POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

# Applying Nonlinear Mixed-Effects Modeling to Model Patient Flow in the Emergency Department

Tesi di Laurea Magistrale in
Biomedical Engineering - Ingegneria Biomedica

Author: **Umberto Rosamilia**

Student ID: 944938
Advisors: PhD Adam Darwich, PhD Jayanth Raghothama, MSc Luca Marzano
Co-advisor: Associate Professor Enrico Gianluca Caiani
Academic Year: 2021-22

# Abstract

Emergency departments are fundamental for providing high-quality care, and their operations directly impact the logistics of the hospitals in their entirety. Poor emergency department performance leads to delays, prolonged hospitalization, and improper allocation of resources, reducing the quality of the provided care and increasing costs. Describing the variability embedded in real clinical data in a useful way is essential for improving the organization of hospitals in the near future. However, it is a challenging task due to clinical complexity and the lack of an established bridge between logistic systems and the clinical insights of the hospital. Therefore, this work aims to design and implement a simplified process model describing patient flow within an emergency department, which could allow the evaluation of the clinical impact of complex patient characteristics on the system's logistics. To achieve this, a novel nonlinear mixed-effects approach with hospital medical records was applied to design patient flow within the emergency department in the form of a multi-state Markov process. Four independent training data samples were extracted from the main dataset. For each of them, the set of covariates that could lead to the most significant improvement in the values of the employed likelihood indicators was selected. Through statistical tests, analysis of the outputs, and a validation process carried out on a fifth and independent dataset, it was possible to obtain a final model containing the most relevant and significant covariates for describing each of the modeled state transitions and confirming their clinical meaningfulness and relevance. The results achieved in this thesis can lead to future improvement of the healthcare logistics systems by extending the use of nonlinear mixed-effects approaches to the estimation of the covariate impact on emergency department flows.

**Keywords:** Emergency Department, Nonlinear Mixed-Effects Modeling, Healthcare Logistics, Patient Flow Modeling, Patient Pathways, Markov Process.

# Abstract in lingua italiana

I reparti di Pronto Soccorso sono fondamentali per la fornitura di assistenza sanitaria di alta qualità, ed il loro funzionamento ha un impatto diretto sulla logistica degli ospedali nella loro interezza. Prestazioni carenti di un reparto di Pronto Soccorso comportano ritardi assistenziali, prolungamenti della durata dell'ospedalizzazione, assegnazione erronea delle risorse, peggioramento della qualità dell'assistenza fornita, ed incremento dei costi. Al fine di migliorare l'organizzazione degli ospedali in un prossimo futuro, è fondamentale descrivere la variabilità contenuta nei dati clinici reali. Tuttavia, si tratta di un compito gravoso per via della elevata complessità clinica e della carenza di una preesistente interfaccia tra sistemi logistici e comprensione clinica degli ospedali. Pertanto, questa tesi si prefigge il compito di progettare ed implementare un modello di processo semplificato in grado di descrivere il flusso di pazienti attraverso un reparto di Pronto Soccorso, che possa consentire di valutare l'impatto clinico delle caratteristiche complesse dei pazienti sulla logistica del sistema. Al fine di raggiungere tale obiettivo, un approccio innovativo di modellizzazione ad effetti misti non lineari a partire dalle cartelle cliniche ospedaliere è stato impiegato per progettare il flusso di pazienti attraverso un reparto di Pronto Soccorso, sotto forma di processo Markoviano a stati multipli. Quattro set indipendenti di dati per il training del modello sono stati campionati dal set di dati principale. Per ciascuno di essi, è stato selezionato il set di covariate la cui introduzione nel modello fosse in grado di comportare il più significativo miglioramento del valore degli indicatori di "likelihood" utilizzati. Mediante l'impiego di test statistici, l'analisi dei risultati, ed il processo di validazione, effettuati su un quinto set indipendente di dati, è stato possibile ottenere un modello finale contenente le covariate più rilevanti e significative per la descrizione di ciascuna delle transizioni tra stati che sono state modellizzate, ed è stato possibile giustificare la loro significatività e rilevanza. I risultati conseguiti in questa tesi hanno il potenziale di portare a futuri miglioramenti dei sistemi di logistica healthcare, mediante l'estensione dell'uso di approcci di modellizzazione ad effetti misti non lineari alla stima dell'effetto delle covariate sui flussi di pazienti dei dipartimenti di Pronto Soccorso.

**Parole chiave:** Pronto Soccorso, Modellizzazione ad Effetti Misti Non Lineari, Lo-

gistica Healthcare, Modellizzazione del Flusso di Pazienti, Patient Pathways, Processo Markoviano

# Contents

# Introduction

Emergency departments (EDs) strive to provide high-quality 24/7 emergency care to severely ill or injured patients. ED performance and overcrowding have been shown to affect the functioning of other parts of the hospital and, indirectly, the "*healthcare systems and communities at large*" [1]. Poor performance of the ED and overcrowding lead to delays, prolonged hospitalization, and improper allocation of resources, which reduce the quality of the provided care and increase costs. Moreover, these negative consequences can "*compromise the patient health outcomes and lead to high admission and re-admission rates*" [1] or produce adverse outcomes for the providers, the healthcare system, and the community [1]. When the providers are exposed to intense workload, for instance, timely service provision and clinical decision-making are hindered, thus increasing the length of stay (LOS) [1]. This consequence is particularly relevant since longer lengths of stay increase the risk of contracting hospital-acquired infections [2] and "*are associated with higher patient mortality and worse outcomes*" [3]. Therefore, to guarantee the proper functioning of the hospitals in their entirety and, thus, improve patient outcomes, it is crucial to monitor and enhance the performance of the EDs continuously.

To achieve this, it is necessary to find a suitable way to evaluate the clinical impact of complex patient characteristics on the logistics of an emergency department and to support hospital management in better understanding and better intervening regarding the problems leading to excessive LOS within the ED. In this perspective, patient flow modeling based on real-world data can help find which factors have the highest impact on the system performance in given situations, support decisions concerning resource allocation and utilization, and help improve the pathways for a process and perform patient stratification [4].

## Comprehensive healthcare framework

Healthcare systems are complex organizations whose primary goal is to provide high-quality health services efficiently [5]. However, their physical facilities and resources are limited [4, 6, 7], the level of variability and uncertainty is high [6], and the performance goals to

be met are several and often conflicting. To enumerate some of these goals, Bhattacharjee and Ray [4] mentioned: "*minimizing the cost of healthcare, maximizing the utilization of physical and human resources, improving the quality of care by providing efficient diagnostic systems, handling an increasing number of patients effectively within a limited time span, arranging varieties of healthcare facilities in a single location, and improving overall healthcare system performance within limited and predetermined budget and time*". In short, such systems are supposed to provide health service effectively, efficiently, and without compromising the quality of care [5], while handling the rising cost of operations and maintenance [4].

Within the healthcare systems, one of the leading roles is covered by hospitals, which are healthcare sub-systems that work as "*integrated service units attending to the needs of the patients under treatment*" [4]. They usually include various departments and sub-units, such as the ED and diagnostic imaging services, e.g., the radiology department, all located within the same organization. Each department is characterized by its specialization and operational issues. Therefore, the overall hospital operational performance results from the interaction between the operational performances of all its departments and sub-units [4].

Among the numerous and, in most cases, interconnected hospital units, the emergency departments (EDs) constitute the specific area in this thesis's focus. Such departments are essential since they provide 24/7 emergency care to severely ill or injured patients, whose health conditions would likely worsen too rapidly for non-emergency healthcare to be effective on them. Furthermore, the emergency departments are responsible for processing the patients before their potential admission to a hospital ward. Modeling both patients' arrival and process flow is challenging due to high patient volumes and clinical variability. However, the importance and tight coupling of the EDs with many other hospital departments make continuous monitoring and improvement of their performance advisable and necessary to guarantee the proper functioning of the hospitals in their wholeness and improve patient outcomes.

## Problem

This section aims to briefly provide a broader definition of the problem addressed in this thesis and an introduction to the encountered engineering issues.

## ED Operational issues

The operations of an emergency department are central to providing high-quality emergency care and depend on clinical, economic, regulatory, and cultural factors, some of which are often also affected by local influences, e.g., a potential region-specific behavior of the population towards alcohol and drugs [8]. EDs face access blocks and overcrowding issues enhanced by increased cost [4], complexity, [9] and patient demand [3, 7, 8]. Moreover, the aging of the population [3, 9] increases the pressure on the system by increasing the number of patients needing care [10]. The resulting overcrowding of the EDs is acknowledged as the most severe problem regarding their patient throughput. It often leads to long waiting times (especially for the triage process), patient dissatisfaction, increased medical errors, increased rate of patients who leave without having been seen (LWBS), delay of care, poorer patient health outcomes, and increased patient length of stay (LOS) [11–13], see figure 1. When these factors increase the acuity of a patient's illness, the amount of consumed time per patient rises even further [14].

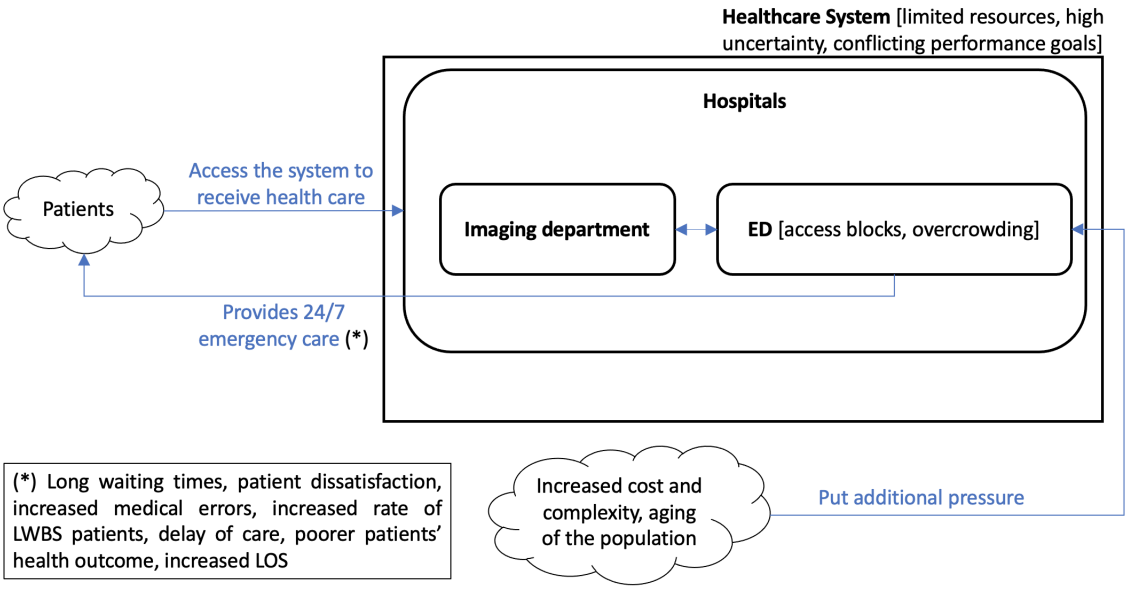



Figure 1: Simplified representation of the operational issues affecting an ED.

Both internal and external factors introduce delays in patient care [11] which usually, in turn, lead to overcrowding. Among them, particularly relevant are highly heterogeneous patient characteristics, unoptimized ED staffing patterns, complexity for patients in accessing healthcare providers, irregular patient arrival patterns, inappropriate management practices, and inappropriate chosen strategies for testing and treatment [11]. Moreover, the delay also derives from general inpatient crowding within the hospital [15]. Indeed, it was found that there exists a nonlinear correlation between higher inpatient bed occupancy and

longer ED waiting times [3]. This link exists since the increase in bed occupancy acts as an access block against the ED patients who need inpatient care but cannot access hospital beds in the correct department or any hospital bed at all within a reasonable time frame [3]. This specific factor is particularly tricky to tackle, mainly because it can often be out of reach for intervention for the stakeholders in the ED [15]. Furthermore, the effect of the increase of the ratio between non-elective admissions and discharges was found by S. Paling et al. [3] to affect the waiting time in the ED for two consecutive days after its occurrence, thus highlighting the importance of maintaining proper discharge levels also during the days of the week in which the bed occupancy is lower.

Although some of the sources of delay and inefficiency mentioned above cannot be eliminated, many others can be addressed with proper design and optimization [8]. Such elements that introduce delays to the otherwise potentially regular patient flow and thus lead to increased ED crowding often combine, leading to considerably significant increases in the LOS. Therefore, given the importance of ensuring efficient, effective, and high-quality care to patients, and given the difficulties in assessing the performance of emergency departments directly due to their high complexity, the systematic measuring and understanding of the factors underlying ED operational performance thus become a task of primary importance for at least two main reasons. In the first place, the care providers need to be informed and updated about system performance, so to be able to identify which elements in the system satisfy reasonable quality criteria and which others of them do not meet such standards. In the second place, it is essential for them to actively measure performance and what affects it, to be able to plan, test, and implement improvements in the system over time. However, contextually, quantification and measurement of the performance of such a complex system as the emergency department and its interactions with the rest of the hospital are particularly complicated since service providers and patients (customers) can have very different objectives and criteria around which to evaluate the performance of the system. It is usually challenging to find a proper trade-off between the interests of such two kinds of stakeholders.

Moving from a general overview to a specific case, the ED of Akademiska sjukhuset, Uppsala University Hospital, is facing significant issues related to prolonged length of stay (LOS) of its patients, also partially caused by bottlenecks related to different hospital wards, given that a considerably high portion of incoming patients commonly needs to be referred to the department of medical imaging or other specialistic care.

## Purpose

This thesis aims to find a suitable way to evaluate the clinical impact of complex patient characteristics on the logistics of an emergency department to support hospital management in better understanding and better intervening regarding problems leading to excessive LOS within the emergency department. The attainment of such an aim is significant since "*longer lengths of stay are associated with higher patient mortality and worse outcomes*" [3]. This correlation also applies to EDs, and it could be argued that LOS can indirectly cause worse health outcomes through more prolonged waiting times and treatment delays. However, it could be equally true that more severe patients to be treated, e.g., patients treated in the so-called Emergency Room of the ED, can require a more extended stay within such a department, thus a longer LOS. Delimiting precisely to what extent longer LOS is the cause of higher patient mortality and to what extent it could be seen as its consequence is not trivial, and the sole information contained in the available datasets for this thesis does not allow for performing such a distinguo confidently. Moreover, since in the present work the real-world case mentioned above was studied, i.e., the emergency department of the hospital "Akademiska sjukhuset", primarily due to the sample size, it was not possible to explicitly include in the model a state describing patient death that could lead to satisfactory results. Therefore, it was decided to aggregate the patients who died within the ED (see table 2.3) with the category "Other, unspecified". This latter aspect is better addressed in later parts of this thesis.

The primary benefit of the achievement of the goals of this thesis is that the employed approach for evaluating the impact of clinical covariates on logistical outcomes could become the starting point for future operational research studies aiming to test length of stay optimization procedures in the emergency department. Consequently, it could become easier to avoid the discussed negative consequences of longer LOS on patients, staff, and management. For what concerns the focus on “clinical factors”, which is introduced in this first chapter and can then be appreciated throughout this thesis, the reader could wonder whether there is a difference between clinical and operational factors in a healthcare facility and to what extent logistics can be considered “clinical”. An answer to these queries is provided in section 4.1.

## Goals

The main goal of this thesis is to design and implement a simplified and empirical process model describing the ED system of Akademiska sjukhuset that could allow evaluation of

the clinical impact of complex patient characteristics on the logistics of the ED system. This model would thus become the framework for future studies aiming at testing LOS optimization procedures that could be suitable for Akademiska sjukhuset. This thesis was divided into the following three phases, each of them represented by some sub-goals:

- Phase 1: population analysis, data pre-processing, exploration of the state of the art regarding modeling methods and techniques for process modeling in healthcare.
- Phase 2: choice of a general modeling approach (i.e., "nonlinear mixed-effects modeling"), testing and evaluation of possible modeling design options, final selection of a modeling technique (i.e., Markov Chains modeling) within the selected general modeling approach.
- Phase 3: model optimization in terms of the number of states and the selection of suitable covariates, parameter extraction for the designed process model, validation and discussion of the model and the results, analysis of the limitations and the potential future developments of this work.

## Research methodology

The methodology choices regarding this thesis can be divided into "choice of a general modeling approach" and "choice of a modeling technique within the selected general modeling approach". Both are briefly introduced in the following two sub-sections and then discussed in further detail in later chapters of this thesis.

### Choice of a general modeling approach

For what concerns the choice of a general modeling approach that could allow evaluating the clinical impact of complex patient characteristics on the logistics of the ED system, it is important to keep into consideration that patient flow is affected by numerous factors both within such characteristics and within the logistics of the specific ED in question. Among such mentioned factors, relevant examples can be the patient arrival patterns, the existing connection between ED and the imaging department, the discharge mechanism, and "patient-intrinsic" factors such as age, sex, chief complaint, triage, ICD-10 main diagnosis. However, it is necessary to discriminate which parameters are the most significant and helpful in allowing for a proper explanation of the high variability in patient characteristics without excessively increasing the complexity of the model and in a way that allows for better management of patient volume. Nevertheless, the modeling techniques that are traditionally applied to hospital medical records tend to select a pool of patient and

system characteristics and apply them to macro sections of the model without allowing neither for much differentiation among the parameters that are relevant and significant for each of the modeled state transitions nor for the consideration of potential random errors. Consequently, after a careful review of the current state of the art for what concerns process modeling approaches, a decision was made to utilize an approach that is not commonly employed to evaluate the impact of clinical covariates on logistical outcomes and for which one main field of application resides in pharmacometrics. The approach in question is the so-called "nonlinear mixed-effects modeling", whose goal is to "*model the relationship between a set of independent variables to some dependent variable*", with functions whose model parameters are nonlinear [16]. Moreover, a nonlinear mixed-effects model allows extracting insight from the data using a population approach [16].

Alternatively to nonlinear mixed-effects modeling, a more conventional approach for evaluating the clinical impact of complex patient characteristics on the logistics of the ED system could have been chosen. Such a conventional approach is the so-called "Canonical variate analysis" (CVA), which is "*mathematically equivalent to a one-way multivariate analysis of variance*" and is extensible to longitudinal data [17]. However, it was decided not to tackle the problem in question with a method from the family of "Canonical variate analysis" and instead to employ nonlinear mixed-effects modeling. This choice was made due to the belief of the latter being able to better describe the variability embedded into the data, also in terms of individual-specific variability, and to differentiate among the parameters that are relevant and significant for each of the modeled state transitions.

### Choice of a modeling technique within the selected general modeling approach

For what concerns choices of methodology for modeling the patient flow, several modeling techniques can be exploited to tackle healthcare applications and it is of extreme importance to identify and choose a suitable method for the application in question. To facilitate the reader's understanding of the performed method selection process, which is explained in further detail in later chapters of this thesis, it is important to underline here that this work was determined to design a process model with reduced complexity. Such a decision was made to make it possible to achieve a higher understanding of the underlying relationship between the modeled elements and the output of the system. Therefore, not all the services provided by the emergency department of Akademiska sjukhuset were included in the modeling process, but only those judged as pertaining and necessary.

Among the several available methods, which are better described in section 3.1, it was

chosen to select the so-called analytical approaches. This choice results from a process of testing and evaluating possible modeling design options, which can be better appreciated in the later chapters. However, in short, analytical approaches resulted particularly easy to frame within the "nonlinear mixed-effects" modeling approach for this thesis due to their ease of implementation and ability to handle causal dependencies naturally [18]. Within the category of the analytical approaches, in particular, it was chosen to select "Continuous-Time Markov Chains modeling". Even for what concerns this choice, the reasoning behind it is better addressed in later parts of this thesis. However, in short, this modeling technique is really valuable to the application in question, mainly due to its capability of modeling both clinical and operational patient flow [4] and to its ease of validation [19].

## General ethical and moral perspective

Despite the overcrowding of the EDs being a primarily important issue in the last decades, no comprehensive and effective solution has yet been implemented [8]. This scenario might partially depend on the role played by hospital-specific and ED-specific issues, which peculiarly characterize each case, and it might stack up with the more general and distinct nation-wise reasons for overcrowding [8]. However, the problem also depends on how little the general commitment of many hospitals to reducing crowding has been through the years [8].

From a moral point of view, this situation constitutes a challenging problem since patients, healthcare providers, and the healthcare system are significantly negatively affected by overcrowding [8]. J. Joseph and B. White (2020) exemplified some of the possible negative consequences of overcrowding in ED on patients. Namely, they listed "*delayed time to the administration of antibiotics in pneumonia and treatment of myocardial infarction, decreased compliance with core measures for sepsis, decreased analgesia for patients with acute pain*", and an increased rate of LWBS patients [8].

From an ethical point of view, according to the framework of ethical principles called "principlism", patients should be granted reasonably easy access to care and a reasonably short length of stay in the emergency department, which also depends on the crowding levels of such department, especially in order not to hinder the so-called principle of justice.

## Structure of the thesis

1. In chapter , which ends with this section, the importance of the treated topic is stated, relevant introductory knowledge is provided, the terms and scope of the topic of this thesis are displayed, the current scenario is outlined and evaluated, the importance of the proposed research is identified, the research questions and objectives are stated, and general ethical and moral issues are introduced.
2. In chapter 1, the theoretical background is presented in all its relevant subareas, providing the readers with the required knowledge to understand this thesis and the methodology choices that follow in the subsequent chapter.
3. In chapter 2, the followed research procedure is presented. In particular, it includes information about the data acquisition, the preliminary data analysis, the initial pre-processing, additional processing before implementing the Markov Chains model, the selection of a general modeling approach, the performed research and comparison of modeling techniques, the choice of potential modeling techniques within the selected general modeling approach, the parameter extraction, the attempted and then dismissed modeling ideas, the exploration of the remaining modeling techniques and setups, the design of a procedural protocol for the chosen technique and setups, the automatic covariate model building, and the validity assessment.
4. In chapter 3, the results of this thesis are presented, and their validity is analyzed.
5. In chapter 4, an interpretation of the obtained results is provided, and a discussion about limitations, technical implementation challenges, and future work on the topic in the question of this thesis is performed.
6. In chapter 5, conclusions on what is discussed in this thesis are given.
7. Thereafter, after including this thesis' bibliography, appendices regarding the datasets and the Monolix code are organized and reported at the end of this thesis.

# 1 | Background

Firstly, this chapter presents a short recap of what needs to be considered when unfolding the theoretical background of this thesis. Secondly, an overview of the system modeling approaches most commonly applied to healthcare is introduced. However, the main advantages and disadvantages of such techniques are presented later in this thesis, in section 3.1. To conclude, previous related work is addressed.

## 1.1. Healthcare needs and critical issues

Healthcare is a complex system based on "*multiple interactions between many different components*" [20], whose primary goal is to provide high-quality health services efficiently [5]. Due to being complex, healthcare is subject to a dynamic equilibrium that makes it challenging to obtain good knowledge about the system in its wholeness. Moreover, resource limitedness [4, 6, 7], rising costs for operations and maintenance [4], high variability and uncertainty [6], dynamicity of technical development, of socio-economic pressure, and of laws and guidelines, as well as conflicting performance criteria to be met, can hinder the ability of healthcare systems to work effectively, efficiently, and without compromising the quality of care.

### 1.1.1. Hospitals and ED critical issues

Hospitals are healthcare sub-systems that work as "*integrated service units attending to the needs of the patients under treatment*" [4]. They are composed of a heterogeneous and complex network of departments and sub-units that are particularly difficult to characterize and optimize. The operations of the emergency departments (EDs), in particular, are coupled with the activity of almost all the other units within the hospital and are dependent on clinical, economic, regulatory, and cultural factors. The high complexity deriving from the factors above makes it difficult to assess ED performance directly. Still, it does not diminish the primary importance of performing such measurement and assessment, both for what concerns evaluating the level of compliance with the preset performance and quality standards and for planning, testing, and implementing improvements in the system

over time. Such a performance measurement is usually carried out with the employment of healthcare-specific performance metrics. Hence, the introduction of these metrics, the introduction of modeling methods, the employment of the latter to improve the performance described by the former, and scientific and engineering issues towards optimization of ED performance are all addressed in the following parts of this chapter.

## Introduction to performance metrics

In the hospitals in general, and specifically in such complex and dynamic environments like Emergency Departments, multiple variables and objectives must be considered to achieve the preset goals. This need is especially true since the modern healthcare reimbursement system is conceived to prioritize reducing unnecessary costs, maximizing operational efficiency, and contextual conservation or improvement of quality [7]. To develop better methods, policies, and decision tools meant for improving hospital systems and achieving the objectives of an efficient healthcare system, it is essential to perform analyses of the hospital processes [4]. Several metrics have been designed and are employed nowadays to assess and monitor the status of the system and provide such information to the decision-makers. According to previous studies, the most widely used metrics for measuring the performance and assessing the care delivery processes in various hospital sub-units are patient wait time by process step, average door to provider time, average waiting time for activities supporting diagnoses, length of stay, inpatient throughput, patient readmission rate, LWBS rate, staff utilization rate, and bed occupancy rate [4]. Among these, the most important for this thesis is the length of stay, in other words, the amount of time that goes from patients' arrival to their disposition [15]. LOS was used in this thesis to model the transition between the state regarding patients' stay in the emergency department of Akademiska sjukhuset and the exit state. Furthermore, LOS allows for evaluating the flow of patients through the system throughout the whole process of care. Consequently, it is an indicator of crucial importance concerning the throughput of emergency departments and is a marker of overcrowding [11].

"*Patient Flow is the movement of patients through the whole process of care*" [4], and its rate can be affected by numerous factors, including seasonal and local ones. Such patient movement through a hospital starts with the patient's arrival at the hospital facility, in most cases happening through the outpatient department or the emergency one. For this thesis, the focus is on the latter one. After arrival, the complex route of the patients through their care process is personalized according to their health conditions and needs and influenced by external factors, such as resource limitedness, which can lead to the generation of queues and consequent waiting times [4]. In addition, several other factors

can add to the pool of uncertainties and complexities, in particular, patients' arrival patterns, the randomness in service times, the evolution of patient's health status, the variability and length of the pathway that a patient can undertake, the uncertainties and delays in physically transferring patients among different departments, the existence of priority rules [4].

## 1.2. Healthcare and models

Measuring system performance and assessing patient flow can be highly challenging. Moreover, it is difficult to navigate performance indicators to extract helpful insight. In addition, the pressure to which the leaders in healthcare institutions are exposed regarding monitoring and improving the system is constantly heavy[7]. Given these factors, and given that the deployment of appropriate resource management strategies is needed to "*avoid preventable high resource utilization that might cause access blocks*" [9], the computation of a model of the system is often necessary or at least advisable. Modeling approaches exist that can inform decision-making based on quantitative data provided by the performance metrics. Using such supportive techniques and methodologies for aiding decision-making is crucial to healthcare leaders in the decisional process [21].

In particular, patient flow modeling can help determine which factors have the highest impact on the system performance in given situations. It can support decisions concerning resource allocation and utilization, help improve the pathways for a process and perform patient stratification [4]. Moreover, tackling patient flow and capacity issues in the whole hospital, such as high bed occupancy, is proven to reduce waiting times in the emergency department and improve its patient health outcomes [3]. Lastly, patient flow modeling can be used to inform operational research, which is significant in improving the planning and management of the hospitals [5] and their departments. Proper planning and organization, together with the employment of Operational Research techniques to understand and model patient flows, can support healthcare managers in performing the optimizations required to reduce the idle times of resources and servers at each stage of the patient flow [4, 5]. Many are the aspects of patient flows that can be modeled, such as arrival distributions and transition probabilities, and it can often be challenging to isolate among services offered by the hospital or by one specific of its departments, given that many of such services interact with each other or at least share some resources. However, despite the difficulties in isolating services, and since objectives, detailedness, and generality of a model are interrelated [5], it is crucial to model only activities and services relevant to the goals of the specific modeling case in question [5].

