affected by amyotrophic lateral sclerosis (ALS) from two Italian regions. It consists in a collection of dynamically acquired data over subsequent screening visits, one visit at a time. PARALS represents a typical instance of complex clinical dataset constituted of both static/dynamic and mixed-type variables and, coherently with its real-world nature, is inevitably subject to missing data.

ALS is a fatal neurodegenerative disorder characterised by progressive muscle paralysis caused by the degeneration of motor neurons in the brain and spinal cord [211]. It is a rare disease, with incidence in Europe and in populations of European descent of 2.6 cases for 100 000 people per year and prevalence of 7–9 cases per 100 000 people [85]. This implies that the available patient data collected in clinical registers is of inestimable importance for furthering the translational research on the disease, and that missing values cannot be treated with simple curing techniques. Compared to clinical trial datasets, epidemiological registers better characterise the general ALS population, since clinical trial populations must fit a stringent set of criteria [61]. This database was selected as case study for the developed technique with the aim to build a complete dataset that can be used for the subsequent application and development of ML algorithms (an example of which is also given in this work as further validation of the implemented imputation procedure).

As mentioned above, the developed imputation method is based on the assumption that subjects with a similar disease progression over a period of time share similar feature values and can therefore be cross-exploited to impute missing values. For the specific case of this dataset, a period of three months of screening visits was selected as time interval. Such span has been chosen accordingly to the threshold selected in two DREAM (Dialogue for Reverse Engineering Assessments and Methods) Challenges on ALS [110, 109]; these challenges required participants to develop algorithms to predict the disease progression and to stratify the patients into meaningful subgroups, respectively, by employing only the clinical information of the first three months of patients' visits. For this aim, different ALS datasets were provided, with the PARALS register used in our work being partially included in the datasets of the second challenge.

Besides being able to adequately characterise the temporal evolution of the disease course [110], the selected time interval is short enough to allow the imputation of subjects with few available visits. Moreover, the information of newly added subjects can be promptly used for the imputation of others. Finally, by focusing on a reduced observation interval, only a relatively small number of visits (and thus a relatively small number of features) is considered. In a k-NN setting, having a small number of features prevents the methods from incurring in the curse of

dimensionality: in general, as the number of dimensions (features) increases, the closest distance among samples tends to the average distance and the predictive power of the algorithm decreases [82].

The dataset used in this work was extracted from the PARALS register as follows. The cohort of patients with first visit from January 1st, 2001 and follow-up up to July 18th, 2017 was first selected. Then, the ones having an onset that predated the first visit by five years or more (average ALS prognosis) were excluded, in order to filter out clinical outliers. The selected cohort includes 700 patients, resulting in a dataset containing the information assessed over their subsequent screening visits, for a total of 6 726 visits.

The 25 variables collected in the dataset include some clinical features recorded during the first visit – the *static* ones – that are: patient sex, body-mass index (BMI) both premorbid and at diagnosis, a measure of respiratory functionality (forced vital capacity, FVC) at diagnosis, familiality of ALS, the result of a genetic screening over the most common ALS-associated genes, presence of frontotemporal dementia (FTD), site of disease onset (spinal/bulbar), age at onset, diagnostic delay (time from ALS onset to diagnosis); the remaining features – the *dynamic* ones – are collected over visits and consist of: the presence/absence up to the current visit of non-invasive ventilation (NIV) and percutaneous endoscopic gastrostomy (PEG), that are two guideline-recommended interventions for symptom management in ALS, and the revised ALS Functional Rating Scale (ALSFRS-R) [27, 28], which is a 12-item questionnaire rated on a 0–4 point scale evaluating the observable functional status and change for patients with ALS over time.

The time of the visit for each patient is expressed in months and set to zero in correspondence of the first visit, resulting in negative values for the onset delta. These variables are detailed in Table 2.5, according to their data type (continuous, ordinal, or categorical), with the percentage of native missing values and the static (S) or dynamic (D) nature of the feature. In this summary, for the NIV and PEG variables the total number of patients who were administered these interventions is reported.

In order to develop and validate the imputation algorithms on independent data, the dataset was split in training (80% = 560 subjects, 5 507 visits) and test (20% = 140 subjects, 1 219 visits) sets, by stratifying the dataset over all variables.

2.6.2 Methods: Adaptive k-NN Sample Construction and MI-Weighted k-NN Imputation

As anticipated, the developed weighted k-NN approach presented in this section was used on the case-study dataset to impute the missing values in the first span of screening visits of each patient. The algorithm is based on the assumption that patients with similar characteristics share the same disease course over time. Patient similarity is assessed by using an apposite distance metric over their features.

According to the k-NN methodology, given a patient with a missing value to be imputed and a pool of other patients having that feature, the algorithm searches for the k-closest subjects in terms of disease progression similarity and infers the estimate for the missing value. First, the

Table 2.5: *Dataset. The feature type, either static (S) or dynamic (D), is defined. For the continuous and ordinal features, percentage of native missing values and inter-quartile range (IQR) values at 25%, 50% and 75% are reported; for the categorical features, levels and corresponding percentage of instances are reported; for the NIV and PEG variables, we reported the total number of patients who were administered these interventions.*

Continuous features			
Feature	Type	% NA	IQR
BMI premorbid [kg/m^2]	S	2.08	23/25/28
BMI diagnosis [kg/m^2]	S	0.91	22/24/27
FVC diagnosis [%]	S	4.12	83/98/108
age at onset [years]	S	0	56/64/70
diagnostic delay [months]	S	0	5/9/14
onset delta [months]	S	0	-18/-11/-6

Ordinal features			
Feature	Type	% NA	IQR
ALSFRS-R 1	D	0	2/3/4
ALSFRS-R 2	D	0	3/4/4
ALSFRS-R 3	D	0	2/3/4
ALSFRS-R 4	D	0	2/3/4
ALSFRS-R 5	D	0	1/2/3
ALSFRS-R 6	D	0	1/2/3
ALSFRS-R 7	D	0	1/3/3
ALSFRS-R 8	D	0	2/2/3
ALSFRS-R 9	D	0	0/1/3
ALSFRS-R 10	D	0	3/4/4
ALSFRS-R 11	D	0	3/4/4
ALSFRS-R 12	D	0	4/4/4

Categorical features			
Feature	Type	Levels	%
sex	S	Female	47.6
		Male	52.4
		NA	0
familiality	S	No	91.4
		Yes	8.1
		NA	0.5
genetics	S	C9orf72	7.1
		FUS	0.3
		SOD1	1.4
		TARDBP	1.6
		wild type	83.6
		NA	6.0
FTD	S	No	53.0
		Yes	13.0
		NA	34.0
onset site	S	Bulbar	34.4
		Spinal	65.6
		NA	0
NIV	D	No	59.6
		Yes	40.4
		NA	0
PEG	D	No	31.9
		Yes	25.0
		NA	43.1

distance between the current patient and each of the other candidate subjects from the pool is computed. Then, a weighted average of the corresponding values in the k most similar patients is obtained and used as plausible estimate of the missing one. To impute the whole dataset, the procedure is iterated for each missing value of the given patient and then for each patient with missing values in their visits. The algorithm takes into account the temporal evolution of the data over visits and handles both the mixed nature of the data and the presence of missing values in the distance computation.

2.6.2.1 Adaptive k-NN Sample Construction

To capture the temporal evolution of the features over subsequent visits, for a given patient i with missing data to be imputed, the algorithm builds a feature vector (k-NN sample) that contains the information recorded during his/her first span of screening visits. Let's express this time interval, that is defined by the user according to the dataset characteristics and/or analysis scopes, as a generic number of months N (or weeks, or days, according to the data granularity). The feature vector is created by binding the static information for that patient (constant throughout all his/her visits) to the dynamic ones in the $[0, N-1]$ months from the first visit in chronological order (with 0 being the first month).

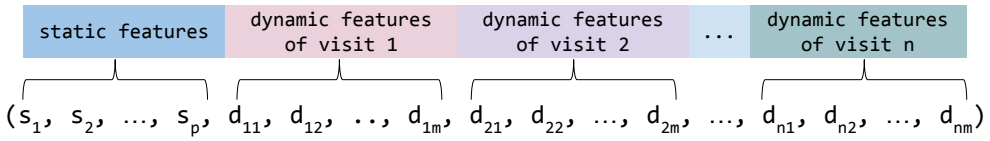
For convenience, let's introduce the sample construction with a practical example based on the current case of study. For the specific dataset used in this work, we considered as time span of interest the first 3 months of screening visits, for the reasons anticipated in 2.6.1. In this time interval, all the patients included in the dataset have between 1 and 4 visits: the algorithm adaptively builds k-NN samples whose length depends on the number of available visits for each subject to be imputed. Figure 2.4a illustrates the sample construction for subject i, with p being the number of static features, m the number of the dynamic ones, and n the number of his/her visits in the first three months of screening.

To identify the subjects in the pool of candidates having disease progression similar to subject i, the algorithm builds an analogous feature vector for each candidate neighbour with an available value in correspondence to the feature to be imputed. In more detail, each candidate neighbour j is temporally mapped over the current subject i, adaptively building a sample according to their matching time points. The feature vector of j is initialised with the subject's static features. Let $\mathbf{t}_i = (t_{i,1}, t_{i,2}, \ldots, t_{i,n})$ be the time points of the visits in the first three months of screening for subject i. For each visit time point $t_{i,l}$ of subject i, the closest-in-time visit of subject j within one month is selected. If no matching visit is found, candidate j is excluded from the k-NN search. Otherwise, the dynamic features of the matching visit are extracted and stacked to the feature vector of subject j; possible missing values in the matching visits of subject j are passed on his/her feature vector. Please notice that a candidate subject j may have repeated blocks of dynamic features in his/her feature vector corresponding to the same visit matching with multiple visits of subject i. Also notice that the feature vectors of the candidate subjects include the dynamic information of visits in the $[0, N]$ months time interval from the first visit (that is, in the case of the example, of the first four months of screening visits). Figure 2.4b schematically depicts the candidate sample construction procedure.

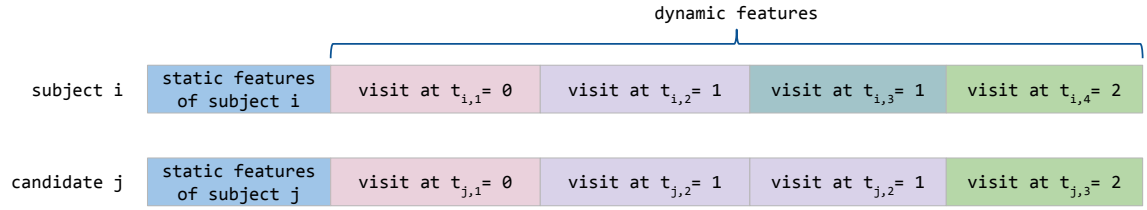
2.6.2.2 Weighted k-Nearest Neighbours Imputation

For a subject i with a missing value to be imputed, the wk-NN algorithm proceeds as follows. The features of the subject sample, together with his/her candidate samples, are normalised to the $[0, 1]$ interval in order to account for the difference among the ranges. Then, the distance between subject i and each candidate j is computed according to the following metric.

Let $\mathbf{v} = (v_1, v_2, \ldots, v_N)$ and $\mathbf{u} = (u_1, u_2, \ldots, u_N)$ be the feature vectors of, respectively, subject i and candidate j. Let $N_{\text{stat}}(\mathbf{v}, \mathbf{u})$ and $N_{\text{dyn}}(\mathbf{v}, \mathbf{u})$, be, respectively, the number of common



(a) *Sample construction for each patient with missing data to be imputed.*



(b) *Candidate sample construction procedure. In this example, subject i has $n = 4$ visits in the first three months of screening (one in the first month, two in the second and one in the third) while candidate j has 3 visits in this interval (one visit per month). Since the visit at $t_{j,2}$ matches both visits at $t_{i,2}$ and $t_{i,3}$, its dynamic feature values are repeated twice in the resulting feature vector (sample).*

Figure 2.4: *Adaptive sample construction for imputation.*

non-missing static and dynamic features in $\mathbf{v}$ and $\mathbf{u}$. Also, let S_{categ}, S_{ord}, S_{cont}, D_{categ}, D_{ord}, and D_{cont} be the sets of indices of, respectively, the static categorical, the static ordinal, the static continuous, the dynamic categorical, the dynamic ordinal, and the dynamic continuous features in $\mathbf{v}$ and $\mathbf{u}$. The distance between $\mathbf{v}$ and $\mathbf{u}$ is given by:

$$d(\mathbf{v},\mathbf{u}) = \frac{n \cdot \left(\sum_{l\in S_{\text{categ}}} I(v_l,u_l) + \sum_{l\in S_{\text{ord}}\cup S_{\text{cont}}} |v_l-u_l|\right)}{n\cdot N_{\text{stat}}(\mathbf{v},\mathbf{u}) + N_{\text{dyn}}(\mathbf{v},\mathbf{u})} + \frac{\sum_{l\in D_{\text{categ}}} I(v_l,u_l) + \sum_{l\in D_{\text{ord}}\cup D_{\text{cont}}} |v_l-u_l|}{n\cdot N_{\text{stat}}(\mathbf{v},\mathbf{u}) + N_{\text{dyn}}(\mathbf{v},\mathbf{u})}, \tag{2.7}$$

where n is the number of visits in the selected first time window of screening for subject i and $I(v_l,u_l)$ is 0 if $v_l = u_l$ and 1 otherwise. If either v_l or u_l, or both, are missing, the feature at index l does not contribute to the distance. The numerator is divided by the number of comparable features in u and v to normalise the distance on the number of common non-missing values. Because of the sample building procedure, each dynamic feature appears n times in the feature vectors: to re-balance the contribution of all the features to the similarity metric, both the distance between static features and the count $N_{\text{stat}}(\mathbf{v},\mathbf{u})$ are multiplied by n.

At this point, a filtering step is performed: candidates with a number of comparable features with subject i smaller than the 90% of the total number of non-missing features in sample i (both computed with the same adjustment for the static features) are dropped.

Once the distances to all the candidates have been computed, the k nearest ones are selected and their values in correspondence to the feature to be imputed are used for the imputation: for continuous and ordinal features, after removing possible outliers (values outside 1.5 times the interquartile range above the upper quartile and below the lower quartile), the missing feature

in i is imputed with the average of the selected values, each weighted by the inverse of the corresponding candidate distance; for categorical features, the missing feature in i is imputed with the mode of the selected values.

The procedure is repeated over all features with missing values in subject i. In our implementation, values previously imputed in i are not used for the subsequent imputations.

2.6.2.3 Weighted k-Nearest Neighbours Imputation with Mutual Information

As a further implementation (hereinafter referred to as wk-NN MI), the wk-NN algorithm was improved by including in the similarity metric the cross-information among the features given by the MI statistic (see Eq. 2.1 in Section 2.5.2.2). The MI among features is computed, for each subject with at least one missing value, over the pool of his/her candidate samples.

We decided in this implementation to use the MI instead of the MIC as the weight. Indeed, with respect to the previous section methodology, where the MIC was computed only once over all the training set, here the MI computation is required each time a subject needs to be imputed. Being the MIC obtained through an optimization algorithm, its use would have implied higher computational costs. Considered the already enhanced complexity of the current imputation methodology, also due to the adaptive sample construction, the MI was therefore selected.

In this implementation, we computed the MI using the *infotheo* R package v1.2.0 [144]. First, the continuous variables (X) are discretised into $i = \sqrt[3]{N}$ intervals of equal width $w = (\max(X) - \min(X))/i$, where N is the number of samples of X.

Let f be the index of the feature currently being imputed in subject i, and let $\mathbf{MI}_f = (\mathrm{MI}_{f,1}, \ldots, \mathrm{MI}_{f,f}, \ldots, \mathrm{MI}_{f,N})$ be the MI values between the feature at index f and all the features in the sample. The MI values are then employed as weights for the distance computation in the wk-NN algorithm:

$$d_f(\mathbf{v}, \mathbf{u}) = \frac{n \cdot \left(\sum_{l \in S_{\text{categ}}} \mathrm{MI}_{f,l} \cdot I(v_l, u_l) + \sum_{l \in S_{\text{ord}} \cup S_{\text{cont}}} \mathrm{MI}_{f,l} \cdot |v_l - u_l| \right)}{n \cdot N_{\text{stat}}(\mathbf{v}, \mathbf{u}) + N_{\text{dyn}}(\mathbf{v}, \mathbf{u})} + \frac{\sum_{l \in D_{\text{categ}}} \mathrm{MI}_{f,l} \cdot I(v_l, u_l) + \sum_{l \in D_{\text{ord}} \cup D_{\text{cont}}} \mathrm{MI}_{f,l} \cdot |v_l - u_l|}{n \cdot N_{\text{stat}}(\mathbf{v}, \mathbf{u}) + N_{\text{dyn}}(\mathbf{v}, \mathbf{u})} \,. \tag{2.8}$$

Please notice that here the distance among samples depends on the missing feature value currently being imputed, which means that the candidates chosen as nearest neighbours may change when imputing different features. An outline of the proposed imputation procedure is given in Fig. 2.5 and thoroughly described in Algorithm 1.

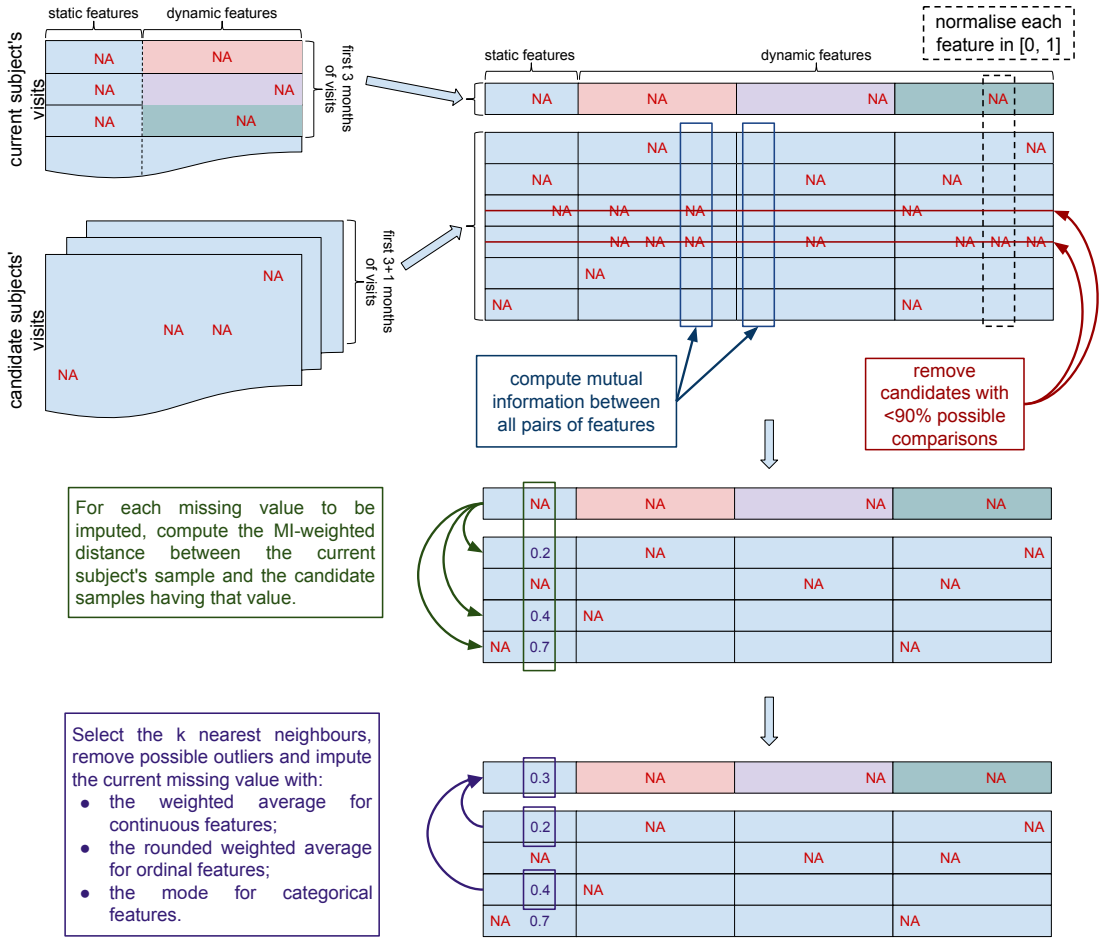


Figure 2.5: *Algorithm workflow of the wk-NN MI imputation method.*

Algorithm 1 wk-NN MI imputation algorithm.

N $\leftarrow$ set of subjects with missing values
w $\leftarrow 3$ ▷ time window (in months) for the visits to be imputed
k $\leftarrow 20$ ▷ number of nearest neighbours to select as candidates
for each subject *i* in **N do**
    select the visits of i in **w** for the sample construction procedure
    if i has at least one missing value in **w then**
        $n \leftarrow$ the number of visits of subject i in **w**
        v $\leftarrow$ k-NN sample for *i*
        $F \leftarrow$ features in **v** with missing values
        $N_v \leftarrow$ number of non-missing features in **v**
        J $\leftarrow$ **N**$\backslash\{i\}$ ▷ pool of candidate subjects for the imputation
        U $\leftarrow$ empty matrix of candidate samples
        for each subject *j* in **J do**
            select the visits of j in **w**$+1$ for the sample construction procedure
            U$[j,\ ] \leftarrow$ k-NN sample for *j*
        end for
        for each feature h of **v do**
            normalise **v**$[h]$ and **U**$[,\ h]$ in $[0, 1]$
        end for
        compute the MI of all pairs of features of **U**
        for f in F **do**
            for each candidate sample **u** in **U do**
                if u$[f]$ is `NA` **then**
                    continue
                end if
                $N_{\text{comparable}} \leftarrow$ number of non-missing features in both **v** and **u**
                if $N_{\text{comparable}} < 0.9 \cdot N_v$ **then**
                    continue
                end if
                compute the MI-weighted distance between **u** and **v**
            end for
            $K_f \leftarrow$ list of values of feature f of the **k** nearest neighbours of v
            if f is continuous **then**
                remove possible outliers from K_f
                $f_{\text{imputed}} \leftarrow$ inverse-distance-weighted average of K_f
            else if f is ordinal **then**
                remove possible outliers from K_f
                $f_{\text{imputed}} \leftarrow$ rounded inverse-distance-weighted average of K_f
            else if f is categorical **then**
                $f_{\text{imputed}} \leftarrow$ mode of K_f
            end if
        end for
    end if
end for

2.6.2.4 Imputation Evaluation Metrics

To evaluate the performance of the developed imputation methods, the normalised root-mean-square deviation (nRMSD) was employed for the continuous and ordinal features and the proportion of falsely-classified (PFC) for the categorical ones.[2]

Let f be the index of a feature imputed in T patient visits: $\mathbf{v}_f^{\text{imp}}$ is the vector of imputed values for that feature and $\mathbf{v}_f^{\text{true}}$ is the vector of true measured values. If f is the index of a continuous or ordinal feature, the corresponding nRMSD is calculated over the T patient visits as:

$$\text{nRMSD}_f = \frac{\sqrt{\frac{\sum_{i=1}^{T}\left(v_{i,f}^{\text{true}} - v_{i,f}^{\text{imp}}\right)^2}{T}}}{\max(\mathbf{v}_f^{\text{true}}) - \min(\mathbf{v}_f^{\text{true}})} \ . \tag{2.9}$$

Otherwise, if f is the index of a categorical feature, the corresponding PFC is calculated over the T patient visits as:

$$\text{PFC}_f = \frac{\sum_{i=1}^{T} I(v_{i,f}^{\text{true}}, v_{i,f}^{\text{imp}})}{T} \ , \tag{2.10}$$

where $I(v_{i,f}^{\text{true}}, v_{i,f}^{\text{imp}})$ equals 0 if $v_{i,f}^{\text{true}} = v_{i,f}^{\text{imp}}$, and 1 otherwise.

Similarly to Section 2.5.2.3, the normalised absolute error (nAE) of each imputed continuous or ordinal value was also computed to better assess the distribution of the error. The nAE for the imputed feature f of a given patient visit is given by:

$$\text{nAE}_f(i) = \frac{|v_{i,f}^{\text{true}} - v_{i,f}^{\text{imp}}|}{\max(\mathbf{v}_f^{\text{true}}) - \min(\mathbf{v}_f^{\text{true}})} \ . \tag{2.11}$$

In all cases, the closer these metrics are to zero the better the imputation.

2.6.2.5 Selection of the Optimal Number of Nearest Neighbours *k*

The proposed wk-NN and wk-NN MI imputation methods require the user to select an adequate k (number of nearest neighbours) hyperparameter. In general, this can be achieved by performing a Cross Validation scheme to test out different k values and select the best one, as described in Section 2.5.2.2. Such procedure also permits to assess the performance of the methods limiting possible overfitting issues.

Please notice that, differently from the k-NN algorithm of 2.5.2.2 where the CV setting only implied the recalculation of the MIC on a different subset, the inter-patients nature of the k-NN of the current implementation implies that, for each fold, the missing values of the subjects belonging to the internal testing set are imputed by employing a different pool of candidates, corresponding to the internal training set.

[2]The definitions of nRMSD and nAE given in this section slightly differ from those introduced in Equations 2.5 and 2.6, being the current imputation method defined over all the population and not intra-patient.

Specifically, here a Leave-One-Out Cross-Validation (LOOCV) was performed on the subjects of the training set, that is, one subject at a time was imputed using as candidates all the other subjects of the training set itself. Although being potentially computationally expensive, LOOCV reduces the likelihood that a split will result in sets that are not representative of the full data set [119]. For a given k value, for each non-missing feature of the patient to be imputed, all the measured values corresponding to that feature are first removed at the same time from his/her k-NN sample, and then imputed by using all the other subjects from the training set as candidates. By repeating this procedure over all the patients, an imputed value is obtained for each known measurement, and the imputation quality for the current value of k can be assessed by using the chosen performance metric. It is worth noticing that, by removing the values of only one feature at a time, the distribution and pattern of missing values in the dataset is generally preserved, which ensures the plausibility of the imputation performance results.

This procedure has been repeated for several values of k in order to determine the best performing one to be finally used to impute the whole dataset.

For the k parameter of both wk-NN algorithms, the values $\{1, 5, 10, 15, \ldots, 45, 50\}$ were tested: the best average error values were obtained with $k = 10$ for wk-NN, and with $k = 20$ for wk-NN MI. Figures 2.6 and 2.7 give the imputation error in terms of average nRMSD for the continuous and ordinal variables and in terms of average PFC for the categorical ones, for each k. The "average error" over all features (continuous green line) was computed by simply averaging the error measure obtained for each feature.

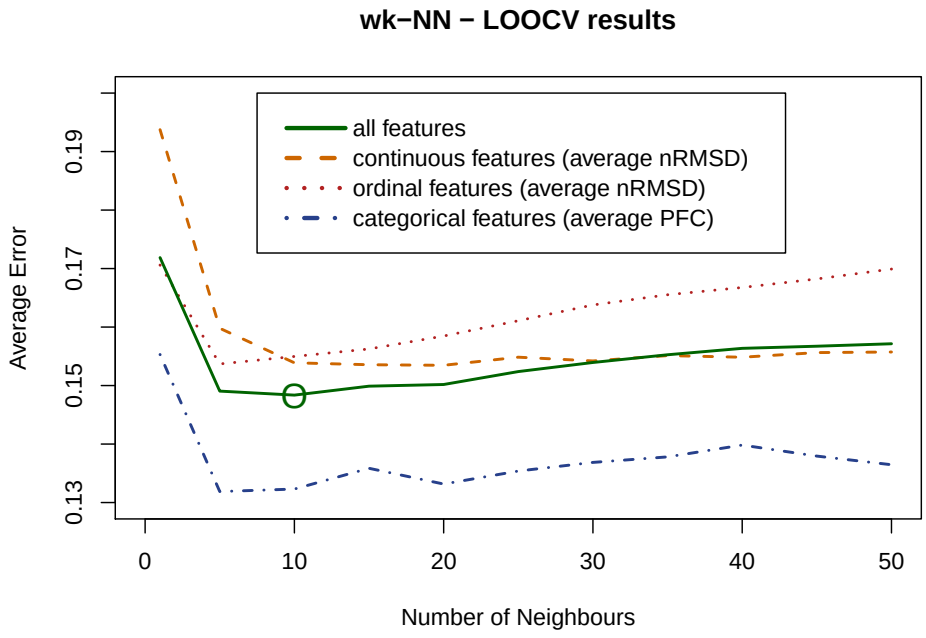


Figure 2.6: *Optimal number of neighbours to use for the imputation procedure with wk-NN. The best results are obtained for* $k = 10$.

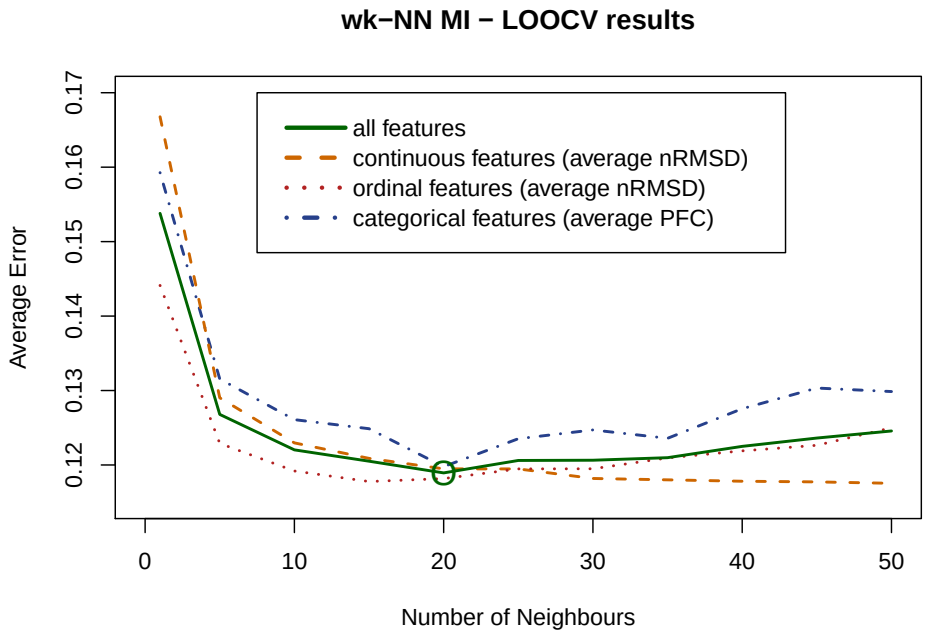


Figure 2.7: *Optimal number of neighbours to use for the imputation procedure with wk-NN MI. The best results are obtained for* $k = 20$.

2.6.2.6 Comparison with Other Imputation Methods

The proposed algorithm was compared with the three state-of-the-art imputation methods, namely Amelia II (*Amelia* R package v1.7.5), missForest (*missForest* R package v1.4) and MICE (*mice* R package v3.6.0). A random version of the proposed algorithm, k-random neighbours (k-RN), was also introduced to be used as a baseline for the imputation performance assessment: this implementation randomly samples a subset of k subjects from the pool of available candidates and uses them as neighbours. The optimal hyperparameter values for the employed competing imputation methods were set as follows.

The default settings were used for missForest and MICE, which correspond to the optimal ones. In detail, missForest uses a random forest trained on the observed values of a data matrix to predict the missing values by automatically managing continuous and/or categorical data including complex interactions and non-linear relation. MICE allows the specification of the imputation method to be used for each column in data and, by selecting the default setting, it sets the optimal associations: predictive mean matching for continuous features, logistic regression imputation for binary data (factors with two levels), polytomous regression imputation for unordered categorical data with more than two levels, and proportional odds model for ordinal features with more than two levels.

Amelia II requires the specification of the categorical and ordinal variables, the cross section variable (which was set to the patient ID variable) and, since it can handle time series data, the time variable (which was set to the visit time). Another important parameter is *polytime*, which

indicates what power of polynomial should be included in the imputation model to account for the effects of time, which accepts integer values from 0 to 3. After testing, the value of 1, which indicates linear time effects, was selected as the one yielding the best imputation results for the training data in a LOOCV setting.

For both MICE and Amelia II, which allow multiple imputation, the number of repetitions was set to 1, in order to perform a single-imputation in accordance with the k-NN-based proposed methods.

Finally, for the baseline k-RN, we used both $k = 10$ and $k = 20$, that correspond to the optimal k values assessed through the LOOCV for the wk-NN and the wk-NN MI methods, respectively.

2.6.3 Results: Imputation Performance Assessment

2.6.3.1 Performance Comparison on the Training Set

On the training set, the imputation performance was evaluated with the LOOCV setting described earlier: for each subject, all the measured values of his/her features were removed one feature at a time, and were then imputed using the competitor methods. The imputed values obtained by each method were compared to the true ones, and the average error was evaluated for each feature.

Tables 2.6, 2.7 and 2.8 show the average error (in terms of nRMSD or PFC) obtained on the training set for each continuous, ordinal and categorical feature, respectively. The proposed wk-NN MI imputation method outperforms the competitors on average and on the majority of the features. For the continuous features, the average nRMSD score obtained by wk-NN MI with the optimal $k = 20$ is 0.1195 against 0.1539 of wk-NN with the optimal $k = 10$, 0.1651 of Amelia II, 0.1572 of MICE, and 0.1784 of missForest. For the ordinal features, the average nRMSD score obtained by wk-NN MI is 0.1182 against 0.1550 of wk-NN, 0.1751 of Amelia II, 0.1521 of MICE, and 0.1728 of missForest. For the categorical features, the average PFC score obtained by wk-NN MI is 0.1198 against 0.1323 of wk-NN, 0.2589 of Amelia II, 0.1761 of MICE, and 0.1900 of missForest. In the three tables, we also report the performance for the k-RN baseline, computed for $k = 10$ and $k = 20$: the obtained performance outperforms the baseline.

To verify that the performance improvement was in fact statistically significant, the nAE distributions and PFC values obtained by wk-NN MI and MICE (the best performing among the competitor methods) on, respectively, the continuous/ordinal and categorical features, were analysed. Figure 2.8 shows the nAE distributions obtained on the training set for the continuous features. The plots show that wk-NN MI yields lower nAE values in all features. A two-tailed Wilcoxon signed-rank test [222] was also performed to assess the difference between the distributions: the obtained p-values are all smaller than 0.001, confirming that the difference is statistically significant. Here, this non-parametric test was employed to assess whether there was any statistically significant difference between the nAE distributions (which are very skewed and cannot be assumed to be normally distributed) obtained on continuous and ordinal data by different imputation methods.

Figure 2.9 shows the nAE distributions obtained on the training set for the ordinal features.

The plots show that wk-NN MI yields lower nAE values on 10 out of 12 features (ALSFRS-R scores 1 to 10). The two-tailed Wilcoxon signed-rank tests with Pratt's correction [158] (since the nAE values on the ALSFRS-R variables can only assume values in $\{0, 0.25, 0.5, 0.75, 1\}$, the signed-rank test has many "ties") was also performed to assess the difference between the distributions: the obtained p-values are smaller than 0.001 for the *ALSFRS-R* scores 1 to 10 which confirms that the difference is statistically significant for these features. Lastly, the tests showed that for *ALSFRS-R* 11 and 12 there was no statistically significant difference between wk-NN MI and MICE.

Figure 2.10 compares the PFC values obtained by wk-NN MI and MICE. The plots show that wk-NN MI outperforms MICE in all the categorical features, resulting in a significant difference in 6 out of 7 of them, namely in *sex*, *familiality*, *genetics*, *FTD*, *onset site*, and *NIV*, while showing no significant improvement for *PEG*. The McNemar's Chi-squared test [140], a statistical test used on paired categorical data, was also performed. This test is applied to 2×2 dichotomous contingency tables with paired samples, to determine whether there is "marginal homogeneity", that is, the row and column marginal frequencies are equal. When comparing two classifiers, each sample can be either be classified correctly or misclassified by each classifier, and thus a 2×2 dichotomous contingency table can be built. The null hypothesis of "marginal homogeneity" would mean there is no difference between the two classifiers. The imputation of categorical data can be seen as a classification task, and thus, McNemar's Chi-squared test can be used to determine if the difference between two imputation methods is statistically significant. The results of the McNemar's Chi-squared test applied to our training set confirms that the difference is statistically significant in the 6 features above.