### 1.2.1. Modeling employed to improve ED performance

Modeling in healthcare can be organized into conceptual modeling and actual model implementation [5]. Conceptual modeling is substantially independent of the chosen simulation software but dependent on the simulation methodology and is meant as a blueprint of the model that is supposed to be built [5]. Among the methods in conceptual modeling that are the most common, process flow diagrams are seen as the most suitable for being applied to the modeling of patient-related processes in a hospital. This suitability is due to their ease of build and understandability for both experts and non-experts [5].

Concerning the actual model implementation, numerous modeling approaches can be exploited to tackle healthcare applications. Each modeling approach and possible optimization technique presents its peculiar advantages and disadvantages. The choice of an optimal approach to be used is determined by many factors, such as the general area of the problem (e.g., emergency medical systems or epidemics models), the level of aggregation of the input data, the length of the simulation horizon, and the goals that are sought to be achieved [22]. In particular, according to Bhattacharjee and Ray [4], modeling methods can be divided into three main categories: analytical approaches, simulation modeling, and statistical or empirical modeling. It is possible to identify "queuing theoretic models" and the so-called "Markov Chains" and compartmental models among the analytical approaches. Simulation modeling can be subdivided into sub-categories as well: "discrete event simulation (DES)", "system dynamics (SD)", "agent-based simulation (ABS)", and "Monte Carlo methods (MC)", but hybrid approaches are not uncommon. Statistical or empirical modeling is not divided into sub-categories. The possible modeling methods are addressed in detail in section 3.1.

### Scientific and engineering issues towards optimization of ED performance

Several scientific and engineering issues underlie the modeling and the optimization of ED performance.

1. When dealing with healthcare-related systems, understanding the current underlying risks and how operational changes would modify these is more difficult than in systems that are not healthcare-related. This difference is due to two main factors: 1) risks in healthcare are inhomogeneous since the sources of risks for patients and operators vary broadly from field to field [23]; 2) risks that are directly related to the disease itself, risks deriving from medical diagnosis and decisions, and risks related to the way the chosen therapy is carried out, can combine in unpredictable ways [23].

2. Another ground problem is given by the little correspondence between measured outcomes in clinical and health research and the results achieved with operational research [8]. This last important engineering issue that J. Joseph and B. White (2020) [8] mentioned is related to the absence of a standardized operational workflow among different EDs. Accordingly, applying the same intervention to two EDs with comparable sizes, volumes, and patient populations could lead to different and contradictory outcomes [8].

## 1.3. Previous related work

The two following sub-sections (1.3.1 and 1.3.2) briefly explain the prior work that was applied to or used in this thesis and mention which prior work was instead not used.

### 1.3.1. Data collection from Akademiska sjukhuset

Structured data regarding all the patients who sought care from Akademiska sjukhuset's emergency department during 2019 were collected by the hospital and organized into two datasets. One, addressed as "D1" in the rest of this thesis, contains the information regarding all the patients accessing the ED during the selected year. The other one is instead addressed as "D2" and contains information related to the sessions of medical imaging performed on such ED patients. After patients' names and surnames were anonymized, both the datasets were made available for working on this project.

### 1.3.2. Complex model design in AnyLogic

A logistic model of the whole Akademiska sjukhuset, i.e., not only focused on the activities that are strictly related to the emergency department, was already developed in AnyLogic by a team from the same research department in which this thesis was produced. At its current development, the model describes the ED system of Akademiska sjukhuset in its high complexity, including the description of elements such as interactions with physicians and nurses. Real clinical data and clinicians' feedback were used to tune and validate such a model. However, due to its complexity, the model did not allow estimating the impact of patient characteristics on model parameters from the data.

# 2 | Materials and Methodology

The first part of this chapter addresses what concerns the healthcare production data used in this thesis. The latter includes the choice of a general modeling approach, the performed research and comparison of modeling techniques, the choice of potential modeling techniques within the selected general modeling approach, the parameter estimation, the attempted and then dismissed modeling ideas, the exploration of the remaining modeling techniques and setups, the design of a procedural protocol for the chosen technique and setups, the automatic covariate model building, and the validity assessment (see figure 2.13). In section 2.5, to facilitate the reader in following the methodological approach of this thesis, the experimental protocol for selecting the best set of covariates for each training data sub-set is summarised in a scheme (see figure 2.12)

## 2.1. Materials

This section aims to introduce the reader to what concerns the datasets employed in this thesis. Such an explanation is done in terms of data acquisition and datasets layout, patient characteristics, pre-processing, and secondary data processing.

### 2.1.1. Data acquisition

Data sampling in this thesis was performed by Akademiska sjukhuset itself, as mentioned in sub-section 1.3.1. The two datasets provided for this study contain anonymized data regarding all the patients who sought care from Akademiska sjukhuset's emergency department during 2019. The link between these two datasets resides in the so-called "*contact_id*", which works as a "case ID" for the patients. To be more specific, each patient was assigned a new *contact_id* at each new time they visited the ED through 2019. Accordingly, it was possible to associate the medical imaging sessions contained in the latter dataset with the visit to the ED during which corresponding patients underwent such sessions.

### 2.1.2. Preliminary data analysis

Before performing any study on the data and any of the steps described in the following sections and sub-sections, the datasets were checked for duplicated entries. Two duplicates were found in dataset D1 and therefore eliminated. Moreover, a brief study was performed to define each variable's meaning and investigate the possible values that each of them can take. Afterward, the characteristics of the population from which ED data was sampled were analyzed from several perspectives through Python programming. In particular, some statistics regarding patients who sought care from the ED more than once during the year 2019 were calculated, as well as statistics related to the performed sessions of medical imaging. The results of this analysis are reported here.

Despite the number of different patients visiting the ED during 2019 being equal to 33 866, the number of entries contained in dataset D1 is equal to 49 936. This mismatch is due to 8772 patients visiting the ED more than once throughout the year. Whereas most of these returning patients visited the department a small number of times in 2019 (e.g., 5442 patients twice and 1776 three times), it is noteworthy to point out that a minority of patients visited the ED numerous times during the same period, with a maximum of 65 times for one individual. For each patient, a variable for tracking the number of times they entered the ED during 2019, named "*times1Year*", was computed and added to dataset D1. For what instead concerns dataset D2, this counts 53 552 entries, where the count is heavily affected by two main reasons. The first one is that most of the performed imaging exams were recorded twice, once referred to as "partial decision" and once as "final decision" (see "*Radiology_Status*" in appendix A.1). The second reason is that some patients performed more than one kind of imaging exam. In total, 18 245 of the cases treated by the ED during 2019 included at least one medical imaging procedure for the corresponding patient, i.e., 35,54% of the total cases.

Information about "patient intrinsic characteristics" and "patient pathways characteristics" were extracted. For what concerns the extracted information regarding **patient intrinsic characteristics**, age distribution, and the count of patients by gender and by the municipality of origin were all calculated. Moreover, the distribution of the most represented "chief complaints" was plotted, where what is meant by chief complaint "*is a concise statement* [...] *of the symptoms that caused a patient to seek medical care*", which is recorded at the beginning of the medical care process [24]. Despite the chief complaint being an easily readable description of the patients' clinical conditions, it is just a preliminary categorization based on a summary evaluation of the patient during the triage process. Therefore, the formal classification that the main diagnosis of each

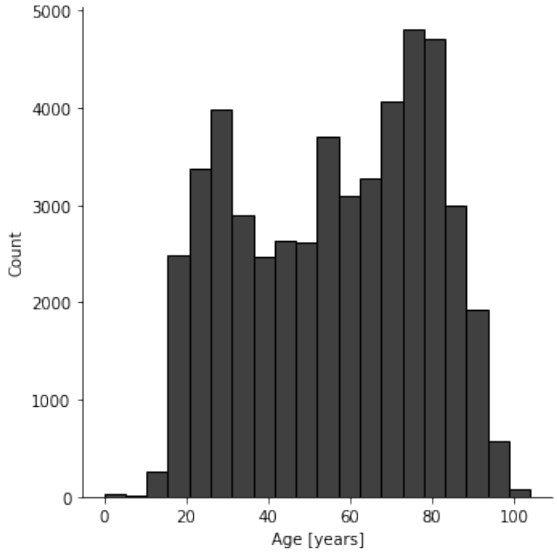
patient received at a later stage, contextually to meeting a doctor, was also analyzed. Such a formal classification is reported in dataset D1 by the employment of ICD-10, the $10^{th}$ revision of the International Classification of Diseases, a medical classification list maintained by the World Health Organization. Nevertheless, due to this classification standard's high level of detail, it would have been purposeless to calculate any distribution or count of patients by their ICD-10 labeling. However, since all the diagnoses of each principal macro diagnostic area are grouped under the same first letter in the ICD-10 terminology, it was possible to simplify the ICD-10 label for each entry in dataset D1 to its first letter. Therefore, the count of patients by ICD-10 macro diagnostic area could be calculated. Table 2.1 in this sub-section provides a list of such simplified codes, their meaning, and the result regarding the computation of patient count for each code.

For what instead concerns the extracted information regarding **patient pathways characteristics**, length of stay distribution, the count of patients reaching the ED by ambulance, the count of patients by their assigned unit of medical alarm, and the count of patients by mode of discharge, were all calculated. The outcome of all the above calculations is reported hereunder for both kinds of patient characteristics.

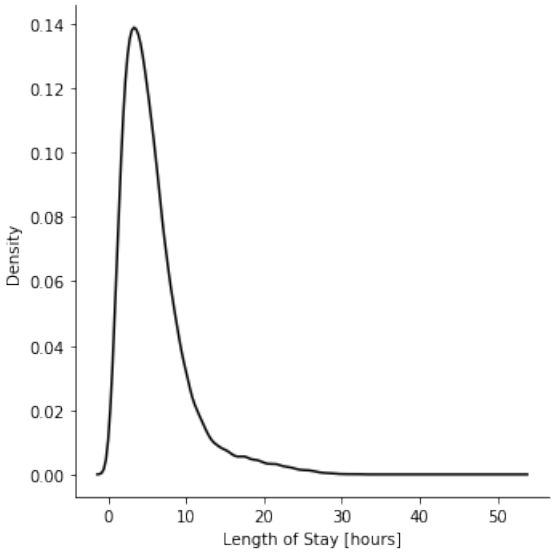
### Patient intrinsic characteristics

In figure 2.1a it is possible to see the age distribution of the ED population. Two peaks seem to characterize the ED population, one around the age of 20 and another between the ages of 70 and 80. For what instead concerns the gender of the patients who visited the ED during the year, 25 433 were females, 24 502 were males, and such information is not available for only one patient. Among the 49 936 total patients, 33 424 were residents in the municipality of Uppsala, thus constituting 66,93% of the total.

Figure 2.2 shows the distribution of the so-called chief complaint (CC) among the patients, giving the reader a general overview of the population. For readability purposes, the plot in figure 2.2 shows only the chief complaints that appeared at least 100 times within dataset D1.




(a) Age distribution of the ED population.




(b) Length of Stay distribution of the ED population.

Figure 2.1: Plots of age distribution and LOS distribution among the patients.

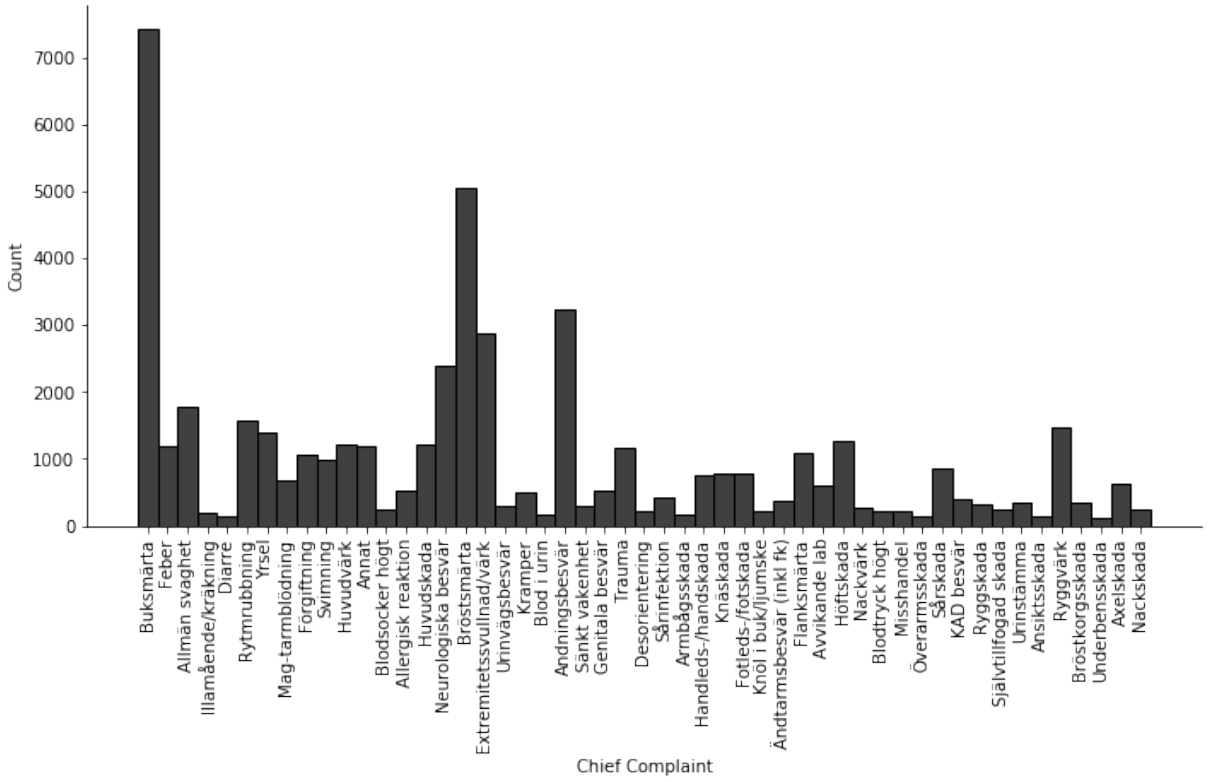



Figure 2.2: Distribution of the main chief complaints in the ED population.

Over a total of 49 936 entries, the five most represented chief complaints resulted in being abdominal pain ("Buksmärta") with 7412 entries (14.84%), chest pain ("Bröstsmärta") with 5052 entries (10.12%), respiratory problems ("Andningsbesvär") with 3229 entries (6.47%), extremity swelling or pain ("Extremitetssvullnad/värk") with 2871 entries (5.75%), and neurological disorders ("Neurologiska besvär") with 2378 entries (4.76%). It is noteworthy to highlight that the visits to the ED for declared poisoning ("Förgiftning") accounted for 1069 entries (2.14%), placing such a chief complaint within the 20 most represented ones.

Over 49 936 entries in dataset D1, for 49 649 entries, a diagnosis coded with ICD-10 was recorded. Table 2.1 shows a grouping of the patients by macro diagnostic areas thanks to simplifying the ICD-10 codes to their first letter. Most patients were assigned a code belonging to the category "R", i.e., "symptoms, signs of disease and abnormal clinical and laboratory findings not elsewhere classified". However, it is noteworthy to highlight that category S, i.e., "Trauma injuries", was assigned to 7634 entries (15.29% of the assigned ICD-10 codes). The category "X" was assigned to only one entry; it is thus reasonable to report its complete ICD-10 code, X6499, and the related description (*): "intentional self-destructive action through poisoning by exposure to other and unspecified drugs, medicaments and biological substances [...]".

## Patient pathways characteristics

Figure 2.1b shows the length of stay distribution among the ED population. Only one peak of the length of stay can be identified, located between the four and six hours of permanence within the ED. Concerning instead the count of patients reaching the ED by ambulance, this is equal to 13 751 over a total of 49 936 entries (27,54%). This information was employed in the design of the process model.

Table 2.2, which was constructed from the value of the variable named "$MA\_unit$" contained in dataset D1 and explained in appendix A.1, shows the count of patients by their assigned medical alarm unit. In addition, table 2.3 shows the count of patients by mode of discharge. Even this latter information was employed in the design of the process model.

Table 2.1: Counts of patients by first letter of the ICD-10-SE medical classification list.

| Code | Description | Number of patients |
|---|---|---|
| R | Not elsewhere classified | 21 686 |
| S | Trauma injury | 7634 |
| M | Musculoskeletal, connective | 4008 |
| K | Digestive system | 3073 |
| I | Circulatory system | 2934 |
| Z | No specific disorder but warranted treatment | 2205 |
| N | Genitourinary system | 1717 |
| T | Intoxication | 1392 |
| J | Respiratory system | 1071 |
| F | Mental, behavioral disorders | 776 |
| A, B | Infectious, parasitic diseases | 689 |
| G | Nervous system | 676 |
| L | Skin, subcutaneous tissue | 580 |
| E | Endocrine, nutritional, metabolic | 481 |
| C, D | Neoplasms | 375 |
| H | Eye and its adnexa, ear, and mastoid process | 330 |
| O | Pregnancy and other obstetric conditions | 11 |
| Q | Malformations, abnormalities | 6 |
| P | Perinatology | 4 |
| X | (*) | 1 |

Table 2.2: Counts of patients by unit of Medical Alarm.

| Most suitable care team | Number of patients |
|---|---|
| Emergency medicine | 22 707 |
| Acute surgery | 14 611 |
| Acute orthopedics | 8984 |
| Acute infection | 1445 |
| Heart disease | 1308 |
| Trauma surgery | 713 |
| Lung and allergy diseases | 101 |
| Hand surgery | 64 |
| Neurology | 2 |
| Infectious diseases | 1 |

Table 2.3: Counts of patients by Mode of discharge.

| Mode of discharge | Number of patients |
|---|---|
| Sent home | 30 773 |
| Admitted to another hospital ward | 12 982 |
| Other, unspecified | 4746 |
| Redirected | 1273 |
| Death of the patient | 103 |
| Taken in charge by consultants | 61 |

### 2.1.3. Pre-processing

Additional pre-processing was carried out; this included:

- The number of accesses to the emergency department during 2019 for each patient was computed and added to dataset D1 as a variable named "$times1Y$".

- A binary variable representing whether a patient who reached the ED was resident in Uppsala or not was created and computed for all the patients. The variable was added to the dataset D1 under the name "*UppsalaYN*".
- A binary variable describing whether a patient reached the ED by ambulance was created and computed for all the patients. The variable was added to dataset D1 under the name "*AmbulYN*".
- A bridging between dataset D1, containing the general ED patient information, and dataset D2, containing the information related to the performed medical imaging on the same population of patients, was conducted through Python programming. Consequently, the following columns were added: scan yes/no (named "*ScanYN*"), the number of scans performed on each patient (named "*number_of_scans*"), time spent by the patient while waiting for the imaging to be performed (named "*LOS_reqTOperf*").
- The other computable time contributions to the total length of stay were identified and added to the dataset, and their first quartiles, medians, and third quartiles were calculated. "*LOS_docTOreq*" represents for each patient the time that passed from their first meeting with a doctor to the request of their first session of medical imaging. "*LOS_perfTOdisch*" represents for each patient the time that passed from the execution of the last session of medical imaging to the patient's discharge or admission to a ward. "*LOS_docTOout*" represents for each patient the time that passed from the first meeting with a doctor to the patient's discharge or admission to a ward.
- A new categorical variable, containing for each patient a simplification of their ICD-10 code to its first letter, was added to dataset D1 under the name "*simple_diag.*
- Three variables for taking into consideration the crowding of the ED were computed through Python programming and added to dataset D1, under the names "*countIN*", "*countOUT*", and "*countAVG*". To do it, it was first necessary to convert patients' arrival and discharge times into the corresponding "$n^{th}$" hour from the beginning of the year 2019". The converted arrival times were saved as a new variable named "*hours_in*". The converted discharge times were saved as "*hours_out*". Thirteen patients were censored since they were discharged from the ED only at the beginning of 2020.
- Ten patients whose age was unknown or unspecified were also censored from the dataset since "age" is a variable used in the design of the process model, and

having an empty cell for some patients for a continuous variable would have been troublesome.

- The ED was subject to comparable pressure throughout the whole year 2019 in terms of average patient count within the department since the standard deviation of the monthly-averaged patient count was equal to 1,84 patients, whereas the mean of such a monthly-averaged patient count was equal to 38,12 patients, thus implying a score of 4,83% in terms of Coefficient of Variation (CV). Contextually, however, the staffing and resource levels of the ED undergo a significant drop during June, July, and August. Therefore, to study the system in its normal working conditions, it was decided to exclude from the study all the patients who entered the emergency departments during any of those three months.
- Among the remaining variables, the ones that were considered a priori irrelevant for the modeling process or unusable at this stage, as well as the ones whose usefulness was only temporary for performing the preliminary data analysis (sub-section 2.1.2) or the pre-processing, were eliminated from the dataset. In particular, this was the case for the variables "*municipality*", "*first_doctor_contact_date*", "*contact_type*", "*arrival_method*", "*main_diagnosis*", "*priority*", "*triage*", "*team_care_contact*", "*triage_level*", and "*last_doctor_contact_date*", all from dataset D1, and for the variables "*ExaminationDate*" and "*RegistrationDate*", that had been imported from dataset D2.

Relevant results of the above-described data pre-processing are shown in this same sub-section in the paragraph named "relevant pre-processing outputs".

Further modifications of the final pre-processed dataset were performed at later stages through Python programming to adapt the data to how the employed software interprets information. Such a dataset was reformatted multiple times according to the method put to the test at each stage.

### Relevant pre-processing outputs

In dataset D1, the value of the variable "*ARRIVAL_DATE*" is available for all the entries (49 936), but the value of the variable "*first_doctor_contact_date*" is available only for or 47 594 of the entries, i.e., 95,31% of the total. Therefore, it was possible to compute the value of the variable "*LOS_inTOdoc*" only for these 47 594 entries. As a result, the first quartile for the variable in question was equal to 00:32:14, the median was equal to 01:13:31, and the third quartile was equal to 02:33:05.

In dataset D1, the value of the variable "*RegistrationDate*" is available for 18 245 entries, and for only 17 935 of these the value of the variable "*first_ doctor_ contact_ date*" is also available. Therefore, it was possible to compute the value of the variable "*LOS_ docTOreq*" only for these 17 935 entries, i.e., 35,92% of the total. As a result, the first quartile for the variable in question was equal to 00:09:41, the median was equal to 00:30:08, and the third quartile was equal to 01:05:18.

In dataset D1, both the variable "*ExaminationDate*" and the variable "*RegistrationDate*" are available for the same 18 245 entries, i.e., 36,54% of the total. Therefore, it was possible to compute the value of the variable "*LOS_ reqTOperf*" for 18 245 entries. As a result, the first quartile for the variable in question was equal to 00:48:00, the median was equal to 01:27:00, and the third quartile was equal to 02:36:00.

In dataset D1, the value of the variable "*DISCHARGE_ DATE*" is available for all the entries (49 936), but the value of the variable "*ExaminationDate*" is available only for 18 245 of the entries, i.e., 36,54% of the total. Therefore, it was possible to compute the value of the variable "*LOS_ perfTOdisch*" only for these 18 245 entries. As a result, the first quartile for the variable in question was equal to 01:13:14, the median was equal to 02:08:42, and the third quartile was equal to 03:46:24.

In dataset D1, the value of the variable "*DISCHARGE_ DATE*" is available for all the entries (49 936), but the value of the variable "*first_ doctor_ contact_ date*" is available only for 47 594 of them, i.e., 95,31% of the total. Therefore, it was possible to compute the value of the variable "*LOS_ docTOout*" only for these 47 594 entries. As a result, the first quartile for the variable in question was equal to 01:28:50, the median was equal to 02:59:56, and the third quartile was equal to 05:19:53.

The final pre-processed dataset, sometimes addressed as dataset "D3" in the next parts of this thesis, includes 37495 entries that refer to the patients who visited the ED during nine months of 2019, where June, July, and August were excluded. This dataset contains the following 26 columns: *contact_ id*, *person_ id*, *sex*, *age*, *cause_ of_ visit*, *simple_ diag*, *reason_ for_ discharge*, *ARRIVAL_ DATE*, *DISCHARGE_ DATE*, *MA_ unit*, *times1Year*, *UppsalaYN*, *AmbulYN*, *ScanYN*, *number_ of_ scans*, *countIN*, *countOUT*, *countAVG*, *hours_ in*, *hours_ out*, *LoS_ hours*, *LOS_ inTOdoc*, *LOS_ docTOreq*, *LOS_ reqTOperf*, *LOS_ perfTOdisch*, and *LOS_ docTOout*. The explanation of their meaning is provided above in this same sub-section for the variables created while performing the pre-processing. Concerning instead the variables that were already present in the datasets D1 and D2, these are described in appendix A.1.

### 2.1.4. Dataset processing for Markov Chains modeling

To interpret the process as a continuous-time Markov Chain for performing the parameter estimation from the data, the approach described in section 2.5, it was necessary to reshape dataset D3 to introduce a time coordinate and an observation (state variable) for each patient. In this framework, the variable "*contact_id*" (see appendix A.1) was used as an identifier in Monolix. In contrast, the time coordinate was specifically created (named "*Time*") so to have all the patients in one of the two initial states at time 0, in state 3 after one minute, and in one of the final states at the time "*Length of stay + 1 minute*". For describing the observations, another variable was specifically introduced, in this case to keep track of the number of the state at which each patient is associated at a given time. Accordingly, figure 2.3 shows a sample of two patients from one of the processed datasets, in the way the employed software groups the *contact_id* for this implementation. No covariates are shown.

| | ID | TIME | OBSERVATION |
|---|---|---|---|
| LINE NUMBER | contact_id | Time | State |
| − | 1018581 | | |
| 10064 | 1018581 | 0 | 1 |
| 2 | 1018581 | 0.0166667 | 3 |
| 3 | 1018581 | 7.08333 | 5 |
| − | 1031398 | | |
| 10065 | 1031398 | 0 | 2 |
| 4 | 1031398 | 0.0166667 | 3 |
| 5 | 1031398 | 1.36667 | 4 |

Figure 2.3: Two samples from one of the five dataset sub-samples for "Markov Chains" modeling, without showing any covariate.

To obtain a suitable dataset for performing Markov Chains modeling like the one shown in figure 2.3, starting from the pre-processed dataset (D3), the following elaborations were achieved through Python programming.

After having grouped four modalities of discharge as previously described, a column for storing the value of the state was added with the name "*State*" and initialized for each patient with the corresponding value of the rearranged (as explained in section 2.5) modality of discharge, conveniently translated into a numerical identifier as shown in figure 2.11.

Afterward, from the complete dataset, five **independent** sub-samples were extracted with five different random seeds, to be able to later perform the validity assessment as described in section 2.7. Since the variable "*simple_ diag*" is the only potentially useful categorical covariate that can take more than two different values, since its meaning is, at least theoretically, particularly relevant for the kind of process model that this thesis was determined to design, and since only 0.55% of the entries of the pre-processed dataset lacks a value for such a variable, the aforementioned extraction of the five sub-samples was performed with a stratified sampling by proportionate allocation of the values taken by *simple_ diag*, excluding the patients having no such value recorded. Four of the extracted samples contained 933 or 934 patients each and were used to extract sets of potentially meaningful covariates as described in section 2.6; the fifth sample contained 5031 patients and was used to assess the model's validity as described in section 2.7.

For each of the five samples, all the entries were replicated to have two exactly equal rows for each *contact_ id*, and then a variable for storing the time information (named "*Time*") was created and initialized with 0 and, for each *contact_ id*, the value of such a variable was changed for one of the two entries to the value of the variable "*LoS_ hours*" for the same *contact_ id*. Then, for each of the five samples, a copy of the only entries with $Time = 0$ was saved and, on the copy itself, the state related to patients for which $AmbulYN = 1$ was set to 1, whereas the state related to patients for which $AmbulYN = 0$ was set to 2. Going back to the original datasets, for all their entries, the time was increased by one minute, and, after that, the state related to entries with $Time$ equal to 1 was set to 3. At this point, each of the five original datasets was merged with its corresponding copy, all the times expressed in minutes were converted into hours, the value of the variable "*number_ of_ scans*" was set to 0 for the patients for which its value was previously null, and the variable "*simple_ diag*" was converted into several binary variables, one per each of the possible letters that it can assume.

Once all these steps were completed, the variables in the datasets that were not needed anymore for implementing the process model were removed. Then the order of the remaining columns was rearranged. The complete list of the removed variables includes: "*person_ id*", "*LOS_ inTOdoc*", "*LOS_ docTOreq*", "*LOS_ reqTOperf*", "*LOS_ perfTOdisch*", "*LOS_ docTOout*", "*ARRIVAL_ DATE*", "*DISCHARGE_ DATE*", "*hours_ in*", and

"*hours_out*". Finally, from dataset D3, the five independent data sub-sets were sampled. Their description is shown in the last part of this sub-section.

| | SEED n°1 | | SEED n°2 | | SEED n°3 | | SEED n°4 | | TEST SET | |
|---|---|---|---|---|---|---|---|---|---|---|
| Covariate | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| A | 924 | 9 | 924 | 9 | 925 | 9 | 925 | 9 | 4984 | 47 |
| AmbulYN | 670 | 263 | 656 | 277 | 685 | 249 | 710 | 224 | 3588 | 1443 |
| B | 929 | 4 | 929 | 4 | 930 | 4 | 930 | 4 | 5011 | 20 |
| C | 932 | 1 | 932 | 1 | 933 | 1 | 933 | 1 | 5026 | 5 |
| D | 927 | 6 | 927 | 6 | 928 | 6 | 928 | 6 | 4999 | 32 |
| E | 924 | 9 | 924 | 9 | 925 | 9 | 925 | 9 | 4984 | 47 |
| F | 918 | 15 | 918 | 15 | 919 | 15 | 919 | 15 | 4950 | 81 |
| G | 920 | 13 | 920 | 13 | 921 | 13 | 921 | 13 | 4960 | 71 |
| H | 927 | 6 | 927 | 6 | 928 | 6 | 928 | 6 | 4999 | 32 |
| I | 876 | 57 | 876 | 57 | 877 | 57 | 877 | 57 | 4726 | 305 |
| J | 912 | 21 | 912 | 21 | 913 | 21 | 913 | 21 | 4919 | 112 |
| K | 875 | 58 | 875 | 58 | 876 | 58 | 876 | 58 | 4715 | 316 |
| L | 922 | 11 | 922 | 11 | 923 | 11 | 923 | 11 | 4974 | 57 |
| M | 860 | 73 | 861 | 72 | 861 | 73 | 861 | 73 | 4640 | 391 |
| N | 901 | 32 | 901 | 32 | 902 | 32 | 902 | 32 | 4858 | 173 |
| O | 933 | 0 | 933 | 0 | 934 | 0 | 934 | 0 | 5030 | 1 |
| Q | 933 | 0 | 933 | 0 | 934 | 0 | 934 | 0 | 5030 | 1 |
| R | 524 | 409 | 523 | 410 | 524 | 410 | 524 | 410 | 2820 | 2211 |
| S | 791 | 142 | 791 | 142 | 792 | 142 | 792 | 142 | 4266 | 765 |
| ScanYN | 586 | 347 | 600 | 333 | 592 | 342 | 601 | 333 | 3174 | 1857 |
| T | 907 | 26 | 907 | 26 | 908 | 26 | 908 | 26 | 4889 | 142 |
| UppsalaYN | 338 | 595 | 286 | 647 | 289 | 645 | 320 | 614 | 1681 | 3350 |
| Z | 892 | 41 | 892 | 41 | 893 | 41 | 893 | 41 | 4809 | 222 |
| Sex_man | 485 | 448 | 457 | 476 | 474 | 460 | 510 | 424 | 2605 | 2426 |
| Handkirurgi | | 0 | | 1 | | 1 | | 2 | | 7 |
| Hjärtsjukdomar | | 24 | | 23 | | 32 | | 23 | | 135 |
| Infektion akut | | 17 | | 22 | | 19 | | 19 | | 129 |
| Kirurgi akut | | 263 | | 285 | | 273 | | 283 | | 1449 |
| Kirurgi trauma | | 17 | | 14 | | 9 | | 6 | | 66 |
| Lung- och allergisjukdomar | | 1 | | 2 | | 3 | | 2 | | 11 |
| Medicin akut | | 444 | | 430 | | 421 | | 429 | | 2346 |
| Ortopedi akut | | 167 | | 156 | | 176 | | 170 | | 887 |

Figure 2.4: Categorical covariates distribution across modalities for the four training sub-samples and the testing sub-sample.

## Dataset sub-sampling

Figure 2.4 regroups in one table the distribution of the categorical covariates across modalities for all the four randomly sampled independent "training" data sub-sets and the "testing" sub-set. Despite the sampling being stratified only by proportionate allocation of the values taken by *simple_diag*, the distribution of the other categorical covariates across modalities is also consistent among the sub-sets. Furthermore, even the distribution of the continuous covariates is consistent among all the sub-sets, as shown in figure 2.5.

| | SEED n°1 | | | | | | SEED n°2 | | | | | | SEED n°3 | | | | | | SEED n°4 | | | | | | TEST SET | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MIN | Q1 | MEDIAN | Q3 | MAX | SD | MIN | Q1 | MEDIAN | Q3 | MAX | SD | MIN | Q1 | MEDIAN | Q3 | MAX | SD | MIN | Q1 | MEDIAN | Q3 | MAX | SD | MIN | Q1 | MEDIAN | Q3 | MAX | SD |
| age | 3 | 34 | 58 | 75 | 100 | 22.75 | 5 | 37 | 58 | 75 | 99 | 22.42 | 0 | 36 | 59 | 75 | 98 | 22.44 | 9 | 35 | 57.5 | 76 | 99 | 22.52 | 9 | 35 | 57.5 | 76 | 99 | 22.52 |
| countAVG | 8 | 31.5 | 39 | 46 | 67.5 | 10.55 | 8 | 31.5 | 39.5 | 46.5 | 73.5 | 10.88 | 9 | 32 | 39.5 | 46.5 | 77 | 10.34 | 10 | 32 | 39.5 | 46 | 73.5 | 10.22 | 10 | 32 | 39.5 | 46 | 73.5 | 10.22 |
| countIN | 4 | 30 | 39 | 48 | 70 | 12.61 | 7 | 30 | 40 | 49 | 77 | 12.77 | 8 | 31 | 39 | 48 | 77 | 12.48 | 9 | 32 | 39 | 47 | 77 | 11.88 | 9 | 32 | 39 | 47 | 77 | 11.88 |
| countOUT | 7 | 30 | 39 | 47 | 76 | 12.02 | 7 | 29 | 39 | 49 | 74 | 12.94 | 8 | 31 | 39 | 48 | 77 | 12.15 | 6 | 30 | 39 | 47 | 77 | 12.31 | 6 | 30 | 39 | 47 | 77 | 12.31 |
| number_of_scans | 0 | 0 | 0 | 1 | 7 | 0.91 | 0 | 0 | 0 | 1 | 7 | 0.96 | 0 | 0 | 0 | 1 | 7 | 0.98 | 0 | 0 | 0 | 1 | 6 | 0.88 | 0 | 0 | 0 | 1 | 6 | 0.88 |
| times1Year | 1 | 1 | 1 | 3 | 51 | 4.22 | 1 | 1 | 2 | 3 | 65 | 3.22 | 1 | 1 | 2 | 3 | 51 | 2.82 | 1 | 1 | 1 | 3 | 51 | 3.79 | 1 | 1 | 1 | 3 | 51 | 3.79 |

Figure 2.5: Statistics on continuous covariates distribution for the four training sub-samples and the testing sub-sample.

## 2.2. Modeling approaches and techniques

In this section, sub-section 2.2.1 covers the chosen general modeling approach, i.e., nonlinear mixed-effects modeling (NLMEM), explaining its fundamental principles, its main advantages, and thus the reasons why it was chosen over more traditional approaches. Sub-section 2.2.2 describes what was done to compare the possible modeling techniques. The results of such research and comparison of modeling techniques are reported in section 3.1 and lead to the choice of potential modeling techniques within the selected general modeling approach.

### 2.2.1. General modeling approach: nonlinear mixed-effects modeling

The modeling approach employed in this thesis is "nonlinear mixed-effects". The choice of a nonlinear approach among the mixed-effects modeling ones was made since it is well-known from previous literature that the distribution of the length of stay is, in turn, nonlinear [25]. Additionally, such a choice was backed up by the plot of the LOS from the data available for this thesis, which shows how this metric follows a log-normal distribution (see figure 2.1b), i.e., it is nonlinear as well.

Less specifically, the choice of using nonlinear mixed-effects modeling as the general modeling approach for this thesis was deduced from several factors. In particular, importance was given to the possibility it provides for working with longitudinal data, i.e., with data collected from the same sample (patients) at several distinct points in time. Alternatively to a longitudinal study design, a cross-sectional study is often the applied approach for hospital medical records [26]. In such studies, however, the regressor coefficients do not keep track of changes in the population at the individual level in case of changes in the covariate values [26].

To provide the reader with a sufficient explanation of the principles of "nonlinear mixed-effects modeling", this approach is hereunder introduced more in detail, using the nomenclature presented in the book "Mixed-Effects Models in S and S-PLUS" by J. Pinheiro and D. Bates [27]. As previously stated, a "nonlinear mixed-effects" model includes both fixed and random effects, where fixed effects represent typical population values, the same for all the individuals, and random effects represent inter-individual, intra-individual, and residual variability [28]. Its "*purpose is to describe a response variable as a function of the predictor (independent) variables*", while recognizing "*correlations within sample subgroups, providing a reasonable compromise between ignoring data groups entirely, thereby losing valuable information, and fitting each group separately, which requires significantly more data points*" [28]. A general nonlinear mixed-effects model can be described by the following equation:

$$y_{ij} = f(\boldsymbol{\phi}_{ij}, \boldsymbol{v}_{ij}) + \epsilon_{ij}, \qquad i = 1, ..., M,\ j = 1, ..., n_i, \tag{2.1}$$

where $M$ is the number of individuals and $n_i$ is the number of observations on the $i^{th}$ individual, $f$ is a "*general, real-valued, differentiable function of a individual-specific parameter vector $\boldsymbol{\phi}_{ij}$ and a covariate vector $\boldsymbol{v}_{ij}$, and $\epsilon_{ij}$ is a normally distributed*" within-individual "*error term*" [27]. When equation 2.1 describes a mixed-effects model that is nonlinear, then at least for one component of the individual-specific parameter vector $\boldsymbol{\phi}_{ij}$ the function $f$ must be nonlinear [27]. Such an individual-specific parameter vector $\boldsymbol{\phi}_{ij}$ is modeled as follows:

$$\boldsymbol{\phi}_{ij} = \boldsymbol{A}_{ij}\boldsymbol{\beta} + \boldsymbol{B}_{ij}\boldsymbol{b}_i, \qquad \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}), \tag{2.2}$$

where $\boldsymbol{\beta}$ is the vector of fixed effects and $\boldsymbol{b}_i$ is the vector of the random effects for the $i^{th}$ individual (with no dependence on $j$), which is associated to the variance-covariance matrix $\boldsymbol{\Psi}$ [27]. Both the matrices $\boldsymbol{A}_{ij}$ and $\boldsymbol{B}_{ij}$ are individual-dependent and "*possibly dependent on the values of some covariates at the $j^{th}$ observation*" [27].

To be taken into consideration is that in the hitherto described general nonlinear mixed-effects model, the observations are assumed to be independent of other individuals, and the intra-individual errors $\epsilon_{ij}$ are assumed to be independently distributed as $\mathcal{N}(0, \sigma^2)$, and independent of the random effects vector $\boldsymbol{b}_i$ [27]. However, the assumption of independence and homoscedasticity for the intra-individual errors can be relaxed [27] if needed.

### 2.2.2. Research and comparison of modeling techniques

This sub-section focuses on the methodology underlying research and comparison among different possible modeling techniques to select the most suitable ones for realizing the actual model implementation. The data analysis step served to understand which variables could have been used to describe patient characteristics and get a first glance at the general structure of the actual patient flow through the ED of Akademiska sjukhuset. Then, since numerous different modeling approaches can be exploited to tackle healthcare applications, and each of such modeling approaches and possible optimization techniques presents its peculiar advantages and disadvantages, in a second step, an actual comparison between relevant modeling techniques was performed, so to construct the theoretical basis on which the choice of one or more modeling methods to be used for this thesis could be achieved. This comparison between modeling techniques included methods from the categories of analytical approaches, simulation modeling, and statistical or empirical modeling, all of them presented according to their specific sub-categories. The corresponding results were organized in tables which show the main advantages and disadvantages deriving from the employment of each of them, and are reported in section 3.1.

### 2.2.3. Choice of potential modeling techniques within the selected general modeling approach

The purpose of this section is to briefly introduce which modeling techniques were, at this point of the model design, framed as potentially promising for being applied to this thesis within the modeling approach of the nonlinear mixed-effects modeling.

Given the choice of NLMEM as a general modeling approach and the intention to exploit the longitudinalization of the data, but also given the results of the preliminary data analysis (see section 2.1.2), and the results of the comparison among modeling techniques (see section 3.1), it was reasoned that the potentially suitable modeling techniques could be Time-To-Event modeling (TTE), or a longitudinal model on day-wise time of arrival, or longitudinal count data on hour-wise yearly time of arrival, or Markov Chains modeling. Therefore, the former three techniques are addressed in section 2.4, whereas the latter is described in section 2.5.

## 2.3. Parameters estimation

Since the chosen approach for data analysis and parameter estimation is to employ mixed-effects modeling, which is not commonly applied to healthcare data and for which one main

field of the application resides in pharmacometrics, it was decided to try to adapt to the purpose of this thesis a software that is designed and optimized for pharmacometrics. The software in question is Monolix, which is, according to its developers, "*the most advanced and simple solution for nonlinear mixed-effects modeling (NLME) for pharmacometrics*" [29]. It is based on the "Stochastic Approximation Expectation-Maximization" (SAEM) algorithm, which is proven to provide reliable convergence for all types of data [29, 30]. Among the advantages of employing such software for this thesis, it is noteworthy that it automatically generates interactive diagnostic plots, in which the population can be easily split into subgroups or stratified by any included variable of interest [29].

In its native field of application, Monolix is mainly used by academic research institutes, the pharmaceutical industry, and the pharmaceutical regulatory agencies, to perform pre-clinical and clinical population pharmacokinetic and pharmacodynamic modeling, and for systems pharmacology [29]. However, since the software can cover a wide range of data types, models, and statistical features, it was reasoned that it could be compatible with the patient flow data contained in the hospital datasets.

### 2.3.1. General setup before parameter estimation

Before parameter estimation, four steps must be completed.

**The first step** consists of importing the dataset and defining the variables intended to be used in the model. At least one variable whose purpose is to function as an identifier, one that serves as a time, and one that serves as observation, are mandatory for the application of this thesis. For what concerns the observation, also its nature must be defined (e.g., continuous or categorical). Additionally, several other variables can be added and defined (e.g., categorical covariate, occasion, event ID, ...). For what concerns this thesis, the only variable types that were used are ID, time, observation, several categorical and continuous covariates, and the label "IGNORE", which allows importing the whole dataset without necessarily having to use all its variables for the computation.

**The second step** consists in defining a structural model. In this model, it is possible to design which are the parameters to be estimated, which characteristics these must have, whether these parameters must have some kind of inter-dependence between each other, what is meant to be the designated output, what is the relationship between the parameters to be estimated and the output, and which constraints define the simulation environment (e.g., if a specific probability must be forced to 0). For what concerns this thesis, the structural model was rewritten several times according to the modeling technique that was put to the test.

**The third step** consists in specifying the initial values for the parameters to be estimated. The values that can be initialized are the fixed effects, the standard deviations of the random effects, the dependency of the population parameters on the covariates, and the residual error parameters. Still, the choice of each initial parameter can be constrained by the statistical model, set up in the fourth step, that is selected for that specific parameter. Only the former three kinds of values are relevant for what concerns this thesis. If this step is skipped, Monolix applies an initial default value equal to 1 for the fixed effects and the standard deviations of the random effects, equal to 0 for the dependency of the population parameters on the covariates. Moreover, in this step, it is possible to choose for each parameter to be estimated whether to estimate it with the "fixed" method, to use the "Maximum Likelihood Estimation", or to use the "Maximum A Posteriori" estimation. These three methods are discussed more in detail in sub-section 2.3.2.

**The fourth step** includes setting up the statistical model and the tasks to be performed. These consist in specifying which parameters to be estimated are meant to be described by a distribution, selecting which distributions to use, defining potential correlations between different sources of random effect, defining for which parameters to include a dependency on which covariates, defining which tasks to perform and defining the settings for each of them. Changes in the statistical model can change the settings regarding the third step, i.e., the choice of some of the initial values for parameter estimation. This happens, for instance, when a new dependence on a covariate is added to the statistical model, which leads to the creation of a new parameter to be initialized, i.e., the dependency of the population parameter(s) on that covariate.

### Distributions and automatic initialization of the parameters

For each parameter meant to be described by a distribution, it is required to select the option "RANDOM EFFECTS" in the Monolix tab used for the above-described "fourth step" of general setup. Such a "random effect" is the random variable used to describe the inter-individual variability of each parameter for which the option is enabled [31]. The most commonly used types of distribution to describe the parameters in nonlinear mixed-effects modeling are normal, log-normal, logit-normal, and probit-normal distribution. Assuming the existence of a Gaussian transformation of the parameters to be estimated, i.e., a monotonic function $h$ such that $h(\psi)$ is normally distributed, then there exists a standard deviation $\omega$ such that for each individual $i$ [31]:

$$h(\psi_i) \sim \mathcal{N}(h(\bar{\psi}_i), \omega^2), \tag{2.3}$$

where $\bar{\psi}_i$ is the predicted value of $\psi_i$ [31]. If no covariate is included in the estimation, then the predicted value of $\psi_i$ is $\bar{\psi}_i = \psi_{\text{pop}}$ and, therefore, equation 2.3 can be rewritten as follows [31]:

$$h(\psi_i) \sim \mathcal{N}(h(\psi_{pop}), \omega^2).$$

Accordingly, the transformation $h$ defines the distribution of $\psi_i$ [31].

**Normal distribution in $(-\infty, +\infty)$** if $h(\psi_i) = \psi_i$:

$$\psi_i \sim \mathcal{N}(\bar{\psi}_i, \omega^2) \iff \psi_i = \bar{\psi}_i + \eta_i, \qquad \text{where } \eta_i \sim \mathcal{N}(0, \omega^2). \tag{2.4}$$

**Log-normal distribution in $(0, +\infty)$** if $h(\psi_i) = \log(\psi_i)$. A log-normally random variable can only take positive values and it can be represented as follows [31]:

$$\begin{aligned} &\log(\psi_i) \sim \mathcal{N}(\log(\bar{\psi}_i), \omega^2) \iff \log(\psi_i) = \log(\bar{\psi}_i) + \eta_i \iff \\ &\psi_i = \bar{\psi}_i e^{\eta_i}, \qquad \text{where } \eta_i \sim \mathcal{N}(0, \omega^2) \text{ and } \bar{\psi}_i \text{ is the median.} \end{aligned} \tag{2.5}$$

**Logit-normal distribution in $(0, 1)$** if $h(\psi_i) = \log(\frac{\psi_i}{1-\psi_i})$. A logit-normally random variable can only take positive values in the interval (0,1) and it can be represented as follows [31]:

$$\begin{aligned} \text{logit}(\psi_i) = \log(\frac{\psi_i}{1-\psi_i}) \sim \mathcal{N}(\text{logit}(\bar{\psi}_i), \omega^2) \iff \text{logit}(\psi_i) = \text{logit}(\bar{\psi}_i) + \eta_i, \\ \text{where } \eta_i \sim \mathcal{N}(0, \omega^2). \end{aligned} \tag{2.6}$$

**Probit-normal distribution in $(0, 1)$**, which is the "*inverse cumulative quantile function* $\Phi^{-1}$ *associated with the standard normal distribution* $\mathcal{N}(0, 1)$", and can be represented as follows [31]:

$$\text{probit}(\psi_i) = \Phi^{-1}(\psi_i) \sim \mathcal{N}(\Phi^{-1}(\bar{\psi}_i), \omega^2). \tag{2.7}$$

In this thesis, a feature for a first setup of the initial parameters, the so-called "auto-init", was employed after having chosen which kinds of distributions to use to describe each of the parameters. Thanks to this feature, some preliminary initial values were automatically produced for all the parameters, not to find the perfect values but rather to have all the parameters in a starting range that is good enough for performing the first estimation. This initialization was computed without inter-individual variability and depended on a random seed ("123456" was employed), using the data from the 12 first individuals and all the observations mapped to a model output [32].

After the population parameters were estimated at least once, it became possible to use the last estimates as initial values for a new estimation [33]. This could be applied to all the last estimates or only to the fixed effects [33].

### Introduction of covariate effects

The modeling approach allowed for the inclusion of covariates on one, several, or all fixed-effect parameters. Each time a covariate was added, a $\beta$ term was added to the individual model in a way that differed according to whether such covariate was continuous or discrete. In the first case, the covariate was "*added linearly to the transformed parameter, with a coefficient* $\beta$" [33]. In the second case, the initial value for the reference category was set to the value of the fixed effects. In contrast, for all other categories, it was set to the initial value for the fixed effect plus the initial value of the $\beta$, in the transformed parameter space [33].

In this thesis, log-normal distributions were selected for all the estimated parameters for the reasons reported in section 2.5. Therefore, the example provided here shows how continuous and discrete covariates were introduced to estimate a parameter to be described by a log-normal distribution. Let a parameter named $q34$ be defined as a transition rate between two model states and let it be decided to describe it with a log-normal distribution, then the basic formula to model such a parameter without including covariates will be the following:

$$\log(q34) = \log(q34\_\text{pop}) + \eta\_q34, \tag{2.8}$$

where $q34\_\text{pop}$ represents the value of the so-called "fixed effects", and $\eta\_q34$ represents the random effects. If a continuous covariate is added, e.g., $age$, the consequently added $\beta$ term will introduce an exponential relationship between the covariate and the parameter $q34$. Therefore, the formula shown in equation 2.8 will be rewritten as follows:

$$\log(q34) = \log(q34\_\text{pop}) + \beta\_q34\_age \times age + \eta\_q34. \tag{2.9}$$

Conversely, in case a categorical covariate is added, e.g., $sex$, the consequently added $\beta$ term will represent the difference between the typical population value for the reference group, i.e., "man", and the value for the other group, i.e., "female", on the log-transform space [34]. Therefore, the formula shown in equation 2.8 will be instead rewritten as follows:

$$\log(q34) = \log(q34\_\text{pop}) + \beta\_q34\_sex\_Man \times [sex = Man] + \eta\_q34. \tag{2.10}$$

Accordingly, if both the continuous covariate *age* and the categorical covariate *sex* are added, the formula will take into consideration both the contributions and will be the following:

$$\begin{aligned}\log(q34) = \log(q34\_\text{pop}) + \beta\_q34\_age \times age \\ +\beta\_q34\_sex\_Man \times [sex = Man] + \eta\_q34.\end{aligned} \tag{2.11}$$

### 2.3.2. Probability distributions and parameter estimation

The first task performed on dataset D3 was the estimation of the population parameters, which was carried out using the SAEM algorithm, i.e., "Stochastic Approximation Expectation-Maximization". The latter was chosen due to its rigorously proven convergence[30] and its capability to work efficiently with categorical data models, count data models, and time-to-event (TTE) models [30], thus potentially offering a suitable framework for this thesis.

SAEM comprises two phases: an exploratory phase over a vast parameter space, where the goal is to move towards a neighborhood of the maximum likelihood, and a smoothing phase, aiming to converge with greater precision towards the maximum likelihood. The algorithm does not compute the likelihood explicitly; therefore, it does not know beforehand where local maxima of the likelihood are located in the parameter space to be explored. This implies that the SAEM algorithm "*converges under quite general hypotheses to a maximum [...] of the likelihood*", [35] and the probability of converging to the global maximum after a small number of iterations is high, but this probability is not equal to 1 [35].

Nevertheless, it is possible to take procedural precautions to help the algorithm escape from local maxima when it falls into them [30]. Indeed, the probability of the algorithm ending up being stuck on a local maximum can be drastically reduced, and, contextually, the speed at which the algorithm reaches convergence can be drastically increased by using proper parameter initial estimates [33]. By default, the software sets the initial values for the fixed effects and the initial standard deviations of the random effects to 1. The initial values for the dependency of the population parameters on the covariates are instead set to 0 by default.

Three possible methods can be selected for the estimation of the parameters [33]. One option is the so-called "Fixed" method, which implies that the parameter is not estimated and is kept to its initial value [33]. Another option consists of the "Maximum Likelihood Estimation" method (default option), which implies that the parameter is estimated using the maximum likelihood criterion and employing only the information available from the data [33]. The last possible option is the "Maximum A Posteriori estimation" (MAP)

method, which requires the user to define a typical value and a standard deviation for the parameter to be estimated [33], so to perform Bayesian estimation of the parameter, i.e., produce an estimate that "*maximizes a penalized version of the maximum likelihood*" based on the given prior distribution [36]. It is possible to combine Maximum A Posteriori estimation for specific population parameters with Maximum Likelihood Estimation for other population parameters [36].

Since no certain prior information about the population was provided, and since the primary goal of this thesis, which is stated in detail in section , is to design and implement an empirical model, it was decided not to employ the Bayesian estimation technique for any of the population parameters. Conversely, the so-called "Fixed" method and the "Maximum Likelihood Estimation" method were used. The Maximum Likelihood Estimation (MLE) method maximizes the likelihood:

$$\mathcal{L}_y(\theta) = p(y;\theta) = \int p(y,\psi;\theta)d\psi, \tag{2.12}$$

that is the joint distribution of the observed data $y$ and the population parameters $\theta$, in which $\psi$ represents the individual parameters [36]. The likelihood in equation 2.12 is maximized by finding the set of parameters:

$$\hat{\theta}^{\text{MLE}} = \underset{\theta}{\text{argmax}}\, \mathcal{LL}_y(\theta). \tag{2.13}$$

To further reduce the risk of the algorithm being stuck on a local maximum, a feature that in Monolix is called "simulated annealing" was enabled and employed throughout this thesis. Given that the size of the parameter space that is explored during the "exploratory phase" of SAEM "*depends on the standard deviations of the random effects ($\omega$) and on the standard deviations of the residual error*", by enabling the simulated annealing it was possible to keep the explored parameter space larger for a longer time by constraining "*the variance of random effects and of the residual error to decrease by maximum 5% between two iterations*" [30]. This helped escape local maxima and improve the convergence towards the global maximum because when the standard deviations of the population parameters ($\omega$) and thus also the standard deviations of the conditional distributions were large (through equation 2.15), it remained still possible for the individual parameters sampled at the $k^{th}$ iteration to be far away from the ones sampled at the $(k-1)^{st}$ iteration, which gave the algorithm better margin to escape local maxima [30].