Table 2.6: *nRMSD scores for the continuous features in the training set. The best performance is highlighted in bold.*

Features	Imputation methods						
	Amelia II	MICE	missForest	k-RN $k = 10$	wk-NN $k = 10$	k-RN $k = 20$	wk-NN MI $k = 20$
BMI premorbid	0.1012	0.0960	0.1323	0.1634	0.1286	0.1617	**0.0731**
BMI diagnosis	0.1560	0.1069	0.1476	0.1750	0.1457	0.1687	**0.0965**
FVC diagnosis	0.2466	0.2463	0.2534	0.1970	0.1876	0.1953	**0.1839**
age at onset	0.2355	0.2362	0.2393	0.1855	0.1748	0.1820	**0.1735**
diagnostic delay	0.1150	0.1218	0.1316	0.1484	0.1282	0.1495	**0.0850**
onset delta	0.1362	0.1362	0.1665	0.1848	0.1584	0.1778	**0.1049**
Average	0.1651	0.1572	0.1784	0.1757	0.1539	0.1725	**0.1195**

Table 2.7: *nRMSD scores for the ordinal features in the training set. The best performance is highlighted in bold.*

Features	Imputation methods						
	Amelia II	MICE	missForest	k-RN $k = 10$	wk-NN $k = 10$	k-RN $k = 20$	wk-NN MI $k = 20$
ALSFRS-R 1	0.1959	0.1540	0.1788	0.2454	0.1529	0.2390	**0.1249**
ALSFRS-R 2	0.1644	0.1433	0.1684	0.1904	0.1394	0.1907	**0.1218**
ALSFRS-R 3	0.1768	0.1387	0.1679	0.2175	0.1331	0.2130	**0.1133**
ALSFRS-R 4	0.2173	0.1916	0.2145	0.2516	0.1606	0.2455	**0.1472**
ALSFRS-R 5	0.2183	0.1863	0.2179	0.2812	0.1763	0.2727	**0.1394**
ALSFRS-R 6	0.2064	0.2015	0.2113	0.2864	0.1849	0.2773	**0.1513**
ALSFRS-R 7	0.1953	0.1696	0.1833	0.2645	0.1544	0.2550	**0.1295**
ALSFRS-R 8	0.2021	0.1488	0.1651	0.2460	0.1470	0.2377	**0.1138**
ALSFRS-R 9	0.2655	0.2405	0.2268	0.3744	0.2222	0.3657	**0.1589**
ALSFRS-R 10	0.1060	0.1093	0.1565	0.2523	0.1668	0.2475	**0.0943**
ALSFRS-R 11	0.0854	0.0982	0.1340	0.2446	0.1585	0.2403	**0.0847**
ALSFRS-R 12	0.0682	0.0434	0.0485	0.0933	0.0637	0.0908	**0.0391**
Average	0.1751	0.1521	0.1728	0.2457	0.1550	0.2396	**0.1182**

Table 2.8: *PFC scores for the categorical features in the training set. The best performance is highlighted in bold.*

Features	Imputation methods						
	Amelia II	MICE	missForest	k-RN $k = 10$	wk-NN $k = 10$	k-RN $k = 20$	wk-NN MI $k = 20$
sex	0.4859	0.4416	0.4463	0.5160	0.3974	0.4831	**0.3823**
familiality	0.1646	0.1268	0.1372	0.0842	0.0823	0.0842	**0.0738**
genetics	0.3310	0.1781	0.1751	0.0956	0.0895	0.0956	**0.0815**
FTD	0.3295	0.2642	0.3565	0.2060	0.2003	0.1960	**0.1903**
onset site	0.2957	0.1516	0.1403	0.3672	0.1017	0.3484	**0.0800**
NIV	0.1111	0.0556	0.0537	0.0518	0.0480	0.0518	**0.0235**
PEG	0.0948	0.0150	0.0208	**0.0069**	**0.0069**	**0.0069**	**0.0069**
Average	0.2589	0.1761	0.1900	0.1897	0.1323	0.1809	**0.1198**

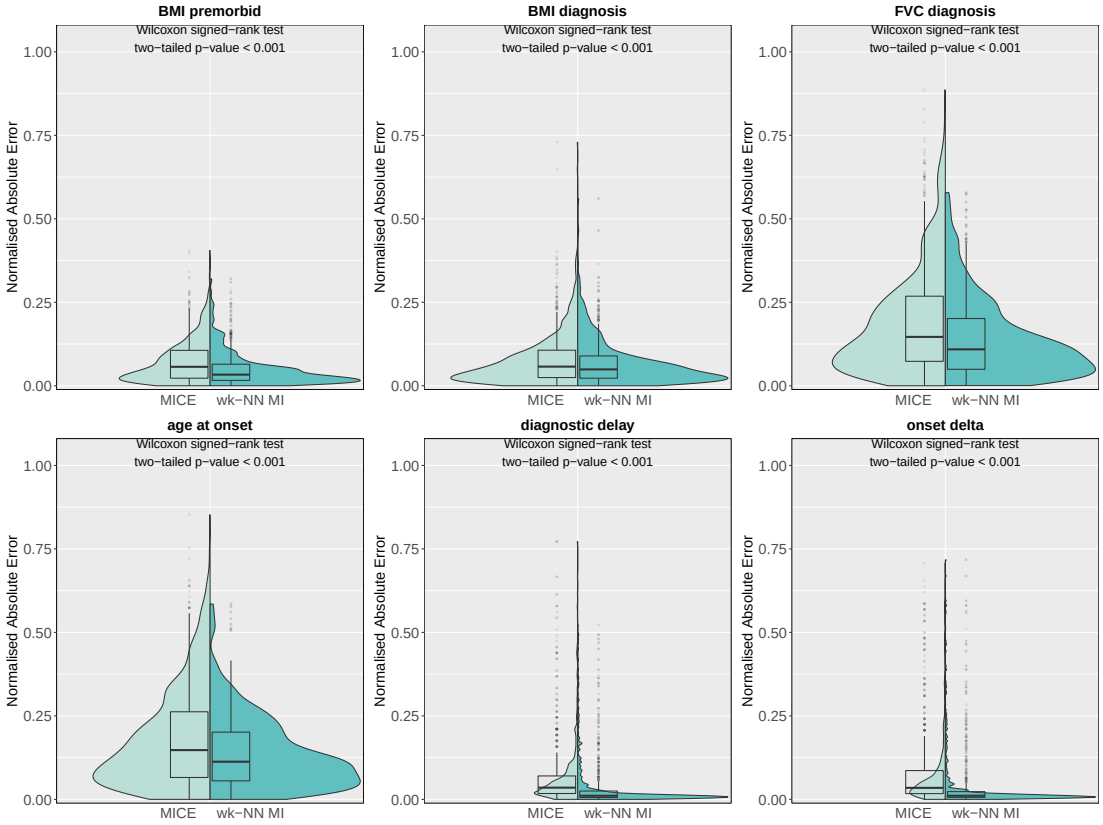


Figure 2.8: *Normalised absolute error distributions obtained with MICE and wk-NN MI (with* $k = 20$*) on the continuous features of the training set.*

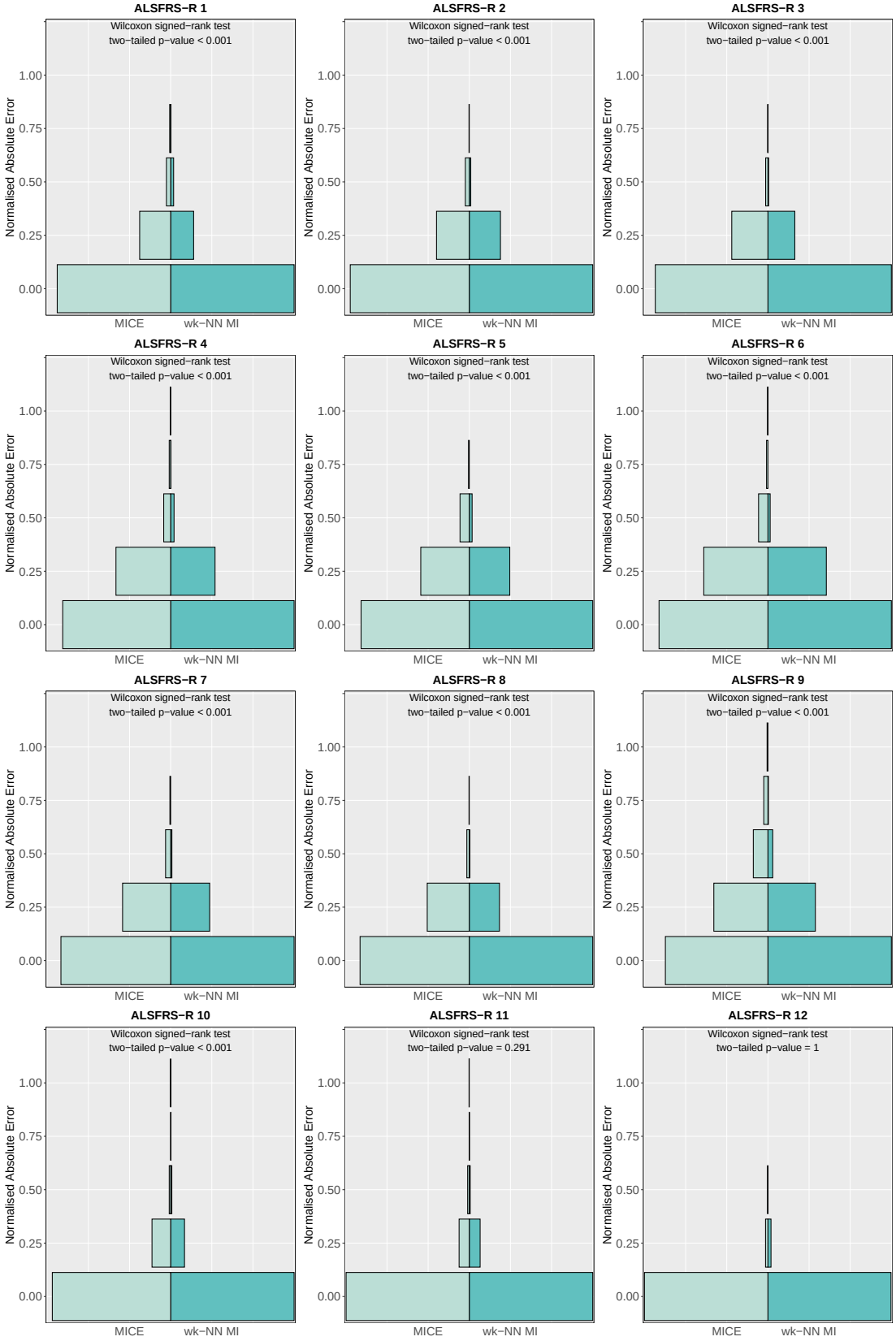


Figure 2.9: *Normalised absolute error distributions obtained with MICE and wk-NN MI (with* $k = 20$*) on the ordinal features of the training set.*

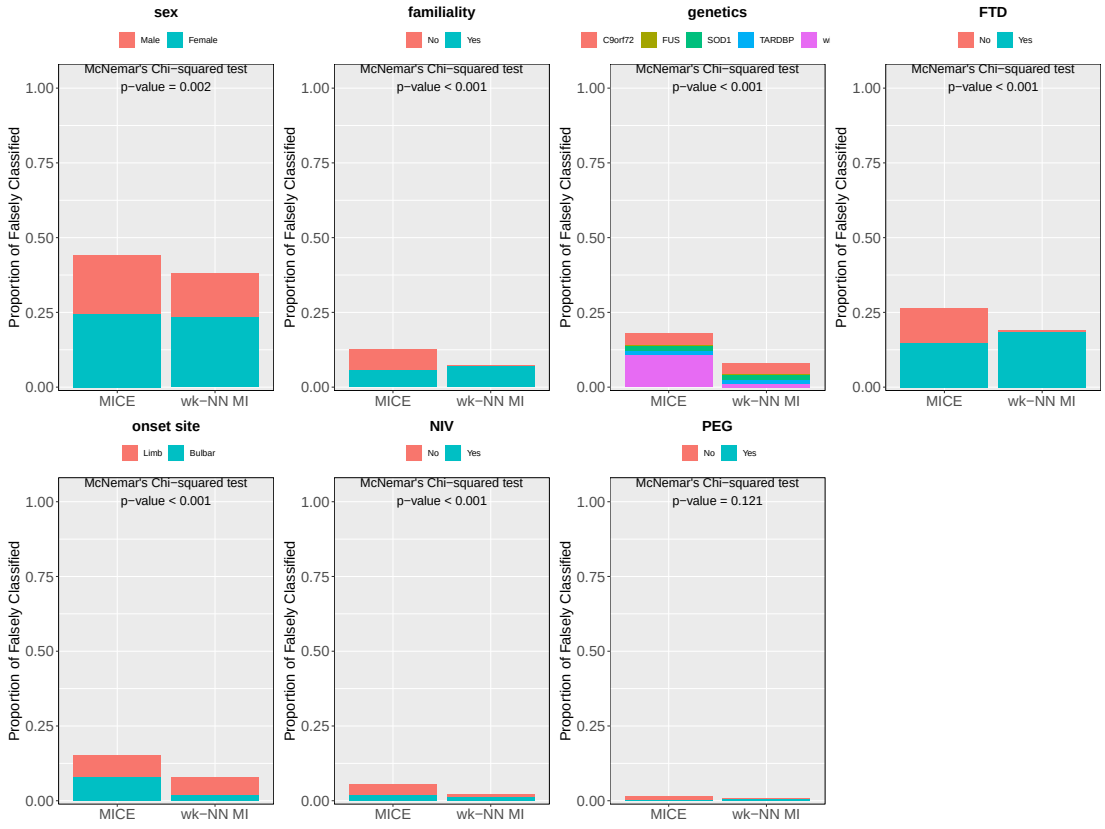


Figure 2.10: *Proportion of falsely classified obtained with MICE and wk-NN MI (with $k = 20$) on the categorical features of the training set.*

2.6.3.2 Performance Comparison on the Test Set

After selecting the methods' hyperparameters on the training set, the performance of the proposed imputation method was compared against those of the competitors on the test set. For each patient in the test set, all the known measurements were removed from his/her visits, one feature at a time, and the missing values were imputed by using all the training set subjects as candidates. The use of the whole training set represents the common situation where new subjects are continuously added to an existing dataset of clinical records and some of their values are natively missing, but there is a pool of previously collected data to use for the imputation. For Amelia II, MICE and missForest, the original format of the data was used for the imputation (as one would do in the original setting, that is, without the implemented adaptive sample construction that characterizes the proposed methodology): the records of the first three months of visits for the given patient in the test set have been bounded with all the information on the training set in a single data frame, which was then used as an input for these imputation algorithms. Finally, the imputed values obtained by each method were compared with the true ones.

The imputation results on the test set are shown in Tables 2.9, 2.10 and 2.11 for each continuous, ordinal and categorical feature, respectively. Results on the held-back test set confirm

that the proposed wk-NN MI imputation method outperforms the competitors on average and on the majority of the features. For the continuous features, the average nRMSD score obtained by wk-NN MI is 0.1332 against 0.1624 of wk-NN, 0.1803 of Amelia II, 0.1731 of MICE, and 0.2011 of missForest. For the ordinal features, the average nRMSD score obtained by wk-NN MI is 0.1274 against 0.1561 of wk-NN, 0.2654 of Amelia II, 0.1542 of MICE, and 0.1740 of missForest. For the categorical features, the average PFC score obtained by wk-NN MI is 0.1303 against 0.1456 of wk-NN, 0.2646 of Amelia II, 0.1900 of MICE, and 0.1966 of missForest. The baseline was also outperformed by the proposed wk-NN approaches.

The nAE distributions and PFC values obtained by wk-NN MI and MICE (the best performing among the competitor methods) on, respectively, the continuous/ordinal and categorical features were also analysed. Figure 2.11 shows the nAE distributions obtained on the test set for the continuous features. The plots and the two-tailed Wilcoxon signed-rank tests show that wk-NN MI yields statistically significant lower nAE values in 5 out of 6 features, namely *BMI premorbid*, *FVC diagnosis*, *age at onset*, *diagnostic delay*, and *onset delta*. The two methods did not obtain statistically significant differences in the imputation of *BMI diagnosis*.

Figure 2.12 shows the nAE distributions obtained on the test set for the ordinal features. The plots and the two-tailed Wilcoxon signed-rank tests with Pratt's correction show that wk-NN MI yields statistically significant lower nAE values on 9 out of 12 features (*ALSFRS-R* scores 1 to 5 and 8 to 11) at the 0.05 level. Lastly, the tests showed that for *ALSFRS-R* 6, 7 and 12 there was no statistically significant difference between wk-NN MI and MICE.

Figure 2.13 compares the PFC values obtained by wk-NN MI and MICE. The plots and the McNemar's Chi-squared tests show that wk-NN MI outperforms MICE in 4 out of 7 categorical features, namely in *sex*, *genetics*, *FTD*, and *onset site*, at the 0.05 statistical significance level. No statistically significant improvements are obtained for *familiality*, *NIV* and *PEG*.

Table 2.9: *nRMSD scores for the continuous features in the test set. The best performance is highlighted in bold.*

Features	Imputation methods						
	Amelia II	MICE	missForest	k-RN $k = 10$	wk-NN $k = 10$	k-RN $k = 20$	wk-NN MI $k = 20$
BMI premorbid	0.1302	0.1353	0.1787	0.2047	0.1692	0.2034	**0.1105**
BMI diagnosis	0.1459	0.1227	0.1653	0.1968	0.1665	0.2033	**0.1145**
FVC diagnosis	0.2481	0.2401	0.2584	0.2036	0.1821	0.1980	**0.1752**
age at onset	0.2799	0.2650	0.2781	0.2024	0.1847	0.2061	**0.1823**
diagnostic delay	0.1286	0.1228	0.1350	0.1481	0.1209	0.1422	**0.0958**
onset delta	0.1489	0.1529	0.1910	0.1785	0.1512	0.1686	**0.1210**
Average	0.1803	0.1731	0.2011	0.1890	0.1624	0.1869	**0.1332**

Table 2.10: *nRMSD scores for the ordinal features in the test set. The best performance is highlighted in bold.*

Features	Imputation methods						
	Amelia II	MICE	missForest	k-RN $k = 10$	wk-NN $k = 10$	k-RN $k = 20$	wk-NN MI $k = 20$
ALSFRS-R 1	0.3148	0.1852	0.1852	0.2467	0.1609	0.2528	**0.1457**
ALSFRS-R 2	0.2680	0.1852	0.2122	0.2197	0.1527	0.2049	**0.1424**
ALSFRS-R 3	0.2663	0.1673	0.1504	0.2443	0.1504	0.2265	**0.1416**
ALSFRS-R 4	0.2832	0.1913	0.1852	0.2770	0.1813	0.2762	**0.1602**
ALSFRS-R 5	0.3012	0.1741	0.2060	0.3039	0.1714	0.2873	**0.1496**
ALSFRS-R 6	0.3035	0.1768	0.1973	0.3141	0.1701	0.2996	**0.1631**
ALSFRS-R 7	0.2873	0.1687	0.1800	0.2762	0.1550	0.2787	**0.1416**
ALSFRS-R 8	0.2910	0.1550	0.1550	0.2645	0.1519	0.2514	**0.1153**
ALSFRS-R 9	0.3189	0.2192	0.2774	0.3709	0.2491	0.3549	**0.1800**
ALSFRS-R 10	0.1845	0.0903	0.1481	0.2410	0.1416	0.2462	**0.0648**
ALSFRS-R 11	0.1938	0.0941	0.1408	0.2316	0.1340	0.2415	**0.0716**
ALSFRS-R 12	0.1728	**0.0432**	0.0506	0.1013	0.0551	0.0990	0.0529
Average	0.2654	0.1542	0.1740	0.2576	0.1561	0.2516	**0.1274**

Table 2.11: *PFC scores for the categorical features in the test set. The best performance is highlighted in bold.*

Features	Imputation methods						
	Amelia II	MICE	missForest	k-RN $k = 10$	wk-NN $k = 10$	k-RN $k = 20$	wk-NN MI $k = 20$
sex	0.4440	0.4813	0.4366	0.5560	0.4366	0.4440	**0.3955**
familiality	0.2724	0.0970	0.1381	**0.0597**	**0.0597**	**0.0597**	0.0821
genetics	0.3166	0.2124	0.1776	0.1506	0.1506	0.1506	**0.1351**
FTD	0.4749	0.3575	0.3911	**0.2179**	0.2626	0.2235	0.2346
onset site	0.2910	0.1418	0.1343	0.4552	0.0896	0.4664	**0.0522**
NIV	0.0485	0.0299	0.0634	0.0410	0.0149	0.0410	**0.0075**
PEG	**0.0050**	0.0101	0.0352	**0.0050**	**0.0050**	**0.0050**	**0.0050**
Average	0.2646	0.1900	0.1966	0.2122	0.1456	0.1986	**0.1303**

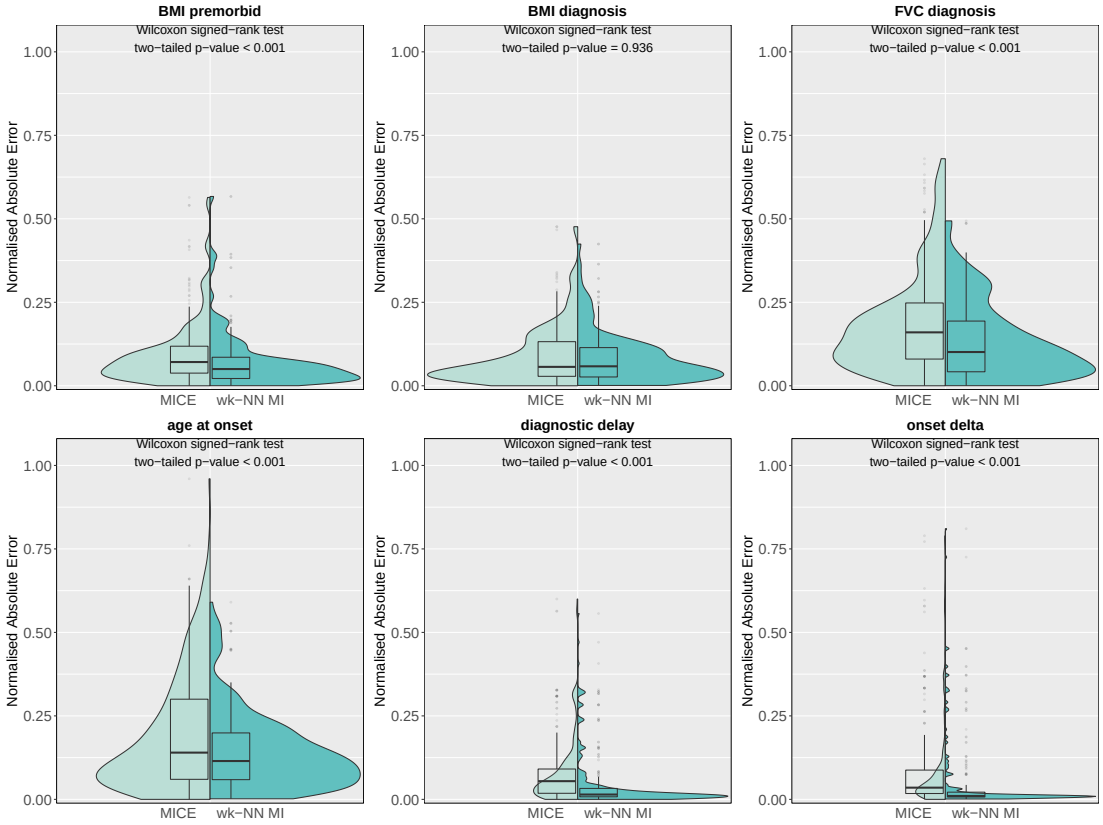


Figure 2.11: *Normalised absolute error distributions obtained with MICE and wk-NN MI (with* $k = 20$*) on the continuous features of the test set.*

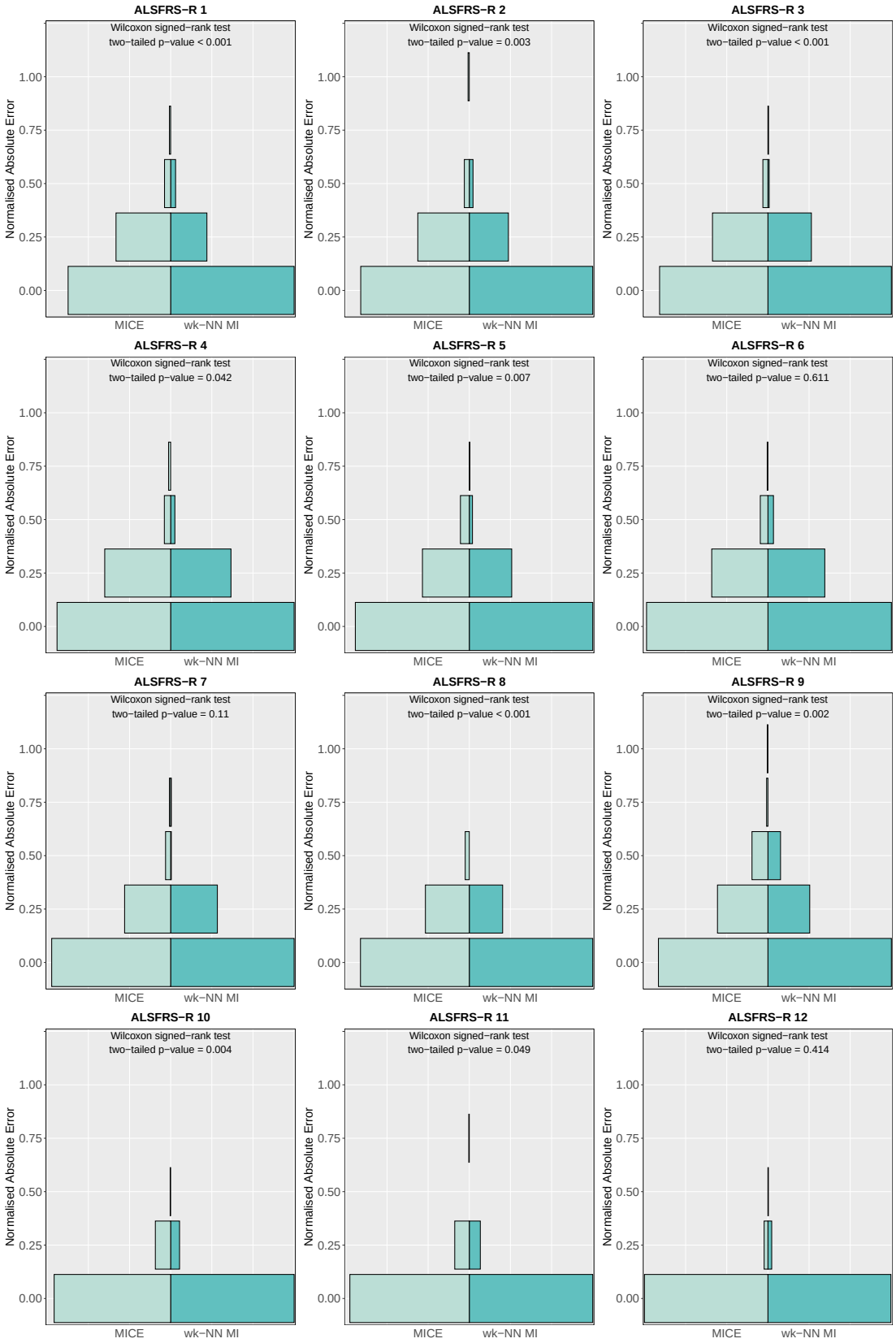


Figure 2.12: *Normalised absolute error distributions obtained with MICE and wk-NN MI (with $k = 20$) on the ordinal features of the test set.*

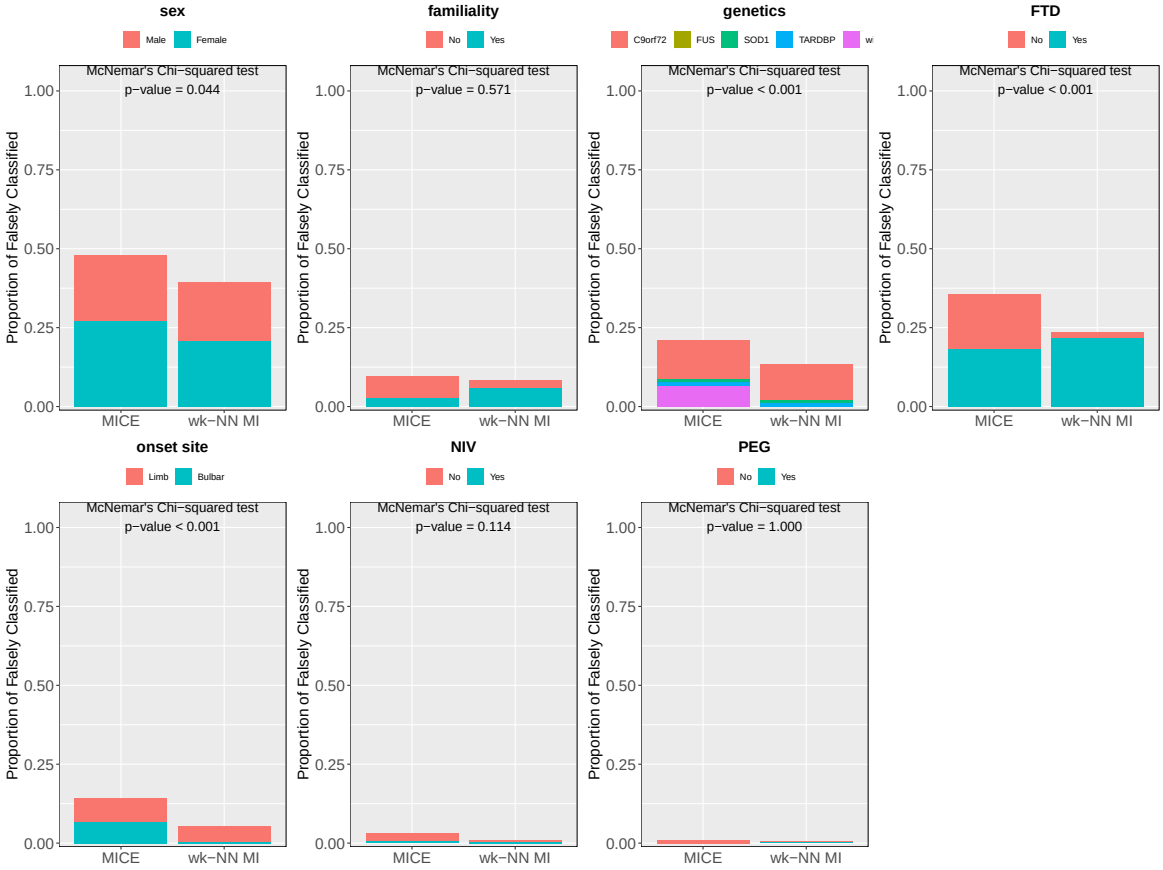


Figure 2.13: *Proportion of falsely classified obtained with MICE and wk-NN MI (with $k = 20$) on the categorical features of the test set.*

2.6.3.3 Computation Time Comparison

The computation times required for imputing the training dataset by using the proposed methods and the competitors were compared. The computations were carried out on a workstation with an Intel® Core™ i7-8700 CPU @ 4.60GHz with 16GB of DDR4 RAM running Linux Mint 19.2. Both wk-NN and wk-NN MI were parallelised by imputing each subject in parallel. The parallelism features were also enabled in Amelia II and missForest so that they could use all the available cores. MICE on the other hand cannot be run in parallel.

The imputation of the training set (560 subjects) was repeated 100 times for each method, and requires on average 60.97 seconds (with 3.19 seconds standard deviation) for wk-NN MI with $k = 20$, 52.97 seconds (with 1.66 seconds standard deviation) for wk-NN with $k = 10$, 0.07 seconds (with 0.04 seconds standard deviation) with Amelia II, 2.88 seconds (with 0.13 seconds standard deviation) with MICE, and 5.94 (with 2.76 seconds standard deviation) seconds with missForest. Even if slightly higher, the computational times of the proposed methodology remain within an acceptable range.

2.6.4 Further Method Validation through an Example of Use of the Imputed Dataset: Enhancing the Performance of a Survival Classification Task with Data Imputation

In order to further validate the developed imputation algorithm while, at the same time, evaluating the enhanced potential of the dataset imputed with the proposed method, a simple survival classification task was implemented.

Specifically, we tested the performance of a NB classifier trained in turn on the training set (i) with its original missing values, (ii) reduced to the only subjects without missing information (complete cases), (iii) reduced to the only features without missing data (complete variables), (iv) imputed with the proposed wk-NN MI algorithm, and (v) imputed with the best imputation competitor MICE.

The accurate prediction of the survival time in ALS patients is of paramount importance, and could aid prognostic counselling, stratification of cohorts for pharmacological trials, and timing of interventions. Nevertheless, this task is not uncomplicated. Patients with ALS exhibit a very high degree of variability in susceptibility, pathogenic mechanisms, and disease evolution. This is one of the main reasons for the negative results of therapeutic trials conducted so far, as statistical variance masks treatment effects [15, 177]. An optimal trial design requires samples size estimation, which, in turn, requires some understanding of the natural progression of the disease.

The PARALS register contains survival information for each patient, either in the form of date of death for the deceased ones or date of the last visit for the censored ones. For each subject, we determined the survival outcome as the binary answer to the question *"Does the subject survive for more than 3 years (36 months) from his/her first screening visit?"*. The patients that were censored before the 36 months threshold were discarded since we were unable to answer the question. The number of patients in the training set was thus reduced to 545 (from the initial 560), and the number of patients in the test set was reduced to 138 (from the initial 140). The 36 months threshold was selected because it ensures that the survival outcome is balanced in both training and test sets.

2.6.4.1 Survival sample construction

In order to develop a model for the survival classification task, there is the need to build, for each patient, a *survival sample*, that is, a feature vector derived from the original/imputed first three months of visits of each patients. It is worth reminding that, in general, this time interval can consist, for different patients, of a different number of acquisitions. Moreover, since the progression of the dynamic variables is expected to be the key point for determining the survival outcome, the survival sample must be able to encode the information on the disease progression that they embody. It is thus necessary, as done in the imputation sample construction, to adequately handle the temporal nature of the data.

As earlier mentioned, each time-varying – or, to maintain the terminology adopted above, dynamic – feature actually constitutes a time series. Feature extraction from time series is an issue common to all those fields where, in general, inputs are signals.

Being the collected information for some variables dynamic, the sample must be able to catch and describe the progression of the variable time series over time. This is achievable by deriving global features that incorporate some of the characteristics of the time series itself, such as the minimum/maximum value reached by the time series, the number of peaks, the slope and/or the intercept.

Here, for constituting the survival sample, we performed the following procedure, also depicted in Figure 2.14. For each dynamic feature in the selected time range, we computed three derived features, namely the minimum, maximum, and the slope. The slope was obtained by fitting a linear regression model on the temporal series constituted by the values of the feature collected over the three months interval. These values were then used together with the static features to construct a fixed-length vector (53 features in total) used as an input sample for our classification task. The survival samples constructed on the original data (that is, before imputation) carry over their missing values. When handling missing static features, the missing values were simply carried over to the constructed samples. In case of missing dynamic features, missing values are reported in the corresponding derived features that could not be computed due to data missingness.

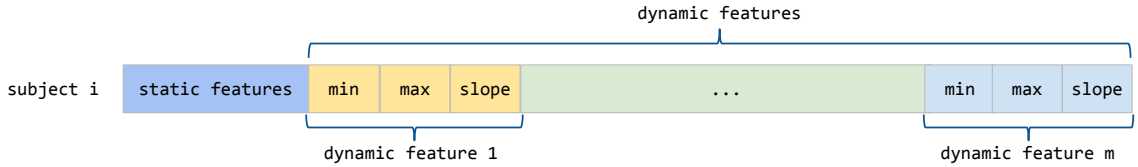


Figure 2.14: *Survival classification sample construction for each patient.*

2.6.4.2 The Naïve Bayes model

For this classification task we employed the naïve Bayes classifier [84] implemented in the *e1071* R package v1.7-2 [143].

NB is a simple learning algorithm that utilises Bayes' theorem in conjunction with the "naïve" assumption that, given the class label, every pair of features is conditionally independent. A NB classifier considers the contribution of each feature to the given class probability as independent, regardless of possible correlations. Although this assumption is often violated in practice, NB classifiers often achieve competitive classification results [229]. Because of their computational efficiency and many other desirable features, NB classifiers are widely used in practice.

More in detail, the naïve Bayes (NB) classifier is a conditional probability model that, for a problem instance $\mathbf{x} = (x_1, \ldots, x_n)$ to be classified, assigns class probabilities $P(y|x_1, \ldots, x_n)$ for each class label y. Bayes' theorem states that, given the class variable y and the feature vector $\mathbf{x}$:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)} \quad . \tag{2.12}$$

By using the independence assumption among features, this relationship can be rewritten as:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)} \ . \tag{2.13}$$

Since the value of $P(x_1, \ldots, x_n)$ is a constant which does not depend on the class label y, the class probability can be written as:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y) \ , \tag{2.14}$$

and the class label can be determined as:

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y) \ . \tag{2.15}$$

$P(y)$ is given by the relative frequency of class y in the training set. If x_i is discrete, $P(x_i \mid y)$ can also be determined by a frequentist approach; otherwise, if x_i is continuous, we assume that its values are sampled from a Gaussian distribution, and the probability is given by:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \ , \tag{2.16}$$

where $\mu_y = E[x_i \mid y]$ is the conditional expectation of x_i given y and $\sigma_y^2 = Var(x_i \mid y)$ is the conditional variance of x_i given y.

A NB classifier can both be trained as well as used to make predictions on a dataset with missing values. The training is carried out by implementing one of two possible strategies: by ignoring the samples with missing values in one or more features (complete case analysis), or by ignoring the missing values in the frequency counts of the probability computations. If we want to make predictions on a sample with missing values, the classification can be carried out by using only the available features. Let M be the set of the indices corresponding to the available features, the decision rule can then be re-written as:

$$\hat{y} = \arg\max_{y} P(y) \prod_{i \in M} P(x_i \mid y) \ . \tag{2.17}$$

2.6.4.3 Application to the case study

In order to evaluate the effect of the different imputation techniques on the classification task and so to further assess the performance of the proposed algorithm, we trained five NB models on five distinct sets of survival samples. First, starting from the original non-imputed training set composed of the first three months of patient visits, we built the corresponding training set of survival samples with their native missing values, from here on referred to as *original dataset*. From this first set we obtained two other sets for the complete case analysis: the *complete cases dataset* obtained by selecting only the survival samples without missing values, resulting in 252 survival samples, and the *complete features dataset* obtained by selecting only the features without missing values, resulting in 44 remaining features in the survival samples. Finally, we built

two other training sets of survival samples for the classification task by imputing the first three months of patient visits from the training set once with the proposed algorithm (wk-NN MI) and once with the best performing competitor.

The models were used to predict the set of test samples obtained from the non-imputed first three months of patient visits in the original test set.

2.6.4.4 Survival classification results

In this section we report the results of the survival classification procedure. Figure 2.15 gives the Precision-Recall (PR) and Receiver Operating Characteristic (ROC) plots of the NB classifiers trained on the five different sets of training samples. These plots were obtained by thresholding on the class label probabilities obtained by the NB classifiers for each survival sample. We also included the PR and ROC plots of a random predictor as a baseline. To ensure that the performance improvement is statistically significant, we computed the absolute classification error of the NB classifiers for each classification sample in the test set. The absolute classification error of each sample was computed as the absolute value of the difference between the class label and the predicted class probability. We performed two-tailed Wilcoxon signed-rank tests to assess the difference between the errors.