The ideal way of "following" convergence while estimating the parameters through the SAEM algorithm would be to compute the likelihood explicitly, which is computationally

not an easy task due to the need to integrate all the possible values of the individual parameters [30]. Therefore, a convergence indicator was used to assess convergence during the parameter estimation rather than explicitly computing the likelihood in the same task. Such a convergence indicator, calculated at each SAEM step, is defined as the "*joint probability distribution of the data and the individual parameters and can be decomposed using Bayes*" law into two terms that have an analytical expression and are easy to calculate [30]:

$$p(y, \psi^{(k)}; \theta^k) = p(y|\psi^{(k)}; \theta^k)p(\psi^{(k)}; \theta^k), \tag{2.14}$$

where the individual parameters $\psi^{(k)}$ are the same ones sampled from the conditional distribution

$$p(\psi_i|y_i, \theta^k) \tag{2.15}$$

at the $k^{th}$ SAEM iteration and used by the algorithm to compute new population parameters at the same $k^{th}$ iteration [30], as is explained hereunder in a more detailed description of the two phases of the algorithm. It is noteworthy that a typical convergence indicator curve progressively decreases with the increase of the iteration number until it stabilizes [30], i.e., until it starts having small oscillations around the same value without ever drifting away from it.

- **The exploratory phase** consists of two steps. In the first step, the individual parameters $\psi_i$ are generated for each individual $i$ in the dataset from the conditional distribution expressed by the equation 2.15, in which $y_i$ consists of the observations for the individual $i$ and $\theta^k$ is the set of population parameters at the $k^{th}$ iteration of the SAEM algorithm. For the first iteration, $\theta^k$ corresponds to the pre-set initial conditions [30]. The conditional distribution in question has no analytical expression; therefore, Monolix employs a Markov Chain Monte Carlo procedure to generate one value per individual $i$.

  In the second step, new population parameters are calculated for the iteration step $(k+1)^{st}$ by averaging over the sampled individual parameters [30]. Equation 2.16 shows how this is computed for the values of the fixed effects.

  $$\theta^{k+1} = \frac{1}{N}\sum_{i=1}^{N}\psi_i \tag{2.16}$$

  After the execution of the two steps that were introduced above, a new iteration can start. Therefore, new individual parameters can be generated according to equation 2.15, where the set of population parameters $\theta^k$ at the $k^{th}$ iteration consists now in the parameter set computed in the previous iteration [30]. A new set of population

parameters can thus be generated according to equation 2.16, and this process can continue until either the auto-stop criterion is met or the pre-set maximum number of iterations is reached [30]. The net purpose of the exploratory phase is to "*converge to a neighborhood of the maximum likelihood*" [30]. However, the user may not have pre-set any auto-stop criterion, or such criterion might not get triggered within the pre-set maximum number of iterations. In these two cases, after the convergence to a neighborhood of the maximum likelihood is achieved, the algorithm keeps performing a "random walk" in such a region of the likelihood up to the maximum number of iterations. At this point, the algorithm automatically shifts to its next phase, i.e., the smoothing phase.

- **The smoothing phase** consists of two steps and is meant to make the algorithm converge with greater precision towards the maximum likelihood. In the first step, as in the exploratory phase, individual parameters $\psi_i$ are generated from the conditional distribution shown in equation 2.15 using a Markov Chain Monte Carlo procedure [30]. However, in this case, $\theta^k$ for the first iteration corresponds to the last parameter set produced in the exploratory phase [30].

  In the second step, new population parameters $\theta^{k+1}$ are computed from the individual parameters $\psi_i$ coming from all the previously computed smoothing iterations [30]:

$$\theta^{k+1} = \frac{1}{k}\left[\frac{1}{N}\sum_{i=1}^{N}\psi_i^{(1)} + \frac{1}{N}\sum_{i=1}^{N}\psi_i^{(2)} + ... + \frac{1}{N}\sum_{i=1}^{N}\psi_i^{(k)}\right]. \tag{2.17}$$

  For computational reasons, since applying equation 2.17 would be extremely demanding due to its requirement of memorizing all the previous individual parameters, and since its first part at the $(k+1)^{st}$ step corresponds to what would be computed at the $k^{th}$ iteration, the equation can be rewritten as follows [30]:

$$\theta^{k+1} = \frac{1}{k}\left[(k-1)\theta^k + \frac{1}{N}\sum_{i=1}^{N}\psi_i^{(k)}\right]. \tag{2.18}$$

  Moreover, when the second step of the smoothing phase is applied to its first iteration, however, only one set of individual parameters $\psi_i$ is available. Therefore, for the iteration in question, the second step of the smoothing phase behaves in the same way as the second step of the exploratory phase, shown in equation 2.16.

After having performed the first task on the data for estimating the population parameters using the SAEM algorithm, four more tasks were performed in this thesis before plotting

the results. These tasks are addressed in detail in sub-sections 2.3.3, 2.3.4, 2.3.5, and 2.3.6.

However, before proceeding any further with explaining the other performed tasks, it is important to provide more details about the functioning of the so-called "Markov Chain Monte Carlo" procedure that the SAEM algorithm employs both during its exploratory and during its smoothing phase. Therefore, this explanation is provided hereunder.

### Markov Chain Monte Carlo procedure

In both the phases of the SAEM estimation, at each iteration, it is necessary to generate individual parameters from the conditional distribution defined by equation 2.15 for each individual $i$ of the population. Such conditional distribution for an individual $i$ "*represents the uncertainty of the individual parameter value*", taking into consideration the observed data for that individual, the covariate values for that individual, and the already estimated population parameters in terms of both fixed effects and standard deviation of the random effects [37]. As stated when describing the phases of the SAEM algorithm, however, the conditional distribution has no analytical expression. Thus, it is impossible to directly calculate the probability for a given set of individual parameters $\psi_i$ for individual $i$, "*but it is possible to obtain samples from the distribution using a Markov Chain Monte Carlo (MCMC) procedure*" [37].

MCMC algorithms "*consist of constructing a stochastic procedure which, in its stationary state, yields draws from the probability distribution of interest*", thus allowing to sample from probability distributions for which it is usually difficult to perform direct sampling [37]. Among the several algorithms within the class of MCMC methods, Monolix uses the so-called Metropolis-Hastings (MH) algorithm, which was used in this thesis also to compute the conditional distributions. Therefore, the functioning of the employed MCMC algorithm is properly described directly in the following sub-section (2.3.3), which is specifically dedicated to explaining the computation of the conditional distributions.

## 2.3.3. Conditional distribution

As briefly mentioned in the previous sub-section, the second task performed on the processed dataset is the estimation of the individual conditional distributions for the individual parameters and includes the sampling of sets of parameter values from these distributions for each individual [37]. This task cannot be performed before the SAEM population parameters estimation is completed since, thanks to it, the population distribution for each parameter is made available[37].

However, in addition to the information provided by the population parameters, it is also important to look at single individuals in the dataset separately and estimate their individual parameter values, which is the primary goal of this second task and is described by a conditional distribution that is different for each parameter [37]. Such distribution, which is shown in equation 2.19, is called "conditional" because it is conditional on the already estimated population parameters, but it also considers the observed data $y_i$ for a specific individual $i$ and the fact that the individual belongs to the population distribution [37]. If covariates are included in modeling a specific parameter, these will also appear for all individuals in their conditional distributions related to such parameter [37].

$$p(\psi_i|y_i;\hat{\theta}) \tag{2.19}$$

Whereas in the "Empirical Bayes Estimates" (EBEs) task, which is described in sub-section 2.3.4, only the most probable value of each parameter for each individual is calculated, i.e., the maximum of each conditional distribution (also called conditional mode), the "conditional distribution" task instead estimates the whole conditional distribution for each individual and each parameter [37]. This is done to obtain detailed information about the uncertainty of the individual parameter values [37]. Anyways, as discussed in sub-section 2.3.2, conditional distributions cannot be computed in close form. Thus, it is impossible to calculate the probability for given parameters directly. Nevertheless, obtaining samples from the distributions is possible using a Markov Chain Monte Carlo procedure. The MCMC algorithm that was used in this thesis is the so-called Metropolis-Hastings algorithm, which allows iteratively simulating a sequence of individual parameters by rewriting the conditional distribution from equation 2.19 as follows [37]:

$$p(\psi_i|y_i) = \frac{p(y_i|\psi_i)p(\psi_i)}{p(y_i)}, \tag{2.20}$$

where $p(y_i|\psi_i)$ is the conditional density function of the data when the individual parameter values are known, and $p(\psi_i)$ is the density function of the individual parameters [37]. Both $p(y_i|\psi_i)$ and $p(\psi_i)$ can be computed, whereas the likelihood "$p(y_i)$ *has no closed form solution but it is constant*" when the goal is to optimize the formula with respect to the individual parameters $\psi_i$, since $p(y_i)$ does not depend on $\psi_i$ [37, 38].

Accordingly, thanks to the Metropolis-Hastings MCMC algorithm, at each $l^{th}$ iteration and for all the individuals, a new vector of random effect values $\eta_i^{(l)}$ is drawn from a **proposal distribution** and new individual parameters $\psi_i^{(l)}$ are calculated from such random effect values [37]. These new individual parameters can be either accepted or rejected according

to the value of their corresponding "acceptance probability" $\alpha$, calculated as follows[37]:

$$\alpha = \frac{p(\psi_i^{(l)})\ p(y_i|\psi_i^{(l)})}{p(\psi_i^{(l-1)})\ p(y_i|\psi_i^{(l-1)})}, \tag{2.21}$$

which depends on the probability of the parameters in the population distribution $p(\psi_i)$, and on the likelihood of the individual data $y_i$ given these parameters, at the current $(l^{th})$ and at the previous $(l-1)^{st}$ iteration. If $\alpha$ results to be greater than 1, i.e., if the combined probability at the $l^{th}$ iteration is greater than the combined probability at the $(l-1)^{st}$ iteration, the new draw of individual parameters is accepted and kept. Conversely, if $\alpha$ results to be smaller than 1, the new draw is kept only with probability $\alpha$ [37]. This so-far described acceptance probability guarantees that the sequence of parameters converges to the individual conditional distributions.

To draw new vectors of random effect values to be evaluated with the above-described criterion, three types of proposal distributions (kernels) were used sequentially with a (2,2,2) turnover pattern to make the Markov Chain more robust [37]. The first proposal distribution to be used is the population distribution for the random effects [37]:

$$\eta_i^{(l)} \sim \mathcal{N}(0, \Omega), \tag{2.22}$$

where $\Omega$ is the estimated variance-covariance matrix for the random effects. The second proposal distribution to be used is an unidimensional Gaussian random walk, in which "*each random effect drawn at the previous iteration is perturbed with a random variable drawn from a normal distriburion*" [37]:

$$\eta_i^{(l)} = \eta_i^{(l-1)} + \xi^{(l)}, \qquad \text{with } \xi^{(l)} \sim \mathcal{N}(0, \theta). \tag{2.23}$$

The third proposal distribution to be used is a multidimensional Gaussian random walk, in which each random effect drawn at the previous iteration is perturbed with a gaussian vector [37]. For both the proposal distributions based on Gaussian random walks, the variance of the Gaussian random variables was automatically adjusted by Monolix at each iteration to reach an optimal acceptance ratio [37].

By extracting new vectors of random effect values with the aforementioned (2-2-2) pattern, at each iteration six parameters are drawn iteratively from the three proposal distributions [37]. At the first iteration, the equation for the acceptance probability (2.21) is rewritten as:

$$\alpha = \frac{p(\psi_i^{(1)})\ p(y_i|\psi_i^{(1)})}{p(\psi_i^{\text{SAEM}})\ p(y_i|\psi_i^{\text{SAEM}})}. \tag{2.24}$$

This means that, for the first iteration, the employed reference value is the value of the parameter estimated by the SAEM algorithm during the first task (sub-section 2.3.2). A candidate value is thus drawn from distribution 2.22 and accepted or rejected according to the acceptance probability $\alpha$ (equation 2.24), then another candidate is drawn from distribution 2.22, and equation 2.21 is used to evaluate it against the previous candidate, using the previous candidate as a reference if this was accepted before testing the new candidate [37]. At this point, a candidate value is drawn from distribution 2.23 and accepted or rejected according to the acceptance probability $\alpha$ (equation 2.21) [37]. The same is applied to the second draw from distribution 2.23 and to both the draws from the last type of proposal distribution. At the end of the sequence, only the last accepted value out of the six is kept for the current iteration [37].

In the second iteration, the value used as a reference is the one accepted at the first iteration, and so on. Eventually, the accepted values will "*cover the whole distribution since the acceptance rate allows enough flexibility to accept some values that may be far from the peak*" [37]. Therefore, after a transition period, the accepted values will follow the conditional distribution, and together they will represent an estimation of the distribution itself [37].

When, for all the parameters, the average conditional means and standard deviations of the last 50 iterations do not deviate by more than 2.5% in each direction from the average and standard deviation values at the $k^{th}$ iteration, the algorithm stops automatically [37] and calculates the conditional mean (equation 2.25) and standard deviation for each parameter for each individual, by averaging over the values drawn at all the iterations [37].

$$\hat{\psi}_i^{\text{mean}} = \frac{1}{K}\sum_{k=1}^{K}\psi_i^k \tag{2.25}$$

In this way, all the individual conditional distributions can be summarised even though they have no explicit formula. Conditional mean and standard deviation for each individual, as well as an average of the conditional mean over the whole population ($E(\psi|y)$, equation 2.26), and the standard deviation of the conditional means over the entire population ($sd(\psi|y)$), are all included in the output of this second task [37].

$$E(\psi|y) = \frac{1}{N}\sum_{i=1}^{N}\psi_i^{\text{mean}} \tag{2.26}$$

Furthermore, among all the samples from the conditional distributions drawn by the algorithm, 10 of them are saved for being used to improve the performance of the diagnostic

plots by including the uncertainty of the individual parameters and for being used to perform statistical tests to diagnose the model [39]. Finally, a table containing a summary of the estimated conditional mean is also generated, including minima, first quartiles, medians, third quartiles, and maxima.

### 2.3.4. Empirical Bayes Estimates (EBEs)

The third task that was performed was the estimation of "*the most probable value of the individual parameters, given the estimated population parameters and the data of each individual*", i.e., the estimation of "*the mode of the conditional parameter distribution for each individual*" (equation 2.27) [38].

$$\hat{\psi}_i^{\text{mode}} = \underset{\psi_i}{\text{argmax}}\ p(\psi_i|y_i;\hat{\theta}) \tag{2.27}$$

Starting from the conditional distribution shown in equation 2.19, where $\psi_i$ are the individual parameters for individual $i$, $\hat{\theta}$ are the estimated population parameters, and $y_i$ are the observations for individual $i$, the mode of the conditional parameter distribution for each individual can be computed according to equation 2.27 [38]. However, since it is not possible to calculate the probability for a given $\psi_i$ directly, it is necessary to employ a Markov Chain Monte Carlo procedure to obtain samples from the conditional distribution [38], which allows rewriting the conditional distribution as in equation 2.20, as described in the last part of sub-section 2.3.2, and in sub-section 2.3.3. Accordingly, equation 2.27 can be rewritten as [38]:

$$\hat{\psi}_i^{\text{mode}} = \underset{\psi_i}{\text{argmax}}\ [p(y_i|\psi_i)\ p(\psi_i)], \tag{2.28}$$

where the first term represents the probability of the data for individual $i$ given the individual parameters $\psi_i$, and the second term represents the probability of the individual parameters $\psi_i$. However, since $\psi$ "*is a multidimensional vector impacting the model prediction, and this prediction may be the solution of an ODE system, for instance*", the computation of $p(\psi_i)$ is complicated and demanding [38]. Therefore, since it does not use derivatives, the so-called Nelder-Mead Simplex algorithm [40] was employed to find for each individual $i$ the $\psi_i$ that maximizes the conditional distribution, i.e., the conditional mode [38]. Once this task was done, its results could be used as individual parameters for individual predictions, e.g., for the plots of the individual fits [39]. Moreover, each individual's minimum, first quartile, median, third quartile, and maximum were also outputted, as well as the predictions based on the conditional modes [39].

By running this task after having run the "conditional distribution" task (sub-section 2.3.3), it was possible to use the mean of the conditional distribution for each individual as a starting point for the Nelder-Mead Simplex algorithm instead of having to use an approximate mean calculated from the last iterations of SAEM algorithm (sub-section 2.3.2).

### 2.3.5. Standard errors

The fourth task that was performed could be carried out either by using the linearization method or by using the stochastic approximation [41]. This task returns the correlation matrix of the estimates and the uncertainty and relative uncertainty of the estimated population parameters, which are calculated by estimating the so-called Fisher Information Matrix (FIM) [41]. Furthermore, this task also computes a Wald test for each beta parameter used for the covariate effect to check if the covariate effect is relevant, which helps detect over-parameterization of the model [41].

For what concerns the evaluation of the uncertainty of the population parameters, the Fisher Information Matrix $\boldsymbol{I}(\hat{\theta})$ was computed as:

$$\boldsymbol{I}(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2}\log(\boldsymbol{\mathcal{L}}_y(\hat{\theta})), \tag{2.29}$$

i.e., as minus the second derivatives of the observed likelihood [41]. The log-likelihood, however, cannot be calculated in closed form and thus applied to the Fisher Information Matrix [41]. Therefore, it was calculated by stochastic approximation [41]. After the calculation of the Fisher Information Matrix is achieved, it is possible to calculate the so-called variance-covariance matrix $\boldsymbol{C}(\hat{\theta})$ as the inverse of the FIM $\boldsymbol{I}(\hat{\theta})$ [41], as shown in equation 2.30:

$$\boldsymbol{C}(\hat{\theta}) = \boldsymbol{I}(\hat{\theta})^{-1}. \tag{2.30}$$

Thereafter, it is possible to calculate the standard errors for each parameter $\hat{\theta}_k$ as shown in equation 2.31:

$$s.e(\hat{\theta}_k) = \sqrt{\tilde{\boldsymbol{C}}_{kk}(\hat{\theta})}, \tag{2.31}$$

i.e., "*as the square root of the diagonal elements of the inverse of the Fisher Information Matrix*" [41]. In Monolix, however, the FIM and the variance-covariance matrix are calculated on the transformed normally distributed parameters and, therefore, the jacobian $J$ had to be used to obtain the variance-covariance matrix $\tilde{\boldsymbol{C}}$ for the untransformed parameters, as in equation 2.32.

$$\tilde{\boldsymbol{C}} = \boldsymbol{J}^T\boldsymbol{C}\boldsymbol{J} \tag{2.32}$$

Concerning the correlation matrix, instead, this is calculated from the off-diagonal element of the variance-covariance matrix as in equation 2.33.

$$corr(\theta_i, \theta_j) = \frac{\tilde{\boldsymbol{C}}_{ij}}{\text{s.e}(\theta_i)\ \text{s.e}(\theta_j)} \tag{2.33}$$

It contains the correlation between each pair of population parameters independently of the correlation of the random effects, and it is essential because it can be used to detect over-parameterization of the model. As a rule of thumb for being confident in the model not being over-parameterized, the software automatically suggests making sure that the so-called "condition number", which is the ratio between the maximum and the minimum eigenvalue of the correlation matrix, results being smaller than 100.

Furthermore, the relative standard errors calculated by this task can be used to perform a Wald test to suggest if any of the added covariates should be removed from the model [41]. It tests the null hypothesis for which the $\beta$ parameter estimated by the SAEM algorithm is equal to 0, and its test statistic is the following:

$$W = \frac{\hat{\beta}}{se(\hat{\beta})}, \tag{2.34}$$

where $\hat{\beta}$ is the beta value estimated by SAEM, and $se(\hat{\beta})$ is the associated standard error calculated during the task described in this sub-section, i.e., "standard errors" [42]. The test statistic described by equation 2.34 is compared to a t-distribution with one degree of freedom, and the higher the p-value, the more likely the tested covariate should be removed from the model [42].

### 2.3.6. Likelihood

The fifth task that was performed is the estimation of the hereunder defined log-likelihood:

$$\mathcal{LL}_y(\hat{\theta}) = \log(\mathcal{L}_y(\hat{\theta})) \triangleq \log(p(y;\hat{\theta})), \tag{2.35}$$

where $\hat{\theta}$ contains the population parameter estimates for the model in use, and $p(y;\hat{\theta})$ represents the probability distribution function of the observed data $y$ given the estimates of the population parameters $\hat{\theta}$ [43]. Although it cannot be computed explicitly for nonlinear mixed-effects models, it is possible to estimate it using the "importance sampling Monte Carlo method", which provides an unbiased estimation of the log-likelihood with controllable variance, even for nonlinear models [43].

Once the estimation of the log-likelihood was completed, two likelihood indicators were computed: -2$\mathcal{LL}_y(\hat{\theta})$, and the "corrected Bayesian Information Criterion" (BICc). Starting from the value of the former, the latter was computed as:

$$BICc = -2\mathcal{LL}_y(\hat{\theta}) + \dim(\theta_{\mathrm{R}}) \log(N) + \dim(\theta_{\mathrm{F}}) \log n_{\mathrm{tot}}, \tag{2.36}$$

where $N$ is the number of subjects. BICc penalizes not only with the logarithm of the number of subjects $\log(N)$, but also with the size of $\theta_{\mathrm{R}}$, and with the size of $\theta_{\mathrm{F}}$, which depend on the number of included covariates and are scaled with the logarithm of the total number of observations ($n_{\mathrm{tot}}$). $\theta_{\mathrm{R}}$ "*represents the random effects and fixed covariate effects involved in a random model for individual parameters*", and $\theta_{\mathrm{F}}$ "*represents all other fixed effects,[...] beta parameters involved in a non-random model for individual parameters, as well as error parameters*" [43].

Due to its penalization dependent on the number of included covariates, it was reasoned that the corrected Bayesian Information Criterion would have been a more reliable indicator to consider while selecting which covariate variables to include in the model. Such a covariates selection process is described more in detail in section 2.6.

## 2.4. Time-To-Event modeling and Longitudinal data modeling

This section addresses the first three of the four modeling ideas that were formerly considered potentially promising for being applied to this thesis within the modeling approach of the nonlinear mixed-effects modeling (sub-section 2.2.3). Conversely, the last one is addressed in section 2.5. The first of the three techniques, discussed hereunder in sub-section 2.4.1, is the so-called "Time-To-Event" modeling. The other two techniques treated in this section are a longitudinal model on day-wise arrival time and a longitudinal count data on hour-wise yearly time of arrival. These two are discussed afterward, in sub-sections 2.4.2 and 2.4.3.

### 2.4.1. Time-To-Event modeling of patients' discharge

The idea behind this Time-To-Event (TTE) approach is to use the "*contact_id*" (see appendix A.1) as the identifier, the variable *hours_in* (see appendix A.1) as the time variable, and as observation the status of the patients in terms of whether they are inside the ED or they have been already discharged in one of the possible ways described by

table 2.3. For describing the observations, a variable was introduced specifically for this technique with the name "*State*". It takes a value equal to 0 for patients who are still in the ED and 1 for the patients who have already been discharged.

For a TTE approach, the functions that play key roles in the analysis are the "survival function", the "hazard function", and the "cumulative survival function" [44]. In this framework, the hazard function $h(t, \psi_i)$ gives the instantaneous rate of the event at time $t$, i.e., "the patient is discharged", for the $i^{th}$ patient, if this has not occurred yet [44]. This function is reported in equation 2.37:

$$h(t, \psi_i) = \lim_{dt \to 0} \frac{S(t, \psi_i) - S(t + dt, \psi_i)}{S(t, \psi_i) dt}. \tag{2.37}$$

The cumulative hazard function $H(t_{\text{start}} + hours\_in_i, t; \psi_i)$ in the interval $[t_{\text{start}} + hours\_in_i, t]$ is defined for the $i^{th}$ patient as in equation 2.38 [44]:

$$H(t_{\text{start}} + hours\_in_i, t; \psi_i) = \int_{t_{\text{start}} + hours\_in_i}^{t} h(t, \psi_i) \, \mathrm{d}t. \tag{2.38}$$

The survival function $S(t, \psi_i)$ gives the probability that the discharge of the patient happens to the $i^{th}$ patient after time $t > t_{\text{start}} + hours\_in_i$, as is shown in equation 2.39 [44].

$$S(t, \psi_i) = \mathbb{P}(T_i > t; \psi_i) = e^{-H(t_{\text{start}} + hours\_in_i, t; \psi_i)} \tag{2.39}$$

Some elaborations were needed to convert the dataset that constitutes the outcome of the pre-processing described in section 2.1.3 into a dataset that could be suitable for performing TTE modeling in Monolix, a sample of which is shown in figure 2.6. These elaborations were performed in Python.

However, after such a reshaping of the dataset, as soon as the process of formulating a structural model had begun, it was reasoned that this technique, even if practically implementable, would have yielded results that would not have been appropriately centered on the actual purpose of this thesis. Thus the study of this approach was discontinued. The reasons behind this discontinuation are discussed more in detail in section 4.4.

| | ID | TIME | OBSERVATION |
|---|---|---|---|
| LINE NUMBER | contact_id | Time | State |
| − | 1000001 | | |
| 2 | 1000001 | 1576.7 | 0 |
| 3 | 1000001 | 1584.86 | 1 |
| − | 1010212 | | |
| 4 | 1010212 | 6080.03 | 0 |
| 5 | 1010212 | 6083.65 | 1 |

Figure 2.6: Two samples from the dataset for TTE modeling.

### 2.4.2. Longitudinal model on day-wise time of arrival

The idea behind this modeling proposal is to use the chief complaint (see "*cause_of_visit*" in appendix A.1) as the identifier instead of using the *contact_id*, to use a day-wise hour of arrival as the time slot to which the observation of each patient is associated, and to employ *LoS_hr* as observation for each patient (see appendix A.1). Concerning the aforementioned day-wise hour of arrival, this variable was introduced specifically for this technique with the name "*ARRIVAL_TIME*", and it represents the time of the day, rounded to the nearest quarter of an hour, at which a given patient reached the emergency department, regardless of which day of the year it happened.

The generation of the patients would have been performed by estimating the parameters of a Poissonian distribution, and the length of stay would have been described as a function of the chief complaint and other potential covariates.

Starting from the dataset that constitutes the outcome of the pre-processing described in section 2.1.3, some elaborations of the dataset were necessary for it to be used to test the technique in question. These elaborations were performed through Python programming, and a sample of the processed dataset obtained in this way is shown in figure 2.7. The latter, from which it is possible to appreciate how the entries are automatically grouped by identifier, i.e., the chief complaints in this case, does not show the possibly usable covariates.

| ☰ | ID ▾ | TIME ▾ | OBSERVATION ▾ |
|---|---|---|---|
| LINE NUMBER ⇅ | cause_of_visit ⇅ | ARRIVAL_TIME ⇅ | LoS_hours ⇅ |
| − | Abstinens | | |
| 3 | Abstinens | 0.25 | 1.07528 |
| 4 | Abstinens | 7.5 | 3.50306 |
| 2 | Abstinens | 17.25 | 4.12306 |
| 6 | Abstinens | 17.25 | 25.3747 |
| 5 | Abstinens | 17.5 | 5.28611 |

Figure 2.7: Sample of one chief complaint from the dataset for "Longitudinal model on day-wise time of arrival".