As a first result, we observe that the proposed method improves the prediction capabilities of a NB classifier: indeed, the PR curve achieves a perfect precision score of 1.0 for wider recall values. Moreover, the proposed method obtains the highest Area Under the Curve (AUC) value of 0.865. The improvement is somewhat less noticeable in terms of ROC curves and ROC-AUCs, although we can see that the proposed method improves the false positive rate which stays at zero for a wider true positive rate interval. The statistical test on the absolute classification error compared to all the other classifiers obtained p-values smaller than 0.001, confirming that the improvement is statistically significant.

Interestingly enough, the complete cases (PR-AUC $= 0.833$ and ROC-AUC $= 0.785$) and complete features analyses (PR-AUC $= 0.840$ and ROC-AUC $= 0.790$) worsen the prediction quality of the classifier with respect to the original dataset (PR-AUC $= 0.850$ and ROC-AUC $= 0.796$). The two-tailed Wilcoxon signed-rank tests' p-value when comparing the complete cases and complete features analyses with the original dataset are < 0.001 and 0.022, respectively, while there is no statistically significant difference between the complete cases and the complete features analyses (p-value$= 0.379$). The loss of information resulting from simply ignoring samples or entire columns with missing data hinders the precision of the classifier. On the other hand, the NB classifier can effectively learn from the survival samples with their native missing values, as reflected by the prediction results.

By comparing the predictions of the NB classifier trained on the original dataset (PR-AUC $= 0.850$ and ROC-AUC $= 0.796$) with the ones trained on the two imputed datasets, we can see how the imputation quality can affect the classification performance: the performance improves when the patient data are imputed with wk-NN MI (PR-AUC $= 0.865$ and ROC-AUC $= 0.816$), while it worsens when using the best competitor for the imputation (MICE), as can be seen from its PR and ROC curves which do not achieve a perfect precision of 1 or a perfect false positive rate of 0 for any interval of recall/true positive rate.

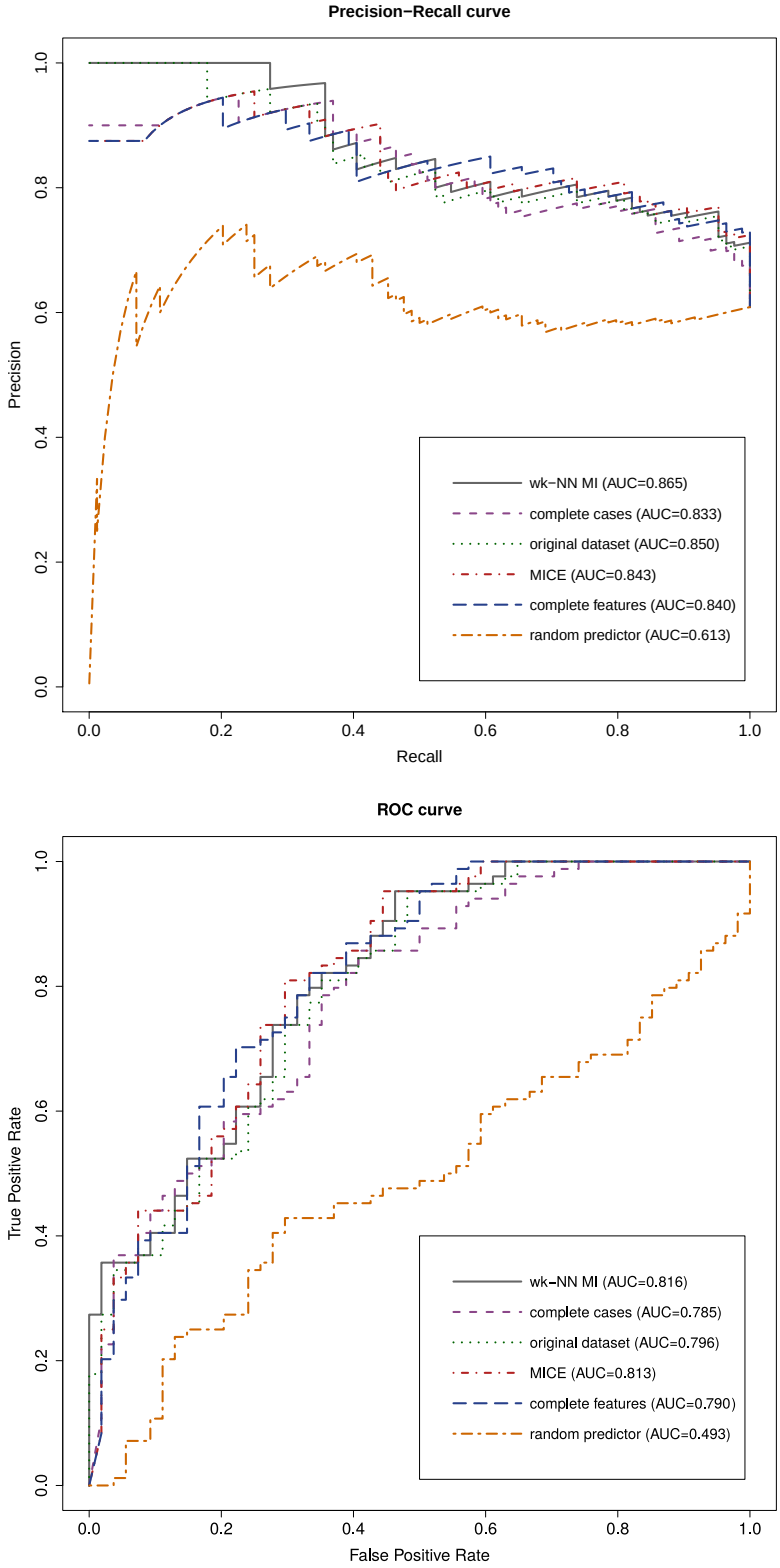


Figure 2.15: *Precision-Recall and ROC plots of the naïve Bayes classifiers. The plots show that the imputation of the training set with the proposed method improves the classification performance of a naïve Bayes classifier.*

2.6.5 Discussion: Applicability and Advantages of the wk-NN MI Imputation Algorithm

While many imputation methods require stringent assumptions on the nature of the missing data, a k-NN-based imputation only requires the presence of some relationship between the variable with the missing value and the other variables. Therefore, when the clinical characteristics of the dataset allow this hypothesis (like in the case-study clinical register), employing a full k-NN oriented approach may prove successful. Moreover, the k-NN guarantees that the imputed values are always in the dynamic range of the existing data and, when a small (with respect to the sample size of the candidates) k parameter is chosen, it allows to preserve the original distribution of the data (as for the k-NN algorithm presented in Section 2.5).

The proposed method employs the MI values between feature pairs as weights in the distance computation of the wk-NN procedure. The results show that wk-NN MI outperforms the wk-NN approach, confirming that the MI can be effectively used to exploit the cross-information of the features for the imputation task.

We showed that the proposed algorithm is able to handle two main characteristics of the clinical data: the heterogeneity of the variable types, that is effectively managed in the similarity metric, and the temporal nature of the longitudinal collection, that is exploited in the sample construction procedure to capture the progression of the patients, and managed in the similarity metric by different weighing the static and dynamic features. Furthermore, our method is also able to handle multiple missing values in the samples being compared, thus not requiring a dataset of complete cases to perform the imputation.

Finally, we provided a simple survival classification task as a further validation of the proposed methodology, as well as a potential application example. Our results show that the imputation of the missing values in the training dataset improves the predictions of a Naïve Bayes classifier. Since the NB represents a very simple classification technique, we believe that more complex and sophisticated analyses could also benefit from our imputation method.

Thanks to its ability to require only a limited number of acquisitions to assess the clinical progression (like the ones referred to the first 3 months of visits, in the case-study above), the methodology lends itself to be applied in real-world scenarios even when few collected time points are available. Moreover, by only using information from the training set to impute the subjects of the test set, it allows to employ the pool of already-available data to impute new subjects that may populate the register a few at a time. Furthermore, we envisage its application in those cases where novel clinical registers covering new clinical variables will become available, where missing values arising from the aggregation with older datasets could be imputed with the proposed approach.

For all these reasons, we believe that our method is potentially applicable in diverse contexts where imputation is needed. The final aim of this work is to provide a tool that can enhance the quality and the quantity of the data employed in analytics tasks, to improve and accelerate translational research. Concretely, the tool will allow clinicians to effectively use the information collected in a limited time interval by curing the possible presence of missing data.

It would also be interesting to extend the algorithm to the imputation of the whole patients' visits history, by modifying it from a fixed-window to a sliding-window approach. Moreover,

other distance metrics with other sophisticated weighting schemes could yield even better imputation results and could thus be conceived.

The proposed algorithm was implemented in the *wkNNMI* R package and is freely available from CRAN at `https://cran.r-project.org/package=wkNNMI`.

2.7 Final remarks

Imputation of missing data is a crucial – and often mandatory – step when working with real-world datasets. The algorithms proposed in this Chapter have been designed to effectively impute clinical datasets with different characteristics, that are exploited both in the working assumptions and in the algorithms' design.

In Section 2.6, a simple task of survival classification was also implemented: besides constituting an example of application of the imputed data, it shows how different choices on how to handle the missing information can have an impact on the classification performance.

The aim of the implemented imputation procedures has been making the best use of the available data, trying to catch their rich informative content even when hard to manage because of the structure or nature of the data.

The combination of linear interpolation with a weighted k-NN algorithm developed as an intra-patient approach provided good performance when applied to a dataset consisting of longitudinal acquisitions of continuous variables where no *a priori* assumptions of trend regularity or relations among features and subject are possible. On the other side, a fully weighted k-NN-based imputation can benefit from hypotheses of clinical similarity among patients, adaptively comparing and exploiting the dynamics of the clinical evolution.

The k-NN implementations of the two methodologies also profit from the cross-information provided by different features, through the use of the MIC and the MI values between pairs of features, respectively. While properly handling the different data types, such measures allow to estimate and integrate the cross-sectional information between features, enforcing the metric used to determine the similarity between samples.

In both the developed approaches, the temporal dimension of the data has been fully exploited: in fact, the employed k-NN and linear interpolation techniques are both based on the hypothesis (but we could also say fact) of an evolution of the clinical situation over time, that enriches the working framework and contributes to the inferable information on the missing values.

As future work, I plan to enhance the proposed methodologies, by refining the above-presented algorithms and by possibly adding new imputation strategies, thus expanding them into a full-fledged ensemble of imputation methods suitable to impute multiple types of clinical and laboratory data. Moreover, it would be very interesting to determine what thresholds of existing missing data and co-dependencies among features would begin to have an impact on the performance of the proposed approaches. I also plan to run these experiments on additional real-world datasets.

Chapter 3

Dynamic Model of Disease Progression

In longitudinal clinical data collections, the multidimensional characterization of the patients' history over consecutive time points constitutes an invaluable basis to study how the clinical condition evolves. By mining the available information through techniques able to deal with its dynamic nature, it is possible to analyze how features modify, interact with each other, and influence clinical evolution as time passes. Moreover, accurate predictive models can be trained, allowing prognosis forecasting at a patient- or population-level.

In this Chapter, a disease progression model of Amyotrophic Lateral Sclerosis based on Dynamic Bayesian Networks is proposed. By learning on the whole dynamics of the data sourced from two clinical registers, this methodology allows to model the disease evolution over time. Moreover, by overcoming the limitations of other "black-box" approaches, the employed technique provides a clear representation of the relationships among variables over time, unveiling their effect on the disease progression. As an outcome, the implemented model supplies the simulation of the progression of a patient or a cohort of patients starting from a given initial clinical characterization of their health status. The ability of the model to simulate the patients' prognosis also allows to perform stratification studies to determine the impact of specific factors on the disease evolution. Finally, specific target variables, such as survival or functional abilities can be computed in terms of probability of occurrence over time.

In general, computational models able to inspect and simulate the evolution of clinical conditions over time can constitute a useful tool to support physicians in planning individually tailored assistance programs or designing clinical trials, and patients in making informed decisions and more effectively managing their own health [141].

3.1 Case Study: Amyotrophic Lateral Sclerosis

As briefly introduced in Section 2.6.1, ALS is a neurodegenerative disorder characterised by the progressive degeneration of motor neurons in the brain and spinal cord [211]. This disease, also known as Lou Gehrig's disease, is progressive and fatal: the symptoms worsen over time and there are no known effective treatments that can effectively halt or reverse its progression which will inevitably result in respiratory failure.

The clinical presentation and the speed of progression of ALS are very heterogeneous. The disease is found to be more common in men than in women [137, 93], it can occur at any adult age and death in average comes 3–4 years from the initial disease onset [93]. However, about 10% of people with ALS survive for 10 or more years; among them, there are very few patients with exceptionally long survival times after diagnosis, such as the British astrophysicist Stephen Hawking (more than 50 years), arguably the most famous patient affected by ALS [159]. Possible symptom onset types include the "spinal onset", *i.e.* extremity muscle deficits, the "bulbar onset", *i.e.* dysarthria or dysphagia and, more rarely, generalized weakness and/or respiratory onset [156]. ALS is also correlated with frontotemporal dementia (FTD) [60] since up to 15% of ALS patients develop FTD and both diseases share common genetic causes [146], although the biology behind this correlation is unknown. The multi-faceted aetiology of the disease is reflected by the fact that only 5–10% of ALS cases are familial with the remaining vast majority being sporadic [92]. More than 30 different genetic conditions have been linked to ALS [166], with the most notable being a hexanucleotide repeat expansion at *C9orf72* which was identified as significantly associated with ALS in both familial and sporadic cases [102].

3.2 Previous Work on ALS disease progression modeling

The enormous social, medical and human costs imposed on ALS patients, their families and the health systems in general [142] are pushing the scientific community towards the development of computational tools to derive predictions for prognostic counselling, stratification of cohorts for pharmacological trials, and timing of interventions. Predicting the progression of ALS patients would improve prognostication and intervention timing in routine clinical practice. Moreover, clinical trials could be more effectively designed, for example by ensuring allocation of equivalent populations to the various intervention arms of a trial. An accurate, validated prediction algorithm could ultimately reduce or remove the need for human control subjects, replacing them with *in silico* simulated controls. Finally, a stratification of ALS patients by their progression or phenotype could give hints on different mechanisms acting in its pathogenesis.

Nevertheless, the inherent heterogeneity and varying progression make diagnosis of ALS and the development of a timely intervention plan tailored to each individual patient very challenging. This variability in disease susceptibility and pathogenic mechanisms is one of the main reasons for the negative results of therapeutic trials conducted so far as statistical variance masks treatment effects [15, 177]. An optimal trial design requires samples size estimation, which, in turn, requires some understanding of the natural progression of the disease. However, there is a paucity of biomarkers (predictors of disease progression) for stratifying patients into more homogeneous groups so that experimental therapies can be tested on patients sharing similar disease mechanisms [76]. Indeed, the potential identification of prognostic markers observable in patients at early stages of the disease progression could help in gaining more insight on its clinical course, in guiding personalised treatment for ALS patients targeting the specific biological pathways indicated by the biomarker, and enable analysis in clinical trials of homogeneous groups [5]. Furthermore, a reliable model of ALS progression could help to explain its manifold nature and predict clinical outcomes.

In order to enhance and accelerate translational ALS research, Prize4Life and the Neurological Clinical Research Institute (NCRI) at Massachusetts General Hospital created the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) platform (`https://nctu.partners.org/ProACT`) [9]. PRO-ACT represents the largest publicly available repository of merged ALS clinical trial data. It contains over 10 700 fully de-identified clinical ALS patient records from multiple completed clinical trials (both publicly- and privately-conducted) and constitutes an invaluable resource for accelerating discovery in the field of ALS.

So far, several computational models of ALS progression have been developed on this dataset to predict the future progression of the disease and to stratify the patients into meaningful subgroups.

In the DREAM-Phil Bowen ALS Prediction Prize4Life challenge [110], competitors developed algorithms for the prediction of disease progression using the PRO-ACT data. The challenge data was comprised of 1 822 patients from ALS clinical trials from the PRO-ACT data set, of which 918 were used to train the models while the rest was used for validation purposes by the challenge managers. As briefly mentioned in Section 2.6.1, by using only information regarding the first 3 months of ALS clinical trial information, participants were asked to predict the future progression of the disease in the subsequent 9 months in terms of slope of decline of the Revised ALS Functional Rating Scale (ALSFRS-R) [27, 28], the clinical questionnaire that measures observable functional status and change for patients with ALS over time (also employed in Section 2.6). As a challenge outcome, several potential biomarkers including uric acid, creatinine and blood pressure were identified, effectively expanding the list of previously reported markers such as time from onset, age, forced vital capacity (FVC), site of onset, sex and weight [110], thus shedding light on the pathobiology of ALS.

In the DREAM Prize4Life ALS Stratification Challenge [109], data from PRO-ACT were integrated with patient information contained in two real-world data sources, namely the Irish National ALS Register [153] and the regional PARALS register [36] (the one partially used in Section 2.6). After aggregation, data were divided into a training (986 patients) and a validation (493) set. Challenge participants were asked to design algorithms able to stratify the ALS patient population into distinct clusters and develop separate predictive models for each subpopulation, including both disease progression and survival as predicted outcome measures. A specific requirement of the task was to limit up to six the number of predictive features used in the models, since this can facilitate the concrete application of predictive algorithms in natural clinical settings [56]. This communal approach revealed a few sub-groups of patients which tended to cluster together across different algorithms, significantly differentiating four patient sub-populations: slow progressing patients and fast progressing patients, as well as patients with an average progression rate which were either early or late in their disease at the beginning of the recorded clinical observation period. The challenge results confirmed the ability of some already well-described features, such as age, gender and respiratory capacity, in stratifying ALS patients. Besides, it interestingly emerged how some features can be most predictive in specific phases of the disease (such as creatinine, that was found to be specifically predictive for patients early in their disease). Among the proposed methods, random forest and Gaussian process regression models with an arithmetic mean kernel performed successfully in most of the sub-tasks. Notably, the comparison of the proposed methodologies showed how data preprocessing can have a high

impact on the final performance: indeed, even if the same machine learning method was used by different teams, the one who obtained best performance effectively represented time-dependent features as a combination of simple summary statistics (for instance minimum, maximum and average of the dynamic feature values, similarly to the method described in Section 2.6.4).

Taylor et al. developed random forest, pre-slope, and generalised linear models (GLM) to predict disease progression [191] in terms of ALSFRS-R scores from the first baseline patient visit. They employed the data of 3 742 patients from the PRO-ACT database: 3 389 records were used for training, and 353 for internal validation. Furthermore, also in this case the clinical trial data were enriched with the records of 630 ALS patients from a clinical population, that were used as an external test dataset. The research showed that ALS predictive models for clinical populations could be developed using only baseline data as predictor variables, demonstrating applicability to a clinical patient care setting.

Another work on PRO-ACT was performed in 2017 by Ong *et al.* [152], who fitted time series data from 8 635 PRO-ACT subjects to exponential models and derived binary classes for total ALSFRS-R score decline (fast/slow decline) and survival (high/low death risk). The authors were able to predict classes of functional decline and survival across the 1–2 year time-frame available in PRO-ACT by using combinations of a small number of variables.

As it appears from the above studies, PRO-ACT represents an invaluable resource for research studies on ALS: its large sample size guarantees high statistical power; moreover, patients participating in clinical trials have more frequent visits, allowing for a better characterisation of disease progression. In addition, the plethora of variables collected within PRO-ACT is much wider than the ones that are commonly measured in a typical clinical setting. Nonetheless, the clinical trial population is not necessarily representative of the general ALS population: patients participating in clinical trials are generally higher functioning and more homogeneous compared to the ones from a typical tertiary care clinic setting [61]. Clinical trial patients tend to be younger, more likely to be male, and are half as likely to have bulbar onset disease, exhibiting therefore slower disease progression with less severe symptoms as compared to a typical clinic population. Furthermore, the duration of their follow-up is often limited [31]. For these reasons, patient data from the clinical context should be included in the development of ALS progression models in order to achieve reliable predictions for the general ALS population.

For these reasons, more recent models have been designed by including only real-world ALS patients. Among these, the model proposed by Van den Berg *et al.* [220] allows to predict the prognosis of individual ALS patients by using an the extensive databases of 11 475 patients from 14 European ALS centres. Eight patient characteristics were identified as prognostic factors: (1) bulbar versus non-bulbar onset, (2) age at onset, (3) definite versus probable or possible ALS, (4) diagnostic delay, (5) forced vital capacity, (6) progression rate, (7) FTD, and (8) the presence of a repeat expansion in the C9orf72 gene. This model can predict the time in months between symptom onset and the survival outcome (defined as non-invasive ventilation for more than 23 hours per day (>23h NIV), tracheostomy, or death).

3.3 Open Issues and Contribution

The majority of the above-described ALS prognostic models can be used to make survival or intervention endpoint predictions, generally in terms of a related risk score. Nevertheless, there is no available tool able to model the entire disease progression over time considering the evolution of all the dynamic variables.

Furthermore, previous models merely capture the associations among the clinical variables and the outcomes, without providing an explicit description of the interactions among variables and how these might change in time, thus not fully exploiting the informative richness of dynamic data. Besides, the lack of an explicit interpretation of the relationships among variables in terms of causes and effects also limits the ability of the models of accurately representing the domain knowledge.

As a further limitation, all the available ALS progression models rely on the ALSFRS-R staging system. Despite being the standard in the clinical practice, this staging system suffers from several limitations in capturing the patient's clinical evolution, such as the inability to effectively catch functional decline in the final and most severe phase of the disease [214, 221].

In addition, most of the described techniques are so-called "black-box" models that do not provide insight into how a certain prognostic prediction is reached, which would be instead a desirable property in clinical decision support systems [216, 19]. Finally, most of the currently available tools were developed on datasets including clinical trial data that are not fully representative of the general ALS population.

Based on these considerations, it would be beneficial to develop tools able to model the entire disease progression over time, considering all the dynamic variables and their relationships. Moreover, instead of only predicting a single survival endpoint it would be opportune to add further outcomes that allow to simultaneously assess the multifaceted evolution of the disease over time. Introducing the prediction of loss of independence in the main functional domains affected by the disease in the models could not only add a dynamic marker of progression, but also help overcoming the limitations of the ALSFRS-R staging system. In addition, the ability to dynamically predict the variations of the risk (probability) of functional impairments in time could aid unveiling the pathobiology of ALS as well as provide valuable insights in the treatment planning perspective. Finally, there is the need to develop models trained on real-wold cohort patients, so that they can be reliably applied in clinical care settings.

As a further step toward exploitation of the potential of artificial intelligence, a model with these capabilities could be used not only in the clinical practice to support clinicians in their decision, but also to generate *in silico* patients mimicking sub-populations of subjects with different characteristics. Medical doctors and researchers could benefit from models able to simulate the natural evolution of the disease in groups of untreated patients with, for instance, different baseline characteristics, in order to mimic disease progression in *in silico* placebo cohorts together with patient stratification.

In the next section I will outline my contribution in addressing these state-of-the-art limitations, consisting in the development of a model of ALS progression based on Dynamic Bayesian Networks (DBNs).

The built tool is able to predict and simulate, in a probabilistic fashion, the evolution of ALS

over time, providing an explicit representation of the temporal nature of the medical problem in terms of changes/loss of independence in the most relevant functional domains impaired by the disease, such as walking/self-care, swallowing, communicating and breathing, besides survival. In addition, the proposed methodology can be used to stratify ALS patients into subgroups of different progression and to assess the effect of different phenotypes at diagnosis on the entire disease course. Furthermore, the model allows an accurate representation of the domain knowledge and describes the dynamics of the ALS course also in terms of interactions among variables across subsequent points in time, unveiling their impact on disease progression.

The model has been developed by employing clinical and genetic data from four Italian real-world datasets, characterized by almost complete follow-up for most patients and high clinical heterogeneity, in order to constitute a tool able to widely represent the ALS population. Notably, in this model we also introduced a methodological novelty to account for the fact that variable dependencies might vary over time due to the long term evolution of the disease.

The model was also implemented as an interactive web application, with the aim to support clinicians with an easy-to-use tool that provides the prediction of a patient prognosis by employing only the data collected during a single visit. Such a tool could help physicians in easier care planning, as well as patients for decision making concerning their future. Another possible application of the model is in the clinical trial setting: thanks to its ability to simulate the disease progression of single patients/cohorts of subjects with specific clinical characteristics, we envisage its use by possibly replacing human control subjects with *in silico* simulated controls.

This work, performed in the context of the *CompALS* project, an Italian–Israeli collaboration, led to the filing of the International Patent *"Method for determining the prognosis of disease progression and survival for patients affected by Amyotrophic Lateral Sclerosis"* (currently pending) [50].

Due to patent-related requirements, the corresponding scientific publications are at the moment limited to conference presentations at the 6th European Academy of Neurology Congress [213], the 50th Congress of the Italian Society of Neurology [37], and the 30th International Symposium on ALS/MND [38].

3.4 A DBN-based Probabilistic Model of ALS Progression

This section describes the design and implementation of a DBN-based model of disease progression developed on a real-world cohort of ALS patients. The proposed methodology allows the accurate inclusion of domain knowledge and clinical predictors of various nature in the model, provides the probabilities of event occurrences over time, and describes the dynamics of the ALS course also in terms of interactions among clinical variables, unveiling their effect on the disease progression. Aside from individual patient prognosis, our model can be used to simulate the ALS disease progression at population level, thus enabling the *in silico* simulation of cohorts of subjects with given characteristics.

The developed ALS prognostic model has been embedded into a web-based tool that will be provided to medical doctors, enabling individual patient-level prognosis in terms of probabilistic evolution of the disease.

3.4.1 Material: Genetic and Dynamic Clinical Data

The cohort of ALS patients included in this study was recruited from two population-based registers (the PARALS Piemonte and Valle d'Aosta ALS register [36], and the ERRALS Emilia Romagna ALS register [129]) and two referral ALS centers, namely the Nemo Clinical Center and Salvatore Maugeri Foundation (Milan)[1]. ALS diagnosis was assessed according to El Escorial revised criteria [23], after excluding other diseases.

The aggregated dataset includes the information recorded over subsequent screening visits of 2 149 ALS patients for a total of 15 767 visits (median follow–up of 34 months, IQR 23–53; median number of visits equal to 5, IQR 3–9). It consists of 25 demographic, genetic, and clinical variables, corresponding to those used in Section 2.6.1.

In particular, for each patient, the following static variables were recorded during the first visit: sex, body-mass index (BMI) both premorbid and at diagnosis, forced vital capacity (FVC) at diagnosis, familiality of ALS, the result of a genetic screening over the most common ALS-associated genes (genes C9orf72, FUS, SOD1 and TARDBP were tested for mutations; if negative, patients were classified as wild type (WT)), presence of frontotemporal dementia (FTD, detected either clinically or through neuropsychological testing), site of disease onset (spinal/bulbar), age at onset, diagnostic delay (time from ALS onset to diagnosis). Moreover, at each visit the following dynamic features were collected: the presence/absence of non-invasive ventilation (NIV) and percutaneous endoscopic gastrostomy (PEG), the ALSFRS-R scores [27, 28], and the survival information (time from ALS onset to the last visit for the censored patients, or time from ALS onset to either death/tracheostomy[2] for the others).

3.4.1.1 Preprocessing

Conversion of the staging system from ALSFRS-R to MITOS

The collected ALSFRS-R staging system [27, 28] constitutes the standard mean to assess disease severity in clinical practice. As previously mentioned (see Section 2.6.1), it consists of a 12-item questionnaire rated on a 0–4 point scale evaluating the progression of disability in ALS patients in specific daily tasks. Being a *de facto* standard, it has been included in most of the available ALS progression models and extensively used to assess treatment efficacy in clinical trials and measure disease progression [178, 42, 16, 116].

Despite its extensive use, this staging system suffers from several limitations. The interpretation of an ALSFRS-R total raw score is hindered by possible different meanings given to the metric depending on the ALS form [128] and by its non-linear relationship with the linear Rasch transformed measures of global function [64, 63]. The scale exhibits multidimensionality, thus it should not be used as a global total score [64]. The ALSFRS-R scale also suffers from the "floor-effect" and is thus unable to capture late-stage clinical changes: patients approaching the bottom of the scale appear to be "slowing down" in their worsening because it becomes increasingly

[1]The study was approved by the ethical committees of the coordinating and participating centres. Informed consent to participate in the study was obtained from all the patients or their legal representatives. The databases were anonymised according to the Italian privacy protection legislation.

[2]Tracheostomy can be considered as an artificial life extension.

difficult for them to lose further raw score points [214, 221]. Finally, there is no agreed-upon threshold at which a change in ALSFRS-R score is viewed as an important transition point in functional status [32].

To overcome these limitations, other ALS staging systems have been introduced. King's staging system [169] is based on disease burden as measured by the involvement of clinical regions and the presence of respiratory or nutritional failure. This system uses five stages, from 1 to 5, with stages 1 to 4 indicating the number of involved clinical regions (stage 1 also indicates the disease onset) and stage 5 being death. Although the King's system is not based on ALSFRS-R scores, it can be estimated from them with 92% concordance [13]. More recently, the Milano-Torino staging (MITOS) system was introduced as a novel tool to measure ALS progression [32]. This system uses six stages, from 0 to 5, with stage 0 representing symptom onset and stage 5 being death, and is based on the assessment of four functional domains (walking/self-care, communication, swallowing and breathing) assayed by the ALSFRS-R. When a domain is not impaired, its MITOS value is equal to 0, whereas the MITOS value is equal to 1 for the domains in which the patient's independence is compromised: the MITOS score corresponds to the total number of functional domains in which the patient has lost independence. This scale was shown to be able to reliably identify relevant stages of disease in patients according to the number of lost functions, to be consistent with sequential disease progression, to overcome the non-linearity and multidimensionality limitations of ALSFRS-R, and to correlate well with the patients' quality of life and health service costs [193].

Based on the above considerations, in this work we decided to employ the MITOS staging system to monitor and assess the functional impairment of patients over time. We chose the MITOS since it tackles all the limitations of the ALSFRS-R scale while at the same time being completely derivable from the latter. King's system, on the other hand, cannot always be fully derived from the ALSFRS-R scale, which represents a limitation. Moreover, unlike King's staging system which summarises the clinical/anatomical spread of the disease, the MITOS system is aimed towards the distinction of functional capabilities during the spread of the disease and is able to differentiate late ALS stages at a higher resolution [58].

Therefore, as a first preprocessing step, we converted the ALSFRS-R scores into MITOS staging system scores encoded in the dynamic variables walking/self-care, breathing, swallowing, and communicating, according to the algorithm proposed by Chiò *et al.* [32].

Coding the temporal dimension of data

As already introduced, the dataset consists of multiple visits for each patient. In order to account for different time-grids and observation windows among subjects, we coded the temporal dimension by adding for each visit the following two temporal variables, derived from the dates: the time between each pair of consecutive visits (time between visits, TBV) and the time of each visit since the patient's disease onset (time since onset, TSO).

Split into training and test sets

We then split the dataset into a training set for developing the DBN, and a test set for validating the model, by stratifying the two sets over all variables. In particular, the dataset was split into a training set of 1 504 patients (11 032 visits), and a test set of 645 patients (4 735 visits).

A complete list of the longitudinally collected data measurements and a detailed overview of the full, training, and test datasets after the preprocessing phase is reported in Tables 3.1 and 3.2.

Table 3.1: *Contingency table for the categorical variables in the full dataset and in its training, test and reduced test sets.*

		Full dataset		Training set		Test set		Reduced test set	
Feature	**Level**	**N**	**(%)**	**N**	**(%)**	**N**	**(%)**	**N**	**(%)**
Subjects	–	2149	–	1504	–	645	–	202	–
Medical centre	Emilia-Romagna	744	(34.6)	516	(34.3)	228	(35.3)	35	(17.3)
	Maugeri Foundation	173	(8.1)	122	(8.1)	51	(7.9)	0	(0.0)
	Nemo Clinical Centre	267	(12.4)	192	(12.8)	75	(11.6)	8	(4.0)
	Piemonte and Valle d'Aosta	965	(44.9)	674	(44.8)	291	(45.1)	159	(78.7)
Sex	Female	976	(45.4)	696	(46.3)	280	(43.4)	83	(41.1)
	Male	1173	(54.6)	808	(53.7)	365	(56.6)	119	(58.9)
Onset site	Bulbar	658	(30.6)	459	(30.5)	199	(30.9)	69	(34.2)
	Spinal	1491	(69.4)	1045	(69.5)	446	(69.1)	133	(65.8)
Survival	Censored	596	(27.7)	427	(28.4)	169	(26.2)	36	(17.8)
	Tracheostomised/Dead	1553	(72.3)	1077	(71.6)	476	(73.8)	166	(82.2)
Familial	No	1951	(90.8)	1364	(90.7)	587	(91.0)	184	(91.1)
	Yes	135	(6.3)	96	(6.4)	39	(6.0)	18	(8.9)
	<NA>	63	(2.9)	44	(2.9)	19	(2.9)	0	(0.0)
Genetics	C9orf72	106	(4.9)	70	(4.7)	36	(5.6)	16	(7.9)
	FUS	9	(0.4)	2	(0.1)	7	(1.1)	4	(2.0)
	SOD1	39	(1.8)	29	(1.9)	10	(1.6)	3	(1.5)
	TARDBP	30	(1.4)	24	(1.6)	6	(0.9)	2	(1.0)
	WT	1429	(66.5)	1019	(67.8)	410	(63.6)	177	(87.6)
	<NA>	536	(24.9)	360	(23.9)	176	(27.3)	0	(0.0)
FTD	No	1564	(72.8)	1094	(72.7)	470	(72.9)	173	(85.6)
	Yes	141	(6.6)	97	(6.4)	44	(6.8)	29	(14.4)
	<NA>	444	(20.6)	313	(20.8)	131	(20.3)	0	(0.0)

Discretization of continuous variables

In this work we employed discrete-space/discrete-time DBNs, which encode probabilistic relationships among discrete variables over a discrete number of time steps. Therefore, in both training and test set we discretized the continuous variables age at onset, diagnostic delay, TBV, BMI premorbid, BMI at diagnosis, and FVC at diagnosis according to their distribution tertiles computed on the training set. Table 3.3 reports the quantization levels for these continuous variables and summarizes the adopted categories for the already-categorical ones. All the variables include the option to be equal to NA, if not recorded in the dataset.

Table 3.2: *Distribution of the continuous variables in the full dataset and in its training, test and reduced test sets. The "Time to ..." features indicate the time from the disease onset to the specified intervention or event. The equality of the distributions of the training–test sets and the training–reduced test sets has been assessed with the Kruskal-Wallis test for each variable.*

	Full dataset		Training set		Test set			Reduced test set		
Feature	**Mean**	**(SD)**	**Mean**	**(SD)**	**Mean**	**(SD)**	**p-value**	**Mean**	**(SD)**	**p-value**
Age at onset [years]	63.43	(11.20)	63.39	(11.15)	63.53	(11.34)	0.76	63.28	(10.88)	0.91
Diagnostic delay [months]	12.56	(12.03)	12.94	(12.65)	11.66	(10.39)	0.22	10.55	(7.86)	0.19
Time between visits [months]	3.39	(3.65)	3.39	(3.66)	3.40	(3.63)	0.25	2.95	(2.39)	0.11
Time since onset [months]	34.55	(31.95)	34.37	(31.03)	34.95	(34.01)	0.21	37.15	(39.31)	0.46
BMI premorbid [kg/m^2]	25.93	(4.01)	26.00	(4.07)	25.79	(3.87)	0.49	25.92	(3.84)	0.97
BMI at diagnosis [kg/m^2]	24.59	(4.07)	24.66	(4.10)	24.42	(4.00)	0.45	24.56	(3.99)	0.98
FVC at diagnosis [%]	87.95	(24.46)	88.45	(24.50)	86.76	(24.36)	0.23	88.31	(25.29)	0.87
Time to NIV [months]	32.01	(27.58)	32.30	(28.30)	31.29	(25.72)	0.68	28.27	(27.38)	0.19
Time to PEG [months]	30.52	(21.78)	30.67	(22.35)	30.19	(20.43)	0.98	30.58	(21.34)	0.91
Time to tracheostomy/death or censoring [months]	43.42	(33.61)	43.75	(33.66)	42.67	(33.52)	0.28	43.49	(35.89)	0.91
Time to MITOS walking/self care impairment [months]	29.65	(24.97)	30.27	(25.99)	28.27	(22.50)	0.15	26.09	(21.61)	0.06
Time to MITOS swallowing impairment [months]	28.93	(19.89)	28.71	(19.67)	29.45	(20.45)	0.53	29.16	(20.60)	0.69
Time to MITOS communication impairment [months]	33.52	(22.50)	33.11	(21.67)	34.52	(24.44)	0.49	33.15	(20.34)	0.82
Time to MITOS breathing impairment [months]	32.33	(27.46)	32.38	(27.59)	32.19	(27.15)	0.69	29.48	(27.44)	0.33

According to specific work hypotheses of the DBNs that will be discussed in the next section, the continuous variable TSO was not quantized according to its distribution percentiles, but developing instead an *ad hoc* temporal slicing procedure that takes into account the evolution of the disease over time (see Section 3.4.2.1).

3.4.2 Methods: Automatic Time Slicing Algorithm and Model Design

Bayesian Networks (BNs) [151] are descriptive models that encode the probabilistic relationships among variables. Given a multivariate dataset, the BNs build a directed acyclic graph (DAG) in which each variable corresponds to a node and the influence of one node (parent) on another (child) corresponds to a directed edge. DBNs [103, 148] are an extension of BNs that describe the dependencies among variables including the temporal dimension: in DBNs, edges between dynamic variables represent the influence of the parent variables at time step t on the child ones at time step $t + 1$.