However, right after having reshaped the dataset so to visualize it in Monolix in a meaningful way, it became clear that this approach would not have had any real meaningfulness for the aims of this study. Consequently, for the reasons discussed in section 4.4, also the study of this approach was discontinued, and no actual result was produced.

### 2.4.3. Longitudinal count data on hour-wise yearly time of arrival

The idea behind this kind of longitudinal count data is to use the so-called "$contact\_id$" (see appendix A.1) as the identifier, the number of patients present in the emergency department at a given time as observations, and the variables "$hours\_in$" and "$hours\_out$" (see appendix A.1) as times at which the observations of each patient are evaluated. Accordingly, the variables that are used as observations are "$countIN$" and "$countOUT$" (see appendix A.1).

In this framework, when the $i^{th}$ patient reaches the emergency department, "$countIN - 1$" patients are already there at the time of arrival. Then, after a time equal to "$LoS\_hours$", the same $i^{th}$ patient leaves the ED, when now the number of patients in the ED is equal to "$countOUT - 1$".

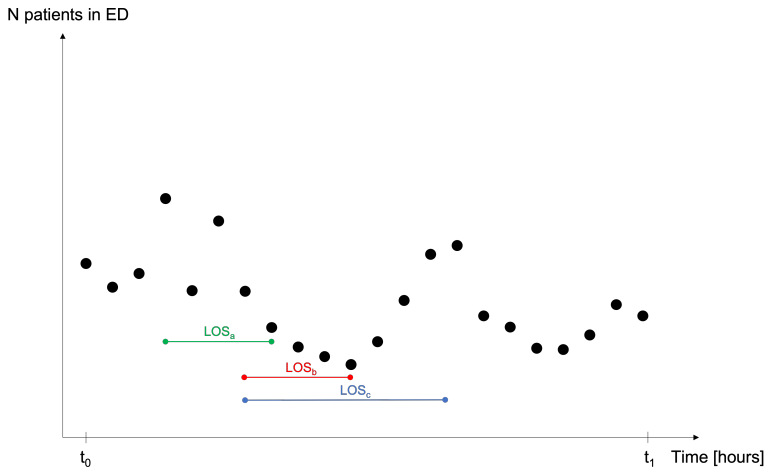



Figure 2.8: "Longitudinal count data on hour-wise yearly time of arrival".

| | ID | OBSERVATION | TIME |
|---|---|---|---|
| LINE NUMBER | contact_id | Count | Time |
| − | 1000001 | | |
| 2 | 1000001 | 47 | 1576.7 |
| 3 | 1000001 | 31 | 1584.86 |
| − | 1010212 | | |
| 4 | 1010212 | 21 | 6080.03 |
| 5 | 1010212 | 36 | 6083.65 |

Figure 2.9: Two samples from the dataset for "Longitudinal count data on hour-wise yearly time of arrival".

The mock conceptual plot presented in figure 2.8, in which the three horizontal lines represent three examples of patient length of stay, graphically shows the idea of creating a model by describing the number of patients in the ED as a function of the arrival time, of the length of stay, and of other potential covariates.

Starting from the dataset that constitutes the outcome of the pre-processing described in section 2.1.3, several elaborations of the dataset were necessary for it to be used to test the technique in question. These elaborations were performed in Python. Figure 2.9 shows a sample of two patients from the so-processed dataset, hiding all the covariate variables, in the way Monolix groups the *contact_id* for this implementation.

By when it became possible to visualize the dataset in Monolix in a meaningful way and start designing the structural model, however, it became clear that this approach would not have had any real meaningfulness for this thesis for the reasons that are discussed in section 4.4. Therefore, the study of this approach was discontinued, and no actual result was produced.

## 2.5. Design of the Markov Chains technique

It was chosen to describe the logistic process as a Markov Chain with "memory 1". In such a framework, the observed data can take values only in a finite and fixed set of nominal categories $\{c_1, c_2, ..., c_k\}$ and the observations $(y_{ij}, 1 \leq j \leq n_i)$ for any $i^{th}$ individual consist in a sequence of random variables [45]. The dependence between observations from the same individual is defined so that, for all $k = 1, 2, ..., K$, to determine the distribution of $y_{ij}$ no older value than the one of the immediately preceding observation $(y_{i,j-1})$ is needed [45]. Accordingly, the probability of the $j^{th}$ observation for the $i^{th}$ individual to be equal to the nominal category $c_k$ can be simplified as in equation 2.40 [45].

$$\mathbb{P}(y_{ij} = c_k | y_{i,j-1}, y_{i,j-2}, ..., \psi_i) = \mathbb{P}(y_{ij} = c_k | y_{i,j-1}, \psi_i) \tag{2.40}$$

In this thesis, it was decided to use the "*contact_id*" (see appendix A.1) as identifier in Monolix, to use as observation the number of the nominal category of the chain, i.e., the state to which the patient belongs at a given time, and to use as time variable the time at which each of the patients starts being in a given associated state. Therefore, the times at which each observation is reported are different for each patient. Consequently, it could not have been possible to use a Discrete-Time Markov Chains (DTMC) approach, which regards the observation times being regularly spaced. Conversely, a Continuous-Time Markov Chains (CTMC) approach was selected since the latter allows to have irregular time intervals between observations. In the CTMC, instead of reasoning in terms of transitioning to a new state or the same one at each time step, as it is done in the discrete case, the system instead "*remains in the current state for some random amount of time before transitioning*" [45]. Therefore, to describe this Markov process, it is necessary to define the so-called "Initial state probability" vector and a matrix of transition rates [45]. The initial state probability vector contains the values $b_k$, which are the probability for the first state in the sequence to be set at the category $c_k$ and are described by equation 2.41 [46].

$$b_k = \mathbb{P}(y_{i,1} = k | \psi_i) \tag{2.41}$$

For what concerns the transition rates, instead, for $k \neq l$ [45]:

$$\mathbb{P}(y_i(t+h) = l | y_i(t) = k, \psi_i) = h\rho_{kl}(t, \psi_i) + o(h), \tag{2.42}$$

and the probability that no transition happens between time $t$ and time $t+h$ is [45]:

$$\mathbb{P}(y_i(s)) = k, \ \forall s \in (t, t+h) | y_i(t) = k, \psi_i) = e^{h\rho_{kk}(t,\psi_i)}, \tag{2.43}$$

given that in a Markov process with $K$ nominal categories, for any $i_{th}$ individual and at any time $t$, the transition rates $\rho_{kl}(t, \psi_i)$ satisfy for any $1 \leq k \leq K$ the property described by equation 2.44.

$$\sum_{k=1}^{K} \rho_{kl}(t, \psi_i) = 0 \tag{2.44}$$

## Choices for implementing the Markov Chains modeling

One of the advantages of using a Markov Chains approach for modeling patient flow in the emergency department is that it allows to easily change the structure of the model in terms of which and how many states to include and in terms of which and how many transitions between states to allow. The first approach that was attempted consisted in defining seven possible states, one of which was common for each patient and represented "being in the ED", whereas the other six steps represented the six possible modalities in which a patient could leave the emergency department (shown in table 2.3). It was reasoned that, despite this being a simple approach, it would have represented a good starting point for modeling the system. Therefore, the pre-processed dataset was re-arranged accordingly, and a structural model in Monolix was designed and implemented. However, due to how Monolix interprets the starting conditions of a Markov chain and due to the software not being initially designed for healthcare logistics, it was found that having a single and common starting state for the chain would have led to the software misinterpreting the starting conditions and therefore produced unreasonable results. This limitation could be overcome in the future by designing ad hoc software for applying nonlinear mixed-effects modeling to healthcare logistics.

Consequently, a second approach was designed, based on splitting the starting state into two different options: "*state 1*" for the patients reaching the emergency department by ambulance ($AmbulYN = 1$) and "*state 2*" for the patients reaching the emergency department by other means of transportation ($AmbulYN = 0$), which are both followed by a common state ("*state 3*") describing that a patient has reached the ED and is

staying within it. From this common state, then, the patients could reach one of the aforementioned discharging states, according to the options shown in table 2.3. In this perspective, the transitions from state 1 to state 3, i.e., associated with the transition rate "*q13*", and from state 2 to state 3, i.e., associated with the transition rate "*q23*", are all dummy transitions that in this thesis serve the sole purpose of solving the problem of the software misinterpreting the starting conditions in a Markov Chain with only one possible starting state. However, since this change introduced one more step in each patient flow through the Markov process, and thus the need for more entries in the dataset and more parameters to be estimated, it was reasoned that the number of states describing the discharge of the patients could be reduced without making the model less effective. Accordingly, the possible output states were reduced from six to four by clumping the 1273 patients who were "redirected to other facilities" and the 61 patients who were "taken in charge by consultants", together with the 4746 patients whose output was classified as "other, unspecified", thus creating a new category called "other" that regrouped the mode of discharge of 6080 patients. The layout of this newly designed chain can be seen in figure 2.10.

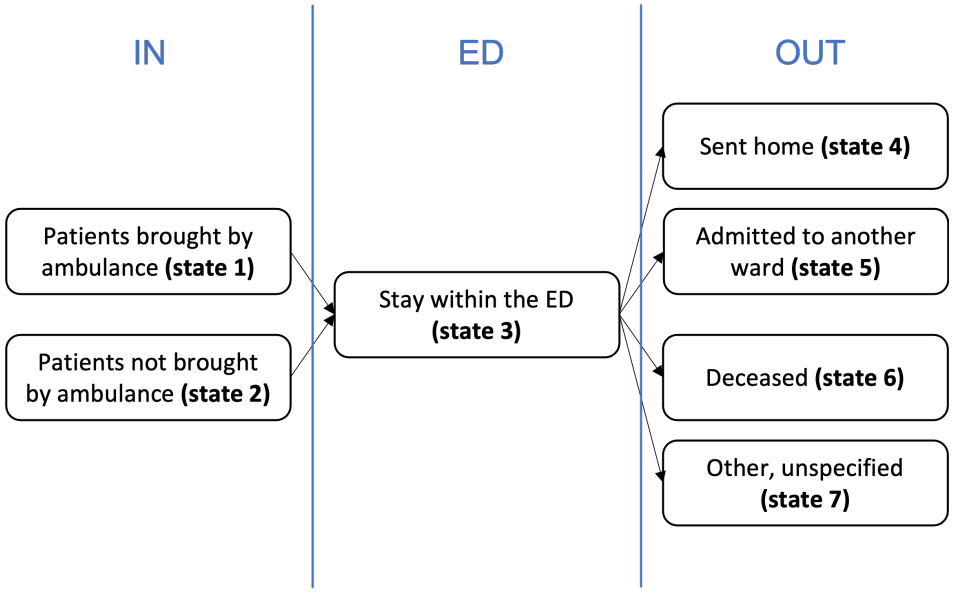



Figure 2.10: 7-states Markov Chain Model.

Since the count of patients who died at the emergency department of Akademiska sjukhuset during 2019 is equal to 103, and these represent 0.21% of all the patients in dataset D1 (49 936), a simplification of the model shown in figure 2.10 was designed, and such simplified model is shown in figure 2.11.

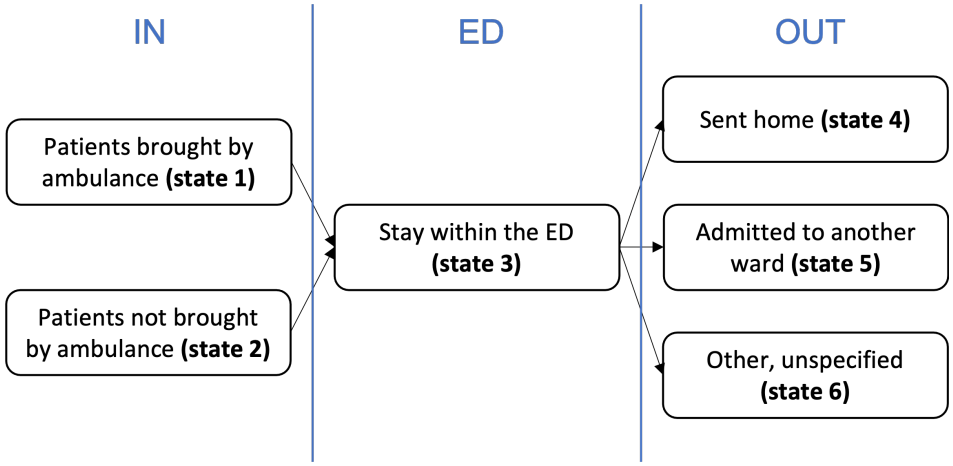



Figure 2.11: 6-states Markov Chain Model.

A complete estimation run, i.e., comprehensive of all the tasks described in the sub-sections from 2.3.2 to 2.3.6, was performed with both the setups, in other words, with three or four output states, on all the four training dataset sub-samples that were produced according to what described in sub-section 2.1.4. Since the corresponding results achieved with both the setups are consistent among the four training dataset sub-samples, and for a matter of shortness, such results are reported in sub-section 3.2.1 only with regards to one of the sub-samples mentioned above, i.e., for "Seed n°1". Due to the evident improvement given by the reduction in the number of states from four to three, that can be appreciated by comparing the outcomes of the use of both the setups (see sub-section 3.2.1), and due to the deriving reduction in computational load and thus in required time for executing a complete estimation run, it was decided to abandon the 7-states MC approach and continue this work only with the 6-states one.

An experimental protocol was thus defined to process the available data and select suitable and meaningful covariates to be included in the process model, with the intent of achieving the goal of describing complex patient characteristics in relation to the length of stay within the ED. The aforementioned protocol is explained by the scheme in figure 2.12, and the extraction of the dataset sub-samples that are mentioned in such a scheme is discussed more in detail in sub-section 2.1.4, together with the other elaborations of the dataset that were necessary before implementing the Markov Chains modeling.

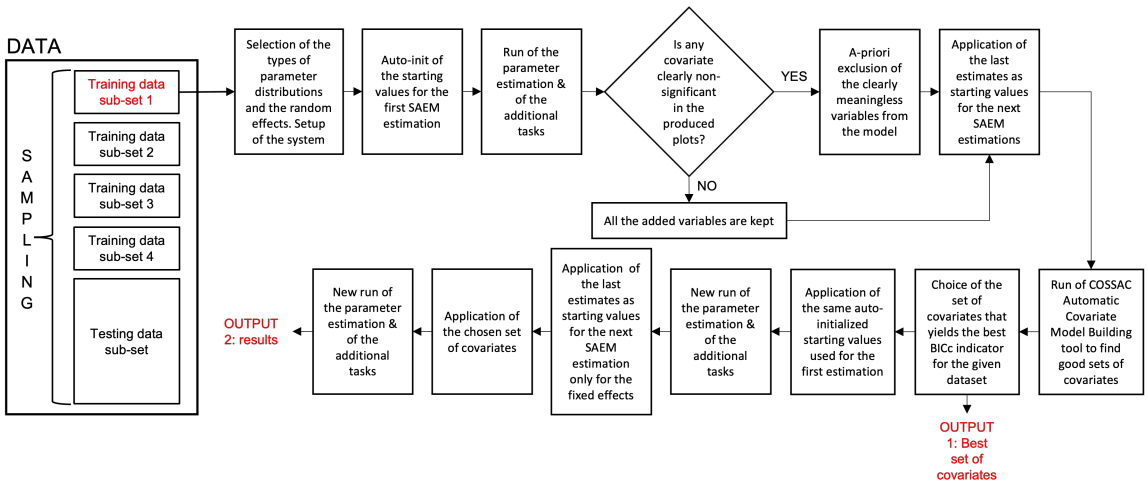



Figure 2.12: Methodological approach for selecting the best set of covariates for each sub-set of data.

Regarding the choice of which of the parameters to be estimated must be described by a distribution, it was decided to do so only with the parameters *q34* (transition from the ED to the home), *q35* (transition from the ED to a hospital ward), and *q36* (transition from the ED to another discharge category), for which the option for the estimation of the random effects was enabled. Concerning instead the choice that was made in terms of which type of distribution to use to describe the parameters of the model (see the available options in 2.3.1), this was the same for all the parameters, i.e., for all of them a log-normal distribution was selected. To clarify this choice, first of all, initial state probabilities and transition rates are all necessarily non-negative, which induced the rejection of the possibility of using a normal distribution. Furthermore, transition rates are supposed to be able to assume values greater than 1, which led to discarding from the available options also the logit-normal distributions and the probit-normal ones. Moreover, the use of a log-normal distribution to describe transition rates is reasonable also from a qualitative point of view since these transition rates are dependent on the length of stay of the patients, and the probability density function of the latter resulted being shaped as a log-normal distribution as well (confirmed by figure 2.1b). Finally, a brief explanation of the working principle behind the automatic covariate model building tool called "COSSAC", the employment of which is also part of the experimental protocol in figure 2.12, is instead provided in section 2.6.

Throughout the execution of the discussed experimental protocol, the software's general settings and the code of its random seed were kept at the default values. The only exception was made for what concerns the maximum number of iterations that the SAEM algorithm

can perform during the exploratory phase, which was raised from the default value of 500 to a custom value of 800 since this seemed to improve the convergence of the estimations for this thesis.

## 2.6. Automatic covariate model building

This section aims to describe the general functioning of the so-called "COSSAC" algorithm, which is the automatic covariate model building algorithm employed in this thesis. COSSAC stands for "COnditional Sampling use for Stepwise Approach based on Correlation tests" and is an innovative covariate search strategy that was validated and published by Lixoft [47].

Instead of blindly "trying" all the covariates as the starting point, as is done in more conventional approaches, this algorithm exploits the information contained in the base model run to choose which covariate to try first. To do so, it uses the correlation between the individual parameters (or random effects) and the covariates as "*hints at possibly relevant parameter-covariate relationships*" [48]. Such values of the correlation between random effects and covariates are calculated using samples from the a posteriori conditional distribution that is produced in the task "Conditional distribution" (see sub-section 2.3.3) [48]. For evaluating the continuous covariates, the Pearson's correlation test is used to derive a p-value, whereas for evaluating the categorical covariates ANOVA is employed [48]. These so-extracted p-values are then used to sort all the possible random effect-covariate relationships according to whether using them in the model or not [48]. After a first initialization, the iterations of COSSAC thus alternate between "forward" and "backward" selection according to the results of the performed correlation tests:

- **Initialization**: the base model is run, so to estimate population parameters, sample from the conditional distributions, and estimate the log-likelihood [48]. Then, the p-values of all the parameter-covariate relationships are calculated with Pearson's correlation tests if the covariate is continuous or with ANOVA if it is discrete [42, 48]. The first step after the initialization is a backward selection.
- **Forward selection**: the covariate with the smallest correlation p-value is added to the model, "*or the next smallest if the smallest has already been tried, and so on until no correlation p-values above threshold remain*" [48]. Then the model is run with the same initial values as the base model, and the relationship is accepted or rejected based on the value of the -2$\mathcal{LL}_y(\hat{\theta})$ or on the value of the BICc (see equation 2.36), i.e., the new model is not accepted if it does not improve the chosen criterion over a certain pre-defined threshold [48]. At this point, all the parameter-covariate

correlation p-values are calculated, and an attempt is made to perform a backward selection in the next step [48].

- **Backward selection**: the currently included covariate with the highest correlation p-value is removed from the model, "*or the next highest if the highest has already been tried, and so until no correlation p-values below a threshold remain*" [48]. Then the model is run with the same initial values as the base model and the "relationship removal" is accepted or rejected based on the value of the -2$\mathcal{LL}_y(\hat{\theta})$ or on the value of the BICc (see equation 2.36), i.e., the new model is not accepted if it does not improve the chosen criterion over a certain pre-defined threshold [48]. Consequently, all the parameter-covariate correlation p-values are calculated, and an attempt is made to perform a forward selection in the next step [48].

The alternation between the forward and the backward selection continues until no selection is possible anymore or until ten new relationships have been tested on the same model [48].

In this thesis, the criterion employed to choose from which of the iterations of COSSAC to extract a set of covariates to be tested on the data was mainly based on the value of the corrected Bayesian Information Criterion" (BICc). This choice was made due to the ability of BICc to penalize not only with the logarithm of the number of subjects $\log(N)$, but also with the size of $\theta_{\mathrm{R}}$, and with the size of $\theta_{\mathrm{F}}$, which depend on the number of included covariates and are scaled with the logarithm of the total number of observations ($n_{\mathrm{tot}}$), as discussed in section 2.3.6. However, it was decided to take into consideration also the value of the -2$\mathcal{LL}_y(\hat{\theta})$ in the cases in which picking the best BICc rather than the second-best would have improved its value really mildly and at the expense of a much worsened corresponding value of the log-likelihood. Accordingly, for these cases mentioned above, the set of covariates associated with the second-best BICc was selected rather than the one associated with the best BICc.

## 2.7. Validity assessment

The purpose of this section is to briefly describe how the process model designed in this thesis was validated, according to figure 2.13. The outcomes of the procedure described hereunder can be seen in section 3.3. These results are then discussed in sub-sections 4.1.1 and 4.1.2.

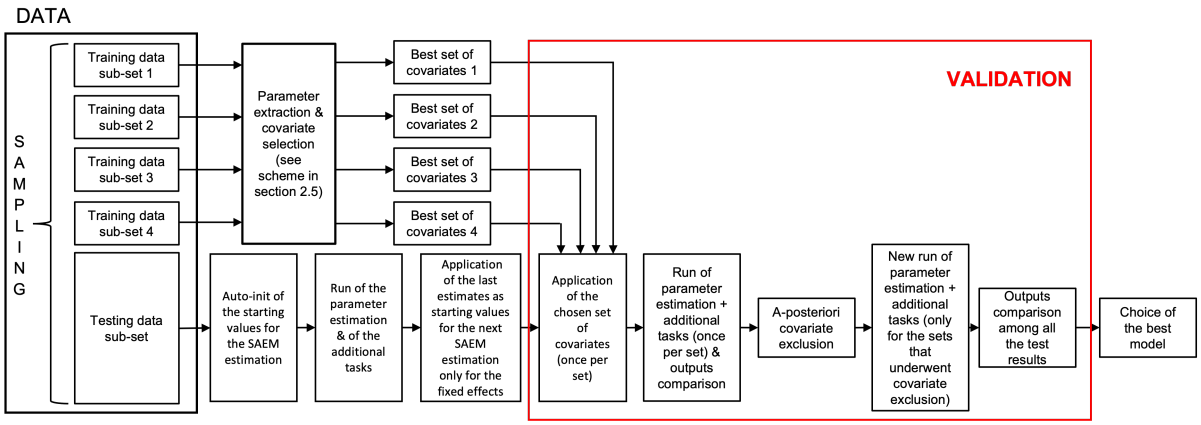



Figure 2.13: Validation protocol.

The extraction of the dataset sub-samples mentioned in the figure above is discussed more in detail in sub-section 2.1.4, together with the other elaborations of the dataset that were necessary before implementing the Markov Chains modeling. The experimental protocol explained by the scheme in figure 2.12 was executed to process the available data, assess the meaningfulness and usability of the base model, and select suitable and meaningful covariates to be included in such a model through the execution of the COSSAC algorithm (see section 2.6) on each of the "training" data sub-samples, as shown in figure 2.12.

Starting from the base Markov Chains model, the initial estimates associated with that model for the testing data sub-sample, and the four sets of covariates selected through the employment of the COSSAC algorithm, each of the sets was then individually applied to such a base model. For each of them, a run of parameter estimation was then executed on the testing data sub-set. As shown in figure 2.13, the four outputs were compared, and a process of a-posteriori covariate exclusion was performed. Afterward, the updated sets of covariates, in other words, the ones which had been subjected to covariate exclusion, were applied again to the base model and tested on the testing data sub-set. A final comparison among all these outputs was performed, the resulting covariates were analyzed under a clinical perspective (see sub-section 4.1.1), and conclusions were drawn accordingly. Lastly, the best-performing final model was determined.

# 3 | Results and Analysis

This chapter presents the outcome of the modeling study described in chapter 2.

## 3.1. Comparison between modeling techniques

This section is meant to answer one of the first goals of this thesis (see section , "Phase 1"), i.e., the exploration of the state of the art regarding modeling methods and techniques for process modeling in healthcare to allow the modeler to find suitable techniques for modeling the system in question. Therefore, the outcomes of a comparison between families of modeling techniques and of a comparison among their sub-categories are presented.

Firstly, table 3.1 shows a schematic comparison between "analytical approaches", "simulation modeling", and "statistical or empirical modeling", which are the three main different options for modeling in healthcare. These are examined in terms of the main advantages and disadvantages deriving from the employment of each of the three.

After that, in sections 3.1.1, 3.1.2, and 3.1.3, the sub-categories of each of the macro families of modeling techniques are described one by one, their main fields of application are briefly introduced, and their advantages and disadvantages are compared in a table format.

Table 3.1: Three different approaches to modeling in healthcare.

| | Advantages | Disadvantages |
|---|---|---|
| **Analytical approaches** | • Less data is required [4]<br>• They can handle causal and time-dependencies [18]<br>• Possibility to find an optimized solution under applied constraints, e.g., budget, resources, ... [49] | • Not great for dealing with nonlinearities [18]<br>• Even though they can handle time dependencies, they do not deal well with them [18]<br>• Non-intuitive inter-variable influence [18] |
| **Simulation modeling** | • Choice of the specific technique according to the required abstraction [4]<br>• Good for modeling time-dependencies [4]<br>• Great handling of complexity, variability, and uncertainty of dynamic systems [22]<br>• Suitable for computing almost any operational performance measure [4]<br>• Suitable for almost any type of analysis [4] | • Generally, a lot of data is required [4]<br>• Despite usually requiring a large amount of modeling time, finding an optimal solution is not even guaranteed [21]<br>• Execution time can be very long |
| **Statistical or empirical modeling** | • Still in a nascent stage – more applications might come [4]<br>• It can model both operational and clinical flows [4]<br>• Possible to describe patients' history as transitions between possible medical conditions or care locations [50]<br>• Possible to add patient frailties as random effects [4] | • Still in a nascent stage – usage is still limited [4]<br>• Cannot be used to determine waiting time measures [4]<br>• Pure statistical modeling is still rarely used. It is more common to find it employed to inform simulation approaches |

### 3.1.1. Analytical approaches

**Queuing theoretic models** compute performance metrics through analytical formulae belonging to queuing theory. They are mainly applied to problems regarding appointment scheduling, bed planning for inward patients, and emergency department resource allocation [4]. In some cases, queueing models have been used to solve problems regarding ED queuing times and to decrease the number of LWBS patients (Left Without Being Seen) [49].

**Markov chains and compartmental models** represent the states of patient flow as the so-called Markov chains. Markov chains are stochastic processes whose state-space is finite or countable and "*in which the conditional distribution of any future state, given the past states and the present state, is independent of the past states and depends only on the present state*" [4]. They are mainly applied to problems regarding modeling inpatient clinical flow and regarding capacity planning in an outpatient environment [4].

The main advantages and disadvantages deriving from employing analytical approaches in healthcare are presented in table 3.2.