A DBN is defined by its structure (set of parent-children dependencies) annotated with a set of conditional probability distributions (CPDs) that define the probabilistic dependency of each node on its parents. Similarly to BNs, nodes in a DBN are still connected through a DAG. However, DBNs allow encoding cycles and feedback between variables when considering their relationships over different time slices.

Table 3.3: *Variable quantization levels*

Feature	**Level**
Age at onset [years]	[21, 58]]58, 67]]67, 89]
Diagnostic delay [months]	[0, 7]]7, 12]]12, 147]
Time between visits (TBV) [months]	[0, 2]]2, 3]]3, 58]
BMI premorbid	[16, 23.9]]23.9, 27.2]]27.2, 44.1]
BMI at diagnosis	[13.8, 23.0]]23.0, 26.2]]26.2, 44.1]
FVC at diagnosis	[10, 84]]84, 102]]102, 162]
Medical centre	Emilia-Romagna Maugeri Foundation Nemo Clinical Centre Piemonte and Valle d'Aosta
Sex	Female Male
Site of onset	Bulbar Spinal
Vital status	Alive Tracheostomised/Dead
Familial	Yes No
C9orf72	Yes No
FUS	Yes No
SOD1	Yes No
TARDBP	Yes No
WT	Yes No
FTD	Yes No
MITOS movement impairment	Yes No
MITOS swallowing impairment	Yes No
MITOS communication impairment	Yes No
MITOS breathing impairment	Yes No

The employed algorithm for DBN structure-learning relies on the following assumptions: (i) two nodes cannot be a deterministic function of a single variable, (ii) variables are related to each other over a discrete number of time steps, also called slices (and thus we say that, for example, variable *A* at time t influences variable *B* at time $t + 1$), and (iii) the CPDs are time invariant. Specifically, the CPDs are usually estimated through techniques like Bayesian estimation or (regularised) maximum likelihood.

DBNs are well suited for describing the evolution of diseases, since they provide an explicit representation of the variable set and their inter-dependencies over time, as well as the means to learn not only from statistical data, but also from domain literature and expert knowledge. For these reasons, DBNs were previously employed for instance in intensive care unit settings [2], for atherosclerosis progression modeling [57], and for the simulation of clinical complications in type 1 diabetes [134]. In the context of ALS, DBNs were already explored in a previous work of the group [227], where a model of disease progression was built on the PRO-ACT dataset. As an enrichment, in this work we switched to fully real-world datasets. Moreover, we added a methodological novelty to account for the fact that variable dependencies might vary over time, according to the disease nature.

3.4.2.1 Time Slicing Algorithm for TSO Discretization

In the learning process, the DBN model infers a set of CPDs for each variable; thus, DBNs are able to identify the combination of factors modulating ALS severity over its course.

As mentioned above (hypothesis (iii) on the time-invariance of the CPDs), this is done under the assumption that such probabilistic relationships among variables do not change over time.

In the reality of clinical data this working hypothesis is not always verified. The disease proceeds by following different possible patterns in its different stages (for instance, it could begin to manifest itself slowly in its early stages, and then speed up while progressing). In this process, some variables can have a different impact depending on the disease phase.

The evolving progression of the disease can be verified by analyzing the outcome occurrence rates over time. Figure 3.1 reports the frequency of the outcome events, defined as the MITOS impairments (the already mentioned walking/self-care, breathing, swallowing, and communicating), and the survival (tracheostomy/death), computed over the training set. The occurrence frequency was computed by employing the Kaplan-Meier estimator as a function of the TSO variable. For the sake of readability, curves in the graph are truncated at 105 months.

The varying slope of each Kaplan-Meier curve details the different rates of progression over the population, with the curves' inflection points corresponding to the instants where disease severity changes. In correspondence of these points, it stands to reason that the relations among variables are changing too. In other words, the CPDs before and after each inflection point might be different.

Therefore, in order to respect the CPDs time-invariance hypothesis of the DBNs, the following procedure needs to be set up: first, an adequate number of "global disease inflection points" should be determined based on the rates of all the outcomes curves; then, the TSO variable that codes the observation grid should be discretized accordingly, each discretization level representing a time slice where the CPD time-invariance assumption is complied.

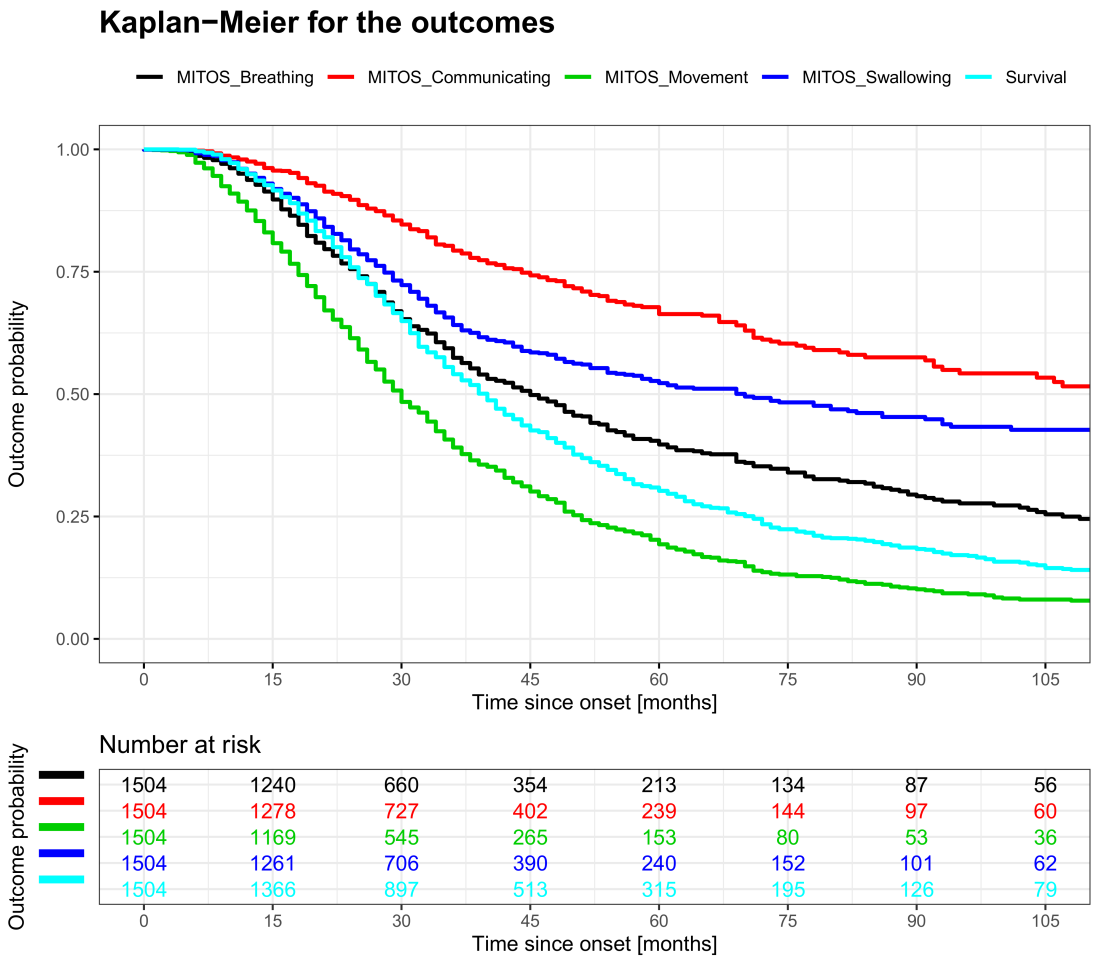


Figure 3.1: *Kaplan-Meier curves of the ALS outcomes (MITOS impairments and survival) computed on the training set and truncated at 105 months.*

When determining the global inflection points constituting the slicing thresholds, two further structural constraints should be imposed.

First, the global inflection points should cut the disease observation interval in such a way that the number of samples falling into each slice is balanced. Indeed, in order to be able to learn a possible relationship among variables, in the training phase the network requires to be fed with an adequate number of occurrences for each level. In this respect, theoretically, the most genuine learning setting would be a dataset consisting in an equal number of samples for each possible combination of the variables (or, at least, for each possible combination of the variables with a relationship). In practice, this condition can of course be only asymptotically met, by both employing a sufficiently high number of samples (here visits) in the training set, and trying to balance the number of occurrences at each variable discretization level. With respect to this second point, the quantiles-based discretization of the other continuous variables in Section 3.4.1.1 ensures that the levels are balanced. In contrast, in the TSO discretization procedure this requirement should be explicitly imposed.

Second, the number of searched global inflection points should be limited. This constraint is related to the ability of the network to learn reliable relationships too: in the CPDs computation, each TSO discretization level is associated with a different set of conditional probabilities. The higher the number of discretization levels for each variable, the lower the cardinality of samples with that combination of values, causing both less reliable computed probabilities and a higher risk to miss interesting relationships. While the choice of employing tertiles for the discretizations of continuous variables in Section 3.4.1.1 ensures by design a low number of levels, for TSO the above requirement should be taken into account when determining the number of global inflection points.

Based on these considerations, the designed algorithm for automatic time slicing proceeds as follows. Given a dataset consisting of a number of samples *N*, a user-defined number of slicing thresholds to identify, named *n_thresh*, and the outcome curves $y_1 ... y_M$, the algorithm starts imposing the requirement of sample balance at each discretization level.

First, the punctual values of TSO, namely $t_1 ... t_{n\_thresh}$, that would exactly split the dataset in (*n_thresh* $+ 1$) quantiles are identified, with each t_i corresponding to the $\frac{100i}{n.thresh+1}$ percentile. For instance, if *n_thresh* was set equal to 3, the algorithm would define t_1 as the 25th, t_2 as the 50th, and t_3 as the 75th TSO percentile.

If these t_i values actually corresponded to the global inflection points and were therefore used as discretization thresholds, they would effectively provide a perfectly balanced number of occurrences at each level. As this is rarely the case, the user can set a confidence interval around each t_i in order to determine a range of sub-optimal (in terms of balancing) TSO values to be inspected for the inflection points search.

Named C this confidence term, expressed as a percentage between 0% and 50%, each global inflection point t_i^* is searched in the percentile range:

$$\left[\frac{100i - C}{n.thresh + 1}, \frac{100i + C}{n.thresh + 1}\right]. \tag{3.1}$$

In the case of the example, if C was set equal to 20%, the 1st inflection point would be inspected in the range of TSO between the 20th and 30th percentiles. Please notice that, being C upper limited to 50%, an overlap between consecutive inspected spans is avoided.

Then, the algorithm considers one search interval at a time starting from the leftmost ($i = 1$). Within each interval, for each outcome curve y_j, the optimal point t_{ij}^* that maximizes the slope difference between the linear models built on the left and right segments of the curve y_j is identified, where the left segment spans the range between t_{i-1}^* (corresponding to the previous global inflection point) and t_{ij}^*, and the right segment spans the range between t_{ij}^* and t_{i+1}[3].

Finally, the global inflection point t_i^* is simply determined as the average value across all the t_{ij}^*'s.

As described above, the algorithm proceeds by first imposing the balancing among the number of discretized samples at each level, with a user-defined confidence interval. Then, it determines the curves' inflection points in these identified time spans. In order to be sure that the

[3] In the first and last iterations, the left segment starts at the first time value of the y_j curve, and the right segment ends at the last time value of the y_j curve, respectively.

global inflection points are effectively caught, it is therefore necessary to accurately explore all the observation time grid, by setting an adequate number of $n.thresh$ and a sufficiently high C. In this process, it must also be kept in mind that the final number of discretization levels should be limited in order to restrict the total number of possible combinations of variables, and thus the set value of $n.thresh$ can not be too high.

To accomplish these constraints while identifying the optimal TSO threshold values on our data, we set up the following iterative procedure. In turn, we set a (relatively low, with respect to the data cardinality) number of $n.thresh$ ranging from 0 to 5. Then, we used the designed time slicing algorithm to identify the TSO thresholds. After discretizing the TSO accordingly, a DBN was trained and its performance was assessed. The optimal $n.thresh$ was finally determined as the one resulting in the best performing DBN. In this procedure, testing in turn different values of $n.thresh$ not only allows to determine the optimal one for the current dataset, but also spans different intervals of the time grid, thus widely looking for the global inflection points over the whole observation range.

Section 3.4.3.1 details how this procedure was carried out in a CV scheme and reports the optimal thresholds identified and employed for discretizing the TSO.

3.4.2.2 Model Development

To build the DBN we employed *bnstruct* [65], an R package that performs structure and parameter learning on discrete/categorical data even in the presence of missing values, which is the case of our data and a common situation in the clinical context.

The DBN model was developed on the training set through a two-step iterative procedure: 1) by inferring the graph topology and 2) learning the parameters of each CPD, *i.e.*, the probability that a variable assumes a specific value conditional to each possible joint assignment of values to its parents.

To infer the DBN structure we used the Max-Min Hill-Climbing (MMHC) algorithm [195], a greedy search-and-score method that starts with an initial graph (empty graph in our case) and searches the complete space of possible graph structures, by adding, reversing or deleting edges. The MMHC runs until a specific score is maximised or a specific number of iterations has been reached. Here, the Bayesian Information Criterion (BIC) scoring was chosen. Thus, the structure-learning phase provided the DBN topology with the highest probability of generating the training data. Subsequently, parameters of CPDs were computed through a maximum *a posteriori* (MAP) estimation for each node. In summary, MMHC detects the dependencies among variables, whereas MAP weights the influence of each variable on the others.

Domain knowledge integration

The employed *bnstruct* tool also allows the encoding of some domain knowledge in the network structure, by applying constraints to the network topology; in this way, clinically or biologically non-sense relations among variables can be forbidden, as well as clinical well-known dependencies can be imposed as mandatory edges. For instance, the dependency of medical centre on patient sex was denied, while the dependencies of the MITOS variables on the time elapsed since

diagnosis, encoded in the TSO variable, was imposed. As another example, the dependency of the BMI premorbid on the TBV was forbidden, since BMI was recorded before the disease (and thus the visits) began. Or, analogously, the dependency of Diagnostic Delay on any variable recorded after the diagnosis was forbidden.

More specifically, when learning the structure of the network from the training set, the following information can be provided:

1. separate (disjoint) grouping of the variables, named *layers*: by default, given a user-defined layer structure, variables in a given layer j can depend only on variables from layers $i \leq j$;
2. specific rules that allow or deny specific dependencies between layers, thus overwriting or integrating the default relationship among layers mentioned above;
3. some mandatory edges, *i.e.*, one of more edges between variables that must be present in the network even if not automatically detected as dependencies.

Here, we defined the following rules. First, the variables were divided into the following layers, thus defining the default possible edges inspected in the learning phase:

- Layer 1: Sex, Genetics (WT, TARDBP, C9orf72, SOD1, FUS), BMI premorbid
- Layer 2: Familiality
- Layer 3: Medical centre
- Layer 4: Age onset, FTD, Onset site, FVC diagnosis, BMI diagnosis
- Layer 5: Diagnostic delay
- Layer 6: MITOS, NIV, PEG variables at time (t)
- Layer 7: TBV
- Layer 8: MITOS, NIV, PEG variables at time (t+1)
- Layer 9: Survival
- Layer 10: TSO

Then, we imposed the following rules between layers:

- Layer 1 can not depend on itself or any other layer.
- Layer 2 can only depend on layer 1.
- Layer 3 can not depend on itself or any other layer.
- Layer 4 can only depend on itself and layers 1 and 2.

- Layer 5 can only depend on layers 1 to 4.
- Layer 6 can not depend on itself or any other layer.
- Layer 7 can only depend on layers 3, 6 or 10.
- Layer 8 can depend on any other layer, except for itself and layers 7 and 9.
- Layer 9 can depend on any other layer, except for itself and layers 7 and 8.
- Layer 10 cannot depend on itself or any other layer.

Finally, we imposed as mandatory the edges representing:

- the dependencies of the variables MITOS at time t on the variable TSO;
- the dependency of the variable Survival on the variable TSO;
- the dependency of the variable TBV on the variable TSO.

3.4.2.3 Simulation Evaluation Metrics

Since the CPDs inferred on the training sets encode the most probable value of a variable given the values of its parents at the previous time point, DBNs allow the simulation of ALS progression starting from the data of the patient at a specific visit.

For assessing the network performance, we therefore set up a simulation framework where the progression of a set of subjects is simulated using the learnt network and then compared with the real one. For each subject, starting from a pre-defined visit, the temporal evolution of the disease can be simulated by sampling the CPDs for a number of consecutive steps. In each simulated time point, automatically determined by the tool according to the time distributions learnt in the training phase, the values of the variables are simulated accordingly with their values at the previous time point, by employing the learnt CPDs.

To assess the prediction accuracy of the simulation – and, therefore, of the DBN model – the simulated prognosis for each patient and the true disease progression were compared using as performance metric the Area Under (AU) the Receiver Operating Characteristic (ROC) curve. This metric allows to assess the ability of the DBN models to rank subjects based on their risk of outcome occurrence.

For a given clinical outcome, the Receiver Operating Characteristic (ROC) curve represents the probability of a patient who has experienced the outcome to be correctly simulated (true positive rate) vs. the probability of a patient who has not experienced the outcome to be incorrectly simulated (false positive rate). The ROC curves can be computed at different time points to assess the performance of the model in simulating distinct phases of the disease.

The Area Under (AU) the Receiver Operating Characteristic (AU-ROC) indicates the probability that a patient who has experienced a certain clinical outcome is assigned a higher risk value by the model than a patient who has not experienced that outcome yet: higher AU-ROC values (range 0–1) correspond to better simulation performance.

To evaluate the accuracy of our model over time, we finally computed, for each clinical outcome, the integral of the AU-ROC (iAU-ROC) across the simulated survival time points. The iAU-ROC can be interpreted as a global concordance index measuring the probability that subjects with a large predicted risk value have a shorter time to clinical outcome than subjects with a small predicted risk value [87].

3.4.3 Results: DBN Implementation and Prognosis Simulation

3.4.3.1 TSO Discretization

The time slicing procedure presented in Section 3.4.2.1 requires the user to set the value of the parameter $n.thresh$, corresponding to the number of thresholds used to discretize the TSO variable.

In order to assess the optimal $n.thresh$ for the current dataset, we tested different values of the parameter ranging from 0 to 5. In particular, $n.thresh = 0$ (corresponding to no time slicing, with the TSO discretized in a single level) was performed to verify that the proposed procedure was both actually required and adequately conceived. In all the cases, the confidence parameter C was set equal to 25%.

We selected the optimal $n.thresh$ by setting up a 3-fold CV procedure. The subjects of the training set were randomly split into 3 partitions, and in turn 2 were used as inner training set and the other one as inner test set.

For each value of $n.thresh$, the optimal TSO thresholds values were identified on the inner training set by employing the automatic time slicing procedure presented in Section 3.4.2.1. The continuous TSO values were then discretized according to these thresholds in both the inner training and inner test datasets. Next, a DBN was trained on the inner training set, and its performance was assessed on the corresponding inner test set, by employing the simulation procedure and the metrics introduced in Section 3.4.2.3[4]. The iAUC was used as performance score: for each $n.thresh$, the iAUC was computed for the four MITOS and the survival outcomes, and then averaged, by obtaining a global simulation performance score of the network for the current value of the $n.thresh$ parameter. The $n.thresh$ performing the best averaged iAUC score was then selected as optimal parameter value.

In order to reinforce the selection, the random split into folds was performed 16 times. Indeed, even if at an added computational cost, repeating the CV multiple times using different splits into folds provides a better Monte-Carlo estimate [104].

The optimal number of thresholds was finally identified by combining the 16 assessments, with $n.thresh = 2$ resulting the best value of the parameter, being voted 9/16 of the times.

Finally, the automatic time slicing procedure was performed on the whole original training set by setting the number of thresholds equal to the optimal value 2, with C=25%. The first TSO threshold was determined in the range 25%-41.66% (corresponding to TSO between 15 and 22 months), while the second one in the range 58.3%-75% (corresponding to TSO between 29 and 42 months).

[4]For the implementation choices adopted in the simulation and metric assessment, please refer to the next section, where the simulation and assessment procedure is fully detailed.

The resulting thresholds, used for the discretization of the TSO variable in both the training and test set, are reported in Table 3.4.

Table 3.4: *TSO quantization levels*

Feature	**Level**
Time since onset (TSO) [months]	≤ 15
	]15, 32]
	>32

Figure 3.2 reports the KM of the MITOS and survival outcomes for the training set over the whole TSO span, together with the ranges of inspection of each threshold and the identified thresholds.

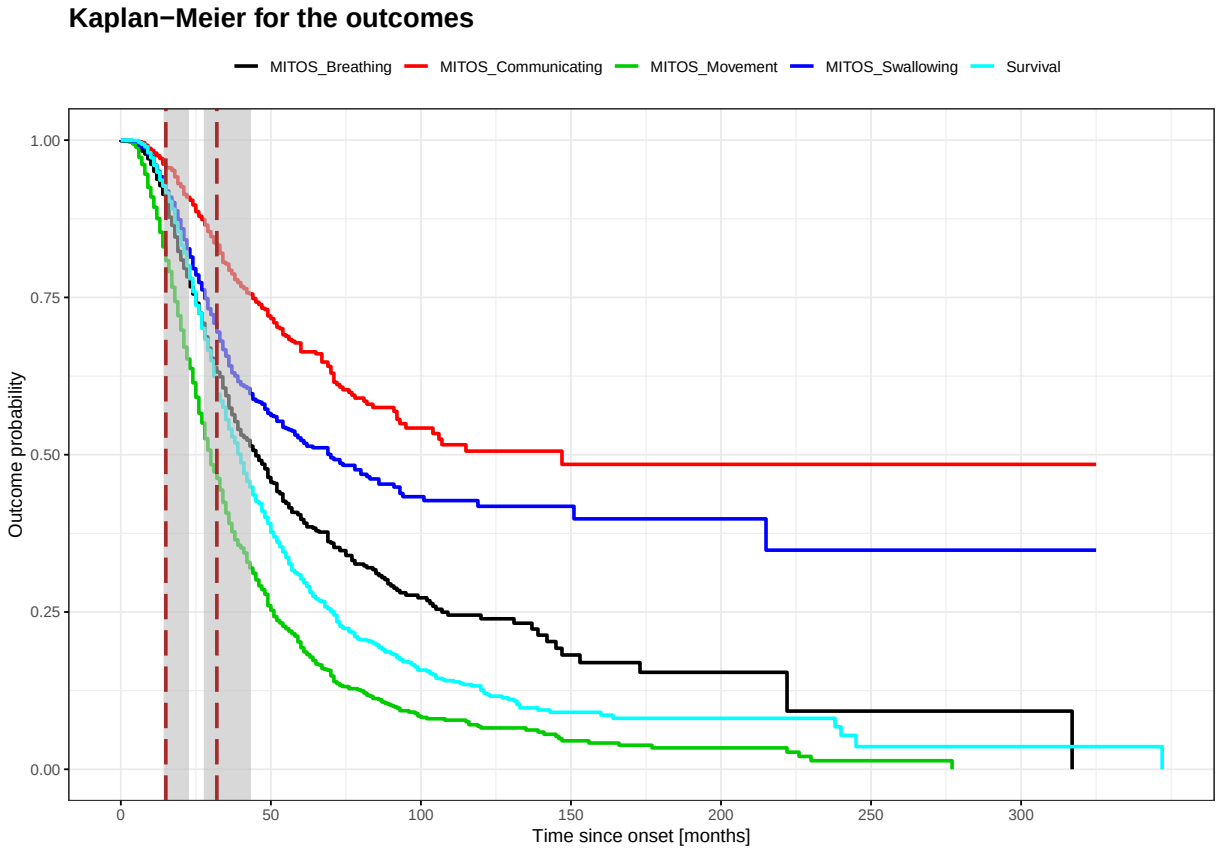


Figure 3.2: *Kaplan-Meier curves of the ALS outcomes. The dashed line indicates the optimal thresholds for TSO discretization; the gray band represents the inspected range of values for each threshold.*

3.4.3.2 DBN-based Simulation and Model Performance Assessment

After discretizing all the variables, a DBN was trained on the training set. Figure 3.3 reports the corresponding graph.

We assessed the model performance through the simulation procedure and the metrics introduced in Section 3.4.2.3, using the subjects of the test set. Since by design the simulation requires a fully-known starting set of variables to run, we extracted from the test set the subsets of patients without missing values in their first visit. This filtering step reduced the sizes of the test set from 645 to 202 patients (for a total of 2004 visits). Again, we made sure that the reduced test set maintained the same distributions over all variables as the corresponding training set. Tables 3.1 and 3.2 report the characterization of the reduced test set.

As the starting point for the simulation, we set the first recorded contact with the medical centre. For each patient, starting from his/her first visit, we simulated the temporal evolution

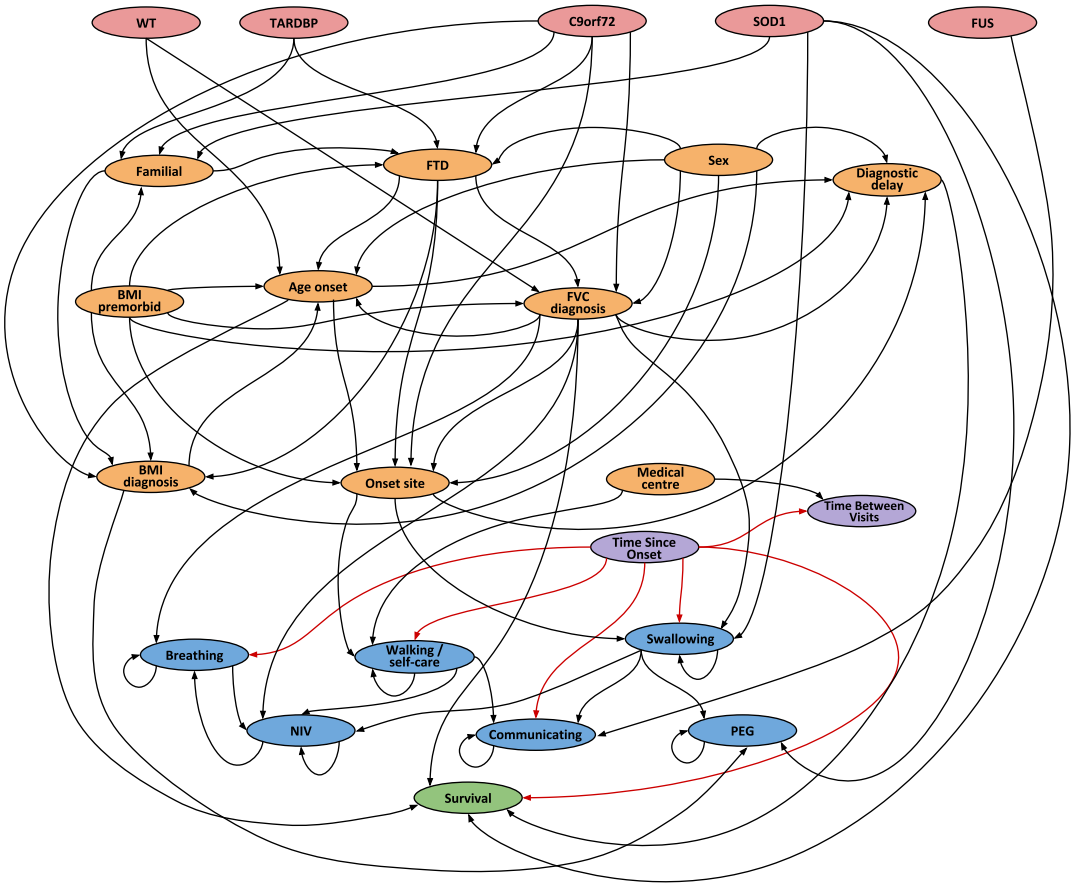


Figure 3.3: *DBN graph obtained on the training set, representing the conditional dependencies among the variables over time. The loops on NIV, PEG and the four MITOS domain variables represent the dependency on the values of the same variable from the previous time-step. The red edges represent the dependencies defined as mandatory in the network learning stage.*

of the disease by sampling the CPDs for 40 consecutive visits or until the simulated death or tracheostomy intervention occurred. The simulation sets the time step between two consecutive visits according to the time steps distribution learnt by the DBNs on the training set, accounting for the variability across patients and stages of the disease. The number of simulated visits was set to a relatively high value (40) so that each patient reaches the tracheostomy/death event with high probability. For each visit, the current values of the variables are simulated, in accordance with their values at the previous time point, by sampling them from the CPDs. Since this process is probabilistic, we performed 100 different simulations of the disease progression for each patient starting from his/her first visit, in order to obtain a statistic on the simulated prognoses: a total of 20 200 simulations were therefore run for the reduced test set subjects.

To assess the prediction accuracy of the DBN models we compared the simulated prognosis for each patient and the true disease progression. As clinical outcomes, we considered the MITOS impairments (walking/self-care, swallowing, communicating, and breathing) and the death/tracheostomy survival event.

Since each patient had a multiple number of simulated outcomes (100, one per simulation), we set the predicted time of occurrence as follows. For each patient with at least 50% of the simulations with a positive outcome, the overall simulated outcome was set to positive (dead/tracheostomised for the survival, impaired for the MITOS), and the time of occurrence was set equal to the median time of occurrence of the positive simulated cases. Analogously, for the patients with less than 50% of the simulations with a positive outcome, the overall simulated outcome was set to negative (censored for the survival, not impaired for the MITOS), and the time of occurrence was set equal to the median time of occurrence of the negative simulated cases. The predicted risk was then defined for each patient as the opposite of his/her median outcome time: a lower median time, usually occurring if the outcome is positive, corresponds to a higher risk.

For each clinical outcome, the ROC curves were computed at subsequent time points from 12 to 84 months, with a 12-months step. We stopped at 84 months since the percentage of deceased patients exceeded 97.5% in the following year. Table 3.5 reports the corresponding AU-ROC and iAU-ROC computed across all the simulated survival time points up to 84 months.

Table 3.5: *AU-ROC and iAU-ROC values on the reduced test set.*

Variables	**AU-ROC at time point [months]:**							**iAU-ROC**
	12	**24**	**36**	**48**	**60**	**72**	**84**	
MITOS Walking/self-care	0.90	0.86	0.85	0.85	0.82	0.79	0.80	0.84
MITOS Swallowing	0.95	0.86	0.83	0.83	0.84	0.84	0.86	0.86
MITOS Communicating	0.99	0.77	0.74	0.79	0.83	0.84	0.77	0.82
MITOS Breathing	0.92	0.87	0.86	0.78	0.82	0.82	0.84	0.86
Survival time	0.99	0.89	0.85	0.83	0.84	0.84	0.87	0.87

The AU-ROC values obtained by the model range from 0.74 to 0.99 for the impairment prediction in the four MITOS domains, and from 0.83 to 0.99 for the prediction of survival time.

The iAU-ROCs range from 0.82 to 0.87, denoting a good concordance of the predictions with the actual ALS evolution. These results confirm the ability of the model to simulate clinically reliable ALS populations by using the first screening visit only, thus validating the model.

In addition, we computed the cumulative probability of outcome occurrence over time. Figure 3.4 reports the cumulative probability of MITOS domain impairment and tracheostomy/death over time describing the true and simulated ALS progression of the reduced test set population. In this case, all the repetitions for each subject have been included in the assessment, in order to obtain a confidence interval of the prediction. The plots show a high concordance between the predicted and actual ALS progression, further confirming that the DBN model provides a precise dynamic simulation of the outcomes.

3.4.3.3 Variable Inter-dependencies

DBNs can be used to detect inter-dependencies among variables in terms of conditional probabilities, that can both qualitatively validate the model or shed a light on new possible interesting relationships.

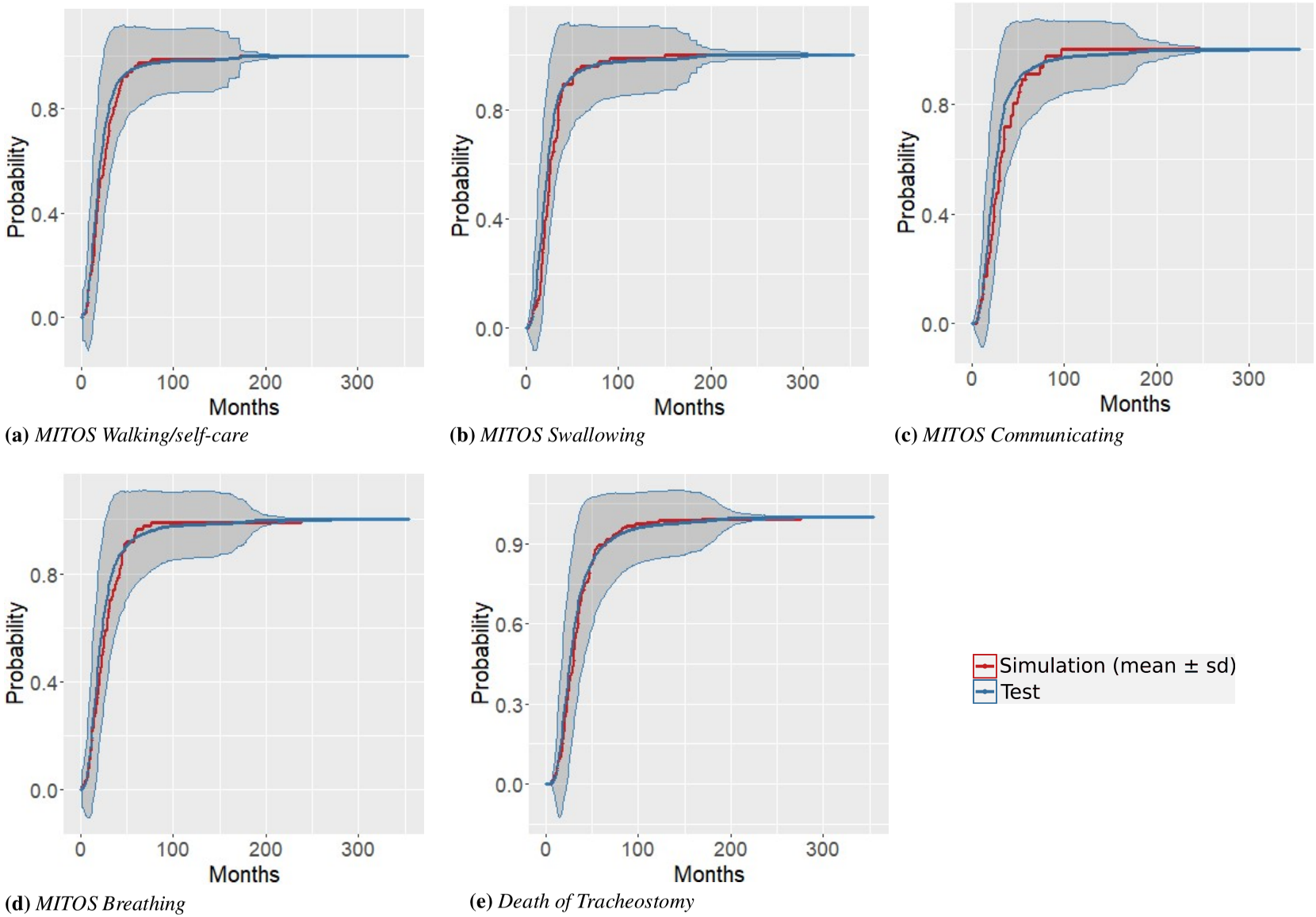
In this work we identified both expected and new dependencies among variables. As said, for a given node (variable), in-going edges represent conditional probability dependencies from the values of its parents at the previous time-point. Thus, in order to infer the state probability of the node at a certain time-point, all the values of its parents at the previous time-point are required. The dependencies among variables should therefore be read in terms of combined effect of the parents on the child variable. Some of these relationships are commented below.

With reference to the trained network of Figure 3.3, we highlighted in red the edges corresponding to the mandatory constraints defined in the learning phase (see Section 3.4.2.2): the TSO variable is a parent to all the MITOS domain variables, as well as to the survival [110, 108], in accordance with the progressive nature of the disease over time. The dependency of TBV from TSO was also imposed, to reflect that the visit frequency could change based on the rate of disease progression.

The graph shows that survival time also depends on age at onset: this dependency is already known in the literature, being a longer survival in younger patients probably correlated to their greater neuronal reserve [157]. Moreover, the survival depends on the diagnostic delay [33], the respiratory functionality at diagnosis reported as FVC at diagnosis [43], and on the SOD1 mutation [62].

As expected, all the variables encoding the MITOS domains, NIV and PEG emerged to depend on their own values at the previous time-point: graphically, this fact is encoded in the loops on the blue nodes. NIV also depends on FVC at diagnosis and breathing, both variables related to the respiratory functionality; PEG depends on BMI at diagnosis and swallowing, both related to the initial and progressive impact of the disease on nutrition ability.

The graph also evidences that the value of the MITOS walking/self-care domain at a given time point impacts on the loss of independence in communication (MITOS communicating) and breathing (with an edge to the NIV node, variable tightly associated with breathing ability) at the next time point. In particular, an impairment in movement abilities increases the probability of experiencing an impairment in communication and a the need for a NIV intervention within the



(a) *MITOS Walking/self-care*

(b) *MITOS Swallowing*

(c) *MITOS Communicating*

(d) *MITOS Breathing*

(e) *Death of Tracheostomy*

Figure 3.4: *Cumulative probability of impairment in the four MITOS domains and of death/ tracheostomy over time in the reduced test set (red line) and in the simulated population (blue line: mean values over population; shaded region: standard deviation), based on probabilities modelled by the DBN.*

next visit.

In addition, the relationship between onset site and swallowing may reflect the direct effect of the bulbar onset on deglutition ability, with anticipated dysarthria and dysphagia occurrence. Also, the direct edge from onset site to diagnostic delay validates the results reported by Kraemer *et al.* [105] and Turner *et al.* [197]. Conversely, Chiò *et al.* [35] and Cellura *et al.* [29] reported the lack of a significant difference in the diagnostic delay between bulbar- and spinal-onset patients, leaving this relationship as an open question.

The genetic etiology of ALS was correctly modelled in the graph, inferring the role of repeat expansion in C9orf72 and mutations in TARDBP and SOD1 on familial ALS [166, 173, 185]. It is also interesting to observe that there is no dependency between familiality and FUS, in line with the fact that the latter is a *de novo* mutation. The graph also evidences that FTD is related to mutations in TARDBP and C9orf72 which were already associated to FTD phenotypes in previous studies [126, 127, 166, 3]: these two genes are among the critical genetic players of both ALS and FTD, neurodegenerative conditions with a known overlapping genetics. The influence of premorbid BMI on ALS familiality also emerges, partially supporting the study by Gorges and colleagues [79], which evidenced a relationship between premorbid BMI and hypothalamus atrophy, a typical ALS signature, in familial ALS patients.

The onset site variable depends on both sex and age at onset, confirming relationships known in the literature: men have a greater likelihood of onset in the spinal regions, while women tend to have higher propensity for bulbar-onset disease [30, 86, 138]; furthermore, bulbar onset is related to a higher age at onset [196, 194].

The role of the medical centre in general on the whole network merits a close examination. This variable is related to walking/self-care and to the time between visits. Similarly to other relationships that involve more than one parent, the distribution of the child values depends on the joint effect of the parent nodes. In this sense, the effect of the medical centre should be interpreted in concert with the other parents' values, resulting in a possible corrective effect. Moreover, it is worth noticing that, in general, different medical centres may take charge of patients with varying disease severity, according to their specialisation level, by implementing different patient care protocols (that may affect the TBV variable) or diverse policies of life support interventions.

Expected relationships among variables can also be found as indirect dependencies. For instance, the effect of the onset site on the survival [30] can be identified from the following path in the graph: onset site $\rightarrow$ diagnostic delay $\rightarrow$ survival. An association between age at onset and SOD1 and C9orf72 is also found as indirect path through FTD in the graph: interestingly, the age-related penetrance of gene mutations is currently an open question in the literature [149, 34].

3.4.3.4 Cohort Stratification: Effect of Risk–Factors on Disease Progression

The DBN-based simulator also allows patient cohort stratification, *i.e.*, the identification of variables whose specific ranges of values could be related to the velocity of disease progression or survival. In particular, we traced how the change in a specific variable (or risk factor) may affect the disease course, by simulating ALS progression of a population with specific phenotypes at onset and comparing how they differentiate in terms of disease severity as well as survival time.

We selected variables with expected and/or documented effects on the disease prognosis, and tested the ability of the DBN models to reproduce the awaited clinical outcome progressions on the reduced test set subjects.

For a given variable of interest, two approaches are possible: (a) the test set can be partitioned in sub-cohorts according to the original value of that variable in the subjects' first visit, or (b) each level of the variable of interest is imposed to the first visit of all the subjects of the test set, in turn. In both cases, the original values of the all the other variables is maintained, thus preserving the population assortment. Then, the evolution of ALS is simulated in each sub-cohort to verify the differentiating effect of the variable on the clinical outcomes.

Before choosing for the (a) or (b) approach, some remarks are needed. In the first case, the cardinality of the sub-cohorts corresponds to that of the test set, and there is therefore no guarantee on its balancing; on the other hand, in the second case the two simulated sub-cohorts have the same cardinality, equal to the dimension of the test set itself, thus possibly providing a more balanced comparison. Nevertheless, in the (b) implementation, imposing in turn the values of the selected variable to all the subjects may cause clinically inconsistent combinations of the variables (for instance imposing sex to male/female to patients that also have other gynecological/andrological variables recorded).

As an example of application, we investigated the effect of the *FVC at diagnosis* (static variable) on the time to MITOS *breathing* impairment.

First, following the approach (a) introduced above, we stratified the patients of the reduced test set according to their original *FVC at diagnosis* discretized values, obtaining the following three partitions: patients with original FVC at diagnosis lower than 84%, between 84% and 102%, and higher than 102%. We then simulated the ALS progression for each partition separately and compared their times to the breathing impairment. Figure 3.5a reports the probability distribution of the simulated impairment times for each sub-cohort. As in Section 3.4.3.2, for each patient 100 distinct simulations were run, and the median outcome time over the majority of his/her repetitions was considered as outcome time.

This analysis shows, as expected by the nature of this MITOS breathing item, that the lower the FVC at diagnosis, the sooner the patients are likely to lose their breathing independence. Indeed the MITOS breathing item records the impairment when the subject either experiences dyspnea at rest, difficulty breathing when sitting or lying, or continuously uses the non invasive positive pressure ventilation (NIPPV) during the night. The worst the situation at diagnosis, the sooner the patient will experience dyspnea. Our model predicted that the breathing impairment would most likely occur at 18.7 months from the disease onset for the patients with an FVC value at diagnosis smaller than 84%, at 22.9 months for the ones with an FVC between 84% and 102%, and at 32.3 months for the ones with an FVC greater than 102%.

Similar results are obtained when imposing in turn the different values of FVC at diagnosis to the whole reduced test set population – according to the (b) approach – and then simulating the progression of the three numerically equivalent sub-cohorts. Figure 3.5b reports the probability distributions. Also in this case, 100 distinct simulations were run and the median was taken as outcome time for every subjects experiencing the impairment in at least 50% of his/her repetitions. Here, the breathing impairment is most likely to occur at 17.8 months from the disease onset for the patients with an FVC value at diagnosis smaller than 84%, at 21.7 months for the

ones with an FVC between 84% and 102%, and at 33.9 months for the ones with an FVC greater than 102%.

Noticeably, in both cases these predicted occurrence times are highly concordant with the real times to breathing impairment experienced by the patients in the reduced test set (13.5 months for the patients with FVC lower than 84%, 25.8 months for the ones with an FVC between 84% and 102%, and 35.6 months for the ones with an FVC greater than 102%). The real impairment times probability distribution is reported in Figure 3.5c.

As another example of assessment of the effect of specific risk-factors on the disease progression, we report below the investigation of the effect of the *onset site* on the time to MITOS *swallowing* impairment.

In this case, the (b) approach was followed, by imposing in turn the two possible levels of the onset site (bulbar/spinal) to all the subjects of the reduced test set. Each time, the disease evolution was then simulated 100 times per patient, and then the median swallowing impairment time was selected.

Figures 3.6a and 3.6b report the probability distribution of the swallowing impairment times for the simulated stratified cohorts and the real ones, respectively.

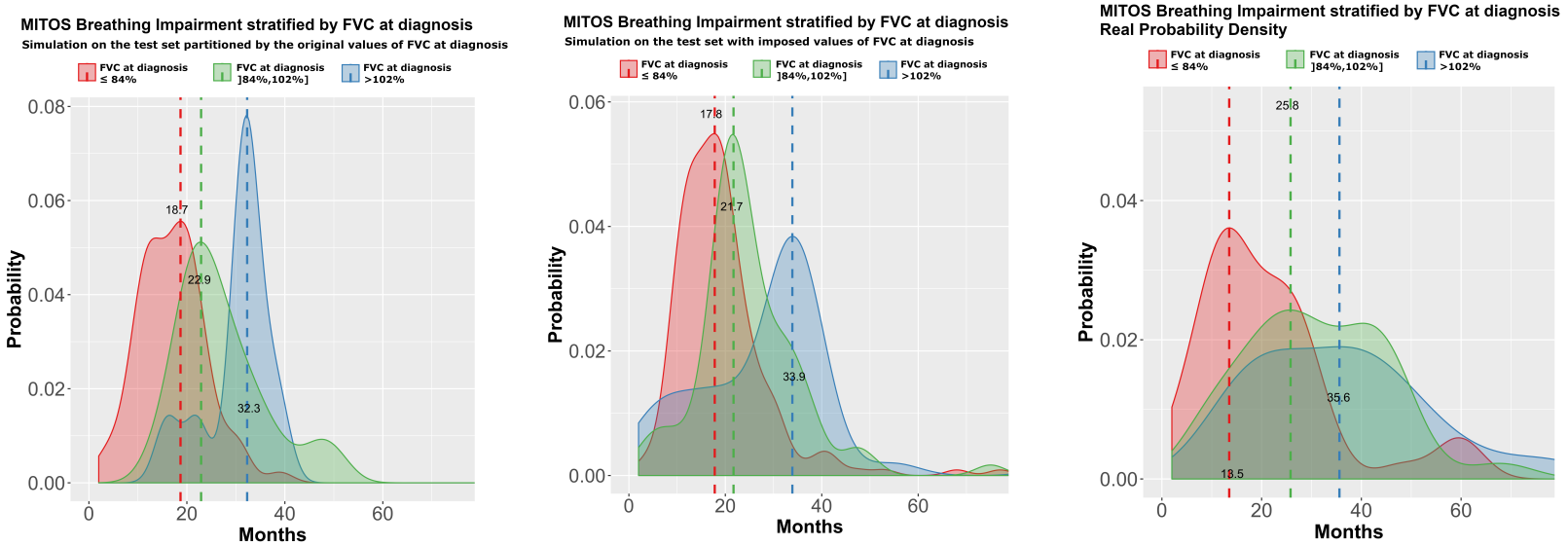
The analysis of the impairment times shows that patients with bulbar onset have higher probability of experiencing swallowing impairment in earlier stages of the disease compared to patients with spinal onset, a result well known in literature [30, 110].

Also in this case, the simulated and real impairment times show a good concordance (21.2 vs 24.2 months for the simulated bulbar and spinal subjects, respectively, and 20.4 vs 25 months for the real bulbar vs spinal subjects.)

3.4.3.5 Dashboard for Clinical Use

Beside allowing population-wide analyses, the model can be used to probabilistically predict the disease progression of a single ALS patient by only using information recorded during his/her first visit (see Section 3.4.2.3) and analyzing his/her simulated repetitions to get a probabilistic characterization of the prognosis.

This single-patient prognosis prediction was implemented as a dashboard using the Shiny framework for R and will be made available to clinicians as an interactive web application. Figure 3.7 shows the GUI of the developed web tool. The physician can enter the clinical data recorded during the first contact with the patient in the left side of the screen, under the "Insert patient data:" label. The variables required by the tool constitute the standard features collected in the clinical practice to assess and monitor the patient's clinical condition over time. After inserting the variable values, the user can start the simulation with up to 1 000 repetitions (100 repetitions were used in the presented example). The tool will produce on the right side of the screen the probability of impairment in each of the four main MITOS domains, computed for each patient over her/his repetitions. In our implementation, different simulations can be run sequentially, allowing the user to decide whether to keep the plots from previous simulations to be viewed alongside with the plots from the last one. This way, it is possible to estimate the effect of one or more biomarkers on the ALS prognosis: for instance, Figure 3.7 compares the effects of the "spinal" and "bulbar" onsets while leaving all other parameters unchanged.

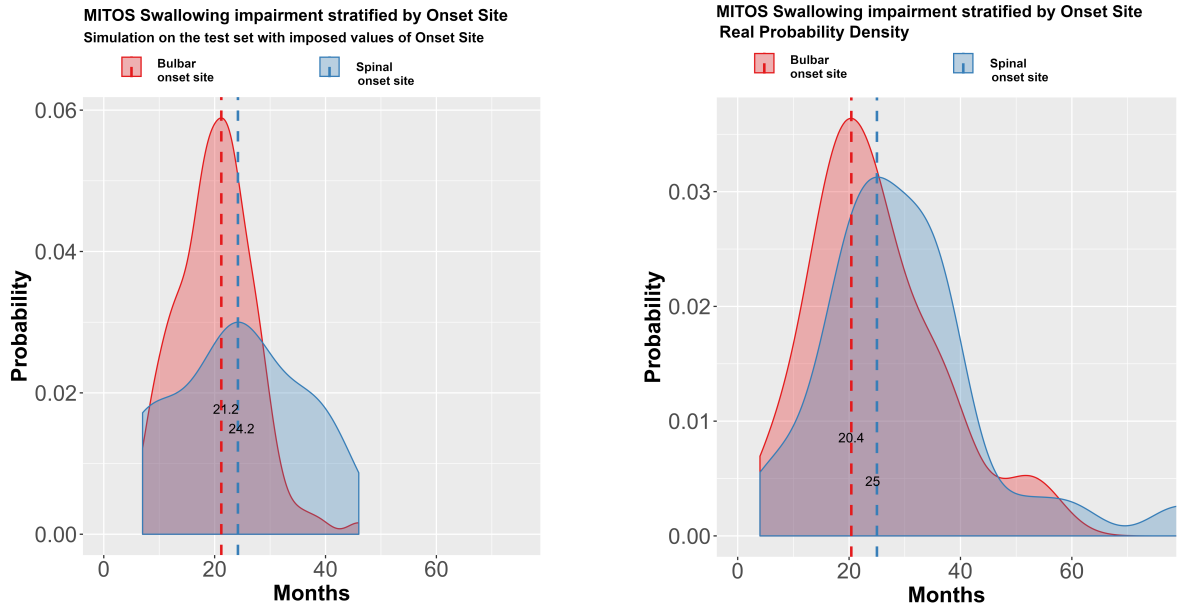


(a) *Simulation after stratification by the original values of FVC at diagnosis*

(b) *Simulation after imposing in turn the different values of FVC at diagnosis*

(c) *Real impairment times*

Figure 3.5: *Density probability plots of the times to MITOS breathing impairment for the patients of the reduced test set stratified by the values of FVC at diagnosis (lower than 84%, between 84% and 102%, and higher than 102%). Most patients experience the impairment in correspondence with the maximum of the probability density curve (mode). In (a) and (b), for each patient 100 distinct simulations of the disease progression were performed starting from the first visit values. The occurrence time was then computed as the median of the impairment times, if the outcome was experienced in at least 50% of the repetitions.*



(a) *Simulation after imposing in turn the different values of onset site*

(b) *Real impairment times*

Figure 3.6: *Density probability plots of the times to MITOS swallowing impairment for the patients of the reduced test set stratified by the values of onset site (spinal or bulbar). Most patients experience the impairment in correspondence with the maximum of the probability density curve (mode). In (a), for each patient 100 distinct simulations of the disease progression were performed starting from the first visit values. The occurrence time was then computed as the median of the impairment times, if the outcome was experienced in at least 50% of the repetitions.*

The developed dashboard can also be used to generate *in silico* populations. For example, it is possible to simulate a population of subjects with bulbar onset by sampling the other variables from real data. Similarly, it is possible to simulate an untreated population, which could serve as control group for clinical trials.

3.4.4 Discussion: Applicability and Advantages of a DBN-based Progression Model

Integrated analyses of large multidimensional datasets by new mathematical and statistical approaches are required to unravel the heterogeneous nature of ALS.

In this work, we developed a probabilistic predictor of the progression of ALS by building a DBN model on real-world data including demographic, genetic and longitudinally-collected clinical variables. Being comprised of patient visits from clinical contexts and partly never investigated before, the dataset employed in this work is more representative of the general ALS population than PRO-ACT or other clinical trial datasets. Moreover, it includes variables that

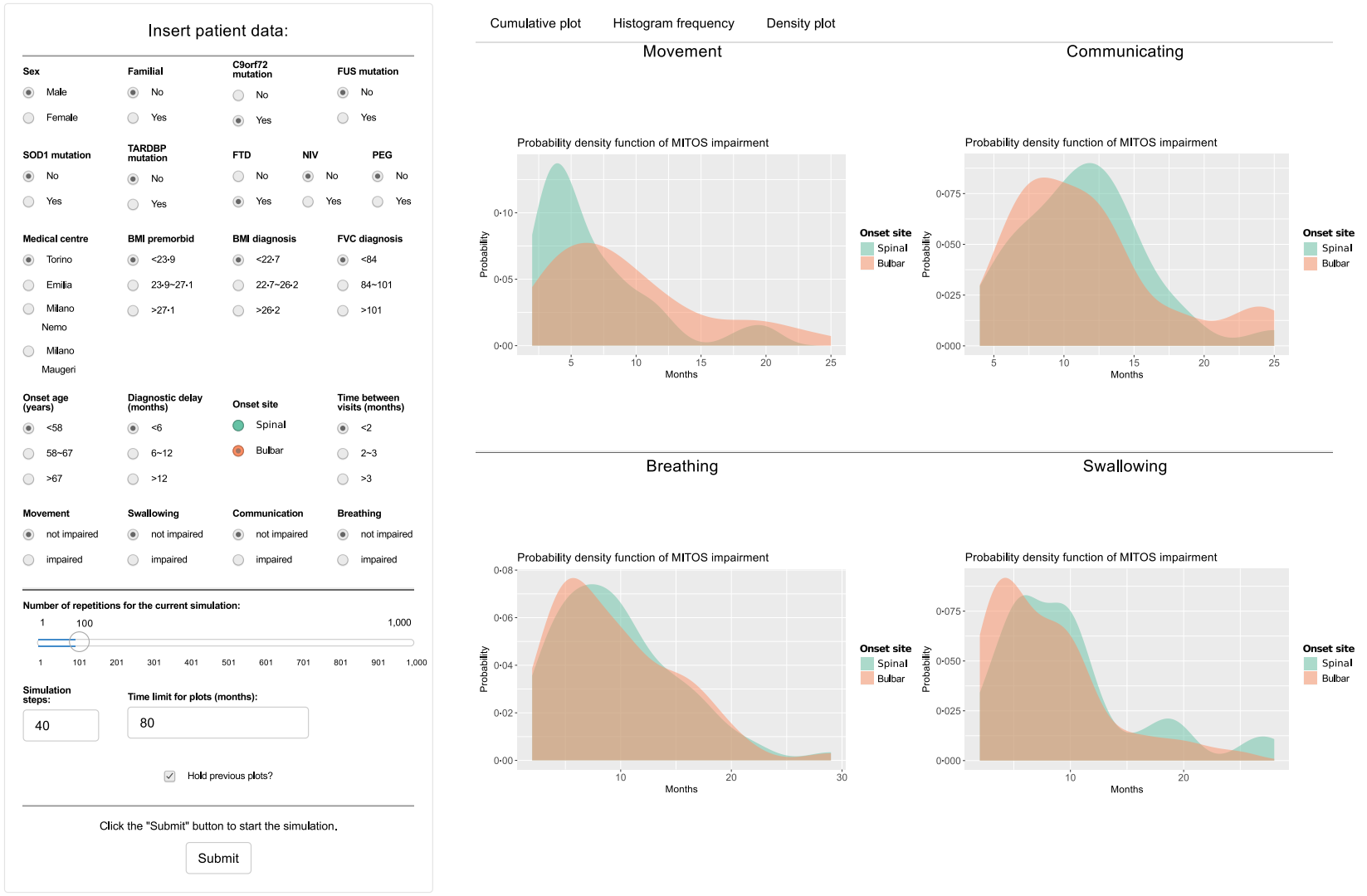


Figure 3.7: *Example of the single-patient ALS prognosis prediction using the web application we developed on the DBN built on the training set. The figure shows the impairment probability evolution in time (months) in each of the four MITOS domains for two hypothetical patients with very similar characteristics, differing only in the onset site of the disease.*

widely characterise the patients' clinical history from premorbidity to the survival endpoint, covering symptom evolution and support interventions.

Relying on DBNs, the implemented model is trained with the entire dynamics of the available data of disease progression, and can be effectively used to simulate, starting from a single time point, the entire disease progression in terms of survival or time to the loss of independence in walking/self-care, swallowing, communicating, and breathing.

The prediction accuracy was assessed by comparing the predicted patients' prognoses with the real data: different performance metrics confirmed that the proposed model offers good performance in terms of both survival and domain impairment prediction. In addition, the model can also be used to stratify ALS patients into subgroups of different progression and to assess the effect of specific phenotypes on the entire disease course.

According to the chosen technique, the method allows the identification and explicit representation of the relationships between the different variables and of the pathways along which they influence the disease evolution. In this work, several notable inter-dependencies among variables were identified and validated by comparison with literature results. Given a specific variable, its parents in the DBN graph can be intended as "composite biomarkers", since the value of the variable at a certain time point can be inferred by their values at the previous one, thus extending the classic "standalone" biomarkers that have been used to date.

A possible limitation of our approach is that the proposed model can only employ discrete variables. This implies that: (1) all continuous variables must be discretized into a finite set of levels before being processed; (2) the model can only predict the most probable range of each variable instead of their actual continuous values. Moreover, in order to correctly predict the prognosis of a given patient, all the information regarding his/her first visit must be available (missing data are not allowed). From this standpoint, it could be beneficial to develop simpler models that employ less variables to predict the patient's prognosis (as we proposed in [213, 37, 38]). On the other hand, a model based on a vast number of variables allows a more detailed characterization of the disease.

Notably, the developed model was also implemented as an interactive web application that can be used by clinicians to simulate the most probable prognosis of a patient already at his/her first visit. An instrument able to simulate patients' outcomes in the main areas of disability will have a strong impact in scheduling the allocation of resources both at individual and health system level, likely reducing the cost of care by improving the provision of pharmacological and non-pharmacological therapies. Furthermore, a reliable model of ALS progression could potentially serve as a control group when the use of placebo may not be appropriate or feasible, or could allow a smaller control group if used in combination [152].

3.5 Final Remarks

The availability of longitudinally-collected clinical data represents an invaluable resource for modeling the clinical progression of patients with conditions.

In this Chapter I presented my experience in building a descriptive and predictive probabilistic model of disease progression, through the employment of a technique intrinsically able to

manage and exploit the dynamic dimension of the data.

In this implementation, I also showed how the data require an appropriate preprocessing phase able to take into account their heterogeneous nature, devoting particular attention to how the temporal dimension should be handled with respect to the working hypotheses of the selected technique. Here, the implemented automatic time slicing algorithm allows training a DBN that respects the time-independence assumptions in each slice.

Effectively preprocessing and structuring the data represents a crucial step when implementing this kind of models on a new dataset or case study. In this process, special attention should be given to how the time variable is coded: first of all, an analysis of how the time dimension is included in the available data should be performed; then, some considerations on the variables to be included in the model with respect to the outcomes of interest should be made. For instance, let's consider a disease for which the age of onset is expected to be a strong predictor of the survival outcome: in this case, it may be convenient to code both a static variable representing the subject's age at onset and a time-evolving variable coding the time passing from the onset itself. In other cases, it could be more effective to include only one variable representing the age of the subject at each measurement point.

Another key point in the model design is the definition of the network layers and the possible rules among them: even here, some domain knowledge should be carefully included in order to effectively "guide" the model toward the inspection of interesting relationships without introducing any bias or artificial effects.

Employing modeling approaches such as DBNs allows to learn upon the whole dynamics of the data, thus fully exploiting the information being collected in the dataset. Such technique is indeed able to inspect the longitudinal information to catch the relationships among variables over time and the pathways along which they influence the disease evolution.

A first valuable output of the employed methodology consists in the structure of the network itself, that provides a very practical way to visualize the emerged dependencies. Moreover, by inspecting the CPD learnt in the learning phase it is possible to further investigate these relationships, thus overcoming the limitation of other "black-box" methods.

In addition, the prediction output consists of a longitudinal simulation of the patient evolution, whose inspection allows to follow the disease progression in terms of survival, as well as time to impairment of functional abilities proper of the domain. By setting up a multiple simulations framework and exploiting its Bayesian nature, the method can provide a probabilistic prediction of the outcome occurrence over time, also allowing an interpretation in terms of confidence interval (for instance by visually inspecting the probability distribution curves).

Noticeably, the tool can be used both at patient level, to predict the evolution of the disease in the next future, and at population level, to compare the prognosis of cohorts with different characteristics or to assess the effect of specific variables on the prognosis by performing stratification analyses.

Chapter 4

Process-Oriented Approaches to Healthcare Analytics

In this Chapter, a fully process-oriented approach is employed to analyze clinical data with a temporal dimension. Process Mining for Healthcare is a discipline that, starting from a set of data structured as successions of events, provides a number of unsupervised and supervised techniques that can easily be employed on dynamically evolving clinical data. Their employment allows to inspect the patients' clinical history by mining the processes generating the data, following the patients' clinical patterns, characterizing the timing between events and the outcome occurrence. Remarkably, the algorithms often provide as outcomes the visual representations of the mined processes, represented for instance in terms of graphs or networks. This additional feature constitutes one of the strengths of this technique, which facilitates access to the information and improves dissemination of the results.

4.1 Process Mining

Process Mining (PM) is a family of process analysis methods that aim at discovering, monitoring and improving the efficiency of real processes by extracting knowledge from a set of executions recorded by an information system. This set of executions, recorded together with their execution times, is also referred to as an Event Log (EL).

Analytic algorithms are applied to ELs with the main goals of: (i) mining the data in order to represent the process able to produce them (*Process Discovery*, PD) [208], (ii) measuring to which extent a given process can represent an input EL or how much an EL complies with a given process (*Conformance Checking*, CC) [206], and (iii) improving process efficiency, by allowing problem diagnosis and delay prediction, recommending process redesigns or supporting decision making (*Process Enhancement*, PE) [201].

PD is acknowledged as the most prominent process mining technique [75]. In PD, process models are mined from an EL without using any *a priori* knowledge, by handling instead data-driven approaches. PD algorithms, such as the α-algorithm [208], the Heuristic Miner [218], or the Fuzzy Miner [83], are employed on the EL to achieve models able to describe the observed

behaviour of processes according to several possible perspectives (such as the control-flow [98], organizational [125, 184], performance [183], or data [174] perspective). The mined processes are then expressed, for instance, in terms of Petri nets [207, 209, 219], Event-driven Process Chains (EPCs) [210], Activity graphs [4, 47], or Control-Flow Graphs [51].

In CC, the EL is used to check if the observed behaviour conforms to a given (discovered or hand-made) process model [175]. Through CC, deviations can be detected and measured both in terms of inability of the model to capture the real data behaviour, or inconsistency of the observed reality with respect to the desired model. The first analysis is taken when the model is supposed to be descriptive, and thus expected to capture or predict reality; the second one when the model is normative, or used to influence or control reality [202]. CC can be used for business auditing and compliance checking, for measuring the performance of process discovery algorithms and for repairing models that are not aligned well with reality [201].

Whereas CC assesses the alignment between reality and an *a priori* given model, PE aims at changing and improving the model itself. By employing the diagnostic provided in CC, the model can be modified to better reflect reality, or extended by adding new information (like costs, risks, or resource usage) and replaying the EL on the model to analyze the additional attributes [202].

In all its forms, PM sits therefore between computational intelligence and data mining on one hand, and process modeling and analysis on the other hand [205]. Since each domain where information can be described as a series of events could potentially be subject to PM analyses, in the last years this data-driven technique has been successfully applied in a variety of research and industrial fields, ranging from manufacturing and logistics, to technology and healthcare.

4.2 Process Mining for Healthcare

Focusing on healthcare, the increasing abundance of clinical and administrative data collected in today's care centres undoubtedly represents a great resource: this precious amount of information is more and more massively exploited by the communities that constitute the healthcare sector (such as the medical, scientific and managerial ones) to constantly move, through the employment of a variety of different analytical approaches, towards the enhancement of the quality of care while dealing with the constant reduction of public spending on health. Among these possible approaches, we find PM.

In PM for Healthcare (PM4HC), processes are meant as a graph of activities which can be performed with the aim of diagnosing, treating and/or preventing diseases to improve the patients' health status. They include both clinical and non-clinical activities (as for instance treatment administrations or medical billings) provided by different stakeholders, and may present different behaviours according to the specific organization [132]. These processes are highly dynamic, highly complex, and increasingly multidisciplinary [88]. Furthermore, processes in Healthcare are often only partly structured and with many exceptional behaviours, due to their intrinsically required flexibility [132]. Not least, most of the activities that compose these processes are often high-cost. All these characteristics make processes in healthcare either crucial to improve and interesting to analyse. Using PM techniques not only ensures that such procedures are deeper

understood, but can also generate benefits associated with process efficiency that, ultimately, will have a direct or indirect impact on the patients' level of assistance. Figure 4.1 shows a general outline of the application of PM in Healthcare.

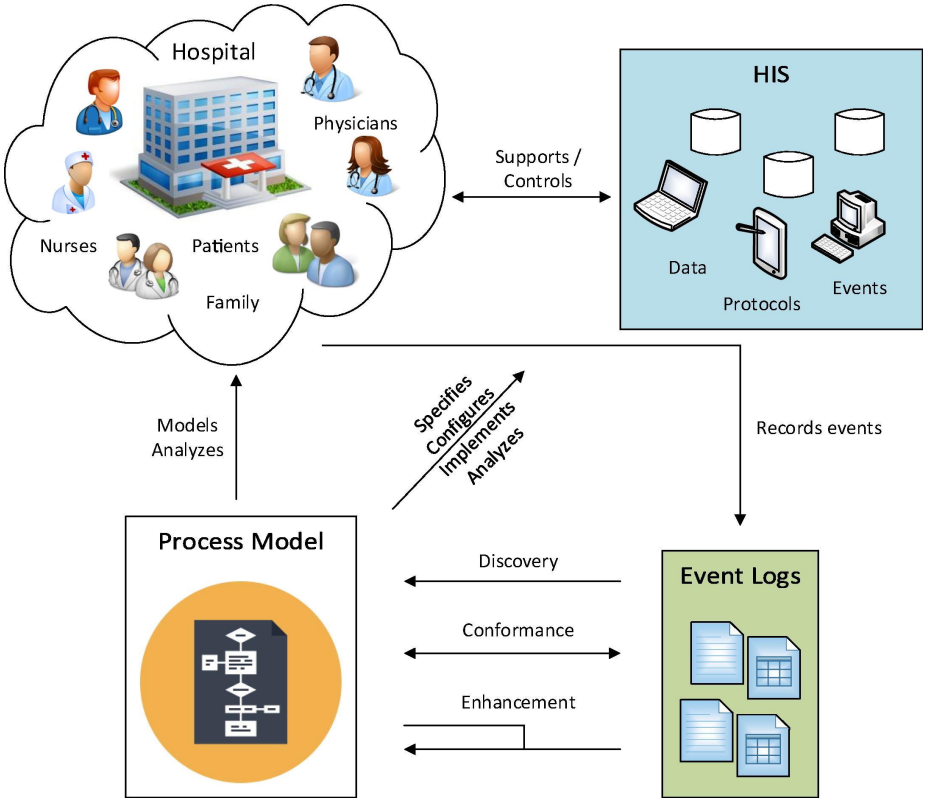


Figure 4.1: *Overview of Process Mining in Healthcare. Taken from [171].*

More specifically, processes in healthcare can be classified into two main classes: *medical treatment processes* and *organizational processes* [54, 118, 165].

Medical treatment processes are the clinical processes related to patient care, including tasks ranging from diagnosis to the execution of actions for treating patients. These processes are often denoted as diagnostic–therapeutic cycle comprising observation, reasoning, and action, with each pass of this cycle aimed at decreasing the uncertainty about the patient's disease or the actual state of the disease progress [150]. This results in a very complex decision process, since medical knowledge includes medical guidelines of various kinds and evidence levels, as well as individual experience of physicians [118]. Notably, their complexity recently increased due to the advent of personalized approaches to care, in which treatments are tailored to the specific profile of the patient and disease, such that the diversity of therapeutic pathways explodes compared to traditional standardized care guidelines.

On the other hand, organizational processes focus on managerial understanding of healthcare activities, capturing the knowledge necessary to coordinate collaborating health care professionals and organizational units [98]. Unlike medical treatment processes, organizational processes do not provide any support for medical decision making [118] and are mainly of a repetitive

nature.

For the purposes of this thesis, in the following sections I will mainly focus on medical treatment processes. It is worth emphasizing how, in this framework, a process-oriented perspective intrinsically allows to manage the temporal dimension of data collected in clinical contexts. PM4HC provides indeed algorithms able to exploit the time dependency among events when inspecting patient care processes and outcomes.

4.3 Previous Work on PM4HC

Pragmatically, PM4HC has shown interesting application in many domains [171] such as Cardiology [112], Oncology [111], Emergency cares [135], Diabetology [46], Anestesiology [98].

Mans *et al.* [133] and Rojas *et al.* [170] outlined five frequently asked questions posed by medical experts which guided the majority of PM4HC performed analyses:

- What are the most commonly followed paths and what exceptional paths are followed?
- Are there any differences between care paths followed by different patient groups?
- Do we comply with internal and external guidelines?
- Where are the bottlenecks in the process?
- What are the roles and the relationships among the users who performed the activities (for instance, do they belong to the same organizational units?).

Through the employment of its methodologies, PM4HC can help in answering these questions. For instance, PD can be an objective way of analyzing care pathways of patients with conditions, without being biased by perceptions or normative behaviors [223]. With this aim, the control-flow perspective, based on discovering the execution order of process activities, has been applied in most of the case studies. Moreover, both PD and CC can support in inspecting the followed paths and detecting patients with infrequent/exceptional behaviours, allowing to further inspect possible differences in terms of outcomes or progression.

CC allows to study how measured ELs comply with protocols or guidelines, in a way that partially overlaps with a similar research topic, called Computer Interpretable Clinical Guidelines: after representing the process model corresponding to the prescribed behaviours, it is possible to monitor how patients flow through it. Conformance between EL and model can thus be checked, easily identifying those groups of patients that did not follow the expected paths (for instance, patients non compliant to a given protocol), in the quest of understanding why that was the case and the related implications [70].

Finally, analyzing execution time of activities and collaboration between resources through PD and PE analyses permits to identify bottlenecks, synchronization and idle time, allowing to design how to improve care pathways and hospital performance. Moreover, some techniques explicitly allow to handle the possible different timing of a same activity across different patients,

by providing for instance formalisms to include the activity in the process investigation only if it occurred within a selected time span.

It is also worth emphasizing how, in general, the graphical representations that PM4HC algorithms provide as outcomes promote dialogue and exchange of views with all stakeholders, allowing to effectively visualize (and thus, in a way, to better "touch") the data.

For a more complete overview on case studies and applications, please refer to the recent review by Rojas *et al.* [171].

4.4 Open Issues and Contribution

Technically, PM4HC can be challenging. When applied to the healthcare domain the traditional process mining approaches may experience difficulties related to the complex and generally unstructured nature of the processes, resulting in process models unable to provide clear insights on the data [130].

Among the possible issues when performing PM4HC analyses we find the *Spaghetti Effect*: healthcare processes are often low-structured, being constituted of many activities performed by a potential high number of actors and often in a variable order. When PD algorithms are applied, this results in wide and sparse model representations with few instances for each branch. Similarly, performing CC and PE can be challenging too, being the process model hard to delineate and to tweak, respectively. In the case of such processes, therefore, only a subset of available process mining techniques is applicable, sometimes requiring the development of new algorithms or an *ad hoc* extension of the existing ones. As an alternative, the EL can be split into smaller yet more homogeneous logs, each corresponding to a limited, but more accessible, process model [200, 223].

A further distinctive feature of PM4HC applications is the necessity to incorporate medical knowledge in basically each step of analysis. Data often require a considerable pre-processing constrained by medical knowledge and medical relations to be eventually structured as informative ELs. This step may for instance consist in the extraction of specific values from continuously monitored time series, with the aim to define specific events that make sense medically. Besides, the employed algorithms require to be adapted to the specific clinical domain, to customize sound analyses and steer the outcomes towards medically relevant results [98].

Another characteristic of logs in the healthcare domain is that they frequently consist in the aggregation of information collected from multiple sources, often constituted by many autonomous, independently developed information systems. This can cause noise or inconsistencies as duplication or incompleteness in the EL, resulting in wrong detected dependencies and models [114].

Altogether, even from a process-oriented perspective particular attention should be payed when approaching the analysis of healthcare data, carefully evaluating the available information, the state-of-the-art algorithms that better fit the research questions, and possibly considering extension of existent methodologies to the specific case study.

However, PM4HC carries great potential in helping to understand different aspects of clinical processes workflows [223]. In [132], Mans *et al.* delineate the possible directions of PM4HC

analyses, mainly corresponding to the descriptive/discovery (what happened/which novelties emerge?), diagnostic (why did it happen?), predictive (what will happen next?), and prescriptive (what are the steps towards improvements?) facets of health analytics described in Chapter 1. However, most of the PM4HC works on clinical data apply process-oriented techniques to address only a few of these questions, possibly combining them with other classical analytics methodologies. In particular, there is a limited number of works (as [117]) where PM4HC techniques were employed for statistical inference, concretely developing process-oriented predictive models that assess the role of covariates in determining disease evolution or the patient's clinical pathway.

Based on these considerations, in my research work I explored the potential of a fully process-oriented approach when performing some of the classical statistical analyses on clinical data with a temporal dimension, namely preprocessing, descriptive, and inferential statistics.

In the following section, I will outline the potential of this approach in analyzing a case-study oncological clinical dataset. The temporal characteristic of the data allowed to identify the sequence of clinical events constituting the care pathway, thus permitting to set up a proper EL. Some covariates were included in the EL with the aim to characterize the subjects in descriptive, inferential, and stratification analyses. In this work, PD and CC tools have been employed to perform the classical steps of the analysis workflow. Survival analyses have been performed to characterize patients following distinct treatment patterns, thus allowing to gain insights not only on the course of the events, but also on the clinical outcomes.

This research resulted in a work presented at the 55th Annual Meeting of the American Society Of Clinical Oncology (ASCO 2020) [73] and in a contribution at the 3rd International Workshop on Process-Oriented Data Science for Healthcare 2020 (PODS4H 2020), part of the 2nd International Conference on Process Mining (ICPM 2020) [190].

4.5 A Process Mining Approach to Statistical Analysis

In this section, I outline my exploration of a process-oriented approach when performing statistical analyses on clinical data with a temporal dimension.

The dataset used in this work as a case study consists in a collection of dynamic information, referred to meaningful clinical events (from diagnosis to survival) and related covariates of a real-world cohort of advanced melanoma patients treated at the Lausanne University Hospital (CHUV). Here, employing a number of PM4HC techniques, from PD to CC, it was possible to delineate how PM can guide and/or assist researchers in three classical steps of a statistical analysis, that is, data preprocessing, descriptive statistics, and inferential statistics. Figure 4.2 summarizes these steps.