Table 3.2: Comparison between queuing theoretic models and Markov chains and compartmental models.

| | Advantages | Disadvantages |
|---|---|---|
| **Queuing theoretic models** | • Waiting time-related measures, congestion, measures of idle time, and server usage are common metrics [4]<br>• Good for determining the relationship between influential parameters and system outcomes [4]<br>• Good for flows with moderate complexity [4] | • Not great for dealing with nonlinearities [18] and with highly complex flows [4]<br>• Not great for dealing with time and causal dependencies [18] |
| **Markov chains and compartmental models** | • Good for both clinical and operational patient flow [4]<br>• Data requirement is usually lower than in simulation modeling (e.g., DES) [19]<br>• Easy to validate [19]<br>• Suitable for computing almost any operational performance measure [4] | • Not possible to analyze performance metrics with regards to waiting time [4]<br>• Queuing theory models are in most cases preferred over Markov Chain models [4]<br>• Limited ability to capture patient's history [19] |

### 3.1.2. Simulation modeling

**Monte Carlo simulations (MC)** are sampling experiments aiming at estimating distributions of some output variables, which are, in turn, dependent on several probabilistic input variables [15]. They are mainly used to perform cohort studies, where the level of detail lies between the high one given by DES and ABS, and the low one provided by the SD approach [22]. The main areas of applications are problems regarding risk management (e.g., identification and analysis of potential dangers and adverse events), health policy, medical decisions, and forecasting of economic and clinical indicators, including evaluating the "*economic effectiveness of a project*" [22]. However, in many applications, Monte Carlo methods are commonly used to inform other families of simulation approaches, such as DES, instead of being used as a self-standing method for the whole simulation. An example of the exploitation of this concept can be found in the work performed by C. Zhang et al. (2021), in which they generated "*hypothetical cohorts of patients*" using Monte Carlo simulation and then used such patients as input for a discrete event simulation model [51].

**Discrete event simulations (DES)** model the system as a process, with sequences of operations that the agents perform. Such operations can include delays and queues if agents compete for limited resources, process branch selection, service by various resources, and so on [18]. Elements such as service times and agents' arrival times are usually stochastic variables drawn from more or less complex probability distributions [18]. The main goal of DES is to assess system effectiveness thanks to the estimation of quantitative parameters (e.g., patient throughput, timeliness of care, and resource utilization [7]) and the development of indicators [22]. Despite being the preferred method in most of the categories of healthcare management problems, the application of discrete event simulations has particularly dominant usage for experiments whose time horizon is within the short or medium range and in the category of "*healthcare system operations and improvements*" [22], e.g., for reducing waiting times, improving patient flow, maximizing staff and resources utilization [7].

**System dynamics simulations (SD)** are based on tracking instantaneous changes in a dynamic system by employing differential equations [5]. They represent, by their quantities, discrete items such as people, products, and events; therefore, it is necessary to identify the stocks and the flows affecting them [18]. Stocks keep track of the accumulated level of the items and flows keep track of the rate of change of such items [5, 22]. The main applications are those that are treatable at a higher and more aggregated level, particularly in the domains of health policy and forecasting. Requiring only one replication, SD simulations are also generally much applied to studies with a long-term horizon or

cases in which a more general perspective is preferable for decision processes at the macro scale [22].

**Agent-based simulations (ABS)** are based on autonomous entities called agents, individual people or groups of people who can interact both with other agents and with the surrounding environment, whose state varies over time, and who can make independent decisions based on pre-defined rules and the current situation [22]. ABS method is bottom-up, starting from defining the environment, the agents, and their relationship; the state of the model depends on the state of the environment and the collective states of the agents populating it [22]. ABS technique is good for enhancing the knowledge about the behavior of the system; it allows to register the history of each entity in detail, and it is usually employed for problems requiring mutual relations between specific entities to be mapped in the model and for cases in which most of the activities to be modeled cannot be described by fully predictable parameters [22]. It is mainly applied to studies where the level of required detail is higher than the one provided by the employment of SD [22]. Common targets of application are problems regarding health behaviors [52] and the prediction of the spread of infective diseases [22] ("non-communicable disease control" [52]) and epidemics [22] ("infectious disease epidemiology" and "social epidemiology" [52]).

The main advantages and disadvantages deriving from the employment of simulation modeling in healthcare are presented in tables 3.3 and 3.4.

### 3.1.3. Statistical or empirical modeling

According to the definition stated by Bhattacharjee and Ray [4], this approach is "*entirely based on observations of the system characteristics and experimentations on the system for analyzing the relationship between the performance-related factors and the influencing variables and parameters related to patient flows*". Despite being an interesting technique, since it is still at a nascent stage, there are still not many studies that capture patient flows by applying an entirely empirical modeling technique [4]. When used on its own, it is mainly applied to problems regarding extracting information related to care pathways [4]. A valid example is provided by S. Adeyemi and T. Chaussalet [53], who used a multinomial logit random-effects model to extract information on patient pathways. In most other cases, statistical modeling is commonly used to inform approaches from the category of simulation modeling.

Table 3.3: Comparison between DES, SD, ABS, and MC methods, part 1.

| | Advantages | Disadvantages |
|---|---|---|
| **Discrete Event Simulation** | • It follows individual, dynamic entities, described by attributes [5, 22]<br>• It supports both low and medium abstraction [18]. Flexible response to scale change [5]<br>• Easy to model queues<br>• Able to relate risks, activities, and interventions, with patients having individual traits [22]<br>• Good for systems with plenty of observable random factors [5, 22]<br>• Typical output includes time spent by agents in the system [18]<br>• Reusable components [5]<br>• Patient flow is represented in a visual way [5] | • Not very feasible for experiments with a long time horizon [22]<br>• In models in which real and simulated performance are compared, operational validity is particularly critical [7]<br>• A lot of data is required [22]<br>• If more than 2 or 3 specific objectives are defined (e.g., LOS, bed occupancy rate), the model might create an unfeasible set of tasks to complete the simulation [7]<br>• In many cases, it is required to have a high level of detail in such input data [22]<br>• Need to carefully assess the detailedness that the gathered data must have [7] |
| **System Dynamics approach** | • It represents cohorts, not individuals [5]<br>• Very helpful to formalize a mental model of a given problem [22]<br>• Good for systems having high nonlinearities, mutual interactions, circular causality concepts [5]<br>• Less data is required [22]<br>• Can analyze structure-behavior relations after initialization of change [22]<br>• Good at long-term prediction in macro-scale models [5]<br>• Both quantitative and qualitative aspects can be included | • It represents cohorts, not individuals [5]<br>• In complex systems, not possible to forecast output changes by visual inspection [22]<br>• Results are highly dependent on adequately calibrating the parameters for driving internal flows [22]<br>• Typically not designed for extracting exact numerical predictions [22] |

Table 3.4: Comparison between DES, SD, ABS, and MC methods, part 2.

| | Advantages | Disadvantages |
|---|---|---|
| **Agent-based Simulation** | • It follows individual, autonomous, dynamic entities, described by attributes [5, 22]<br>• Good when the overall system behavior is unknown, but the behavior of the agents is known [18]<br>• Wide range of supported abstraction levels [18]<br>• Good for exploring causal mechanisms and testing theories of causation (since able to incorporate multiple interacting causes) [52]<br>• Often possible to use simple rules to describe complex behaviors [5]<br>• "*Provides insight into the underlying mechanisms that give rise to health behaviors and outcomes*" [52]<br>• It can identify the minimum required "dose" of intervention for achieving the sought result [52]<br>• Good for modeling dynamic, autonomous, adaptive systems [5] | • Not very feasible for experiments with a long time horizon [22]<br>• Datasets often lack useful info regarding network influences and strength of interactions between units [52]<br>• A lot of data is required [22]<br>• When exploring causal mechanisms, several configurations might lead to the expected population pattern [52]<br>• Difficult to balance between simplicity and model realism [52]<br>• Computation, validation, and running are significantly costly in terms of time and resources [52]<br>• It can be challenging to validate and parameterize the model [52] |
| **Monte Carlo methods** | • Good for evaluating the impact of policy changes [22]<br>• Good for evaluating the risks in a decision process [22]<br>• Good flexibility [22]<br>• It can estimate the variability involved in the decision process [22]<br>• Easy to handle the modeling of population-based disease [54] | • The validity of the conclusions holds only for specific pre-defined individuals [22]<br>• It usually requires a large number of replications. The number of objects cannot be too high [22]<br>• A lot of data is required [22] |

## 3.2. CTMC modeling

This section presents the modeling of the patient flow within the emergency department. Sub-section 3.2.1 shows the comparison between a 7-states Markov Chain and a 6-states one, which led to the choice of the latter approach over the former, whereas the other sub-sections in this section show the results related to various steps of the experimental protocol shown in figure 2.12. As discussed in sub-section 2.5, this protocol was designed to process the available data and select meaningful covariates to be included in the model.

### 3.2.1. Comparison between 7-states and 6-states chain

This sub-section addresses the comparison between the Markov Chains model with four output states (figure 2.10) and the one with only three output states (figure 2.11). As stated in section 2.5, this comparison was performed for all the four random dataset samples, and it resulted consistent from sample to sample. Therefore, for a matter of shortness, this sub-section only shows the results for "random seed n°1".

Figures 3.1a and 3.1b show the population parameters estimated by using the SAEM algorithm (see sub-section 2.3.2), as well as the standard errors and the relative standard errors, calculated from the Fisher Information Matrix as discussed in sub-section 2.3.5, respectively for the CTMC model with four output states and the one with three output states. Excluding the value of the initial state probability $p\_pop$, i.e., the probability of having state 1 as the initial state, and the transition rates $q13$ and $q23$, which are just the description of a quasi-instantaneous transition from the two input states to state 3 (see section 2.5), it is possible to appreciate how the rates of the relative standard errors (R.S.E.) are for all the other parameters lower in the model with three output states than in the one with four output states. In particular, in figure 3.1a, the two rates of the R.S.E. colored in yellow resulted greater than 50%, i.e., the standard error on their corresponding estimated population parameter was equal to more than 50% of the population parameter itself. Contextually, as shown in table 3.5, a reduction in the number of states from four to three leads to a substantial decrease in the so-called "condition number". According to the rule of thumb described in sub-section 2.3.5, this shifts the model from being probably over-parameterized to being almost surely not over-parameterized.

Moreover, a reduction in the number of states from four to three leads also to an improvement in the likelihood, which is reported in table 3.5, according to both the considered indicators. Lastly, concerning the residuals, normality and symmetry around 0 are greatly achieved by both approaches.

| | VALUE | STOCH. APPROX. | |
|---|---|---|---|
| | | S.E. | R.S.E.(%) |
| **Fixed Effects** | | | |
| p_pop | 0.3 | 0.015 | 5.04 |
| q13_pop | 436.7 | 140.1 | 32.1 |
| q23_pop | 437.83 | 88.51 | 20.2 |
| q34_pop | 1.3 | 0.27 | 20.9 |
| q35_pop | 0.49 | 0.06 | 12.1 |
| q36_pop | 0.0029 | 0.00079 | 27.6 |
| q37_pop | 0.27 | 0.022 | 8.27 |
| **Standard Deviation of the Random Effects** | | | |
| omega_q34 | 0.22 | 0.13 | 58.9 |
| omega_q35 | 0.55 | 0.16 | 29.2 |
| omega_q36 | 1.33 | 0.24 | 18.4 |
| omega_q37 | 0.61 | 0.42 | 68.3 |

(a) CTMC model with four output states.

| | VALUE | STOCH. APPROX. | |
|---|---|---|---|
| | | S.E. | R.S.E.(%) |
| **Fixed Effects** | | | |
| p_pop | 0.28 | 0.015 | 5.23 |
| q13_pop | 439.56 | 103.64 | 23.6 |
| q23_pop | 439.61 | 103.53 | 23.5 |
| q34_pop | 1.33 | 0.058 | 4.38 |
| q35_pop | 0.52 | 0.034 | 6.66 |
| q36_pop | 0.16 | 0.022 | 13.6 |
| **Standard Deviation of the Random Effects** | | | |
| omega_q34 | 0.24 | 0.033 | 14.0 |
| omega_q35 | 0.41 | 0.077 | 18.9 |
| omega_q36 | 0.92 | 0.14 | 15.5 |

(b) CTMC model with three output states.

Figure 3.1: Estimated population parameters on the dataset sub-sample generated with "random seed n°1".

Table 3.5: Log-likelihood, corrected Bayesian Information Criterion, and condition number from a CTMC model with four or three output states, both on the dataset sub-sample generated with "random seed n°1".

| Output states | -2LL | BICc | Cond. number |
|---|---|---|---|
| 4 | 2962.52 | 3045.43 | 186.86 |
| 3 | 2815.59 | 2883.73 | 5.3 |

### 3.2.2. First estimation and covariates check

The results concerning the first complete estimation run are reported in this first part of sub-section 3.2.2. No covariates were included in it. Figure 3.2 shows the population

parameters estimated with the SAEM algorithm (see sub-section 2.3.2), along with the standard errors and the relative standard errors, which were calculated from the Fisher Information Matrix (see sub-section 2.3.5). Aside from the transition rates $q13$ for the $2^{nd}$ and $3^{rd}$ data sub-samples, which are anyways just the description of a quasi-instantaneous transition from input state 1 to state 3 (see section 2.5), the relative standard error for all the parameters is reasonable across all the data sub-samples.

| | | S.E. | R.S.E.(%) | | S.E. | R.S.E.(%) | | S.E. | R.S.E.(%) | | S.E. | R.S.E.(%) | | S.E. | R.S.E.(%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **FIXED Effects** | **Seed 1** | | | **Seed 2** | | | **Seed 3** | | | **Seed 4** | | | **TEST** | | |
| p_pop | 0.28 | 0.015 | 5.23 | 0.3 | 0.015 | 5.04 | 0.27 | 0.014 | 5.43 | 0.24 | 0.014 | 5.83 | 0.29 | 0.0064 | 2.22 |
| q13_pop | 439.56 | 103.64 | 23.6 | 440.38 | 320.4 | 72.8 | 450.63 | 405.57 | 90.0 | 442.08 | 174.86 | 39.6 | 437.29 | 76.44 | 17.5 |
| q23_pop | 439.61 | 103.53 | 23.5 | 441.77 | 88.94 | 20.1 | 451 | 111.46 | 24.7 | 442.77 | 92.35 | 20.9 | 437.07 | 48.24 | 11.0 |
| q34_pop | 1.33 | 0.058 | 4.38 | 1.23 | 0.056 | 4.60 | 1.09 | 0.047 | 4.29 | 1.24 | 0.03 | 2.46 | 1.34 | 0.026 | 1.97 |
| q35_pop | 0.52 | 0.034 | 6.66 | 0.52 | 0.033 | 6.21 | 0.49 | 0.03 | 6.06 | 0.56 | 0.034 | 6.01 | 0.58 | 0.014 | 2.32 |
| q36_pop | 0.16 | 0.022 | 13.6 | 0.19 | 0.031 | 16.2 | 0.19 | 0.025 | 13.6 | 0.22 | 0.017 | 7.46 | 0.23 | 0.0086 | 3.78 |
| **RAND. Effects** | | | | | | | | | | | | | | | |
| omega_q34 | 0.24 | 0.033 | 14.0 | 0.22 | 0.033 | 15.0 | 0.26 | 0.1 | 38.8 | 0.13 | 0.026 | 20.4 | 0.24 | 0.019 | 8.03 |
| omega_q35 | 0.41 | 0.077 | 18.9 | 0.37 | 0.054 | 14.7 | 0.39 | 0.16 | 39.9 | 0.4 | 0.066 | 16.6 | 0.27 | 0.018 | 6.64 |
| omega_q36 | 0.92 | 0.14 | 15.5 | 0.99 | 0.19 | 19.1 | 0.7 | 0.23 | 32.1 | 0.4 | 0.1 | 25.8 | 0.57 | 0.039 | 6.90 |

Figure 3.2: Estimated population parameters for the four training data samples and the testing sample with no covariates.

Table 3.6: Statistical tests on random effects and individual parameters, likelihood indicators, and condition numbers with no covariates in the model.

| | **Seed 1** | **Seed 2** | **Seed 3** | **Seed 4** | **Test** |
|---|---|---|---|---|---|
| S.W. R.E. P-val>0.05? | Yes | Yes | Yes | Yes | Yes |
| T-test R.E. P-val>0.05? | Yes | Yes | Yes | Yes | Yes |
| S.W. I.P. P-val>0.05? | Yes | Yes | Yes | Yes | Yes |
| -2LL | 2815.59 | 2940.84 | 2911.28 | 2814.77 | 15 720.54 |
| BICc | 2883.73 | 3008.98 | 2979.42 | 2882.92 | 15 803.85 |
| Cond. number | 5.30 | 9.10 | 8.65 | 3.38 | 2.55 |

Table 3.6 regroups and compares the results of a Shapiro-Wilk normality test ("S.W. R.E.") and a T-test ("T-test R.E.") on the random effects, as well as the results of a Shapiro-Wilk normality test on the transformed individual parameters ("S.W. I.P."), two likelihood

indicators, and the condition number, computed on all the four "training" data sub-sets (seeds from 1 to 4) and on the "testing" sub-set for the basic model without covariates. Both the Shapiro-Wilk normality tests show a p-value ($p$) greater than 0.05 for all the random effects and all the transformed individual parameters on all the data sub-sets, thus confirming their normality. Contextually, the T-test for the correlation between random effects also shows a p-value greater than 0.05 on all the data sub-sets, thus implying no correlation between random effects in any of them. Furthermore, the values of the condition number indicate good confidence in the model not being over-parameterized for all the data sub-samples since they are largely smaller than a value of 100 (see sub-section 2.3.5).

Lastly, figures 3.3, 3.4, and 3.5, show how the normalized prediction distribution errors (NPDE), a nonparametric version of the population-weighted residuals, are normally distributed and symmetric around 0 for all the data sub-samples. In each of the five blocks in the figure, the comparison between the empirical and theoretical probability density function is plotted on the left, whereas the comparison between the empirical and theoretical cumulative distribution function is plotted on the right.

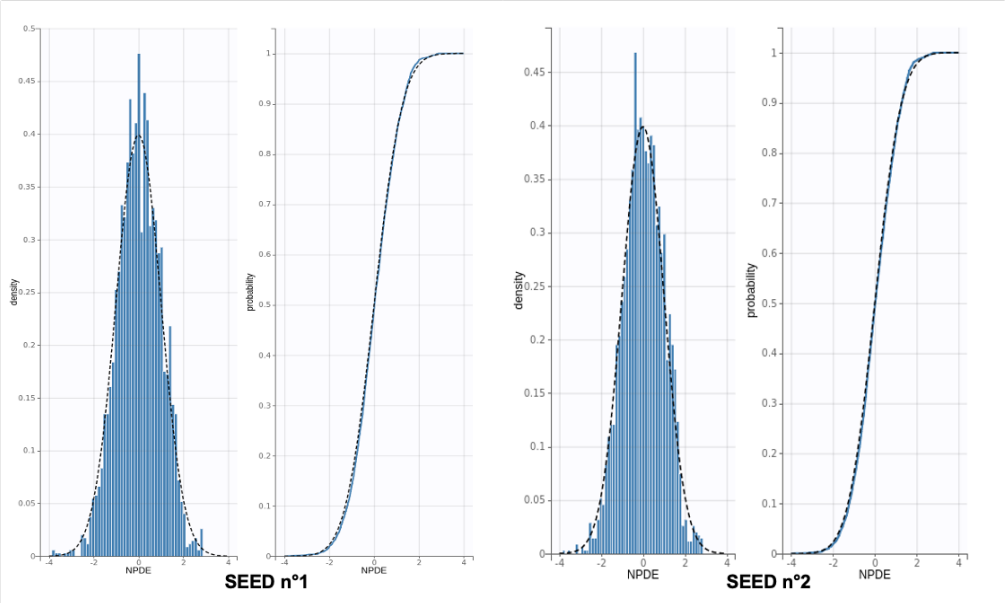



Figure 3.3: Probability Distribution Function and Cumulative Distribution Function of the NPDE in the model without covariates for data sub-samples 1 and 2.

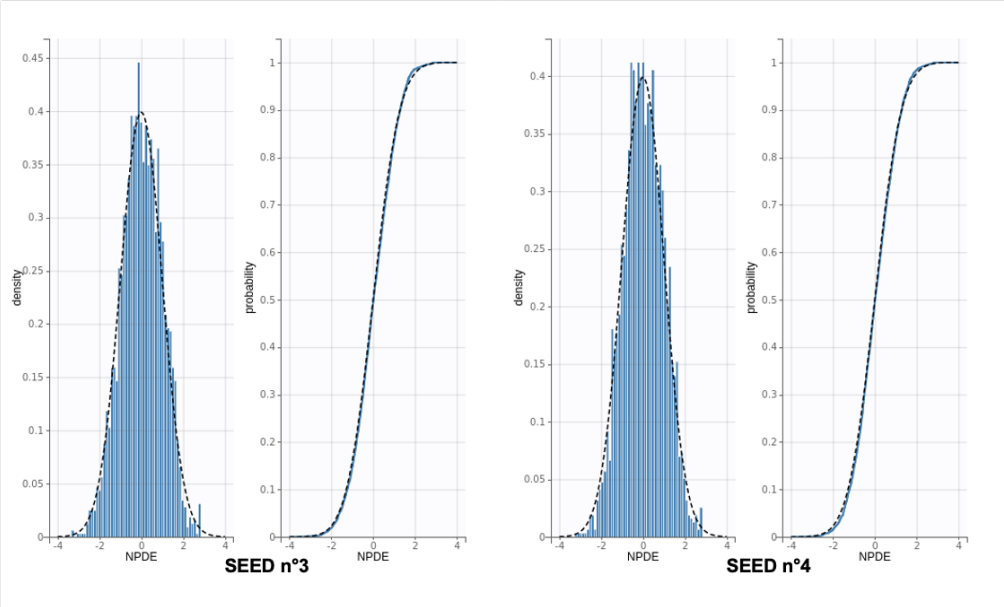



Figure 3.4: Probability Distribution Function and Cumulative Distribution Function of the NPDE in the model without covariates for data sub-samples 3 and 4.

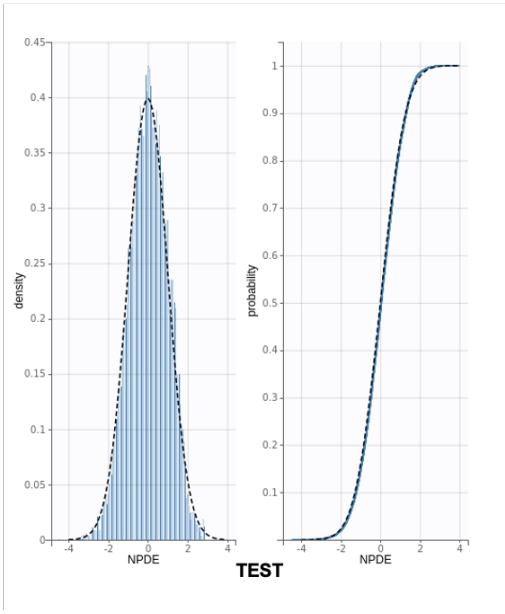



Figure 3.5: Probability Distribution Function and Cumulative Distribution Function of the NPDE in the model without covariates for the testing data sub-sample.

## Covariate check

This last part of sub-section 3.2.2 is dedicated to showing which covariates, visible in figures 3.6, 3.7, and 3.8, were a priori excluded before running the tool for automatic

covariate model building (see section 2.6).

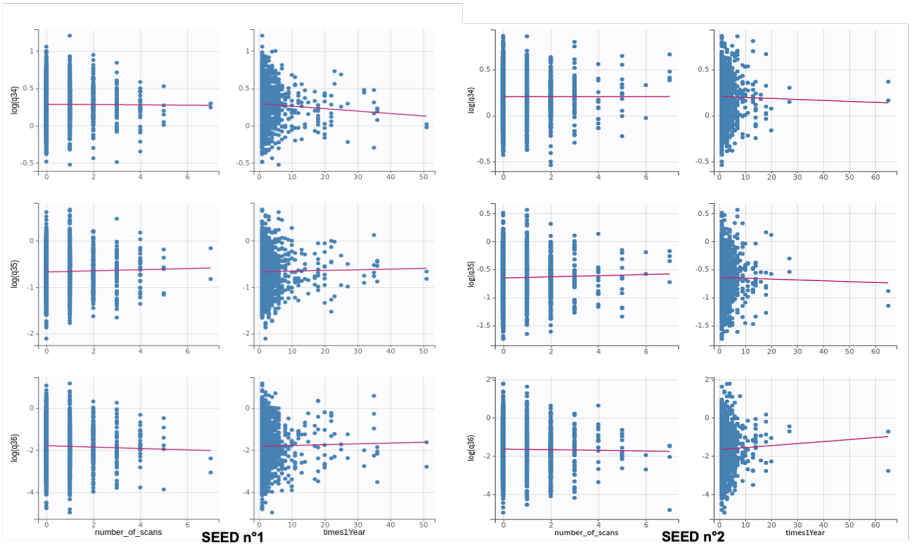



Figure 3.6: I.P. plotted against the covariates "number_of_scans" and "times1Year" for the training data sub-samples 1 and 2.

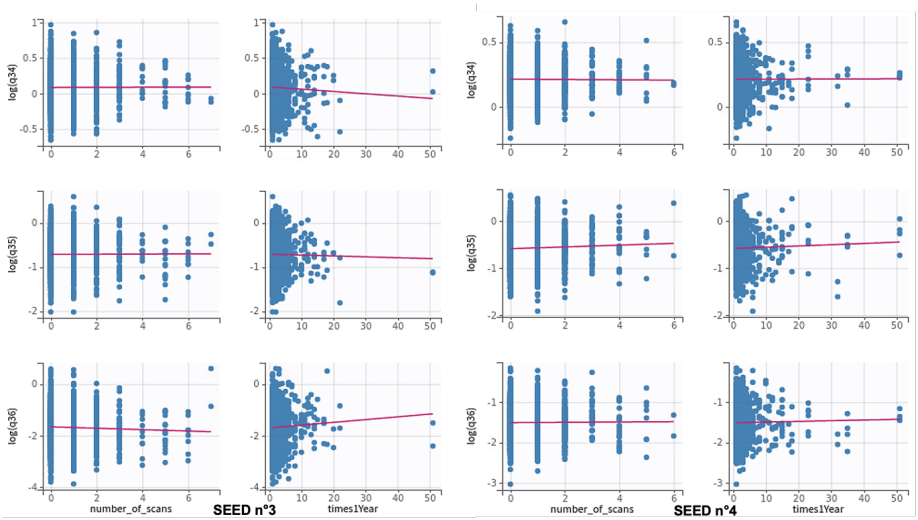



Figure 3.7: I.P. plotted against the covariates "number_of_scans" and "times1Year" for the training data sub-samples 3 and 4.

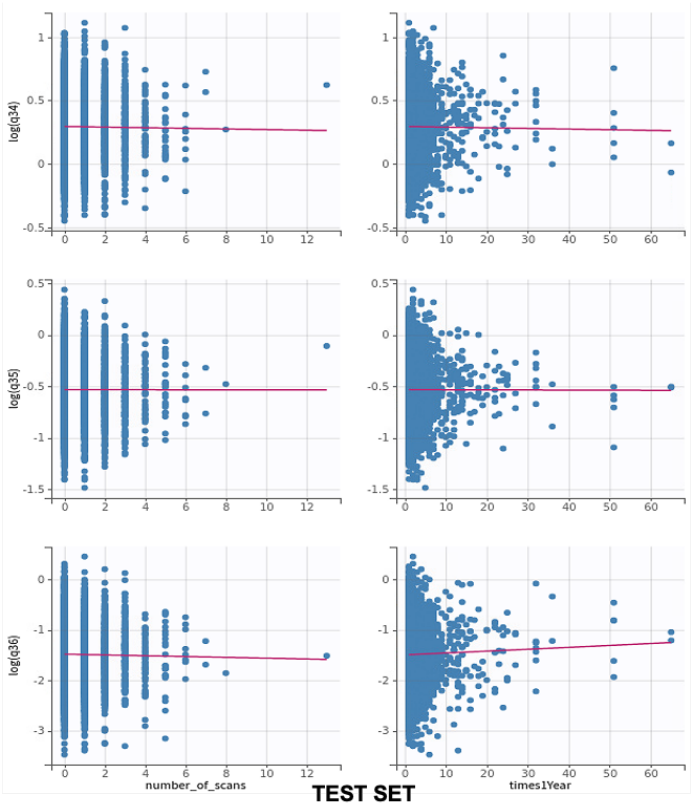



Figure 3.8: I.P. plotted against the covariates "number_of_scans" and "times1Year" for the testing data sub-sample.