In the preprocessing step, we approached the data inspecting their structure, their information content, and their quality: after identifying the clinical milestones of interest (like diagnosis, treatments, survival outcome), data were first shaped as an EL. We then employed the visualization tools provided by PM to detect data inconsistencies due to input errors or missing values. This allowed us to go back to the data sources, recheck and correct the recorded information, thus recursively improving the data quality.

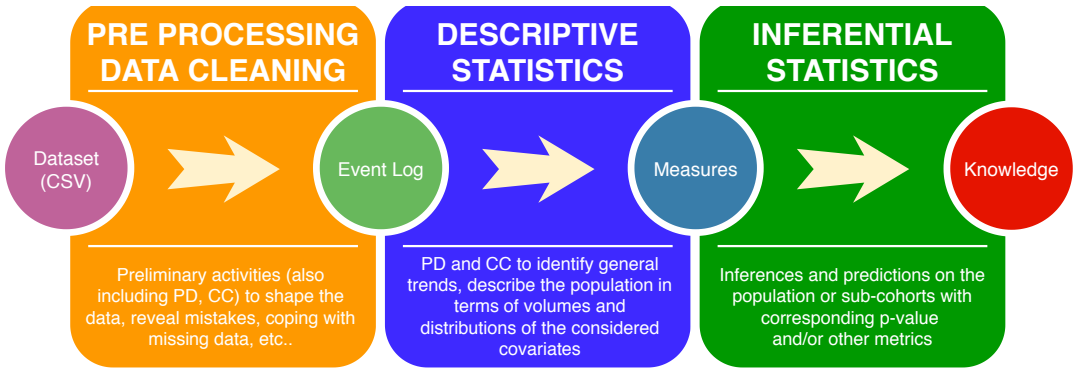


Figure 4.2: *Workflow of the classical steps of a statistical analysis, here implemented exploiting a process-oriented approach.*

In the descriptive analysis step, we first employed the EL time-oriented structure to inspect cardinality and order of the pharmacological treatments administered to patients. Then, we implemented both unsupervised and supervised methods to capture the flow of the patients' pathways over data-driven graphs (PD approach) or user-defined graphs (CC approach), respectively. In this part of the analysis, the graphical outputs provided by PM allow a fast access to the design and/or interpretation of the models, and an immediate assessment of the treatments in terms of type, order and timing of consecutive administrations.

Finally, in the inferential statistics step, we built upon the processes constructed in the previous step to quickly select sub-cohorts of patients characterized by similar patterns of care and/or clinical attributes. The cohorts were then compared in terms of time-to-event outcome and overall survival (OS), using Kaplan-Meier analysis and the log-rank test.

4.5.1 Material: Longitudinal Data of a Real-World Advanced Melanoma Cohort

In this work, we analyzed the data of a cohort of patients treated at the CHUV and diagnosed with advanced melanoma.

Melanoma is an aggressive cancer that arises from melanocytes (pigment cells). Cutaneous melanoma is the most common type. Additionaly, uveal and mucosal melanomas can occur in the eye and in the mucosa (such as the mouth or the vulva), respectively. The primary risk factor of cutaneous melanoma is ultraviolet light exposure. As outdoor activities are a way of life in Switzerland, melanoma incidence is high in the country [24]. The extent of the disease progression is described by the staging system of the American Joint Committee on Cancer (AJCC), 8th edition [74], with stages ranging from 0 to IV: stage 0 represents a localized and not diffused tumor; stage I includes small primary tumors that have not spread to lymph nodes; stages II and III indicate larger or more extensive primary tumors without or with, respectively, melanoma extending to lymph nodes; stage IV indicates metastatization of melanoma cells to distant organs. Surgery is the most common and resolutive approach for the lowest stages, but, when the disease is more extensive, systemic treatments such as Immunotherapy are required. In

addition, Radiotherapy can also be used as palliative or local treatment.

Specifically, Immunotherapies constitute new revolutionary treatments, but their administration has only recently entered standard of care, and the guidelines are still shifting. In addition, some of the patients included in the case study dataset were treated in clinical trials, *i.e.* outside of standard guidelines. In the performed analyses, therefore, a specific focus was given on the inspection of the patterns of care including this kind of treatments.

The study cohort includes 184 patients diagnosed with advanced melanoma between March 18th, 2008 and November 17th, 2019, with follow-up up to 2019, December 30th.[1] Data were sourced from the EHRs available at CHUV and curated by trained oncologists.

Data includes: sex, date of birth, primary tumor type (among conjunctival, cutaneous, melanoma of unknown primary, mucosal, and uveal), stage and diagnosis date, advanced tumor diagnosis date and mutation type (among BRAF-V600, BRAF-nonV600, NRAS, wild type (wt)), pharmacological treatments, and survival information (date of death or last follow-up). In this study, only the medications administered after the stage IV diagnosis were considered.

A brief description of the data is reported in Table 4.1.

4.5.2 Methods: Automatic and Supervised Process Mining Techniques

In Oncology, PM4HC was previously successfully applied to identify the most common pattern of cares for many kind of tumors, even though the purpose remained exploratory. Rectal cancer [72], gynecological cancer [131], breast cancer [46], and melanoma [168] were investigated both in terms of PD and CC: many works were addressed to measure how protocols or guidelines were respected, while the application of PD remained maily descriptive of the general trends [111]. Specifically on melanoma, further analytics approaches (similar to but not explicitly declared as process-oriented) were also applied for treatment patterns inspection [40].

In this work, we implemented the classical statistical analysis pipeline shown in Figure 4.2 by fully employing PM4HC techniques to achieve the goals of each step. To perform the analyses, we used pMineR, an open source R library implementing PM4HC functionalities [69]. By handling data in the form of ELs, this tool allows, among its features, to implement PD and CC analyses.

We started with the raw data set, which we first assumed to be *clean* from mistakes. First, we cast the data in the form of ELs, by selecting the main clinical events of interest for the analysis and defining the rules to cope with missing values. Then, we implemented a PD algorithm based on First Order Markov Models (FOMMs) [69], to provide a fast and easy-to-understand representation of the subsequent events. This representation allowed us to visually identify some unexpected links between clinical events (*e.g.* due to mistakes in some dates). With the help of a physician, we iteratively reviewed the data and rerun the PD algorithm in order to increasingly approach the expected graph and thus refine the data quality.

To describe the general statistics of the population and quantify the flux of patients though

[1] This study was approved by the Research Ethical Committee of Canton de Vaud (CER-VD) and includes only patients who did not oppose usage of their data, and was conducted according to the Swiss Federal Act on Research involving Human Beings.

Table 4.1: *Description of the cohort of advanced melanoma patients used in this work.*

Variables	**Subjects (n=184)**	**(%)**
Gender		
Females	71	(38.6)
Males	113	(61.4)
Age at primary diagnosis		
mean +- SD (years)	58.1	±16.5
Stage primary tumor		
0	1	(0.5)
I	23	(12.5)
II	44	(23.9)
III	52	(28.3)
IV	27	(14.7)
<NA>	37	(20.1)
Subtype of melanoma		
Conjunctival	1	(0.5)
Cutaneous	134	(72.8)
Melanoma of unknown primary	25	(13.6)
Mucosal	10	(5.4)
Uveal	12	(6.5)
<NA>	2	(1.1)
Age at IV stage diagnosis		
mean +- SD (years)	62.0	±15.4
Mutations		
BRAF mutated	89	(48.4)
NRAS mutated	46	(25.0)
wt	32	(17.4)
<NA>	17	(9.2)
Survival		
Alive	87	(47.3)
Dead	97	(52.7)

different patterns of cares (the second step in Figure 4.2), we exploited both PD and CC techniques. A first unsupervised PD analysis was based on the same FOMM model as described above. A following supervised CC approach was based on a pre-defined representation of the different treatment lines implemented with the Pseudo-Workflow formalism (PWF) available in the software tool. After performing both PD and CC, patients were grouped according to their paths through the graphs using the selection language provided by the tools. Finally, Kaplan-Meier survival curves and log-rank tests were used to quantify statistical differences between the

groups, considering time-to-event and OS as end-points in PD and CC, respectively.

4.5.2.1 Process Discovery

PD methods allow users to automatically mine processes based on an EL of observed events, often providing graphical visualizations [203]. Among the possible Process Discovery algorithms (see [49] for a complete review), we selected for our analysis first order Markov Models (FOMMs).

Markov Models (MMs) are stochastic models act at describing a randomly changing system that satisfies the Markov property, that is, the assumption that a system's future state only depends on its previous states in a number that can be fixed (*fixed-order MMs*, with the number being the order of the MM) or variable (*variable-order MMs*) [68]. The changes of state of the system are called transitions. When the probability of any transition is independent of time, they are named *time-homogeneous MMs* and can be visualized with a labeled directed graph, with nodes representing the states and for which the sum of the labels of any nodes' outgoing edges is 1.

In healthcare analytics, MMs are well suited for representing disease processes that evolve over time being, as in our case, the patients' paths of care modeled as sequences of transitions over a set of discrete states (events) of health [179, 192].

In this work we built first order time-homogeneous MMs, that is, we worked on the assumption that the transition probability to a state only depends upon the previous state attained by the system, with transition probabilities constant over time. In other words, we inspected the relations occurring in the EL over consecutive events. These FOMMs correspond to one of PD's most diffused process representations, named directly-follows graphs (DFGs). As a difference, in the pMineR FOMM implementation a cutoff can be applied on the maximal number of edges per pathway to reduce the "Spaghetti Effect", in which complex ELs give rise to overcrowded graphs with very long branches and few subjects in each branch [201] (see Section 4.4). Even if DFGs have some well-known limitations [204], they are very intuitive and can be helpful to share with clinicians a first representation of the data.

4.5.2.2 Conformance Checking

CC was performed by using the PWF, designing a diagram that describes the expected flow of events in terms of diagnoses, treatment lines, and survival events. Graphically, this results in a set of nodes, representing the *status* that the subjects can assume, and a set of conditions (*triggers*) which fire transitions between status [71]. This representation allows to count which triggers/status are activated while automatically running down the events of each subjects, thus capturing the population behaviours through the diagram.

4.5.3 Results: Process-Oriented Statistical Analysis

4.5.3.1 Data preprocessing

Event Log

In order to proceed with a PM analysis, the data needs to be structured as an EL, that is, sequences of records each constituted by the tuple: *ID*, *event*, *timestamp*, and, where appropriate, one or more *attributes* that describe specific characteristics of the event itself.

Therefore, we identified among the available patients' data the main events constituting their clinical history. Specifically, for each patient, we built the EL with the following events, each associated with a time stamp:

- *Primary Stage*: the primary diagnosis, with melanoma type, tumor stage at the diagnosis, and somatic mutation harboured by the tumor as attributes;
- *Stage IV*: the diagnosis of stage IV;
- *T-Begin*: the begin of a line of treatment, with the type of the given drug(s) as attribute;
- *T-End*: the end of a line of treatment, with the type of the given drug(s) as attribute;
- *Dead, Censored*: the survival information, consisting in death of the patient or in his/her last follow-up date.

The collected treatments belong to the following categories:

- *Immunotherapy (IO)*: anti-CTLA4, anti-PD1, anti-CTLA4 + anti-PD1 (in combination), or other IO;
- *Chemotherapy (Chemo)*;
- *Targeted therapy*: tyrosine kinase inhibitors (TKI), other targeted therapy (TT).

In this study, only the treatments after stage IV diagnosis were considered.

Please notice that in the EL patients who received more than one line of treatment present multiple consecutive begin-end treatment instances. The patients included in this dataset record up to 7 lines of treatments.

Missing data

During the preprocessing, we had to handle the issue of missing values in the data. Uncollected data can in general have multiple causes, from poor handwriting, missing recordings, or measurements being documented in inconsistent locations. *Ad hoc* choices have to be made when facing this issue, as for example deciding to consider for the analyses only the complete-cases records, to discard the variables with at least one missing value, to impute the gaps with plausible information, or even to select analytic tools that admit the presence of unrecorded information (see Chapter 2).

In time-oriented analyses, missing information can consist either in unrecorded events or in missing dates associated to the events themselves. The main cause of missed collection in this case was the non-availability of some data in the expected locations (like visit notes).

PM uses all and only the information made available in the given EL, thus intrinsically dealing with the missing values issue. According to the adopted preprocessing, if one or more of the above-described milestones/dates are missing in the original data, then the corresponding events were not created in the EL. This results in a shorter sequence of available events for those subjects with unrecorded information, but with no flag indicating that something is missing.

Designed to focus on treatment patterns, timing and effects, the performed analyses requires complete treatment lines (here corresponding to both the treatment beginning and the treatment end events). Therefore, we additionally chose to explicitly manage the cases of treatment lines with missing start or end event, in order to preserve the clinical information. Indeed, given a patient with one of these events missing, the choice of completely ignoring the line with partial information would have caused possible misinterpretations of the effects of the other administered treatments on survival (the sequence would have been incomplete and thus inconsistent with reality). As an alternative, imputing the missing date with a plausible time point obtained from the adjacent events would have introduced an artifact in the treatment duration. Therefore, we decided to truncate the patient EL records to the last available certain information, artificially introducing a *Censored* event before the line with missing information.

Specifically, if the beginning date of a treatment was missing, the censoring event was set to the day after the last previously recorded information (stage IV diagnosis or end of another line). The choice of adding one day in the date calculation is to avoid the conjunction of more than one event in the same time point, a circumstance that, if possible, should be avoided in time-oriented analyses to preserve the consequentiality of the events. If, instead, the missing date was the treatment end – because unrecorded or still on-going at the last follow-up – then the patient was censored the day before the corresponding treatment start date. In both cases, we set the censoring event date in such a way that any possible effect induced by the incomplete line on the outcomes is left out. As a trade-off, this introduced more and/or earlier censored patients.

Data Cleaning

To detect mistakes in the data, we adopted an iterative approach: a FOMM process was discovered and visually analyzed to detect inconsistencies on unexpected edges. Then, the data were updated and the procedure repeated until no more mistakes were found.

With this approach we revealed some previously uncaught mistakes in the data format in the input csv file, inconsistency in dates representation (*e.g.* dd/mm/yy vs dd/mm/yyyy), temporal event inversion (*e.g.* cancer treatment begin before a tumor diagnosis).

To give a practical example of detection, we report in Figure 4.3a) the FOMM resulting from an intermediate version of the dataset, where unexpected edges emerge because the beginning of the first line of treatment was erroneously dated before the stage IV diagnosis for one patient in the source data.

When reading a FOMM graph, we find on each edge the transition probability to shift from a node to another, computed over all the instances of the first node. The BEGIN and END

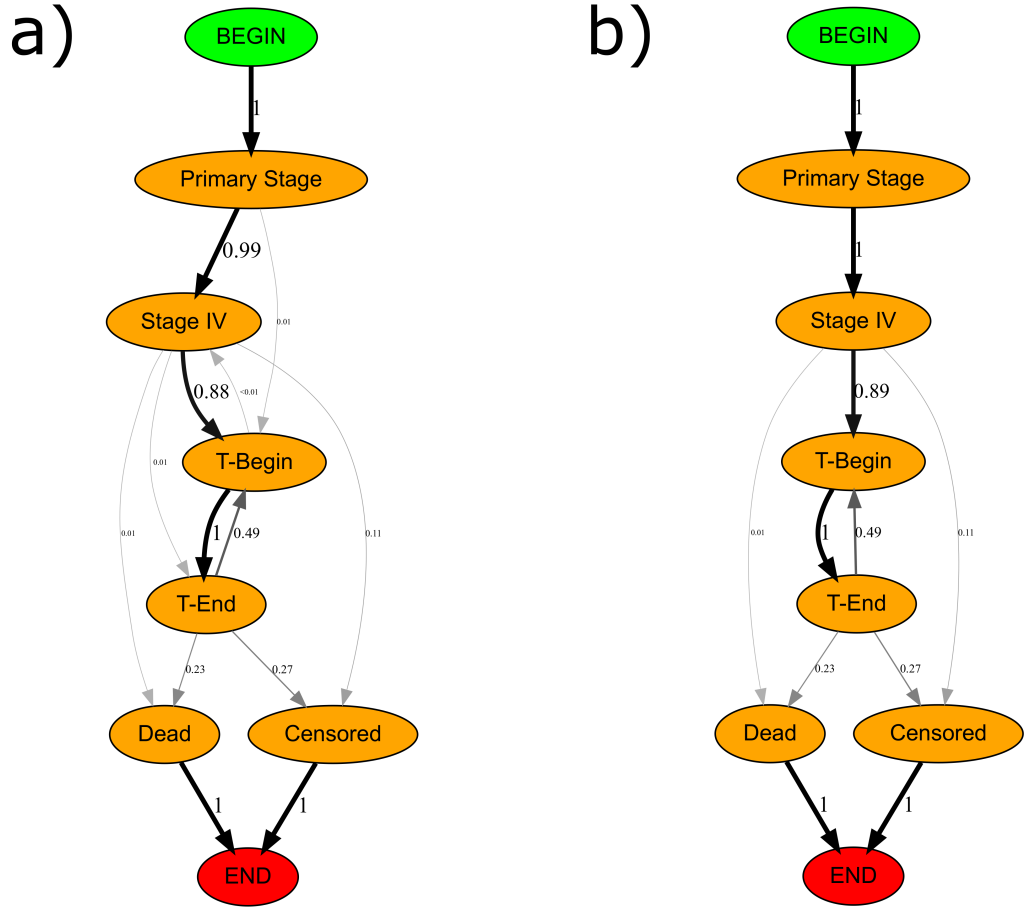


Figure 4.3: *First Order Markov Models obtained on all the events constituting the EL: a) before cleaning the information of a subject with an error in the dates, b) after data cleaning.*

nodes are artificially added to each patient's path by the algorithm. Such representation allows to describe the whole sequence of events constituting the data, depicting a general overview of the patients' most common paths. In the graph of Figure 4.3a) we can observe how the majority of the subjects gets, as expected due to the nature of the dataset, a primary stage diagnosis followed by a stage IV diagnosis, and then begins a pharmacological treatment. As previously mentioned, in this version of the dataset, because of an input error, one patient presents the begin of the first line of treatment dated before the diagnosis of stage IV (that is, his/her sequence of events is: Primary Stage $\rightarrow$ T-Begin $\rightarrow$ Stage IV $\rightarrow$ T-End $\rightarrow$ ...). An user with a proper domain (and data design) knowledge can detect this issue by observing the FOMM's graph and noticing the following unexpected edges: from Primary Stage to T-Begin, from T-Begin to Stage IV, and from Stage IV to T-End. The low transition probabilities associated with these edges are a further clue that something unusual (but, in general, not necessarily wrong) appeared in the data. Please

notice that the transition probability from T-Begin to Stage IV is even lower (< 0.01) than the other two (0.01), due to the higher total number of occurrences of the T-Begin event (one per line, with possible multiple lines per patient) with respect to Primary Stage and Stage IV (one per patient).

In Figure 4.3b) we can observe the FOMM after correction of the inaccurately collected information. This updated graph, that will be further discussed in Section 4.5.3.2, conversely presents only relations fully compliant with the nature (and the collection design) of the data.

At the end of the preprocessing, the obtained EL consists of 1196 records referred to the 184 patients, with 6 distinct events labelled as: Primary Stage, Stage IV, T-Begin, T-End, Dead, Censored.

4.5.3.2 Descriptive statistics

A first descriptive analysis was performed by querying the input EL: its structure easily allowed us to explore in the first instance cardinality, timing, and order of the administered treatments. Then, we delved into the data by using the FOMMs, to obtain agnostic data representations, and a PWF diagram, to verify the consistency of the process with respect to the expected behaviour.

Event Log querying

By analysing the EL it was possible to perform some first descriptive investigations. Specifically, we focused on the treatments administered to the patients. Considering the events of all the patients, regardless of the position in the path of care, we extracted a total of 322 administered treatments. Table 4.2 reports, for each treatment category, its absolute and relative frequency of occurrence, and its duration in terms of median and inter-quartile range (25%-75%).

By exploring the sequence of consecutive events that chronologically delineate the patients' paths of care, as structured through the preprocessing, we could also extract the possible patterns of treatment over all the population. Out of 163 patients that received at least one recorded line of treatment, we identified 49 distinct patterns of treatment sequence. The most frequent ones are reported in Table 4.3.

Table 4.2: *Occurrences and duration (in days) of the administered treatments collected in the data. The inter-quartile ranges (IQR) are computed at 25% and 75%.*

Drug category	**Occurrence (n=322)**	**(%)**	**Median (IQR) duration [days]**
TKI	76	(23.6)	122 (76.5–228.0)
anti-CTLA4 + anti-PD1	70	(21.7)	46.5 (0.0–167.8)
anti-PD1	66	(20.5)	84.0 (33.0–253.2)
anti-CTLA4	66	(20.5)	61.5 (31.0–63.0)
Chemo	29	(9.0)	44.0 (22.0–67.0)
Other IO	13	(4.0)	92.0 (22.0–203.0)
TT	2	(0.6)	461.5 (300.7–622.2)

Table 4.3: *Most frequent patterns of treatment recorded in the data. The relative frequency of occurrence is computed over the total number of patients with at least one recorded treatment.*

First line	**Second line**	**Occurrence (n=163)**	**(%)**
anti-CTLA4 + anti-PD1	-	36	(22.1)
anti-PD1	-	22	(13.5)
anti-CTLA4	-	11	(6.7)
anti-CTLA4 + anti-PD1	TKI	11	(6.7)
Chemo	anti-CTLA4	9	(5.5)
anti-CTLA4	anti-PD1	8	(4.9)
TKI	anti-CTLA4	6	(3.7)
anti-CTLA4	TKI	5	(3.1)
TKI	-	3	(1.8)

Process discovery on all the events

As introduced above, in the Data Cleaning step recursive implementations of the FOMM were performed on all the events constituting the EL until obtaining the FOMM of Figure 4.3b). When reading it, we can get a first high-level description of the clinical history evolution of the study cohort.

Since the data were designed as a collection of the treatments prescribed to cure advanced melanoma – and thus administered to patients that already got stage IV cancer diagnosed – in this graph all the lines properly follow the stage IV diagnosis chronologically. By analyzing edges and their probabilities, we can gain further information on the dataset. For instance, by observing the couple of edges between T-Begin and T-End with their transition probabilities, we can catch that a number of patients undergo a multiple number of lines. Moreover, accordingly with the EL design, all the lines are complete (transition probability from T-Begin to T-end equal to 1). Finally, patients can experience the survival event at the end of a line of treatment, or after the Stage IV diagnosis. In this latter case, the absence of recorded treatments can be due to the unavailability of any treatment information in the original data, or, for some of the censored subjects, correspond to the introduction of the "artificial" censoring events in the preprocessing when the first line resulted incomplete (see Section 4.5.3.1).

When reading FOMM graphs in general, it is important not to trip up in the interpretation of the probabilities. Let's consider for instance the edge from T-End to T-Begin: being the FOMM computed over all the consecutive first-order couples of events, 0.49 is the probability of having another line after the previous one. Thus, 0.49 should not be interpreted as the probability of having a second line of treatment after finishing the first one and, similarly, 49% is not the percentage of patients with exactly two lines. As an extreme case, the multiple lines could all belong to a same subject performing consecutive treatments, with all the others experiencing a single line. If we want to inspect the probability of passing from one specific line to another, we have to extend the analysis to a model with a higher memory order.

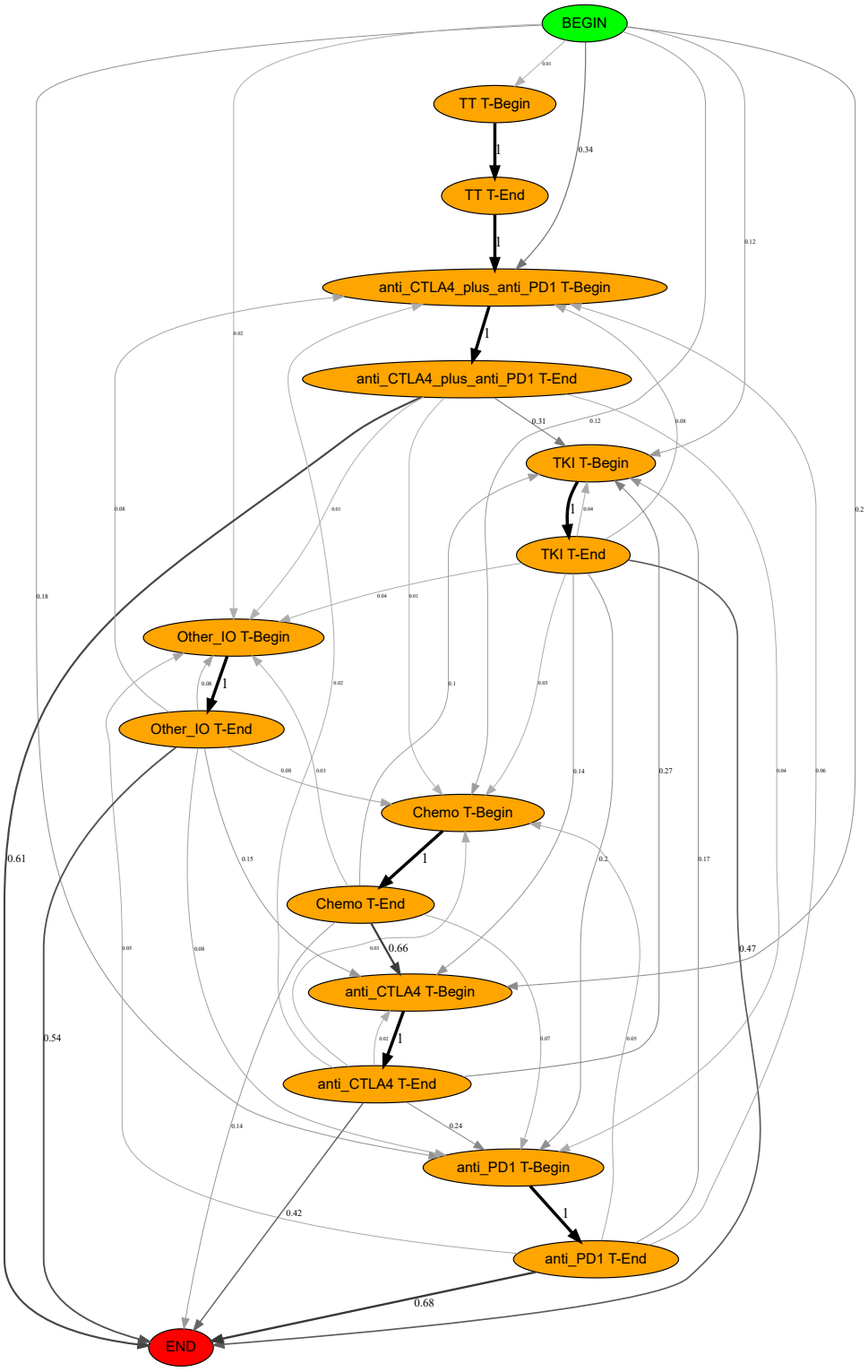


Figure 4.4: *First Order Markov Model obtained on the treatments.*

Process discovery on treatment sequences

We then built a second FOMM focusing on the administered treatments available in the EL only. Figure 4.4 reports the obtained graph.

Such a process model allows to inspect the temporal causality of the single treatments, highlighting the most frequent connections over all the population. It also provides a first overview of the position of the treatments in the paths: due to limits of the FOMMs of catching only the one-step successions, however, such considerations mainly restrict to (possibly) stand-alone treatments and to the ones sourcing in the BEGIN or terminating in the END nodes.

Conformance checking on treatment sequences

We designed a PWF able to capture the chronological order of the events: at the top, we represented the status related to the staging, and then the different treatment lines.

The PWF was designed as follows. First, since the processes are expected to present at first the events corresponding to the primary stage and stage IV diagnoses, we introduced two nodes that activates respectively when the first event is the Primary Diagnosis and when there is a Stage IV Diagnosis event after it. Then, a sequence of one or more lines of treatments is expected in the processes: for each line, we defined a set of nodes representing the treatments that turn on if the patient presents, after the previous event (fixed as Stage IV Diagnosis or end of another line), the beginning of a drug administration. A number preceding the drug category name marks the corresponding line. In order to be able to define treatment paths at different levels of granularity we added a further status for each treatment line, that is, *IO* (immunotherapy). Each time a patient presents in his/her event log an immunotherapy treatment among anti-CTLA4, anti-PD1, the combination of anti-CTLA4 and anti-PD1, or other IO, both the corresponding specific status and the node *IO* turn on. This is doable thanks to the possibility in the PWF formalism to define simultaneous activation of multiple status, thus allowing the inspection of the data at different levels of abstraction. We then introduced status to mark the end of each line of treatment, that is reached if the patient's data present a T-End event while a treatment status is active from the previous step. Finally, we introduced two additional status to catch the survival outcomes, namely *Dead* and *Censored*, that can be activated without constraints on the previous status, as soon as a survival event is read in the EL. The activation of the survival status terminates the inspection of the flow of events for that patient.

Figure 4.5 reports the result of the run of our cohort on the PWF graph. Nodes and boxes report the number of times that a status/trigger was reached/fired. For the sake of readability, the reported plot is limited to the first two lines of treatment, even if the designed PWF included all the 7 lines available in the data.

By inspecting the graph, it is possible to follow the population's paths and read the corresponding number of subjects that run specific patterns. For instance, we can observe that all the patients included in the dataset (and thus starting in the automatically-introduced BEGIN status) had a Stage IV diagnosis (expected by design), that only 163 over 184 patients had a first line recorded, followed in 89 cases by a second line, or that the most frequent first line of treatment was the combination of anti-CTLA4 and anti-PD1 with a total of 56 occurrences.

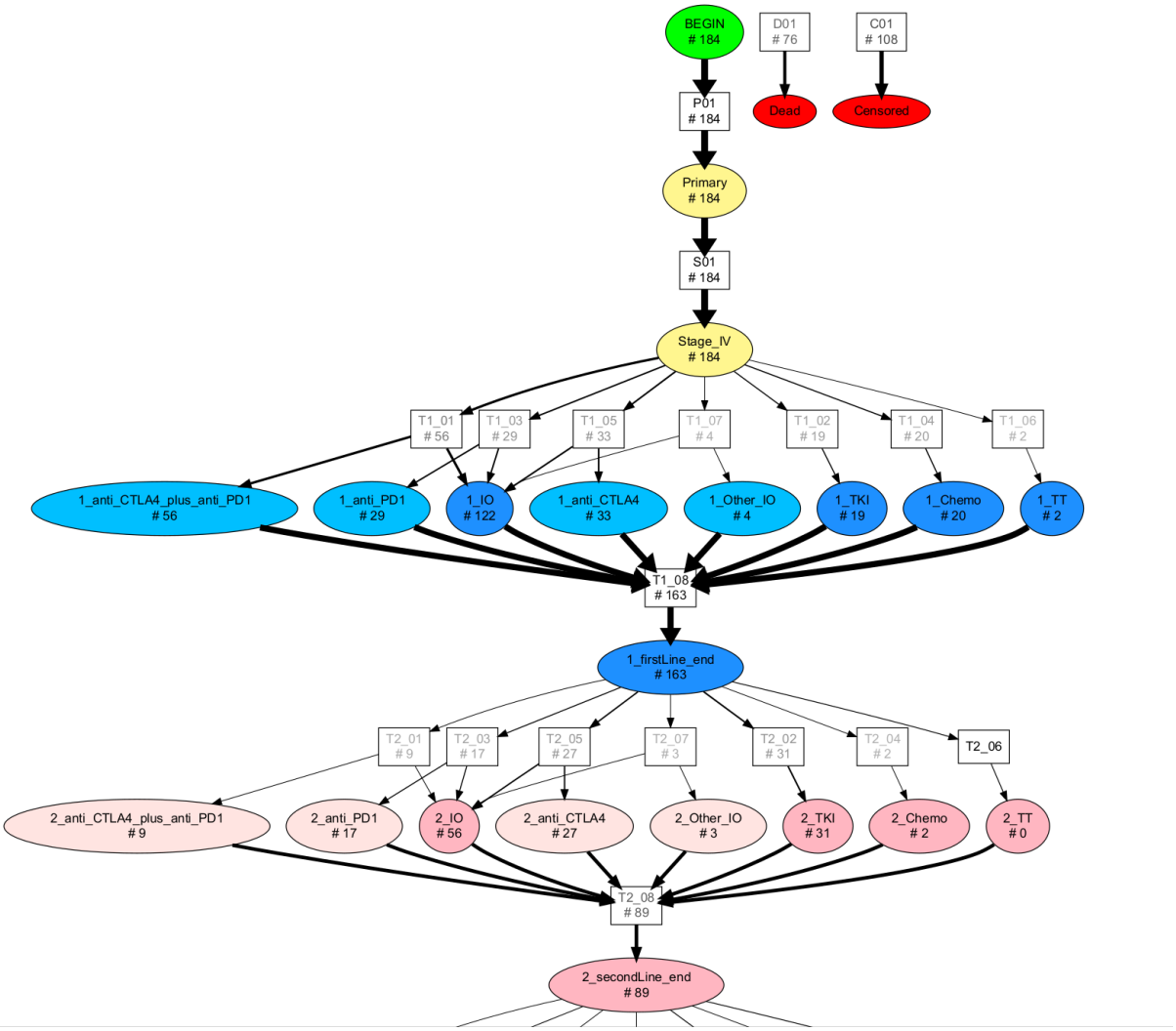


Figure 4.5: *Conformance Checking model (limited to the first two lines of treatments) reporting the status activated by the patients' processes over the used-defined PWF.*

We can also observe how the rule of double-status activation defined for the immunotherapy treatments is graphically translated into two edges outgoing from the box that represents these triggers (*e.g.* triggers T1_03 or T1_07). The presence of the IO node allows to easily count the cumulative number of subjects that experienced immunotherapy at each line, while the single drug categories maintain a higher level of detail.

The survival nodes (*Dead* and *Censored*) have been graphically separated from the others in order to limit the number of edges in the graph. However they can be reached from any point in the graph, and the available query tool can inspect at what precise point they were activated.

4.5.3.3 Inferential statistics

By exploiting the EL, the FOMM and the PWF diagrams of the previous analyses, we could easily select cohorts characterized by specific patterns of interest and perform survival analyses. While the FOMM strongly reflects (and is limited to) the events and the information present in the EL, the PWF represents an abstraction where the user has the opportunity to provide additional knowledge in the definition of the PWF status and structure itself. This enhanced semantic expressiveness is one of the main reasons why PWF was previously used in structuring Clinical Guidelines [117]. Descriptive statistics can help in suggesting hypotheses: in our case, the previous PWF and FOMM diagrams allowed to easily identify and query cohorts for statistical inference analyses. We report below two examples of the investigations we performed.

First, we inspected the relationship between type of somatic tumor mutation and time between primary and Stage IV diagnosis. Here, we consider the following mutation status: BRAF V600 mutated, BRAF non-V600 mutated, NRAS mutated, and wt. For this study, we limited the cohort to cutaneous melanoma patients, exploiting a filtering tool to easily query the EL attributes.

We implemented a survival analysis by first using the FOMM structure of Figure 4.3b) to query the path of interest (between the nodes Primary Stage and Stage IV) and obtain the time between the two events. Then, the Kaplan-Meier estimator was computed, with patients stratified by mutation status, as shown in Figure 4.6. Even if a difference between the BRAF v600 mutated and the NRAS mutated sub-cohorts seems to emerge, the log-rank test computed between all the survival distributions pairs reports no significant differences (all p-values were >0.05) for any combinations.

To demonstrate the potential of the analysis – even if in this case limited by sample cardinality – we performed a further stratification of the data, distinguishing patients by their primary stage. pMineR facilitates this step too, by allowing direct selection on the patient attributes. Figure 4.7 reports the plot of the corresponding Kaplan-Meier estimator. Even if, as expected, no statistically significant clinical evidence emerges from this analysis, mainly due to the low number of subjects per category, it is interesting to observe how rapidly this approach allows to enrich the analysis' level of detail.

The second survival analysis exploits the PWF defined in Figure 4.5. We queried the data in order to identify any differences in terms of OS based on the following patterns of interest: (1) only IO (any BRAF status), (2) IO $\rightarrow$ TKI, (3) TKI $\rightarrow$ IO, (4) only TKI. In defining the rules, we grouped together consecutive lines belonging to the same category. Patterns interspersed with TT or Chemo treatments were excluded. Upon the suggestions of clinicians, in case of sequences with multiple treatment lines, only the first occurring pattern was considered (*e.g.* a patient with IO $\rightarrow$ TKI $\rightarrow$ IO falls into the sub-cohort IO $\rightarrow$ TKI). The resulting OS survival curves are shown in Figure 4.8.