The figures show the plot of the individual parameters (I.P.) against the two excluded covariates, "*number_of_scans*" and "*times1Year*", for all the training data sub-samples and the test sub-sample. Clearly, for none of the parameters in any of the sub-samples, there exists any dependence of the individual parameter on such covariates.

### 3.2.3. 6-states CTMC Covariate model building

This sub-section aims to show the results produced by applying the COSSAC algorithm to the four "training" data sub-sets and then by applying the selected sets of covariates to the corresponding data sub-sets from which these were extracted.

#### Seed n°1

Tables 3.7 and 3.8 present the outcome of the COSSAC algorithm in terms of the selected covariates and the likelihood of the Markov Chains model including such covariates, for the data sub-set sampled in Python with random seed n°1.

Table 3.7: COSSAC Results for random seed n°1 - Part 1.

| It. | Introduced Covariates | -2LL | BICc |
|---|---|---|---|
| $1^{st}$ | None | 2815.12 | 2883.26 |
| $2^{nd}$ | q35_K | 2806.43 | 2881.41 |
| $3^{rd}$ | q34_sex, q35_K | 2806.09 | 2887.91 |
| $4^{th}$ | q35_AmbYN, q35_K | 2729.77 | 2811.58 |
| $5^{th}$ | q34_sex, q35_AmbYN, q35_K | 2730.24 | 2818.89 |
| $6^{th}$ | q35_AmbYN, q35_K, q36_ScanYN | 2728.71 | 2817.36 |
| $7^{th}$ | q35_AmbYN, q35_K, q36_MA | 2710.91 | 2833.76 |
| $8^{th}$ | q34_T, q35_AmbYN, q35_K, q36_MA | 2706.00 | 2835.69 |
| $9^{th}$ | q34_sex, q35_AmbYN, q35_K, q36_MA | 2712.67 | 2842.35 |
| $10^{th}$ | q35_AmbYN, q35_K, q36_B, q36_MA | 2711.08 | 2840.77 |
| $11^{th}$ | q35_AmbYN, q35_K, q36_MA, q36_ScanYN | 2714.77 | 2844.46 |
| $12^{th}$ | q34_R, q35_AmbYN, q35_K, q36_MA | 2711.46 | 2841.14 |
| $13^{th}$ | q34_countAVG, q35_AmbYN, q35_K, q36_MA | 2712.37 | 2842.05 |
| $14^{th}$ | q35_AmbYN, q35_K, 36_MA, q36_Z | 2698.93 | 2828.62 |
| $15^{th}$ | q34_T, q35_AmbYN, q35_K, q36_MA, q36_Z | 2690.04 | 2826.56 |
| $16^{th}$ | q34_T, q34_sex, q35_AmbYN, q35_K, q36_MA, q36_Z | 2693.09 | 2836.45 |
| $17^{th}$ | q34_J, q34_T, q35_AmbYN, q35_K, q36_MA, q36_Z | 2694.46 | 2837.82 |
| **$18^{th}$** | **q34_T, q35_AmbYN, q35_K, q35_ScanYN, q36_MA, q36_Z** | **2669.89** | **2813.25** |
| $19^{th}$ | q34_T, q35_AmbYN, q35_K, q35_ScanYN, q36_B, q36_MA, q36_Z | 2672.81 | 2823.02 |

Table 3.8: COSSAC Results for random seed n°1 - Part 2.

| It. | Introduced Covariates | -2LL | BICc |
|---|---|---|---|
| $20^{th}$ | q34_T, q34_sex, q35_AmbYN, q35_K, q35_ScanYN, q36_MA, q36_Z | 2670.01 | 2820.20 |
| $21^{st}$ | q34_T, q35_AmbYN, q35_K, q35_ScanYN, q36_MA, q36_ScanYN, q36_Z | 2672.08 | 2822.28 |
| $22^{nd}$ | q34_T, q35_AmbYN, q35_K, q35_ScanYN, q36_G, q36_MA, q36_Z | 2672.07 | 2822.27 |
| $23^{rd}$ | q34_T, q35_AmbYN, q35_D, q35_K, q35_ScanYN, q36_MA, q36_Z | 2670.37 | 2820.57 |
| $24^{th}$ | q34_R, q34_T, q35_AmbYN, q35_K, q35_ScanYN, q36_MA, q36_Z | 2666.36 | 2816.56 |
| $25^{th}$ | q34_T, q35_AmbYN, q35_K, q35_ScanYN, q36_I, q36_MA, q36_Z | 2674.59 | 2824.79 |
| $26^{th}$ | q34_T, q34_countAVG, q35_AmbYN, q35_K, q35_ScanYN, q36_MA, q36_Z | 2669.66 | 2819.86 |
| $27^{th}$ | q34_J, q34_T, q35_AmbYN, q35_K, q35_ScanYN, q36_MA, q36_Z | 2672.88 | 2823.08 |

According to the tables 3.7 and 3.8, the best BICc was scored at the $4^{th}$ iteration. However, since picking the best BICc rather than the second-best would have improved its value really mildly at the expense of a much worsened corresponding value of the $-2\mathcal{LL}$, the covariates producing the second-best BICc were selected, in other words, the ones from the $18^{th}$ iteration: "T" for $q34$; "AmbulYN", "K", and "ScanYN" for $q35$; "MA_unit" and "Z" for $q36$.

A second estimation run on the same sub-set, this time including the set of covariates mentioned above, resulted in good R.S.E. on the estimated population parameters, normally distributed NPDEs with symmetry around 0, and better $-2\mathcal{LL}$ and BICc than the ones in the run without covariates with the same dataset ("Seed 1" in table 3.6). However, after including such covariates, it became impossible to compute the standard error and R.S.E. for the fixed effect of the *MA_unit* called "Infektion akut" on $q36$, as well as the standard deviation of the random effect on $q36$, and the condition number. The correlation test

between individual parameters and covariates revealed a p-value $> 0.01$ for the correlation between $q36$ and being assigned to the $MA\_unit$ "Kirurgi akut" (P-val $= 0.0115$), and for the correlation between $q36$ and being assigned to the $MA\_unit$ "Ortopedi akut" (P-val $= 0.0292$), suggesting the possibility to remove these covariates from the model.

## Seed n°2

Table 3.9 presents the outcome of the COSSAC algorithm in terms of selected covariates, and the likelihood of the Markov Chains model including such covariates, for the data sub-set sampled in Python with random seed n°2.

Table 3.9: COSSAC Results for random seed n°2.

| It. | Introduced Covariates | -2LL | BICc |
|---|---|---|---|
| $1^{st}$ | None | 2940.38 | 3008.51 |
| $2^{nd}$ | q36_Z | 2904.47 | 2979.44 |
| $3^{rd}$ | q35_K, q36_Z | 2893.17 | 2974.98 |
| $4^{th}$ | q35_K, q35_age, q36_Z | 2760.95 | 2849.60 |
| $5^{th}$ | q34_countOUT q35_K, q35_age, q36_Z | 2761.32 | 2856.81 |
| $6^{th}$ | q34_countAVG, q35_K, q35_age, q36_Z | 2761.80 | 2857.29 |
| $7^{th}$ | q34_J, q35_K, q35_age, q36_Z | 2759.37 | 2854.86 |
| **$8^{th}$** | **q34_R, q35_K, q35_age, q36_Z** | **2752.86** | **2848.35** |
| $9^{th}$ | q34_R, q35_K, q35_age, q36_B, q36_Z | 2753.51 | 2855.84 |
| $10^{th}$ | q34_R, q35_K, q35_N, q35_age, q36_Z | 2754.79 | 2857.12 |
| $11^{th}$ | q34_R, q34_countAVG, q35_K, q35_age, q36_Z | 2758.40 | 2860.73 |
| $12^{th}$ | q34_R, q35_K, q35_age, q36_Z, q36_sex | 2751.08 | 2853.41 |
| $13^{th}$ | q34_R, q34_countOUT, q35_K, q35_age, q36_Z | 2754.12 | 2856.45 |
| $14^{th}$ | q34_J, q34_R, q35_K, q35_age, q36_Z | 2750.76 | 2853.09 |
| $15^{th}$ | q34_R, q35_K, q35_age, q36_K, q36_Z | 2754.17 | 2856.50 |
| $16^{th}$ | q34_R, q35_D, q35_K, q35_age, q36_Z | 2754.58 | 2856.91 |

According to table 3.9, the best BICc was scored at the $8^{th}$ iteration, so the selected

covariates are: "R" for $q34$; "K" and "age" for $q35$; "Z" for $q36$. A second estimation run on the same sub-set, this time including such covariates, resulted in 65.19 as condition number, good R.S.E. on the population parameters, NPDEs normally distributed around 0, and better $-2\mathcal{LL}$ and BICc than in the run without covariates with the same dataset ("Seed 2" in table 3.6).

## Seed n°3

Tables 3.10 and 3.11 present the outcome of the COSSAC algorithm in terms of selected covariates and the likelihood of the Markov Chains model including such covariates, for the data sub-set sampled in Python with random seed n°3.

Table 3.10: COSSAC Results for random seed n°3 - Part 1.

| It. | Introduced Covariates | -2LL | BICc |
|---|---|---|---|
| $1^{st}$ | None | 2910.61 | 2978.76 |
| $2^{nd}$ | q34_age | 2818.32 | 2893.30 |
| $3^{rd}$ | q34_age, q36_T | 2816.61 | 2898.44 |
| $4^{th}$ | q34_age, q35_I | 2814.11 | 2895.93 |
| $5^{th}$ | q34_age, q35_MA | 2799.23 | 2922.10 |
| $6^{th}$ | q34_age, q35_MA, q36_T | 2799.43 | 2929.13 |
| $7^{th}$ | q34_age, q35_A, q35_MA | 2796.65 | 2926.35 |
| $8^{th}$ | q34_M, q34_age, q35_MA | 2791.38 | 2921.08 |
| $9^{th}$ | q34_M, q34_age, q35_MA, q36_T | 2788.95 | 2925.50 |
| $10^{th}$ | q34_E, q34_M, q34_age, q35_MA | 2789.68 | 2926.22 |
| $11^{th}$ | q34_M, q34_age, q35_A, q35_MA | 2787.56 | 2924.10 |
| $12^{th}$ | q34_M, q34_age, q35_MA, q36_ScanYN | 2771.34 | 2907.88 |
| $13^{th}$ | q34_M, q34_age, q35_A, q35_MA, q36_ScanYN | 2768.30 | 2911.68 |
| $14^{th}$ | q34_M, q34_age, q35_MA, q36_ScanYN, q36_T | 2773.32 | 2916.70 |
| $15^{th}$ | q34_M, q34_age, q35_MA, q36_K, q36_ScanYN | 2765.45 | 2908.84 |
| $16^{th}$ | q34_M, q34_T, q34_age, q35_MA, q36_ScanYN | 2773.40 | 2916.78 |
| $17^{th}$ | q34_E, q34_M, q34_age, q35_MA, q36_ScanYN | 2772.43 | 2915.81 |

Table 3.11: COSSAC Results for random seed n°3 - Part 2.

| It. | Introduced Covariates | -2LL | BICc |
|---|---|---|---|
| $18^{th}$ | q34_M, q34_age, q35_MA, q36_A, q36_ScanYN | 2755.92 | 2899.31 |
| $19^{th}$ | q34_M, q34_age, q35_MA, q36_A, q36_K, q36_ScanYN | 2749.33 | 2899.55 |
| $20^{th}$ | q34_M, q34_T, q34_age, q35_MA, q36_A, q36_ScanYN | 2758.56 | 2908.78 |
| $21^{st}$ | q34_M, q34_age, q35_A, q35_MA, q36_A, q36_ScanYN | 2756.50 | 2906.72 |
| $22^{nd}$ | q34_E, q34_M, q34_age, q35_MA, q36_A, q36_ScanYN | 2754.89 | 2905.11 |
| $23^{rd}$ | q34_M, q34_age, q35_MA, q35_Z, q36_A, q36_ScanYN | 2751.26 | 2901.49 |
| $24^{th}$ | q34_M, q34_age, q35_MA, q36_A, q36_ScanYN, q36_UppsalaYN | 2756.48 | 2906.70 |
| $25^{th}$ | q34_M, q34_age, q35_MA, q36_A, q36_ScanYN, q36_T | 2757.66 | 2907.88 |
| $26^{th}$ | q34_M, q34_age, q35_I, q35_MA, q36_A, q36_ScanYN | 2754.82 | 2905.04 |
| $27^{th}$ | q34_M, q34_age, q35_MA, q35_age, q36_A, q36_ScanYN | 2735.07 | 2885.29 |
| **$28^{th}$** | **q34_M, q34_age, q35_MA, q35_age, q36_A, q36_K, q36_ScanYN** | **2727.48** | **2884.54** |
| $29^{th}$ | q34_M, q34_age, q35_A, q35_MA, q35_age, q36_A, q36_K, q36_ScanYN | 2730.75 | 2894.65 |
| $30^{th}$ | q34_M, q34_T, q34_age, q35_MA, q35_age, q36_A, q36_K, q36_ScanYN | 2729.10 | 2893.00 |
| $31^{st}$ | q34_M, q34_age, q35_MA, q35_Z, q35_age, q36_A, q36_K, q36_ScanYN | 2726.53 | 2890.43 |
| $32^{nd}$ | q34_E, q34_M, q34_age, q35_MA, q35_age, q36_A, q36_K, q36_ScanYN | 2728.64 | 2892.54 |
| $33^{rd}$ | q34_M, q34_age, q35_MA, q35_age, q36_A, q36_G, q36_K, q36_ScanYN | 2722.86 | 2886.76 |
| $34^{th}$ | q34_M, q34_age, q35_MA, q35_age, q36_A, q36_K, q36_ScanYN, q36_UppsalaYN | 2726.64 | 2890.54 |

According to the tables 3.10 and 3.11, the best BICc was scored at the $28^{th}$ iteration, so the selected covariates are: "M" and "age" for $q34$; "MA_unit" and "age" for $q35$; "A", "K", and "ScanYN" for $q36$. A second estimation run on the same sub-set, this time including the aforementioned set of covariates, resulted in good R.S.E. on all the estimated population parameters but the fixed effect of $q13$ (clinically meaningless transition rate), in normally distributed NPDEs with symmetry around 0, and in better $-2\mathcal{LL}$ and BICc than in the run without covariates with the same dataset ("Seed 3" in table 3.6). However, after including such covariates, it became impossible to compute the condition number, and the standard error and R.S.E. for the fixed effect of $q23$, for the fixed effect of the *MA_unit* "Hjärtsjukdomar" on $q35$, and for the standard deviation of the random effects on $q36$. Moreover, a Wald test (see sub-section 2.3.5) shows a p-value = 0.0434, i.e., greater than 0.01, for the effect on $q34$ of having "M" as *simple_diag*.

## Seed n°4

Tables 3.12 and 3.13 present the outcome of the COSSAC algorithm in terms of selected covariates, and the likelihood of the Markov Chains model including such covariates, for the data sub-set sampled in Python with random seed n°4.

Table 3.12: COSSAC Results for random seed n°4 - Part 1.

| It. | Introduced Covariates | -2LL | BICc |
|---|---|---|---|
| $1^{st}$ | None | 2815.01 | 2883.16 |
| $2^{nd}$ | q35_A | 2809.66 | 2884.65 |
| $3^{rd}$ | q35_I | 2806.20 | 2881.19 |
| $4^{th}$ | q35_A, q35_I | 2800.28 | 2882.11 |
| $5^{th}$ | q35_I, q35_K | 2787.60 | 2869.42 |
| $6^{th}$ | q35_A, q35_I, q35_K | 2780.37 | 2869.04 |
| $7^{th}$ | q34_M, q35_A, q35_I, q35_K | 2773.31 | 2868.82 |
| $8^{th}$ | q34_M, q35_A, q35_I, q35_K, q36_G | 2773.06 | 2875.41 |
| $9^{th}$ | q34_M, q34_ScanYN, q35_A, q35_I, q35_K | 2736.07 | 2838.42 |
| $10^{th}$ | q34_M, q34_ScanYN, q35_A, q35_I, q35_K, q36_G | 2742.29 | 2851.48 |

Table 3.13: COSSAC Results for random seed n°4 - Part 2.

| It. | Introduced Covariates | -2LL | BICc |
|---|---|---|---|
| $11^{th}$ | q34_E, q34_M, q34_ScanYN, q35_A, q35_I, q35_K | 2738.29 | 2847.47 |
| $12^{th}$ | q34_M, q34_ScanYN, q34_countOUT, q35_A, q35_I, q35_K | 2739.04 | 2848.22 |
| $13^{th}$ | q34_K, q34_M, q34_ScanYN, q35_A, q35_I, q35_K | 2740.36 | 2849.54 |
| $14^{th}$ | q34_M, q34_ScanYN, q35_A, q35_I, q35_K, q36_E | 2730.79 | 2839.97 |
| $15^{th}$ | q34_M, q34_ScanYN, q35_A, q35_I, q35_K, q36_Z | 2712.52 | 2821.71 |
| $16^{th}$ | q34_M, q34_ScanYN, q35_A, q35_I, q35_K, q36_G, q36_Z | 2707.09 | 2823.11 |
| $17^{th}$ | q34_K, q34_M, q34_ScanYN, q35_A, q35_I, q35_K, q36_Z | 2704.29 | 2820.31 |
| $18^{th}$ | q34_E, q34_K, q34_M, q34_ScanYN, q35_A, q35_I, q35_K, q36_Z | 2695.51 | 2818.37 |
| $19^{th}$ | q34_E, q34_K, q34_M, q34_ScanYN, q34_age, q35_A, q35_I, q35_K, q36_Z | 2657.54 | 2787.25 |
| **$20^{th}$** | **q34_E, q34_K, q34_M, q34_ScanYN, q34_age, q35_A, q35_I, q35_K, q35_Z, q36_Z** | **2645.86** | **2782.40** |
| $21^{st}$ | q34_E, q34_K, q34_M, q34_ScanYN, q34_age, q35_A, q35_H, q35_I, q35_K, q35_Z, q36_Z | 2646.42 | 2789.80 |
| $22^{nd}$ | q34_E, q34_K, q34_M, q34_ScanYN, q34_age, q34_countOUT, q35_A, q35_I, q35_K, q35_Z, q36_Z | 2655.44 | 2798.82 |
| $23^{rd}$ | q34_E, q34_K, q34_M, q34_ScanYN, q34_age, q35_A, q35_I, q35_K, q35_Z, q36_G, q36_Z | 2646.67 | 2790.05 |
| $24^{th}$ | q34_E, q34_K, q34_M, q34_ScanYN, q34_age, q35_A, q35_I, q35_K, q35_Z, q36_T, q36_Z | 2652.96 | 2796.34 |

According to the tables 3.12 and 3.13, the best BICc was scored at the $20^{th}$ iteration, so the selected covariates are: "E", "K", "M", "ScanYN", and "age" for $q34$; "A", "I", "K",

and "Z" for $q35$; "Z" for $q36$. As in the previous case, a second estimation run on the same sub-set, this time including the aforementioned set of covariates, resulted in good R.S.E. on all the estimated population parameters but the fixed effect of $q13$ (clinically meaningless transition rate), in normally distributed NPDEs with symmetry around 0, and in better $-2\mathcal{LL}$ and BICc than in the run without covariates with the same dataset ("Seed 4" in table 3.6). However, after including such covariates, it became impossible to compute the condition number and the standard error and R.S.E. for the fixed effect of having "A" as *simple_ diag* on $q35$. Nevertheless, the Wald test (see sub-section 2.3.5) showed no p-value greater than 0.01, thus not suggesting removing any covariates from the model.

## 3.3. Validity Analysis

This section presents the outcomes of the validation process (section 2.7). Table 3.14 compares, for all the runs on the "testing" dataset, the results of a Shapiro-Wilk normality test and a T-test on the random effects, as well as two likelihood indicators and the condition number. Moreover, it includes the results of a Shapiro-Wilk test on the transformed individual parameters for the model in which these do not depend on covariates and a Kolmogorov Smirnov adequacy test ("K.S. I.P.") for the models in which parameters do depend on covariates.

Table 3.14: Statistical tests on random effects and individual parameters, likelihood indicators, and condition numbers on the testing data sub-set.

| | No cov. | Set 1 | Set 2 | Set 3 | Set 4 |
|---|---|---|---|---|---|
| S.W. R.E. $p$>0.05? | Yes | Yes | Yes | Yes | Yes |
| T-test R.E. $p$>0.05? | Yes | Yes | Yes | Yes | Yes |
| S.W. I.P. $p$>0.05?<br>K.S. I.P. $p$>0.10? | Yes<br>/ | /<br>Yes | /<br>Yes | /<br>Yes | /<br>Yes |
| -2LL | 15 720.54 | 14 963.44 | 14 993.37 | 15 025.46 | 14 850.44 |
| BICc | 15 803.85 | 15 149.03 | 15 110.77 | 15 219.57 | 15 018.98 |
| Cond. number | 2.55 | 4108.50 | 40.74 | 2290.03 | 98.01 |

Only for table 3.14, the notation "$p$" is proposed as a more compact option for representing the p-value. In the same table, the Shapiro-Wilk normality tests show a p-value ($p$) greater

than 0.05 for all the random effects, thus confirming their normality. The same is valid for all the transformed individual parameters in the test with no included covariates. Contextually, the p-value of the Kolmogorov Smirnov adequacy test is $\sim 1$ for all the individual parameters in all the tests that include covariates, thus confirming that the individual parameters are samples from a mixture of transformed normal distributions.

Concerning the T-test for the correlation between random effects, the p-value results greater than 0.05 for the model without covariates but also for all the models with an introduced set of covariates, thus implying no correlation between random effects in any of the cases. Lastly, the values of the condition number indicate good confidence in the model not being over-parameterized only for the model with no covariates and for the one with the second set of covariates since such condition numbers are considerably smaller than 100 (see sub-section 2.3.5). For the model with the fourth set of covariates, the condition number is smaller than 100 only with little margin. In contrast, for the first and third sets of covariates, the condition number takes values that clearly indicate a high risk of overfitting.

Figures 3.9, 3.10, 3.11, and 3.12, show the distribution of the parameters estimated on the model with the four different sets of covariates on the same "testing" data sub-sample.

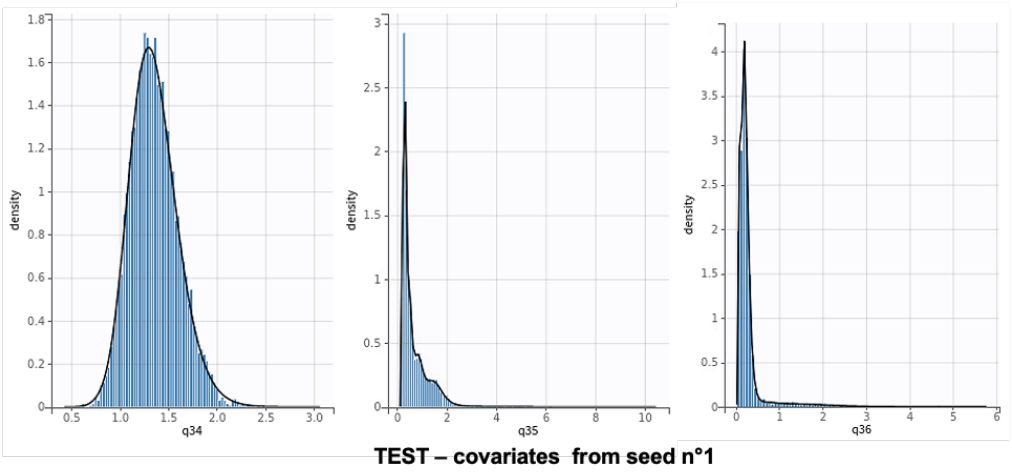



Figure 3.9: Covariates from seed n°1.

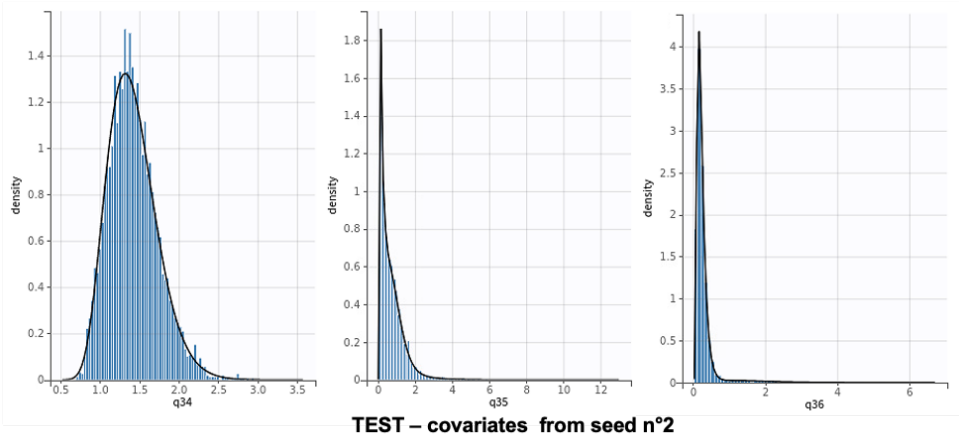



Figure 3.10: Covariates from seed n°2.

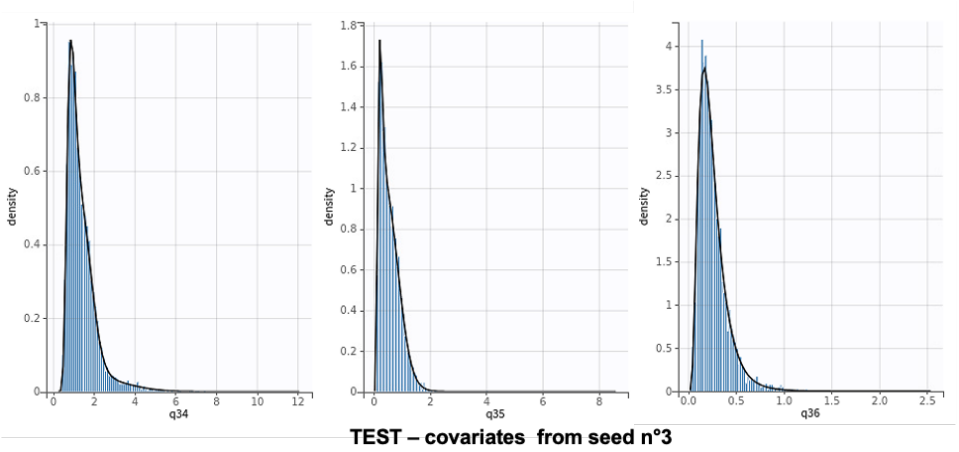



Figure 3.11: Covariates from seed n°3.