Table 4.4 reports the frequency of occurrence of each pattern, the median OS time (in years), and the percentage of patients alive at 1.5 and 3 years (CI at 95%), respectively. Statistical significance of OS differences was assessed with the log-rank test, which turned out to be significant for IO vs IO $\rightarrow$ TKI (p-value<0.0001) and IO vs TKI $\rightarrow$ IO (p-value: 0.012). The difference between IO and IO $\rightarrow$ TKI is expected because patients who receive TKI after IO are those who

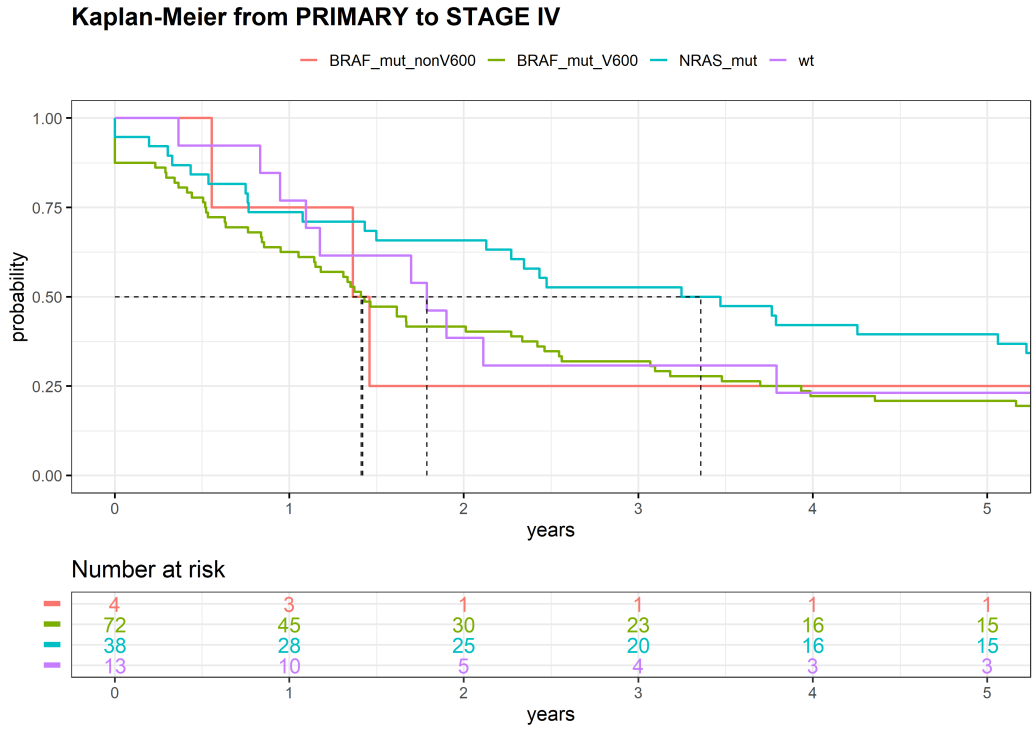


Figure 4.6: *Time-to-event analysis based on a mined FOMM: time from primary to stage IV diagnosis, stratified by mutation.*

did not respond to IO. Knowing that the benefits of TKI are usually only temporary, it is not surprising that these patients have shorter OS. The difference between IO and TKI → IO is interesting, as it may be related to recent biological findings showing that acquired resistance to TKI may hinder IO efficacy.

Table 4.4: *OS for the main treatment patterns of interest.*

Treatment path	Occurrence	Median OS [years]	1.5-year OS % (95% CI)	3-year OS % (95% CI)
all	100 %	3.87	72.7 (66.1 - 80.1)	54.9 (47.1 - 64.1)
IO	45.7 %	NA	76.9 (68.0 - 86.9)	69.4 (59.1 - 81.5)
IO → TKI	17.9 %	1.77	63 (48.3 - 82.1)	18.6 (7.7 - 45.2)
TKI → IO	8.7 %	1.92	57.4 (36.6 - 90.1)	25.1 (9.7 - 65.3)
TKI	1.6 %	1.00	0	0

4.5.4 Discussion: Applicability and Advantages of a Process-Oriented Approach to Statistical Analysis

PM4HC is expected to have an increasingly relevant role in the analysis of healthcare data. Process-oriented representations, together with tools able to query the data in terms of temporal

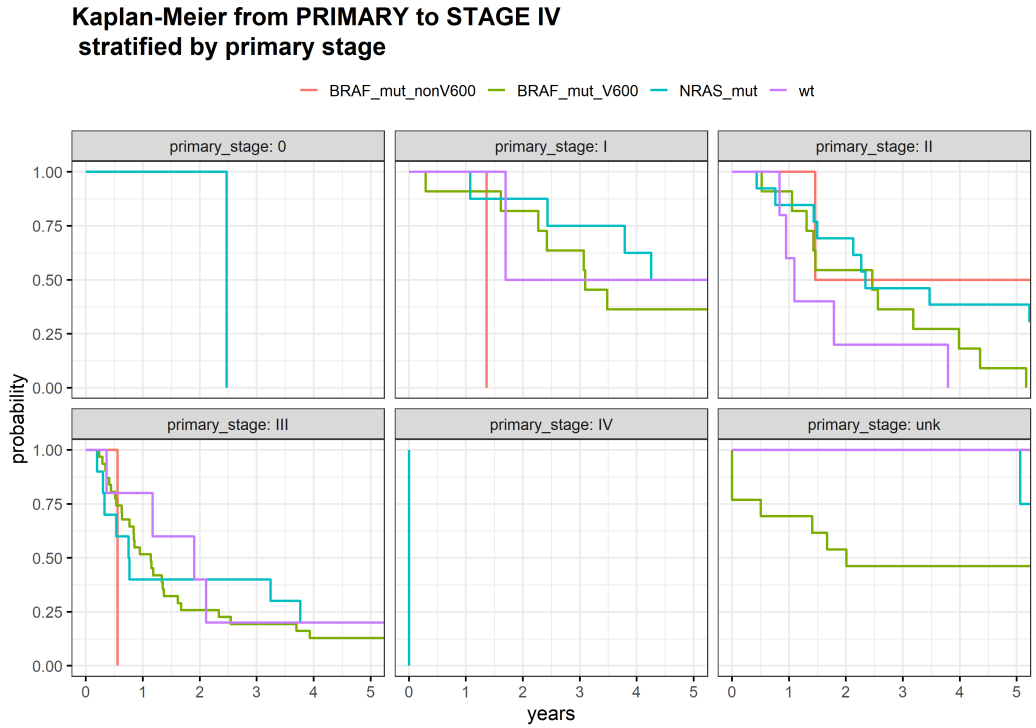


Figure 4.7: *Time-to-event analysis based on a mined FOMM: time from primary to stage IV diagnosis, stratified by mutation and type of primary.*

patterns identified through paths in a workflow, are efficient ways to easily generate clinically-relevant hypotheses and measure statistical significance, in particular in survival analysis.

In this preliminary work, we demonstrated contributions of our process-oriented approach in analyzing a real-world retrospective dataset of patients treated for advanced melanoma at the Lausanne University Hospital. Addressing the clinical questions raised by our oncologists, we integrated PM in almost all the steps of a common statistical analysis. We showed: (1) how PM can be leveraged to improve the quality of the data (data cleaning/pre-processing), (2) how PM can provide efficient data visualizations that support and/or suggest clinical hypotheses, also allowing to check the consistency between real and expected processes (descriptive statistics), and (3) how PM can assist in querying or re-expressing the data in terms of pre-defined reference workflows for testing survival differences among sub-cohorts (statistical inference).

The main remarkable points emerging from this experience are: (a) query languages for EL, PD and CC are efficient tools for data cleaning and preprocessing, by quickly identifying previously unrecognized mistakes; (b) graphical representations can promote dialogue between clinicians and data scientists, suggesting alternative perspectives and possible research questions; (c) PD gives a relevant contribute in representing the data in an agnostic way; on the other hand CC (with formalisms such as PWF) allows implementing multi-scale data abstractions and identifying patterns or inconsistencies of the data in pre-defined workflows; (d) the process representations, both in PD and CC, effectively support survival analysis techniques, allowing rapid

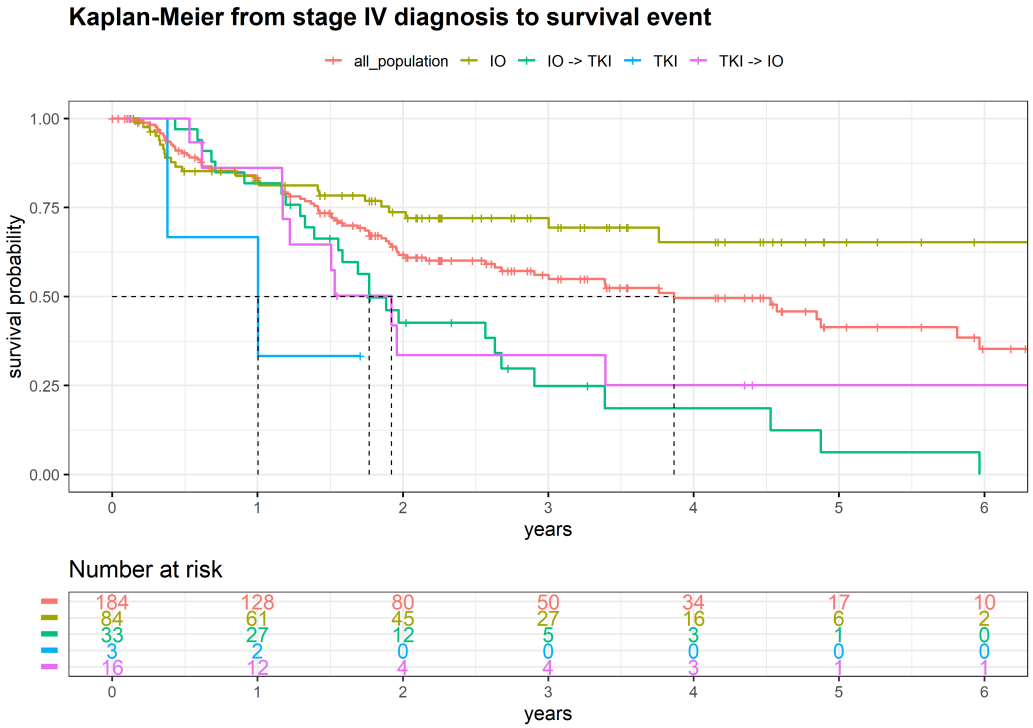


Figure 4.8: *Overall survival analysis based on a CC graph: time from stage IV diagnosis to death, stratified by treatment pattern.*

definition of sub-cohorts of interest and providing immediate statistical measures of differences between various paths of the graph.

It is interesting to observe how the missing information, although managed in both PD and CC, has a different impact on the analysis. In PD, the unsupervised built model does not suffer the presence of any possible gaps in the data, limiting to the description of relations that it actually sees in the data. It is up to the user to detect any emerging non-sense causality among the events caused by the missing information. CC, instead, focuses on testing the adherence of the processes to predetermined models: the absence of some events can cause a discrepancy between the recorded sequences and the expected ones, resulting in a decrease in the number of instances that follow the investigated paths.

With specific reference to the investigated case study, emerging clinical evidences are mainly limited by the low data availability with respect to the richness of different patterns of treatment. However, the current dataset could be expanded by including multicentric studies, in order to balance the sub-cohorts and be able to include more population attributes. Further process models could also be considered. Similar investigations may also be performed in other clinical contexts, with the overall aim to provide real-world insights into future personalized care options.

4.6 Final Remarks

Thanks to its ability to offer an alternative perspective on the events that characterize the patients' clinical history, PM4HC is assuming an emerging role in clinical data analytics.

Its process-oriented approach intrinsically allows to exploit the temporal nature of the data, providing tools to effectively depict and explore the available information. After designing and setting up a – limited, after all – preprocessing to convert the information into an EL form, PM4HC offers a portfolio of techniques to treat the available information either in an agnostic way or integrating *a priori* domain knowledge. Recent tools integrate functions to perform investigations typical of the clinical context, such as survival analysis. In addition, this approach provides important self-consistency checks for data and allows to inspect patterns at different levels of abstraction, together with the associated outcome. Moreover, the visualizations provided when implementing discovery and conformance algorithms constitute an easy-to-understand and communicative resource able to boost dialogue and discussion among the working team, suggesting or confirming worthy research directions.

In general, when considering performing a Process-Oriented approach on a new dataset, it could be useful to first assess the expected variability of the paths among the population. As mentioned in Section 4.4 indeed, PM techniques can experience some limitations when processes are intrinsically characterized by a high variability (because for instance belonging to complex clinical contexts and/or managed through heterogeneous protocols), leading to sparse model representations difficult to interpret and potentially little informative. In this cases, beside considering alternative approaches, some actions can be performed on the events' preprocessing in order to reduce the item variability, such as aggregating events by relying on clinical ontologies. Nevertheless, is also worth noticing how sometimes little represented paths can refer to very interesting group of patients, characterized for instance by unexpected outcomes or experiencing rare adverse events. There is therefore the need to assess case by case if and how PM4HC can be preferred or not to other approaches.

In the future, PM4HC has great potential to be developed further in synergy with classical analytics tools to work on healthcare-related data. In particular, the fast-growing amount of real-world clinical data produced in modern hospitals, each patient's therapeutic journey being by nature a temporal process, represents a formidable opportunity for PM4HC to contribute to the advent of precision medicine.

Chapter 5

Conclusions

Healthcare analytics is increasingly bringing improvements to medical knowledge and patient care, supporting medicine in its transformation towards personalized approaches.

In the development of clinical decision support systems, longitudinally-collected clinical data constitute an important resource for many kinds of investigations, ranging from biomarkers identification to prognosis prediction. Nevertheless, for their employment, techniques able to properly deal with the temporal dimension of data are needed. In addition, further features related to the variables' clinical and heterogeneous nature should be taken into account.

This thesis has explored under a number of aspects the potential as well as the requirements when employing this kind of data. Ranging from data preprocessing procedures to the implementation of computational models for descriptive and predictive purposes, the challenges and limitations encountered when exploring the existing methodologies have been introduced. Hence, innovative techniques developed for addressing these issues have been presented both from a methodological perspective and with practical cases of use, by performing analyses in different clinical contexts, from Neurology to Oncology.

In those cases where data usability is limited by missing values, imputation approaches can constitute a valid tool for curing the missing information. For longitudinal clinical data, imputation approaches based on the similarity assessed among visits or patients can properly exploit the data nature and informative content, as depicted in Chapter 2. In general, based on the potentially vast differences in the data, the most suitable imputation method should be evaluated on a case by case basis.

In data modeling, different procedures can be set up to approach the information collected in the temporally-evolving features: dynamic data can be summarized in derived variables coding their evolution over time, as carried out in the naïve Bayes implementation of Section 2.6.4. Although constituting a rapid access to the dynamic information, this approach is likely to oversimplify the informative richness of the data. For greater robustness, techniques such as feature selection could be employed to effectively identify the derived variables that better preserve the information, even if at an added computational cost. On the other side, models that require statically structured information such as the NB classifier, although being easy to design and rapid to implement, can also limit their prediction ability to static outcomes.

To overcome these limitations, computational approaches able to handle the dynamic nature

of the data can be employed. In Chapter 3, a DBN-based model of disease progression has been presented. Thanks to its intrinsic dynamic nature, such approach allows to fully exploit the temporal information in the data, allowing to catch how the relationships among variables change over time and influence the disease progression. Beside allowing to probabilistically characterize the outcome occurrence over time, this method provides a number of additional remarkable outcomes, such as the graphical description of the mutual relations among variables and the related CPDs, that explicitly describe how the condition evolves over the study population. On the other side, a DBN is a sort of craftwork, due to the many attentions and details it requires for its development, thus constituting a high-demanding approach in terms of time and resources.

An as alternative approach, Process Mining for Healthcare provides a number of techniques properly designed for handling clinical data with a temporal dimension. As presented in Chapter 4, the patient's clinical evolution can be modeled in terms of a sequence of events, allowing to follow the patterns of care and assess their effect on the outcomes. Even if the employment of these techniques in this thesis has mainly been exploratory, a great potential has emerged in terms of alternative to other traditional approaches when performing classical statistical analyses. Noticeably, these methodologies also provide effective visual outcomes, that both get access to the mined information and constitute a communicative mean to present the results. As a limitation, but not limited to this approach, a certain grasp on the data is required since the very first preprocessing steps, in order to identify, follow, and interpret clinical events and paths that can be complex and often require specific medical knowledge.

From a data scientist's point of view, this Ph.D. experience on healthcare analytics provided a number of useful lessons, reported below.

- First of all, real data are *complex*. The more dimensions they have, the more informative they are, but also the more challenging to handle.
- When approaching the analysis, the available data should be carefully evaluated in terms of nature, structure, and informative content, in order to identify the state-of-the-art algorithms that better fit the research questions, or design *ad hoc* methodologies for the specific case study.
- An adequate (and sometimes massive) *preprocessing* is in general required, to aggregate data from different sources, handle the possible data type heterogeneity or temporal nature, or structure them as required by the selected analysis methodology. A few techniques can assist in this process, by providing effective visualizations of the information coded in the data.
- Some *domain knowledge* is required at each analysis step, from data structuring to study design and results interpretation. With this aim, working in a proactive multidisciplinary research team allows to design and implement studies that meet the clinical needs while at the same time making the best use of the analytic tools.
- The adoption of "*non black-box*" methodologies provides a number of advantages: among these, the chance to monitor and adjust the model during its implementation phase, the

access to understanding how results are obtained, and an easier dissemination of the gained knowledge.

- Finally, *technology transfer* should be a goal of every developed tool, to ensure that research not only provides an advancement of the medical knowledge, but also, where possible, constitutes a practical mean to support clinicians and patients.

Most of these lessons have been acquired by studying the state of the art, designing and implementing the methodologies presented in this thesis, and by sharing the progresses with my working teams as well as the scientific community.

5.1 Publications

The work presented in this thesis has produced the following publications.

5.1.1 Journal Papers

- Daberdaku S*[1], Tavazzi E*, and Di Camillo B. *A Combined Interpolation and Weighted K-Nearest Neighbours Approach for the Imputation of Longitudinal ICU Laboratory Data*. Journal of Healthcare Informatics Research, pages 1–15, 2020.

- Tavazzi E*, Daberdaku S*, Vasta R, Calvo A, Chiò A, and Di Camillo B. *Exploiting Mutual Information for the Imputation of Static and Dynamic Mixed-Type Clinical Data with an Adaptive K-Nearest Neighbours Approach.* BMC Medical Informatics and Decision Making, 20(5):1–23, 2020.

5.1.2 Conference Abstract and Short Papers

- Tavazzi E, Gerard CL, Michielin O, Wicky A, Gatta R, and Cuendet MA. *Process Mining approach to statistical analysis: application to a real-world advanced melanoma dataset.* In Lecture Notes in Business Information Processing, ICPM workshops proceedings 2020.

- Gerard CL, Tavazzi E, Gatta R, Delyon J, Cuendet MA, and Michielin O. *A process mining approach to real-world advanced melanoma treatments.*, In Proc. 55th the American Society of Clinical Oncology (ASCO) Conference, 2020.

- Vasta R, Zandonà A, Daberdaku S, Tavazzi E, Nefussy B, Lunetta C, Mora G, Mandrioli J, Grisan E, Tarlarini C, Calvo A, Moglia C, Gotkine M, Drory V, Chiò A, and Di Camillo B. *Functional Impairment and Survival Prediction in Amyotrophic Lateral Sclerosis Patients: a Probabilistic Model of Disease Progression.* In European Journal of Neurology, volume 27, pages 172–172, 2020. Abstracts from the 6th European Academy of Neurology Congress

[1] * = equal contribution

- Tavazzi E, Daberdaku S, Zandonà A, Vasta R, Calvo A, Chiò A, and Di Camillo B. *An Adaptive K-Nearest Neighbours Algorithm for the Imputation of Static and Dynamic Mixed-Type Clinical Data.* In Proc. 16th International Conference on Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB), pages 285–284, 2019.

- Daberdaku S*, Tavazzi E*, and Di Camillo B. *Interpolation and K-Nearest Neighbours Combined Imputation for Longitudinal ICU Laboratory Data.* In Proc. 7th IEEE International Conference on Healthcare Informatics (ICHI), pages 550–552. IEEE Computer Society, 2019.

- Chiò A, Zandonà A, Daberdaku S, Vasta R, Nefussy B, Tavazzi E, Lunetta C, Mora G, Mandrioli J, Grisan E, Gotkine M, Calvo A, Moglia C, Drory V, and Di Camillo B. *Functional Impairment and Survival Prediction in Amyotrophic Lateral Sclerosis Patients: a Probabilistic Model of Disease Progression.* In: Proc. 50th Congress of the Italian Society of Neurology, 2019.

- Chiò A, Zandonà A, Daberdaku S, Vasta R, Nefussy B, Tavazzi E, Lunetta C, Mora G, Mandrioli J, Grisan E, Gotkine M, Calvo A, Moglia C, Drory V, and Di Camillo B. *Functional Impairment and Survival Prediction in Amyotrophic Lateral Sclerosis Patients: a Probabilistic Model of Disease Progression.* In: Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, volume 20, 2019. Issue sup1: Abstracts from the 30th International Symposium on ALS/MND.

5.2 Patents

- Di Camillo B, Zandonà A, Daberdaku S, Tavazzi E, Chiò A, Vasta R, Calvo A, Moglia C, Casale F, D'Ovidio F, Mandrioli J, Lunetta C, Drory V, Mora G, and Gotkine M. *"Method for determining the prognosis of disease progression and survival for patients affected by Amyotrophic Lateral Sclerosis"*. International Patent, Serial number PCT/IT2020/000057, filed on July 22, 2020. *Currently patent pending*.

5.3 Software projects

- Development of the R package *wkNNMI*, an adaptive Mutual Information-weighted k-NN algorithm for the imputation of static and dynamic mixed-type data. Released in January 2020, available on CRAN `https://cran.r-project.org/package=wkNNMI`.

- Development of *PD_impute*, an Interpolation and K-Nearest Neighbours combined imputation algorithm for longitudinal ICU laboratory data. Developed in the context of the 2019 ICHI Data Analytics Challenge on Missing data Imputation (DACMI). Released in May 2019, available on github `https://github.com/sebastiandaberdaku/PD_Impute`.

Appendix

Publications and Side Projects

During my Ph.D., I was involved in further projects and collaborations not included in this thesis, mainly focused on: the development a Value-Based Healthcare (VBHC) approach for delineating new procurement models at a regional level; the role and contribution of PM4HC in representing clinical guidelines; the investigation of adverse events as a result of immunotherapies in melanoma patients; digital signal processing and psychoacoustics in binaural audio rendering.

These projects have led to the following publications.

Journal Papers

- Gatta R, Vallati M, Fernandez-Llatas C, Martinez-Millana A, Orini S, Sacchi L, Lenkowicz J, Marcos M, Munoz-Gama J, Cuendet M, De Bari B, Marco-Ruiz L, Stefanini A, Valero-Ramon Z, Michielin O, Lapinskas T, Montvila A, Martin N, Tavazzi E, and Castellano M. *What Role can Process Mining play in recurrent Clinical Guidelines issues? A Position Paper*, International Journal of Environmental Research and Public Health (2020): 17(18), 6616.

- Comoretto RI, Gasparetto T, Tavazzi E, and Gregori D. *Towards Value-Based Healthcare and the Role of Regional Agencies: the Approach of the Veneto Region*, Epidemiology, Biostatistics and Public Health. 2019 Jun 21;16(2).

- Spagnol S, Tavazzi E, and Avanzini F. *Distance rendering and perception of nearby virtual sound sources with a near-field filter model*, Applied Acoustics 115, pages 61-73, 2017.

Conference Abstracts and Short Papers

- Comoretto RI*[2], Tavazzi E*, Bortolussi G, Gnoato M, and Gasparetto T. *Outlining outcomes for Transcatheter Aortic Valve Implantation (TAVI) patients: towards a definition of "value"*, International Consortium for Health Outcomes Measurement (ICHOM) Conference, (2020) [* equal contribution].

[2]* = equal contribution

- Comoretto RI*, Tavazzi E*, Gnoato M, and Gasparetto T. *Improving value in TAVI patients: insights from the Veneto Region experience*. European Health Economics Association (EuHEA) Conference, (2020).

- Ghisoni E, Wicky AM, Latifyan S, Mederos-Alfonso NN, Özdemir BC, Cuendet MA, Imbimbo M, Marandino L, Delyon J, Gerard CL, Tavazzi E, Gatta R, Valabrega G, Aglietta M, Obeid M, Coukos G, Peters S, Di Maio M, Bouchaab H, and Michielin O. *Long-lasting, irreversible and late-onset immunerelated adverse events (irAEs) from immune checkpoint inhibitors (ICIs): A real-world data analysis*. American Society of Clinical Oncology (ASCO) Conference, (2020).

- Tavazzi E*, Comoretto RI*, Gnoato M, and Gasparetto T. "Value-Based Healthcare: studio pilota della Regione Veneto in ambito cardiochirurgico". XXIV AIES Italian Health Economics Association National Conference, Pisa (2019).

Software projects

- Contribution to the R package *pMineR* v. 0.45, a tool for building and training Process Mining models in the clinical domain. Available on github `https://github.com/robertogattabs/pMiner.v045`.

Bibliography

[1] Ascent of machine learning in medicine. *Nature Materials*, 18(407), 2019.

[2] Ciamak Abkai and Jürgen Hesser. Virtual intensive care unit (ICU): real-time simulation environment applying hybrid approach using Dynamic Bayesian Networks and ODEs. *Stud. Health Technol. Inform*, 142:1–6, 2009.

[3] Yevgeniya A Abramzon, Pietro Fratta, Bryan J Traynor, and Ruth Chia. The overlapping genetics of amyotrophic lateral sclerosis and frontotemporal dementia. *Frontiers in Neuroscience*, 14:42, 2020.

[4] Rakesh Agrawal, Dimitrios Gunopulos, and Frank Leymann. Mining process models from workflow logs. In *International Conference on Extending Database Technology*, pages 467–483. Springer, 1998.

[5] Ammar Al-Chalabi, Orla Hardiman, Matthew C Kiernan, Adriano Chiò, Benjamin Rix-Brooks, and Leonard H van den Berg. Amyotrophic lateral sclerosis: moving towards a new classification system. *The Lancet Neurology*, 15(11):1182–1194, 2016.

[6] Davide Albanese, Michele Filosi, Roberto Visintainer, Samantha Riccadonna, Giuseppe Jurman, and Cesare Furlanello. minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. *Bioinformatics*, 29(3):407–408, 12 2012.

[7] Rebecca R Andridge and Roderick JA Little. A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64, 2010.

[8] Claudia Antunes. Pattern mining over nominal event sequences using constraint relaxations. *Unpublished doctoral dissertation, Instituto Superior Técnico, Lisboa*, 2005.

[9] Nazem Atassi, James Berry, Amy Shui, Neta Zach, Alexander Sherman, Ervin Sinani, Jason Walker, Igor Katsovskiy, David Schoenfeld, Merit Cudkowicz, et al. The PRO-ACT database design, initial analyses, and predictive features. *Neurology*, 83(19):1719–1725, 2014.

[10] Charles Auffray, Rudi Balling, Inês Barroso, László Bencze, Mikael Benson, Jay Bergeron, Enrique Bernal-Delgado, Niklas Blomberg, Christoph Bock, Ana Conesa, et al. Making sense of big data in health research: towards an eu action plan. *Genome medicine*, 8(1):1–13, 2016.

[11] Iman Azimi, Tapio Pahikkala, Amir M Rahmani, Hannakaisa Niela-Vilén, Anna Axelin, and Pasi Liljeberg. Missing data resilient decision-making for healthcare iot through personalization: A case study on maternal health. *Future Generation Computer Systems*, 96:297 – 308, 2019.

[12] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 2011.

[13] Rubika Balendra, Ashley Jones, Naheed Jivraj, I Nick Steen, Carolyn A Young, Pamela J Shaw, Martin R Turner, P Nigel Leigh, Ammar Al-Chalabi, UK-MND LiCALS Study Group, et al. Use of clinical staging in amyotrophic lateral sclerosis for phase 3 clinical trials. *Journal of Neurology, Neurosurgery & Psychiatry*, 86(1):45–49, 2015.

[14] Brett K Beaulieu-Jones, Daniel R Lavage, John W Snyder, Jason H Moore, Sarah A Pendergrass, and Christopher R Bauer. Characterizing and managing missing structured data in electronic health records: data analysis. *JMIR medical informatics*, 6(1):e11, 2018.

[15] Ettore Beghi, Adriano Chiò, Philippe Couratier, Jesùs Esteban, Orla Hardiman, Giancarlo Logroscino, Andrea Millul, Douglas Mitchell, Pierre-Marie Preux, Elisabetta Pupillo, et al. The epidemiology and treatment of ALS: focus on the heterogeneity of the disease and critical appraisal of therapeutic trials. *Amyotrophic Lateral Sclerosis*, 12(1):1–10, 2011.

[16] Ettore Beghi, Elisabetta Pupillo, Virginio Bonito, Paolo Buzzi, Claudia Caponnetto, Adriano Chiò, Massimo Corbo, Fabio Giannini, Maurizio Inghilleri, Vincenzo La Bella, et al. Randomized double-blind placebo-controlled trial of acetyl-L-carnitine for ALS. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 14(5-6):397–405, 2013.

[17] Melanie L Bell, Mallorie Fiero, Nicholas J Horton, and Chiu-Hsieh Hsu. Handling missing data in rcts; a review of the top medical journals. *BMC Medical Research Methodology*, 14(1):118, Nov 2014.

[18] Riccardo Bellazzi, Fulvia Ferrazzi, and Lucia Sacchi. Predictive data mining in clinical medicine: a focus on selected methods and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(5):416–430, 2011.

[19] Riccardo Bellazzi and Blaz Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, 77(2):81–97, 2008.

[20] Lorenzo Beretta and Alessandro Santaniello. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*, 16(3):74, 2016.

[21] Lorenzo Beretta and Alessandro Santaniello. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*, 16(3):74, Jul 2016.

[22] Michele Berlingerio, Francesco Bonchi, Fosca Giannotti, and Franco Turini. Mining clinical data with a temporal dimension: a case study. In *2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007)*, pages 429–436. IEEE, 2007.

[23] Benjamin Rix Brooks, Robert G Miller, Michael Swash, and Theodore L Munsat. El Escorial revisited: Revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*, 1(5):293–299, 2000.

[24] Jean-Luc Bulliard, Renato G Panizzon, and Fabio Levi. Melanoma prevention in Switzerland: where do we stand? *Revue medicale suisse*, 2(63):1122–1125, 2006.

[25] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, pages 6775–6785, 2018.

[26] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

[27] Jesse M Cedarbaum, Nancy Stambler, Errol Malta, Cynthia Fuller, Dana Hilt, Barbara Thurmond, and Arline Nakanishi. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences*, 169(1):13 – 21, 1999.

[28] Jesse M Cedarbaum, Nancy Stambler, Errol Malta, Cynthia Fuller, Dana Hilt, Barbara Thurmond, and Arline Nakanishi. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences*, 169(1):13 – 21, 1999.

[29] Eleonora Cellura, Rossella Spataro, Alfonsa Claudia Taiello, and Vincenzo La Bella. Factors affecting the diagnostic delay in amyotrophic lateral sclerosis. *Clinical neurology and neurosurgery*, 114(6):550–554, 2012.

[30] Adriano Chiò, Andrea Calvo, Cristina Moglia, Letizia Mazzini, Gabriele Mora, et al. Phenotypic heterogeneity of amyotrophic lateral sclerosis: a population based study. *Journal of Neurology, Neurosurgery & Psychiatry*, pages jnnp–2010, 2011.

[31] Adriano Chiò, Antonio Canosa, Sara Gallo, Stefania Cammarosano, Cristina Moglia, Giuseppe Fuda, Andrea Calvo, and Gabriele Mora. ALS clinical trials: do enrolled patients accurately represent the ALS population? *Neurology*, 77(15):1432–1437, 2011.

[32] Adriano Chiò, Edward R Hammond, Gabriele Mora, Virginio Bonito, and Graziella Filippini. Development and evaluation of a clinical staging system for amyotrophic lateral sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 86(1):38–44, 2015.

[33] Adriano Chio, Giancarlo Logroscino, Orla Hardiman, Robert Swingler, Douglas Mitchell, Ettore Beghi, Bryan G Traynor, Eurals Consortium, et al. Prognostic factors in als: a critical review. *Amyotrophic Lateral Sclerosis*, 10(5-6):310–323, 2009.

[34] Adriano Chiò, Letizia Mazzini, Sandra D'Alfonso, Lucia Corrado, Antonio Canosa, Cristina Moglia, Umberto Manera, Enrica Bersano, Maura Brunetti, Marco Barberis, et al. The multistep hypothesis of ALS revisited: The role of genetic mutations. *Neurology*, pages 10–1212, 2018.

[35] Adriano Chiò, Gabriele Mora, Maurizio Leone, Letizia Mazzini, D Cocito, Maria Teresa Giordana, Edo Bottacchi, Roberto Mutani, et al. Early symptom progression rate is related to ALS outcome: a prospective population-based study. *Neurology*, 59(1):99–103, 2002.

[36] Adriano Chiò, Gabriele Mora, Cristina Moglia, Umberto Manera, Antonio Canosa, Stefania Cammarosano, Antonio Ilardi, Davide Bertuzzo, Enrica Bersano, Paolo Cugnasco, Maurizio Grassano, Fabrizio Pisano, Letizia Mazzini, Andrea Calvo, for the Piemonte, and Valle d'Aosta Register for ALS (PARALS). Secular Trends of Amyotrophic Lateral Sclerosis: The Piemonte and Valle d'Aosta Register. *JAMA Neurology*, 74(9):1097–1104, 09 2017.

[37] Adriano Chiò, Alessandro Zandonà, Sebastian Daberdaku, Rosario Vasta, Beatrice Nefussy, Erica Tavazzi, Christian Lunetta, Gabriele Mora, Jessica Mandrioli, Enrico Grisan, Marc Gotkine, Andrea Calvo, Cristina Moglia, Vivian Drory, and Barbara Di Camillo. Functional impairment and survival prediction in amyotrophic lateral sclerosis patients: a probabilistic model of disease progression. In *50th Congress of the Italian Society of Neurology*, 2019.

[38] Adriano Chiò, Alessandro Zandonà, Sebastian Daberdaku, Rosario Vasta, Beatrice Nefussy, Erica Tavazzi, Christian Lunetta, Gabriele Mora, Jessica Mandrioli, Enrico Grisan, Marc Gotkine, Andrea Calvo, Cristina Moglia, Vivian Drory, and Barbara Di Camillo. Functional impairment and survival prediction in amyotrophic lateral sclerosis patients: a probabilistic model of disease progression. In *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, volume 20, 2019. Issue sup1: Abstracts from the 30th International Symposium on ALS/MND.

[39] James W Cortada, Dan Gordon, and Bill Lenihan. The value of analytics in healthcare. *IBM Institute for Business Value IBM, Global Business Service*, 2012.

[40] C Lance Cowey, Frank Xiaoqing Liu, Marley Boyd, Kathleen M Aguilar, and Clemens Krepler. Real-world treatment patterns and clinical outcomes among patients with advanced melanoma: A retrospective, community oncology-based cohort study (A STROBE-compliant article). *Medicine*, 98(28), 2019.

[41] David Crockett and Brian Eliason. What is data mining in healthcare. *HealthCatalyst,[Online]. Available: https://www. healthcatalyst. com/data-mining-in-healthcare*, 2014.

[42] Merit E Cudkowicz, Leonard H van den Berg, Jeremy M Shefner, Hiroshi Mitsumoto, Jesus S Mora, Albert Ludolph, Orla Hardiman, Michael E Bozik, Evan W Ingersoll, Donald Archibald, et al. Dexpramipexole versus placebo for patients with amyotrophic lateral sclerosis (EMPOWER): a randomised, double-blind, phase 3 trial. *The Lancet Neurology*, 12(11):1059–1067, 2013.

[43] Adam Czaplinski, Albert A Yen, and Stanley H Appel. Forced vital capacity (FVC) as an indicator of survival and disease progression in an ALS clinic population. *Journal of Neurology, Neurosurgery & Psychiatry*, 77(3):390–392, 2006.

[44] Sebastian Daberdaku, Erica Tavazzi, and Barbara Di Camillo. Interpolation and K-Nearest Neighbours Combined Imputation for Longitudinal ICU Laboratory Data. In *The Seventh IEEE International Conference on Healthcare Informatics (ICHI)*, pages 550–552. IEEE Computer Society, 2019.

[45] Sebastian Daberdaku, Erica Tavazzi, and Barbara Di Camillo. A Combined Interpolation and Weighted K-Nearest Neighbours Approach for the Imputation of Longitudinal ICU Laboratory Data. *Journal of Healthcare Informatics Research*, pages 1–15, 2020.

[46] Arianna Dagliati, Lucia Sacchi, Carlo Cerra, Paola Leporati, Pasquale De Cata, Luca Chiovato, John H Holmes, and Riccardo Bellazzi. Temporal data mining and process mining techniques to identify cardiovascular risk-associated clinical pathways in type 2 diabetes patients. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 240–243. IEEE, 2014.

[47] Anindya Datta. Automating the discovery of as-is business process models: Probabilistic and algorithmic approaches. *Information Systems Research*, 9(3):275–301, 1998.

[48] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, 6(2):94–98, 2019.

[49] Jochen De Weerdt, Manu De Backer, Jan Vanthienen, and Bart Baesens. A multi-dimensional quality assessment of state-of-the-art process discovery algorithms using real-life event logs. *Information Systems*, 37(7):654–676, 2012.

[50] Barbara Di Camillo, Alessandro Zandonà, Sebastian Daberdaku, Erica Tavazzi, Adriano Chiò, Rosario Vasta, Andrea Calvo, Cristina Moglia, Federico Casale, Fabrizio D'Ovidio, Jessica Mandrioli, Christian Lunetta, Vivian Drory, Gabriele Mora, and Marc Gotkine. Method for determining the prognosis of disease progression and survival for patients affected by amyotrophic lateral sclerosis, July 22 2020. PCT patent PCT/IT2020/000057.

[51] Pedro C Diniz and Diogo R Ferreira. Automatic extraction of process control flow from i/o operations. In *International Conference on Business Process Management*, pages 342–357. Springer, 2008.

[52] Ivo D Dinov. Volume and value of big healthcare data. *Journal of medical statistics and informatics*, 4, 2016.

[53] A Rogier T Donders, Geert JMG van der Heijden, Theo Stijnen, and Karel GM Moons. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10):1087 – 1091, 2006.

[54] Marlon Dumas, Wil MP van der Aalst, and Arthur H Ter Hofstede. *Process-aware information systems: bridging people and software through process technology*. John Wiley & Sons, 2005.

[55] Christo El Morr and Hossam Ali-Hassan. Healthcare analytics applications. In *Analytics in Healthcare: A Practical Introduction*, pages 57–70. Springer International Publishing, Cham, 2019.

[56] Marwa Elamin, Peter Bede, Anna Montuschi, Niall Pender, Adriano Chio, and Orla Hardiman. Predicting prognosis in amyotrophic lateral sclerosis: a simple algorithm. *Journal of neurology*, 262(6):1447–1454, 2015.

[57] Konstantinos P Exarchos, Clara Carpegianni, Georgios Rigas, Themis P Exarchos, Federico Vozzi, Antonis Sakellarios, Paolo Marraccini, Katerina Naka, Lambros Michalis, Oberdan Parodi, et al. A multiscale approach for modeling atherosclerosis progression. *IEEE journal of biomedical and health informatics*, 19(2):709–719, 2015.

[58] Ton Fang, Ahmad Al Khleifat, Daniel R Stahl, Claudia Lazo La Torre, Caroline Murphy, Uk-Mnd LicalS, Carolyn Young, Pamela J Shaw, P Nigel Leigh, and Ammar Al-Chalabi. Comparison of the King's and MiToS staging systems for ALS. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 18(3-4):227–232, 2017. PMID: 28054828.

[59] Luis Fernández-Luque and Teresa Bau. Health and social media: perfect storm of information. *Healthcare informatics research*, 21(2):67–73, 2015.

[60] Raffaele Ferrari, Dimitrios Kapogiannis, E D Huey, and PFTD Momeni. FTD and ALS: a tale of two diseases. *Current Alzheimer Research*, 8(3):273–294, 2011.

[61] Christina Fournier and Jonathan D Glass. Modeling the course of amyotrophic lateral sclerosis. *Nature biotechnology*, 33(1):45, 2015.

[62] Kevin D Foust, Desirée L Salazar, Shibi Likhite, Laura Ferraiuolo, Dara Ditsworth, Hristelina Ilieva, Kathrin Meyer, Leah Schmelzer, Lyndsey Braun, Don W Cleveland, et al. Therapeutic aav9-mediated suppression of mutant sod1 slows disease progression and extends survival in models of inherited als. *Molecular Therapy*, 21(12):2148–2159, 2013.

[63] Franco Franchignoni, Jessica Mandrioli, Andrea Giordano, Salvatore Ferro, and ERRALS Group. A further Rasch study confirms that ALSFRS-R does not conform to fundamental measurement requirements. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 16(5-6):331–337, 2015.

[64] Franco Franchignoni, Gabriele Mora, Andrea Giordano, Paolo Volanti, and Adriano Chiò. Evidence of multidimensionality in the ALSFRS-R Scale: a critical appraisal on its measurement properties using Rasch analysis. *J Neurol Neurosurg Psychiatry*, 84(12):1340–1345, 2013.

[65] Alberto Franzin, Francesco Sambo, and Barbara Di Camillo. bnstruct: an R package for Bayesian Network structure learning in the presence of missing data. *Bioinformatics*, 33(8):1250–1252, 2017.

[66] Nelson B Freimer and David C Mohr. Integrating behavioural health tracking in human genetics research. *Nature Reviews Genetics*, 20(3):129–130, 2019.

[67] Sullivan Frost. Drowning in big data? reducing information technology complexities and costs for healthcare organizations, 2015.

[68] Paul A Gagniuc. *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons, 2017.

[69] Roberto Gatta, Jacopo Lenkowicz, Mauro Vallati, Eric Rojas, Andrea Damiani, Lucia Sacchi, Berardino De Bari, Arianna Dagliati, Carlos Fernandez-Llatas, Matteo Montesi, et al. pMineR: an innovative R library for performing process mining in medicine. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 351–355. Springer, 2017.

[70] Roberto Gatta, Mauro Vallati, Carlos Fernandez-Llatas, Antonio Martinez-Millana, Stefania Orini, Lucia Sacchi, Jacopo Lenkowicz, Mar Marcos, Jorge Munoz-Gama, Michel A. Cuendet, Berardino de Bari, Luis Marco-Ruiz, Alessandro Stefanini, Zoe Valero-Ramon, Olivier Michielin, Tomas Lapinskas, Antanas Montvila, Niels Martin, Erica Tavazzi, and Maurizio Castellano. What role can process mining play in recurrent clinical guidelines issues? a position paper. *International Journal of Environmental Research and Public Health*, 17(18), 2020.

[71] Roberto Gatta, Mauro Vallati, Jacopo Lenkowicz, Eric Rojas, Andrea Damiani, Lucia Sacchi, Berardino De Bari, Arianna Dagliati, Carlos Fernandez-Llatas, Matteo Montesi, et al. Generating and comparing knowledge graphs of medical processes using pminer. In *Proceedings of the Knowledge Capture Conference*, K-CAP 2017, New York, NY, USA, 2017. Association for Computing Machinery.

[72] Gijs Geleijnse, Himalini Aklecha, Mark Vroling, Rob Verhoeven HA, Felice van Erning N, Pauline Vissers A, Joos Buijs CAM, and Xander Verbeek A. Using process mining to evaluate colon cancer guideline adherence with cancer registry data: a case study. In *AMIA*, 2018.

[73] Camille Lea Gerard, Erica Tavazzi, Roberto Gatta, Julie Delyon, Michel A Cuendet, and Olivier Michielin. A process mining approach to real-world advanced melanoma treatments. In *Proceedings of the 55th the American Society of Clinical Oncology (ASCO) Conference*, 2020.

[74] Jeffrey E Gershenwald, Richard A Scolyer, Kenneth R Hess, Vernon K Sondak, Georgina V Long, Merrick I Ross, Alexander J Lazar, Mark B Faries, John M Kirkwood, Grant A McArthur, et al. Melanoma staging: evidence-based changes in the american joint committee on cancer eighth edition cancer staging manual. *CA: a cancer journal for clinicians*, 67(6):472–492, 2017.

[75] Mahdi Ghasemi and Daniel Amyot. Process mining in healthcare: a systematised literature review. *International Journal of Electronic Healthcare*, 9(1):60–88, 2016.

[76] Jonathan D Glass. New drugs for ALS: How do we get there? *Experimental Neurology*, 233(1):112 – 117, 2012. Special Issue: Stress and neurological disease.

[77] NJ Gogtay and UM Thatte. Survival analysis. *Journal of The Association of Physicians of India*, 65:80–84, May 2017.

[78] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, 101(23):e215–e220, June 2000.

[79] Martin Gorges, Pauline Vercruysse, Hans-Peter Müller, Hans-Jürgen Huppertz, Angela Rosenbohm, Gabriele Nagel, Patrick Weydt, Åsa Petersén, Albert C Ludolph, Jan Kassubek, et al. Hypothalamic atrophy is related to body mass index and age at onset in amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry*, 88(12):1033–1041, 2017.

[80] John W Graham. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60(1):549–576, 2009.

[81] Sander Greenland and William D Finkle. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American journal of epidemiology*, 142(12):1255–1264, 1995.

[82] Joel Grus. *Data Science from Scratch: First Principles with Python*. O'Reilly Media, 2019.

[83] Christian W Günther and Wil MP van der Aalst. Fuzzy mining–adaptive process simplification based on multi-perspective metrics. In *International conference on business process management*, pages 328–343. Springer, 2007.

[84] David J Hand and Keming Yu. Idiot's Bayes–not so stupid after all? *International Statistical Review*, 69(3):385–398, 2001.

[85] Orla Hardiman, Ammar Al-Chalabi, Carol Brayne, Ettore Beghi, Leonard H van den Berg, Adriano Chiò, Sarah Martin, Giancarlo Logroscino, and James Rooney. The changing picture of amyotrophic lateral sclerosis: lessons from European registers. *Journal of Neurology, Neurosurgery & Psychiatry*, 88(7):557–563, 2017.

[86] Orla Hardiman, Ammar Al-Chalabi, Adriano Chiò, Emma M Corr, Giancarlo Logroscino, Wim Robberecht, Pamela J Shaw, Zachary Simmons, and Leonard H Van Den Berg. Amyotrophic lateral sclerosis. *Nature Reviews Disease Primers*, 3:17071, 2017.

[87] Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.

[88] Payam Homayounfar. Process mining challenges in hospital information systems. In *2012 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 1135–1140. IEEE, 2012.

[89] James Honaker, Gary King, and Matthew Blackwell. Amelia II: A Program for Missing Data. *Journal of Statistical Software*, 45(7):1–47, 2011.

[90] Tomoaki Hori, David Montcho, Clement Agbangla, Kaworu Ebana, Koichi Futakuchi, and Hiroyoshi Iwata. Multi-task gaussian process for imputing missing data in multi-trait and multi-environment trials. *Theoretical and Applied Genetics*, 129(11):2101–2115, Nov 2016.

[91] Nicholas J Horton and Ken P Kleinman. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1):79–90, 2007.

[92] James Hrastelj and Neil P Robertson. Ice bucket challenge bears fruit for amyotrophic lateral sclerosis. *Journal of neurology*, 263(11):2355–2357, 2016.

[93] Mark HB Huisman, Sonja W de Jong, Perry TC van Doormaal, Stephanie S Weinreich, H Jurgen Schelhaas, Anneke J van der Kooi, Marianne de Visser, Jan H Veldink, and Leonard H van den Berg. Population based epidemiology of amyotrophic lateral sclerosis using capture–recapture methodology. *Journal of Neurology, Neurosurgery & Psychiatry*, 82(10):1165–1170, 2011.

[94] Md Saiful Islam, Md Mahmudul Hasan, Xiaoyi Wang, Hayley D Germack, et al. A systematic review on healthcare analytics: application and theoretical perspective of data mining. In *Healthcare*, volume 6, page 54. Multidisciplinary Digital Publishing Institute, 2018.

[95] Md Saiful Islam, Md Mahmudul Hasan, Xiaoyi Wang, Hayley D Germack, and Md Noor-E-Alam. A systematic review on healthcare analytics: Application and theoretical perspective of data mining. *Healthcare*, 6(2), 2018.

[96] Evans R James. Business analytics: Methods, models and decisions, 2013.

[97] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[98] Uzay Kaymak, Ronny Mans, Tim van de Steeg, and Meghan Dierks. On process mining in health care. In *2012 IEEE international conference on Systems, Man, and Cybernetics (SMC)*, pages 1859–1864. IEEE, 2012.

[99] Basel Kayyali, David Knott, and Steve Van Kuiken. The big-data revolution in us health care: Accelerating value and innovation. *Mc Kinsey & Company*, 2(8):1–13, 2013.

[100] Mohamed Khalifa. Health analytics types, functions and levels: A review of literature. In *ICIMTH*, pages 137–140, 2018.

[101] Mohamed Khalifa and Ibrahim Zabani. Utilizing health analytics in improving the performance of healthcare services: A case study on a tertiary care hospital. *Journal of Infection and Public Health*, 9(6):757–765, 2016.

[102] Evangelos Kiskinis, Jackson Sandoe, Luis A Williams, Gabriella L Boulting, Rob Moccia, Brian J Wainger, Steve Han, Theodore Peng, Sebastian Thams, Shravani Mikkilineni, et al. Pathways disrupted in human ALS motor neurons identified through genetic correction of mutant SOD1. *Cell Stem Cell*, 14(6):781 – 795, 2014.

[103] Uffe Kjaerulff. A computational scheme for reasoning in dynamic probabilistic networks. In *Uncertainty in Artificial Intelligence*, pages 121–129. Elsevier, 1992.

[104] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[105] Markus Kraemer, Melanie Buerger, and Peter Berlit. Diagnostic problems and delay of diagnosis in amyotrophic lateral sclerosis. *Clinical neurology and neurosurgery*, 112(2):103–105, 2010.

[106] Clemens Scott Kruse, Rishi Goswamy, Yesha Jayendrakumar Raval, and Sarah Marawi. Challenges and opportunities of big data in health care: a systematic review. *JMIR medical informatics*, 4(4):e38, 2016.

[107] Stephan P Kudyba. *Healthcare informatics: improving efficiency and productivity*. CRC Press, 2010.

[108] Robert Kueffner, Neta Zach, Maya Bronfeld, Raquel Norel, Nazem Atassi, Venkat Balagurusamy, Barbara di Camillo, Adriano Chiò, Merit Cudkowicz, Donna Dillenberger, et al. Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. *bioRxiv*, page 294231, 2018.

[109] Robert Küffner, Neta Zach, Maya Bronfeld, Raquel Norel, Nazem Atassi, Venkat Balagurusamy, Barbara Di Camillo, Adriano Chiò, Merit Cudkowicz, Donna Dillenberger, et al. Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. *Scientific reports*, 9(1):690, 2019.

[110] Robert Küffner, Neta Zach, Raquel Norel, Johann Hawe, David Schoenfeld, Liuxia Wang, Guang Li, Lilly Fang, Lester Mackey, Orla Hardiman, et al. Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nature biotechnology*, 33(1):51, 2015.

[111] Angelina Prima Kurniati, Owen Johnson, David Hogg, and Geoff Hall. Process mining in oncology: A literature review. In *2016 6th International Conference on Information Communication and Management (ICICM)*, pages 291–297. IEEE, 2016.

[112] Guntur Kusuma, Marlous Hall, and Owen Johnson. Process mining in cardiology: A literature review. *Int. J. Biosci. Biochem. Bioinform.*, 8:226–236, 10 2018.

[113] Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001.

[114] Martin LANGab, Thomas Bürkle, Susanne Laumann, and Hans-Ulrich Prokosch. Process mining for clinical workflows: challenges and current limitations. In *EHealth beyond the horizon: Get it there: Proceedings of MIE2008 the XXIst international congress of the european federation for medical informatics*, page 229, 2008.

[115] Nicole Lazar. The big picture: big data computing. *Chance*, 26(2):28–32, 2013.

[116] Timothée Lenglet, Lucette Lacomblez, Jean-Louis Abitbol, Albert Ludolph, Jesús S Mora, Wim Robberecht, Pamela J Shaw, Rebecca M Pruss, Valerie Cuvier, Vincent Meininger, et al. A phase II-III trial of olesoxime in subjects with amyotrophic lateral sclerosis. *European journal of neurology*, 21(3):529–536, 2014.

[117] Jacopo Lenkowicz, Roberto Gatta, Carlotta Masciocchi, Calogero Casà, Francesco Cellini, Andrea Damiani, Nicola Dinapoli, and Vincenzo Valentini. Assessing the conformity to clinical guidelines in oncology: An example for the multidisciplinary management of locally advanced colorectal cancer treatment. *Management Decision*, 56(10):2172–2186, October 2018.

[118] Richard Lenz and Manfred Reichert. It support for healthcare processes–premises, challenges, perspectives. *Data & Knowledge Engineering*, 61(1):39–58, 2007.

[119] Jake Lever, Martin Krzywinski, and Naomi Altman. Points of significance: model selection and overfitting, 2016.

[120] Shan Li, Liying Kang, and Xing-Ming Zhao. A survey on evolutionary algorithm based hybrid intelligence in bioinformatics. *BioMed research international*, 2014, 2014.

[121] Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, and Xiaojie Yuan. Multivariate time series imputation with generative adversarial networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1603–1614. Curran Associates Inc., 2018.

[122] Yuan Luo. Missing Data Imputation For Longitudinal ICU Laboratory Test Data, 2019.

[123] Yuan Luo, Peter Szolovits, Anand S Dighe, and Jason M Baron. 3D-MICE: integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data. *Journal of the American Medical Informatics Association*, 25(6):645–653, 11 2017.

[124] Yuan Luo, Yu Xin, Rohit Joshi, Leo Celi, and Peter Szolovits. Predicting ICU mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 42–50. AAAI Press, 2016.

[125] Linh Thao Ly, Stefanie Rinderle, Peter Dadam, and Manfred Reichert. Mining staff assignment rules from event-based data. In *International Conference on Business Process Management*, pages 177–190. Springer, 2005.

[126] Ian RA Mackenzie, Rosa Rademakers, and Manuela Neumann. TDP-43 and FUS in amyotrophic lateral sclerosis and frontotemporal dementia. *The Lancet Neurology*, 9(10):995–1007, 2010.

[127] Elisa Majounie, Alan E Renton, Kin Mok, Elise GP Dopper, Adrian Waite, Sara Rollinson, Adriano Chiò, Gabriella Restagno, Nayia Nicolaou, Javier Simon-Sanchez, et al. Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. *The Lancet Neurology*, 11(4):323–330, 2012.

[128] Jessica Mandrioli, Sara Biguzzi, Carlo Guidi, Elisabetta Sette, Emilio Terlizzi, Alessandro Ravasio, Mario Casmiro, Fabrizio Salvi, Rocco Liguori, Romana Rizzi, et al. Heterogeneity in ALSFRS-R decline and survival: a population-based study in Italy. *Neurological Sciences*, 36(12):2243–2252, 2015.

[129] Jessica Mandrioli, Sara Biguzzi, Carlo Guidi, Elisabetta Venturini, Elisabetta Sette, Emilio Terlizzi, Alessandro Ravasio, Mario Casmiro, Fabrizio Salvi, Rocco Liguori, et al. Epidemiology of amyotrophic lateral sclerosis in emilia romagna region (italy): A population based study. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 15(3-4):262–268, 2014.

[130] Ronny Mans, Helen Schonenberg, Giorgio Leonardi, Silvia Panzarasa, Anna Cavallini, Silvana Quaglini, and Wil MP van der Aalst. Process mining techniques: an application to stroke care. In *MIE*, volume 136, pages 573–578, 2008.

[131] Ronny S Mans, Helen Schonenberg, Minseok Song, Wil MP van der Aalst, and Piet JM Bakker. Application of process mining in healthcare - a case study in a dutch hospital. In *BIOSTEC*, 2008.

[132] Ronny S Mans, Wil MP van der Aalst, and Rob JB Vanwersch. *Process mining in healthcare: evaluating and exploiting operational healthcare processes*. Springer, 2015.

[133] Ronny S Mans, Wil MP van der Aalst, Rob JB Vanwersch, and Arnold J Moleman. Process mining in healthcare: Data challenges when answering frequently posed questions. In *Process Support and Knowledge Representation in Health Care*, pages 140–153. Springer, 2012.

[134] Simone Marini, Emanuele Trifoglio, Nicola Barbarini, Francesco Sambo, Barbara Di Camillo, Alberto Malovini, Marco Manfrini, Claudio Cobelli, and Riccardo Bellazzi. A Dynamic Bayesian Network model for long-term simulation of clinical complications in type 1 diabetes. *Journal of biomedical informatics*, 57:369–376, 2015.

[135] Antonio Martinez-Millana, Aroa Lizondo, Roberto Gatta, Salvador Vera, Vicente Traver Salcedo, and Carlos Fernandez-Llatas. Process mining dashboard in operating rooms: Analysis of staff expectations with analytic hierarchy process. *International Journal of Environmental Research and Public Health*, 16(2):199, 2019.

[136] Yuya Matsue, Peter van der Meer, Kevin Damman, Marco Metra, Christopher M O'connor, Piotr Ponikowski, John R Teerlink, Gad Cotter, Beth Davison, John G Cleland, et al. Blood urea nitrogen-to-creatinine ratio in the general population and in patients with acute heart failure. *Heart*, 103(6):407–413, 2017.

[137] Pamela A McCombe and Robert D Henderson. Effects of gender in amyotrophic lateral sclerosis. *Gender medicine*, 7(6):557–570, 2010.

[138] Pamela A McCombe and Robert D Henderson. Effects of gender in amyotrophic lateral sclerosis. *Gender medicine*, 7(6):557–570, 2010.

[139] J Michael McGinnis, LeighAnne Olsen, W Alexander Goolsby, Claudia Grossmann, et al. *Clinical data as the basic staple of health learning: Creating and protecting a public good: Workshop summary*. National Academies Press, 2011.

[140] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[141] Nishita Mehta and Anil Pandit. Concurrence of big data analytics and healthcare: A systematic review. *International journal of medical informatics*, 114:57–65, 2018.

[142] Lisa Meng, Amy Bian, Scott Jordan, Andrew Wolff, Jeremy M. Shefner, and Jinsy Andrews. Profile of medical care costs in patients with amyotrophic lateral sclerosis in the Medicare programme and under commercial insurance. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 19(1-2):134–142, 2018.

[143] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2019. R package version 1.7-2.

[144] Patrick E Meyer. infotheo: Information-Theoretic Measures. R package version 1.2. 0. 2014.

[145] Theophano Mitsa. *Temporal data mining*. CRC Press, 2010.

[146] Sarah Mizielinska and Adrian M Isaacs. C9orf72 amyotrophic lateral sclerosis and frontotemporal dementia: gain or loss of function? *Current opinion in neurology*, 27(5):515, 2014.

[147] Steffen Moritz and Thomas Bartz-Beielstein. imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, 9(1):207–218, 2017.

[148] Kevin Patrick Murphy and Stuart Russell. Dynamic bayesian networks: representation, inference and learning. 2002.

[149] Natalie A Murphy, Karissa C Arthur, Pentti J Tienari, Henry Houlden, Adriano Chiò, and Bryan J Traynor. Age-related penetrance of the C9orf72 repeat expansion. *Scientific reports*, 7(1):2116, 2017.

[150] Mark A Musen and Jan H van Bemmel. *Handbook of medical informatics*. Bohn Stafleu Van Loghum Houten, the Netherlands, 1997.

[151] Thomas Dyhre Nielsen and Finn Verner Jensen. *Bayesian networks and decision graphs*. Springer Science & Business Media, 2009.

[152] Mei-Lyn Ong, Pei Fang Tan, and Joanna D Holbrook. Predicting functional decline and survival in amyotrophic lateral sclerosis. *PloS one*, 12(4):e0174925, 2017.

[153] Orna O'Toole, Bryan J Traynor, Paul Brennan, Colm Sheehan, Eithne Frost, Bernie Corr, and Orla Hardiman. Epidemiology and clinical features of amyotrophic lateral sclerosis in ireland between 1995 and 2004. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(1):30–32, 2008.

[154] Chao-Ying Joanne Peng, Michael Harwell, Show-Mann Liou, Lee H Ehman, et al. Advances in missing data methods and implications for educational research. *Real data analysis*, 3178, 2006.

[155] Adam Perer. Healthcare analytics for clinical and non-clinical settings. In *Proceedings of CHI Conference*, 2012.

[156] Stephen R Pfohl, Renaid B Kim, Grant S Coan, and Cassie S Mitchell. Unraveling the complexity of amyotrophic lateral sclerosis survival prediction. *Frontiers in Neuroinformatics*, 12:36, 2018.

[157] Charles M Poser, Donald W Paty, Labe Scheinberg, W Ian McDonald, Floyd A Davis, George C Ebers, Kenneth P Johnson, William A Sibley, Donald H Silberberg, and Wallace W Tourtellotte. New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 13(3):227–231, 1983.

[158] John W Pratt. Remarks on zeros and ties in the wilcoxon signed rank procedures. *Journal of the American Statistical Association*, 54(287):655–667, 1959.

[159] John Preskill. Stephen Hawking (1942–2018). *Science*, 360(6385):156–156, 2018.

[160] Llorenç Quintó, John J Aponte, Clara Menéndez, Jahit Sacarlal, Pedro Aide, Mateu Espasa, Inacio Mandomando, Caterina Guinovart, Eusebio Macete, Rosmarie Hirt, et al. Relationship between haemoglobin and haematocrit in the definition of anaemia. *Tropical Medicine & International Health*, 11(8):1295–1302, 2006.

[161] Wullianallur Raghupathi et al. Data mining in health care. *Healthcare informatics: improving efficiency and productivity*, 211:223, 2010.

[162] Wullianallur Raghupathi and Viju Raghupathi. An overview of health analytics. *J Health Med Informat*, 4(132):2, 2013.

[163] Carl Edward Rasmussen. *Gaussian Processes in Machine Learning*, pages 63–71. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[164] Evan L Ray, Jing Qian, Regina Brecha, Muredach P Reilly, and Andrea S Foulkes. Stochastic imputation for integrated transcriptome association analysis of a longitudinally measured trait. *Statistical Methods in Medical Research*, page 0962280219852720, 2019. PMID: 31172883.

[165] Álvaro Rebuge and Diogo R Ferreira. Business process analysis in healthcare environments: A methodology based on process mining. *Information systems*, 37(2):99–116, 2012.

[166] Alan E Renton, Adriano Chiò, and Bryan J Traynor. State of play in amyotrophic lateral sclerosis genetics. *Nature neuroscience*, 17(1):17, 2014.

[167] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

[168] Christoph Rinner, Emmanuel Helm, Reinhold Dunkl, Harald Kittler, and Stefanie Rinderle-Ma. Process mining and conformance checking of long running processes in the context of melanoma surveillance. *International journal of environmental research and public health*, 15(12):2809, 2018.

[169] Jose C Roche, Ricardo Rojas-Garcia, Kirsten M Scott, William Scotton, Catherine E Ellis, Rachel Burman, Lokesh Wijesekera, Martin R Turner, P Nigel Leigh, Christopher E Shaw, and Ammar Al-Chalabi. A proposed staging system for amyotrophic lateral sclerosis. *Brain*, 135(3):847–852, 2012.

[170] Eric Rojas, Michael Arias, and Marcos Sepúlveda. Clinical processes and its data, what can we do with them. In *Proceedings of the International Conference on Health Informatics (HEALTHINF 2015), Lisbon, Portugal*, pages 12–15, 2015.

[171] Eric Rojas, Jorge Munoz-Gama, Marcos Sepúlveda, and Daniel Capurro. Process mining in healthcare: A literature review. *Journal of biomedical informatics*, 61:224–236, 2016.

[172] Ines Rombach, Alastair M Gray, Crispin Jenkinson, David W Murray, and Oliver Rivero-Arias. Multiple imputation for patient reported outcome measures in randomised controlled trials: advantages and disadvantages of imputing at the item, subscale or composite score level. *BMC medical research methodology*, 18(1):87, 2018.

[173] Daniel R Rosen, Teepu Siddique, David Patterson, Denise A Figlewicz, Peter Sapp, Afif Hentati, Deirdre Donaldson, Jun Goto, Jeremiah P O'Regan, Han-Xiang Deng, et al. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature*, 362(6415):59, 1993.

[174] Anne Rozinat, Ronny S Mans, Minseok Song, and Wil MP van der Aalst. Discovering simulation models. *Information systems*, 34(3):305–327, 2009.

[175] Anne Rozinat and Wil MP van der Aalst. Conformance checking of processes based on monitoring real behavior. *Information Systems*, 33(1):64–95, 2008.

[176] Philip Russom. Big data analytics. *TDWI best practices report, fourth quarter*, 19(4):1–34, 2011.

[177] Seward B Rutkove. Clinical measures of disease progression in amyotrophic lateral sclerosis. *Neurotherapeutics*, 12(2):384–393, 2015.

[178] Francesco Saccà, Mario Quarantelli, Carlo Rinaldi, Tecla Tucci, Raffaele Piro, Gaetano Perrotta, Barbara Carotenuto, Angela Marsili, Vincenzo Palma, Giuseppe De Michele, et al. A randomized controlled clinical trial of growth hormone in amyotrophic lateral sclerosis: clinical, neuroimaging, and hormonal results. *Journal of neurology*, 259(1):132–138, 2012.

[179] Renato Cesar Sato and Désirée Moraes Zouain. Markov models in health care. *Einstein (São Paulo)*, 8(3):376–379, 2010.

[180] Michael Schroeck, Rebeca Shockley, Janet Smart, Dolores Romero-Morales, and Peter Tufano. Analytics: the real-world use of big data: How innovative enterprises extract value from uncertain data, executive report. *IBM Institute for Business Value and Said Business School at the University of Oxford*, 2012.

[181] Daniel I Sessler. Big data–and its contributions to peri-operative medicine. *Anaesthesia*, 69(2):100–105, 2014.

[182] Allan F Simpao, Luis M Ahumada, Jorge A Gálvez, and Mohamed A Rehman. A review of analytics and clinical informatics in health care. *Journal of medical systems*, 38(4):45, 2014.

[183] Minseok Song and Wil MP van der Aalst. Supporting process mining by showing events at a glance. In *Proceedings of the 17th Annual Workshop on Information Technologies and Systems (WITS)*, pages 139–145, 2007.

[184] Minseok Song and Wil MP van der Aalst. Towards comprehensive support for organizational mining. *Decision Support Systems*, 46(1):300–317, 2008.

[185] Jemeen Sreedharan, Ian P Blair, Vineeta B Tripathi, Xun Hu, Caroline Vance, Boris Rogelj, Steven Ackerley, Jennifer C Durnall, Kelly L Williams, Emanuele Buratti, et al. TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science*, 319(5870):1668–1672, 2008.

[186] Daniel J Stekhoven and Peter Bühlmann. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 10 2011.

[187] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

[188] Erica Tavazzi, Sebastian Daberdaku, Rosario Vasta, Andrea Calvo, Adriano Chiò, and Barbara Di Camillo. Exploiting Mutual Information for the Imputation of Static and Dynamic Mixed-Type Clinical Data with an Adaptive K-Nearest Neighbours Approach. *BMC Medical Informatics and Decision Making*, 20(5):1–23, 2020.

[189] Erica Tavazzi, Sebastian Daberdaku, Alessandro Zandonà, Rosario Vasta, Andrea Calvo, Adriano Chiò, and Barbara Di Camillo. An Adaptive K-Nearest Neighbours Algorithm for the Imputation of Static and Dynamic Mixed-Type Clinical Data. In *Proc. 16th International Conference on Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB)*, pages 285–284, 2019.

[190] Erica Tavazzi, Camille Gerard, Olivier Michielin, Alexandre Wicky, Roberto Gatta, and Michel A Cuendet. A process mining approach to statistical analysis: application to a real-world advanced melanoma dataset. In Sander Leemans and Henrik Leopold, editors, *ICPM workshops proceedings 2020*, Lecture Notes in Business Information Processing, Germany. Springer.

[191] Albert A Taylor, Christina Fournier, Meraida Polak, Liuxia Wang, Neta Zach, Mike Keymer, Jonathan D Glass, David L Ennist, and Pooled Resource Open-Access ALS Clinical Trials Consortium. Predicting disease progression in amyotrophic lateral sclerosis. *Annals of clinical and translational neurology*, 3(11):866–875, 2016.

[192] Dixon Thomas, Mickael Hiligsmann, Denny John, Ola Ghaleb Al Ahdab, and Hong Li. Pharmacoeconomic analyses and modeling. In *Clinical Pharmacy Education, Practice and Research*, pages 261–275. Elsevier, 2019.

[193] Irene Tramacere, Eleonora Dalla Bella, Adriano Chiò, Gabriele Mora, Graziella Filippini, and Giuseppe Lauria. The MITOS system predicts long-term survival in amyotrophic lateral sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 86(11):1180–1185, 2015.

[194] Kim Traxinger, Crystal Kelly, Brent A Johnson, Robert H Lyles, and Jonathan D Glass. Prognosis and epidemiology of amyotrophic lateral sclerosis: analysis of a clinic population, 1997–2011. *Neurology: Clinical Practice*, 3(4):313–320, 2013.

[195] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, Oct 2006.

[196] Martin R Turner, Jessica Barnwell, Ammar Al-Chalabi, and Andrew Eisen. Young-onset amyotrophic lateral sclerosis: historical and other observations. *Brain*, 135(9):2883–2891, 2012.

[197] Martin R Turner, Jakub Scaber, John A Goodfellow, Melanie E Lord, Rachael Marsden, and Kevin Talbot. The diagnostic pathway and prognosis in bulbar-onset amyotrophic lateral sclerosis. *Journal of the neurological sciences*, 294(1-2):81–85, 2010.

[198] Stef van Buuren, Hendriek C Boshuizen, and Dick L Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6):681–694, 1999.

[199] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, Articles*, 45(3):1–67, 2011.

[200] Wil MP van der Aalst. Process mining: discovering and improving spaghetti and lasagna processes. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 1–7. IEEE, 2011.

[201] Wil MP van der Aalst. *Process mining: discovery, conformance and enhancement of business processes*, volume 2. Springer, 2011.

[202] Wil MP van der Aalst. Process mining. *Communications of the ACM*, 55(8):76–83, 2012.

[203] Wil MP van der Aalst. Process mining: Overview and opportunities. *ACM Transactions on Management Information Systems (TMIS)*, 3(2):1–17, 2012.

[204] Wil MP van der Aalst. A practitioner's guide to process mining: Limitations of the directly-follows graph. *Procedia Computer Science*, 164:321–328, 2019.

[205] Wil MP van der Aalst, Arya Adriansyah, Ana Karla Alves De Medeiros, Franco Arcieri, Thomas Baier, Tobias Blickle, Jagadeesh Chandra Bose, Peter van Den Brand, Ronald Brandtjen, Joos Buijs, et al. Process mining manifesto. In *International Conference on Business Process Management*, pages 169–194. Springer, 2011.

[206] Wil MP van der Aalst, Arya Adriansyah, and Boudewijn van Dongen. Replaying history on process models for conformance checking and performance analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(2):182–192, 2012.

[207] Wil MP van der Aalst, Boudewijn F van Dongen, Joachim Herbst, Laura Maruster, Guido Schimm, and Anton JMM Weijters. Workflow mining: A survey of issues and approaches. *Data & knowledge engineering*, 47(2):237–267, 2003.

[208] Wil MP van der Aalst, Ton Weijters, and Laura Maruster. Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1128–1142, 2004.

[209] Jan Martijn EM van der Werf, Boudewijn F van Dongen, Cor AJ Hurkens, and Alexander Serebrenik. Process discovery using integer linear programming. In *International conference on applications and theory of petri nets*, pages 368–387. Springer, 2008.

[210] Boudewijn F Van Dongen and Wil MP van der Aalst. Multi-phase process mining: Building instance graphs. In *International Conference on Conceptual Modeling*, pages 362–376. Springer, 2004.

[211] Michael A van Es, Orla Hardiman, Adriano Chiò, Ammar Al-Chalabi, R Jeroen Pasterkamp, Jan H Veldink, and Leonard H Van den Berg. Amyotrophic lateral sclerosis. *The Lancet*, 2017.

[212] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91, 2006.

[213] Rosario Vasta, Alessandro Zandonà, Sebastian Daberdaku, Erica Tavazzi, Beatrice Nefussy, Christian Lunetta, Gabriele Mora, Jessica Mandrioli, Enrico Grisan, Claudia Tarlarini, Andrea Calvo, Cristina Moglia, Marc Gotkine, Vivian Drory, Adriano Chiò, and Barbara Di Camillo. Functional impairment and survival prediction in amyotrophic lateral sclerosis patients: a probabilistic model of disease progression. In *European Journal of Neurology*, volume 27, pages 172–172. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2020.

[214] Andrei Voustianiouk, Gregory Seidel, Janki Panchal, Mark Sivak, Adam Czaplinski, Albert Yen, Stanley H Appel, and Dale J Lange. ALSFRS and appel ALS scores: discordance with disease progression. *Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine*, 37(5):668–672, 2008.

[215] Akbar K Waljee, Ashin Mukherjee, Amit G Singal, Yiwei Zhang, Jeffrey Warren, Ulysses Balis, Jorge Marrero, Ji Zhu, and Peter DR Higgins. Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, 3(8), 2013.

[216] David S Watson, Jenny Krutzinna, Ian N Bruce, Christopher EM Griffiths, Iain B McInnes, Michael R Barnes, and Luciano Floridi. Clinical applications of machine learning algorithms: beyond the black box. *Bmj*, 364, 2019.

[217] Griffin M Weber, William G Adams, Elmer V Bernstam, Jonathan P Bickel, Kathe P Fox, Keith Marsolo, Vijay A Raghavan, Alexander Turchin, Xiaobo Zhou, Shawn N Murphy, and Kenneth D Mandl. Biases introduced by filtering electronic health records for patients with "complete data". *Journal of the American Medical Informatics Association*, 24(6):1134–1141, 08 2017.

[218] AJMM Weijters, Wil MP van der Aalst, and AK Alves De Medeiros. Process mining with the heuristics miner-algorithm. *Technische Universiteit Eindhoven, Tech. Rep. WP*, 166:1–34, 2006.

[219] Lijie Wen, Jianmin Wang, Wil MP van der Aalst, Biqing Huang, and Jiaguang Sun. A novel approach for process mining based on event types. *Journal of Intelligent Information Systems*, 32(2):163–190, 2009.

[220] Henk-Jan Westeneng, Thomas PA Debray, Anne E Visser, Ruben PA van Eijk, James PK Rooney, Andrea Calvo, Sarah Martin, Christopher J McDermott, Alexander G Thompson, Susana Pinto, et al. Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model. *The Lancet Neurology*, 17(5):423–433, 2018.

[221] Paul Wicks, Michael P Massagli, C Wolf, and James A Heywood. Measuring function in advanced ALS: validation of ALSFRS-EX extension items. *European Journal of Neurology*, 16(3):353–359, 2009.

[222] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[223] Wei Yang and Qiang Su. Process mining for clinical pathway: Literature review and future directions. In *2014 11th International Conference on Service Systems and Service Management (ICSSSM)*, pages 1–5. IEEE, 2014.

[224] Sumanth Yenduri and S Sitharama Iyengar. Performance evaluation of imputation methods for incomplete datasets. *International Journal of Software Engineering and Knowledge Engineering*, 17(01):127–152, 2007.

[225] Jinsung Yoon, William R Zame, and Mihaela van der Schaar. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 66(5):1477–1490, May 2019.

[226] Hsiang-Fu Yu, Nikhil Rao, and Inderjit S Dhillon. Temporal regularized matrix factorization for high-dimensional time series prediction. In DD Lee, M Sugiyama, UV Luxburg, I Guyon, and R Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 847–855. Curran Associates, Inc., 2016.

[227] Alessandro Zandonà, Rosario Vasta, Adriano Chiò, and Barbara Di Camillo. A Dynamic Bayesian Network model for the simulation of amyotrophic lateral sclerosis progression. *BMC bioinformatics*, 20(4):118, 2019.

[228] Jeffrey Zetino and Natasha Mendoza. Big data and its utility in social work: Learning from the big data revolution in business and healthcare. *Social Work in Public Health*, 34(5):409–417, 2019.

[229] Harry Zhang. The optimality of naive bayes. In Valerie Barr and Zdravko Markov, editors, *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*, Miami Beach, Florida, USA, 2004. AAAI Press.

[230] Zhongheng Zhang. Missing data imputation: focusing on single imputation. *Annals of Translational Medicine*, 4(1), 2016.