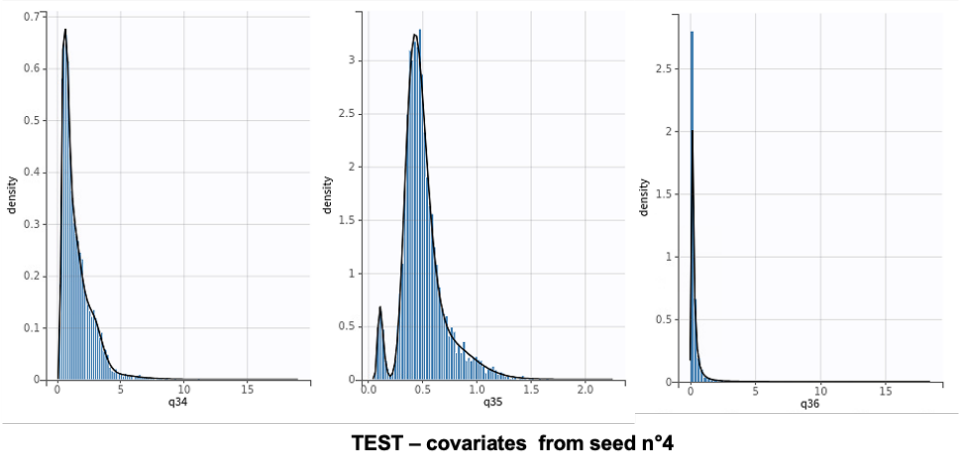



Figure 3.12: Covariates from seed n°4.

It is noticeable how the distribution of the transition rate $q35$ shows two distinct peaks only in figure 3.12, i.e., only when using a model including the covariates extracted from the analysis of the fourth "training" data sub-set. Lastly, the normalized prediction distribution errors are normally distributed and symmetric around 0 for all sets of covariates.

### A-posteriori exclusion of MA_unit from the sets of covariates

The sets of covariates selected from the analysis of the first and the third "training" data sub-sets include the dependence of some parameters on the covariate MA_unit. For these two cases, rerunning the same test, this time excluding only the dependence on such covariate of $q36$ for the former set and of $q35$ for the latter set, leads to a mild worsening of almost all the likelihood indicators, together with a great improvement of the condition numbers. Specifically, $-2\mathcal{LL}$ increases from 14 963.44 to 15 074.56 in the model with covariates set number 1 and from 15 025.46 to 15 071.25 in the model with covariates set number 3. Furthermore, BICc increases from 15 149.03 to 15 200.48 in the first case and, instead, it decreases from 15 219.57 to 15 205.69 in the second case. Contextually, however, the condition number for the former model improves from 4108.50 to 14.26, and for the latter model it improves from 2290.03 to 56.35.

Due to the reasons addressed in sub-section 4.1.2, after a thorough comparison and evaluation of the four resulting models, one for each of the extracted sets of covariates, the second one was selected. Figure 3.13 presents the population parameters estimated with such a model from the testing dataset. Both fixed and random effects for this model show good values of the relative standard error on all the clinically meaningful estimated population parameters. In this context, to be considered is that, for instance, Monolix suggests as good R.S.E. values the ones smaller than 50% [41]. Despite it being still considered a good R.S.E. value, the one given by the effect of having an ICD-10 code beginning with "R" ("beta_q34_R_1" in the figure below) resulted slightly higher than for the other estimations. However, greater errors on such a covariate value are to be expected since this refers to patients with generic symptoms, "Not elsewhere classified" according to table 2.1, which account for a significant portion (43,43%) of the total population.

| | VALUE | STOCH. APPROX. | |
|---|---|---|---|
| | | S.E. | R.S.E.(%) |
| **Fixed Effects** | | | |
| p_pop | 0.29 | 0.0064 | 2.22 |
| q13_pop | 425.95 | 78.67 | 18.5 |
| q23_pop | 432.16 | 43.81 | 10.1 |
| q34_pop | 1.52 | 0.034 | 2.23 |
| beta_q34_R_1 | -0.22 | 0.032 | 14.6 |
| q35_pop | 0.071 | 0.0049 | 6.90 |
| beta_q35_K_1 | 0.88 | 0.078 | 8.91 |
| beta_q35_age | 0.034 | 0.001 | 3.02 |
| q36_pop | 0.2 | 0.0077 | 3.75 |
| beta_q36_Z_1 | 2.1 | 0.13 | 6.10 |
| **Standard Deviation of the Random Effects** | | | |
| omega_q34 | 0.19 | 0.014 | 7.43 |
| omega_q35 | 0.27 | 0.02 | 7.33 |
| omega_q36 | 0.5 | 0.037 | 7.45 |

Figure 3.13: Population parameters estimated with the second model from the testing dataset.

# 4 | Discussion

This chapter provides the interpretation of the obtained results, discusses the value of the approach employed in this thesis and its uncertainties, summarizes the encountered technical implementation challenges, gives a brief overview of the tried and failed approaches, and introduces the future work.

## 4.1. Analysis of the results

From a theoretical point of view, it is challenging to separate clinical and operational factors in a healthcare facility and clearly distinguish between such aspects by analyzing the available data since these factors intertwine. In other words, the output can be affected by innate characteristics of the patients, e.g., age, gender, etc, by characteristics related to their main diagnosis, and by logistic factors relating to the treatment pathways. Furthermore, the LOS, i.e., the metric evaluated in this thesis, constitutes a result of both how the work within the healthcare facility is performed and how the patients are clinically characterized, thus making it impossible to fully separate the two. Nevertheless, it could be argued that logistics can be considered "clinical" to some extent. This is the case since it is possible to employ the found correlations between patients' characteristics and their LOS to know beforehand what is important to consider for an incoming patient with given characteristics to be treated effectively and efficiently. This would help reduce the negative impact of the arrival of such a patient on the logistics of the whole system. Consequently, from another perspective, if the final goal would be to use the approach employed in this thesis to inform operational research, such an approach would allow the modeler to "generate" patients for the simulation more representatively. With these premises, this section aims to frame the results of this thesis into a clinical perspective and discuss the meaningfulness of such results.

Starting from the samples whose stratification is described in sub-section 2.1.4, the designed base process model with no covariates was applied to all of such sub-sets, and the model parameters were estimated. This was done to set a benchmark for assessing how and how much the model's performance would change with the introduction of covariate variables.

Since the design of a "simplified" model is within the goals of this thesis, this operation was useful to objectively evaluate the goodness of the introduction of covariates in the model in order not to a-priori exclude the hypothesis that its base version, i.e., without covariates, could perform better than the version in which these are included. Furthermore, the aim of choosing the optimal set of covariates was also of primary importance in the reasoning process that led to the decision to employ a nonlinear mixed-effects approach and apply the COSSAC algorithm during the model building phase.

Concerning the performance of the 6-states base process model on the sub-sets, in all five cases none of the clinically meaningful parameters showed any large relative standard error (%), the transformed random effects were normal and uncorrelated, thus not dependent on the same source of variability, the transformed individual parameters resulted normally distributed as well, and the NPDEs were normally distributed around 0. Moreover, unlike the 7-states Markov Chains model, this base model showed no signs of overfitting. According to the factors mentioned above, it was thus possible to confidently state that the designed base model was meaningful and could be used as a foundation of this work.

### 4.1.1. Clinical perspective behind covariates selection

Through the execution of the COSSAC algorithm, after the "a-priori exclusion" of two of the potential covariates ("*number_of_scans*" and "*times1Year*") due to their evident independence on any of the parameters for any of the data sub-samples, it was possible to select the four sets of covariates the inclusion of which could lead to the greatest likelihood improvement. Some of the chosen covariates differ from data sub-sample to data sub-sample. However, this is very likely ascribable to sample size issues and to the potential correlation between some covariate effects.

Since the automatic covariate model building tool was applied to four data sub-samples to find covariates able to capture the variability affecting three parameters (*q34*, *q35*, *q36*) in each case, each covariate could be selected at most 12 times, and 28 different covariates were included in the exploration. Among such 28 possible covariates, only 12 (42.86%) were selected at least once by the algorithm for one of the 12 possible slots. In the last part of this sub-section, comments on the clinical perspective behind the covariates selection are reported. Figure 4.1 can be used as a guideline to follow the related discussion. In a future step, however, such comments should undergo a review process by doctors and members of the hospital management as an additional validation step.

| Variable | q34 | q35 | q36 | TOT | % pick over 12 possibilities |
|---|---|---|---|---|---|
| K | 1 | 3 | 1 | 5 | 41,67 |
| age | 2 | 2 | 0 | 4 | 33,33 |
| Z | 0 | 1 | 3 | 4 | 33,33 |
| ScanYN | 1 | 1 | 1 | 3 | 25,00 |
| M | 2 | 0 | 0 | 2 | 16,67 |
| MA_unit | 0 | 1 | 1 | 2 | 16,67 |
| A | 0 | 1 | 1 | 2 | 16,67 |
| T | 1 | 0 | 0 | 1 | 8,33 |
| R | 1 | 0 | 0 | 1 | 8,33 |
| E | 1 | 0 | 0 | 1 | 8,33 |
| AmbYN | 0 | 1 | 0 | 1 | 8,33 |
| I | 0 | 1 | 0 | 1 | 8,33 |

Figure 4.1: Count of the selected covariates.

## Covariates selected three or more times

- The medical imaging covariate *ScanYN* (see appendix A.1) was selected three times, once for each of the parameters (see figure 4.1). This is reasonable since the execution of medical imaging can potentially increase the LOS of the patients, regardless of how they are discharged. However, this result also provides strong proof of how important is the role that such a variable plays in LOS variability. Therefore, it constitutes a hint for where to focus future management improvement.

- The covariates that were selected four times are: *Z*, once on *q35* (ED to ward) and three times on *q36* (ED to other); *age*, twice on *q34* (ED to home) and twice on *q35* (see figure 4.1). *Z* represents conditions in which no specific disorder was found, but treatment was warranted (e.g., due to self-poisoning, dizziness, or abuse), and it was selected by three in four runs of automatic covariates selection for what concerns the transition rate *q36*. Such a result implies a strong capability that the variable *Z* has in explaining the variability of the LOS for patients who are neither sent home nor admitted to any hospital ward. This is reasonable since, on average, around half of the patients having an ICD-10 code beginning with "Z" were discharged in one of the ways included in the category "other" (mostly sent for consultation). To be specific, the data analysis performed on dataset D1 revealed that the ICD-10 code "Z711", i.e., "person with feared health complaint in whom no diagnosis is made", was the fifth by count in the overall ranking of the most represented individual ICD-10 codes. In most cases, the ED staff provided basic and immediate cures for these patients if the health conditions were life-threatening, and then the patients were redirected to specialized clinics.

For what concerns *age*, it was selected by two in four runs of automatic covariates selection for both *q34* (ED to home) and *q35* (ED to ward). One possibly reasonable explanation for this output is that several young patients sought help from the ED because of intoxication and poisoning, thus requiring immediate care and becoming transferable to hospital wards or intensive care units relatively soon (short LOS within the ED) due to the ease in understanding the causes of their health conditions, whereas several elderly patients visited the ED for very simple needs (e.g., disorientation) but could not leave the hospital facility until an ambulance or other special means of transportation would be available for bringing them home or to an elderly care home (prolonged LOS within the ED). This scenario would be compatible with the presence of two age peaks in the age distribution plotted in figure 2.1a.

- Lastly, the covariate *K*, which refers to diseases of the digestive system, was selected five times and at least once by each of the runs of the COSSAC algorithm: once on *q34* (ED to home), three times on *q35* (ED to ward), and once on *q36* (ED to other). It seems to be the variable that most clearly describes LOS variability in the model, especially concerning the patients discharged from the ED for being admitted to a ward. This might be because, on average, more than half of the patients to which an ICD-10 code starting with "K" was assigned had "abdominal pain" as chief complaint, which is reasonably affected by great LOS variability due to how various the causes of such pain could be and to the difficulty to evaluate them.

## Covariates selected twice

Among the 12 selected covariates, three (25%) were selected twice (see figure 4.1): *M* twice on *q34* (ED to home), and both *MA_unit* and *A* once on *q35* (ED to ward) and once on *q36* (ED to other). *M* represents diseases of the musculoskeletal system and connective tissues. *MA_unit* (see A.1) contains the type of medical team whose operators' skills could be the most suitable to treat a specific patient. *A*, together with *B*, refers to infectious and parasitic diseases. *M* is chosen by two in four runs of automatic covariate selection, and for the same transition rate (*q34*). This is reasonable because many of the patients to which this code is assigned experienced pain and injuries or swelling in a musculoskeletal district, and it is probable that, in many of these cases, these injuries did not require monitoring or operations in the ward. In this framework, much variability can arise since the execution of medical imaging may be necessary, or because there may be the need to execute surgeries that do not imply admission to a ward but do indeed increase the LOS.

For what instead concerns *MA_unit*, its impact was never selected as significant for

the transition rate *q34* (ED to home). Without receiving feedback from clinicians and hospital management about this specific result, it would be difficult to discuss whether this is meaningful or not without risking to end up producing foundation-less speculation, especially since the parameter in question is assigned to patients for merely practical reasons and is thus biased accordingly. Furthermore, since it can potentially take eight different values, and five of these are largely underrepresented within the population, it can be difficult to compute the correlation between some values of *MA_unit* and the parameters in the model. This last concept, however, is discussed in greater detail in sub-section 4.1.2.

Lastly, it might be reasonable that the covariate $A$ was never selected for the transition ratio *q34* (ED to home) since it is hard to believe that, among the patients reaching the ED and being sent home, for many the LOS could show significant variability due to infectious or parasitic diseases.

### Covariates selected only once

Among the 12 different selected covariates, five were selected only once (see figure 4.1): $T$, $R$, and $E$ for parameter *q34* (ED to home); $AmbYN$ and $I$ for parameter *q35* (ED to ward). As reported in table 2.1, the covariate $T$ represents cases of poisoning, whereas the covariate $R$ is associated with generic symptoms or not elsewhere classified pathologies, and $E$ refers to endocrine, nutritional, and metabolic diseases. Due to the clinical meaning of these covariates, it can be argued that it is reasonable to find them to be significant only for *q34*, i.e., the transition rate from being in the ED to being sent home. This is because many of the medical conditions to which these codes are applied are diseases for which the patient can either be quickly seen or treated within the ED and then sent home (e.g., in case of high blood sugar, high blood pressure, neck pain, or electrical damage) or, conversely, they are conditions in which the patients show generic symptoms that may induce the doctors to request sessions of medical imaging to better clarify the situation. In the latter case, eventually, many of these patients may have minor issues for which the treatment can be quickly terminated in the ED or continued at home.

Patients with a simplified ICD-10 diagnosis $I$, i.e., presenting diseases of the circulatory system, reaching the ED by ambulance, may very reasonably present symptoms and diseases that in most cases cannot be treated at home in the short term (e.g., cardiac arrest) and, therefore, the corresponding patients would need to be admitted to a ward. Consequently, it is reasonable to find such covariates to be significant in terms of variability of the LOS only for *q35* (ED to ward), i.e., the transition rate from being in the ED to being admitted to a ward.

An important final consideration for what concerns the clinical perspective behind the covariates selection regards the variables *countIN*, *countOUT*, and *countAVG*. It is noteworthy that such indexes of ED crowding were not selected in any of the four proposed sets of covariates. Although it could be interesting to investigate this matter further, a probable reason behind this outcome resides in the dependence of these variables on the crowding of the other wards of the hospital that receive patients from the ED throughout the year, about which no data were given. Furthermore, since these variables assume the same value for $n$ patients at a given time, the variance to which they are associated can be really low, thus leading to a weak statistical power for a potential inference.

### 4.1.2. Final model assessment

After the employment of the COSSAC algorithm with each of the four selected sets of covariates on the corresponding training data sub-sample from which such set had been selected, it was possible to assess the meaningfulness of these aforementioned sets in helping the model describe the high clinical variability embedded into complex patient characteristics. The first set of selected covariates includes: $T$ on *q34*; *AmbYN*, $K$, and *ScanYN* on *q35*; *MA_unit* and $Z$ on *q36*. The second set includes: $R$ on *q34*, $K$ and age on *q35*, Z on *q36*. The third set includes $M$ and *age* on *q34*, *MA_unit* and *age* on *q35*; $A$, $K$, and *ScanYN* on *q36*. Lastly, the fourth selected set includes: $E$, $K$, $M$, *ScanYN*, and *age* on *q34*; $A$, $I$, $K$ and $Z$ on *q35*; $Z$ on *q36*.

With all the sets of covariates, the model showed good R.S.E. on all the clinically meaningful estimated population parameters, normally distributed NPDEs with symmetry around 0, and a considerable improvement in the values of the likelihood indicators when comparing them to the ones in the corresponding run without covariates with the same dataset. However, these models also showed some problems in the computation of the effects of the covariate *MA_unit* for seed n°1 and seed n°3 and a non-computable condition number for the seeds 1, 3, and 4. Nevertheless, the latter issue was only related to the small sample size of the employed "training" data sub-sets since the condition numbers could then be properly calculated when a larger sample was applied to Markov Chains models with the same sets of covariates, as discussed in the last part of this sub-section. Remarkably, according to the output of the estimation runs performed with the selected covariates, again on the training sets of data, it was already possible to easily have a hint of which covariates could probably be removed to improve the model, in other words, *number_of_scans* and *times1Year*, starting from the analysis of the estimated parameters. The ease and accuracy of such early feedback constitute a powerful advantage of the approach employed in this thesis.

## Validation

With the knowledge gathered and discussed so far in section 4.1, it was thus possible to test the four sets of covariates on a larger data sub-sample, containing 5031 patients and independent of the other sets. Once again, the random effects achieved with all the four sets of covariates resulted normal and uncorrelated, and the NPDEs normal and symmetric around 0. On such a larger data sample, it was also possible to compute the condition number in all four cases, but its value clearly indicated overfitting with the first and the third sets of covariates, and the first model also showed high R.S.E. on some of the estimated parameters.

Grouping the information given by the condition number together with the hints that the estimations on the training data sub-sets had produced at an earlier stage regarding the potential need to remove $MA\_unit$ from the sets of covariates led to a confident belief in the need to perform a-posteriori exclusion of such a covariate from the models 1 and 3. Since a new testing execution for these two models, this time without the covariate in question, produced a good condition number in both models 1 and 3, the goodness of the performed covariate exclusion was confirmed.

At this stage, it became possible to evaluate the best performing sets of covariates, so to be able to propose a final model. The model that included the fourth set of covariates was the one achieving the best values of the likelihood indicators but at the cost of a relatively high condition number (98.01), which posed the reasonable question of whether such a model was overfitted or not. By looking at the plot of the individual parameters estimated with this fourth model (see figure 3.12), however, it was possible to see two separated peaks in the estimation of the parameter *q35* (ED to ward), which had never happened in any of the other estimations. Due to the presence of such a second peak, the high condition number, and the highest number of covariates included in the model, it was reasoned that the latter was overfitted, or at least the worst in terms of generalization, thus it was excluded. It was also possible to exclude the third model since it showed the worst values in terms of likelihood indicators and the second-worst condition number after the one scored by the fourth model.

With only two models left, these were compared not only in terms of likelihood values, better in the second model, and in terms of condition number, better in the first model, but also in terms of the number and quality of the covariates included in each of the two models. Specifically, the first model included five covariates (without $MA\_unit$), two of which were not selected in any other model during the execution of the COSSAC algorithm. Conversely, the second model included only four covariates, three of which were exactly

the three most selected ones by COSSAC. Two of such covariates had been selected by the algorithm also in two other models, and one had been selected by the algorithm also in another model. To conclude, it was reasoned that, taking into consideration likelihood, condition number, quantity and quality of the covariates and, thus, level of generalization, the best model was the second one. The population parameters estimated with this model from the testing data sub-set can be seen in figure 3.13. This last discussed aspect is representative of how difficult handling this kind of real-world data can be since insidious effects can arise, and the modeling decisions must be made only after carefully evaluating several factors, which can sometimes be conflicting. Finally, concerning the values of the relative standard error achieved by the best model (figure 3.13) and presented in section 3.3, obtaining a larger R.S.E. (14.60 %) on the effect of having an ICD-10 code beginning with "R" ("beta_q34_R_1") can even be considered as a positive outcome. Indeed, this is positive because it is very likely an indicator of the model being sensitive to the higher generality of a specific group of patients. Moreover, since such a group is the one associated with generic symptoms ("Not elsewhere classified" according to table 2.1), this becomes particularly meaningful. In this context, the extreme over-representation of the class in question should not be seen as an indicator of poor data quality but as additional proof of the entity of clinical complexity and variability.

## 4.2. Value of the approach

Nonlinear mixed-effects modeling is not commonly employed to evaluate the impact of clinical covariates on logistical outcomes in the context of process modeling. However, since the execution of a preliminary data analysis revealed high complexity in patient characteristics (sub-section 2.1.2) and data sparseness on some of the variables, and since the assessment of the conventional approaches showed several relevant limitations that these face when trying to properly describe the variability embedded into the data, it was reasoned that the employment of a mixed-effects modeling approach would have helped to design an effective process model, by allowing the modelers to differentiate among the covariates that are relevant and significant for each of the modeled state transitions, thus creating a bridge between the domain of covariate analysis and the one of logistical simulation modeling. Furthermore, given that nonlinear mixed-effects modeling has no requirements for "rich" or "dense" data, nor for it to have any particularly structured sampling time [16], it was also reasoned that the employment of a mixed-effects modeling approach would have made the designed model less affected by data sparseness than if using more conventional approaches. Moreover, the modeling techniques that are traditionally applied to health logistics data tend to select a pool of patient and

system characteristics and apply them to macro sections of the model without allowing neither for much differentiation among the covariates that are relevant and significant for each of the modeled state transitions nor for the consideration of potential random errors. Conversely, a nonlinear mixed-effects model incorporates both fixed effects and random effects" [28], and allows extracting insight from the data using a population approach [16], which is able to fit one model to data coming from all the subjects without losing the notion of individuals, thus allowing to investigate the sources of variability very specifically by discriminating between inter-individual and intra-individual variability [16]. The ability to estimate variability and covariate effects is relevant and powerful for this area of application.

The employment of the approach used in this thesis also answers the research question related to understanding the impact of complex patient characteristics on ED logistics concerning the effect that the covariates underlying such characteristics produce on the LOS. Due to limitations deriving from the available data and the system's complexity, however, at the current stage of the work, it is not possible to clearly state whether or not such patient characteristics affect specific aspects of the overall ED logistics. Nevertheless, this thesis tries, to a certain extent, to simultaneously study the effect of patient characteristics on the logistics and characterize the logistics according to patient characteristics. On the one hand, it employs the logistic output, in terms of LOS, to trace this back to the underlying patient characteristics. On the other hand, the deriving acquisition of prior knowledge about the probable implications of the arrival of a new patient with specific characteristics has the potential to be employed in the future either for better informing operational research or for practical implementations in the hospital settings. Despite this, some challenges, discussed in section 4.3, are yet to be overcome.

## 4.3. Limitations

In this section, this thesis's boundaries and limits are described, both regarding the chosen approach and for what concerns the encountered technical implementation challenges.

### 4.3.1. Uncertainties in the approach

- The model in this study does not represent the real system in all its fine details and is not meant to emulate the real system in its full complexity. Conversely, an empirical process model was inferred directly from the real clinical data, achieving to model the real system in a simplified way so that it could be suitable for potential future testing of optimization procedures. Moreover, the boundaries of modeling

and analysis were limited to the emergency department and to what concerns the employment of the imaging department for patients belonging to the ED.

- Elements such as the impact of teamwork and subjective personal interactions among clinicians, and the impact of potential biases towards some specific classes of patients, have not been explicitly considered for the inference of the model. To be more specific, the limits of the model in this sense are mostly limited to the data itself since no assumptions regarding other knowledge have been made.
- The recent history proved that sudden and unexpected macro-scale events, e.g., a global pandemic, can drastically change the functioning and the equilibrium of the healthcare systems in all their components, despite these being usually stable and consolidated otherwise. Therefore, since the process model elaborated in this thesis does not introduce any automatic parameter update over time if the conditions of the system change, further study might be needed to compare the post-pandemic behavior of the ED of Akademiska sjukhuset to the pre-pandemic one, which is the one represented in the data employed in this thesis. However, despite the impossibility of studying some nonlinear effects such as overcrowding, an attempt was made to address this specific factor by considering the effect of the covariates *countIN*, *countOUT*, and *countAVG*.
- No real dataset is realistically devoid of outliers among the data it contains. A process for outlier removal was not performed explicitly in this work on the values of LOS, and it could be relevant to include such an activity in the pre-processing procedure for future works with the approach proposed in this thesis. Anyway, the lack of such an outlier removal was partially mitigated by making sure that the data samples employed in this thesis for training, testing, and validation, were stratified by proportionate allocation of the values taken by *simple_ diag* and balanced according to the remaining covariates, as can be seen in figures 2.4 and 2.5. Moreover, as shown by figure 2.1b, the empirical distribution of the LOS in dataset D1 accurately follows a log-normal probability density function. Therefore, it is implausible that the value of the length of stay could take extreme values for a considerably high count of patients in such a dataset. All that said, despite a process for outlier removal could be relevant for future work, at least for what concerns the scope of this thesis, the effect of the outliers on the designed model can be neglected.

### 4.3.2. Technical implementation challenges

The main technical implementation challenge was related to the high computational demand of the employed approach, which directly impacted this thesis in two ways:

- From the former population of patients visiting the ED during 2019, constituted by 49 936 entries, the dataset had to be considerably sub-sampled to study the feasibility of several modeling techniques and their possible variations without an excessive time expense. For this purpose, it was decided to parallelize the approach for the initial model exploration and later training purposes by using four samples of $\sim 933$ patients each, instead of a single and much larger sample, whereas a larger sample, i.e., composed by 5031 patients, was used for testing and validation.
- The number of possible states and possible transitions included in the model had to be carefully designed in order to not further extend computational times. Therefore, the transition from state 1 to state 3 and the one from state 2 to state 3 were not designed in a way that would give them any clinical or logistic meaning, so that it could be possible to avoid estimating random effects and parameters distributions for the parameters corresponding to such transitions.

A second technical implementation challenge was then given by the software chosen for implementing a nonlinear mixed-effects approach. This is due to the considerable differences between the intended use of the software and the readaptation that was operated in this thesis to be able to apply it to modeling patient flow instead of pharmacokinetics. To be specific, the main software-related implementation challenges were concerning the interpretation of the initial conditions for the Markov Chains model and the slowness of the software in dealing with large datasets since the available ones are much larger than what is usually employed in pharmacokinetics. Monolix GUI certainly showed several advantages of implementation. Contextually, however, for what concerns the computational slowness, it also introduced the limitation of not being able to design a parallelization of the runs through its GUI. Therefore, to facilitate the experiments for this thesis, two separated devices were used simultaneously to run the sessions of parameters estimation, so to increase the number of tests that could be successfully carried out on each day of work. The two employed devices are a computer server (Intel(R) Xeon(R) CPU: E5-2630 v2 @ 2.60 GHz, 64 bit processor. RAM: 62GB), and a laptop (MacBook Pro 15' 2016. CPU: Quad-Core Intel Core i7 @ 2,60 GHz, 64 bit processor. RAM: 16GB. GPU: AMD Radeon Pro 450, 2GB VRAM).

Lastly, another important technical implementation challenge resided in the goal of designing a model informed by real data since this implied having to deal with high clinical

variability and with the possibility that there could be missed data points or wrongly registered information in the datasets obtained from the healthcare provider.

## 4.4. Exclusion of modeling techniques

The purpose of this section is to briefly discuss the other modeling techniques and ideas whose implementation was attempted in this thesis (see section 2.4), addressing why they were early discontinued and, consequently, no actual modeling was performed with such techniques.

For what concerns the **Time-To-Event approach**, introduced in sub-section 2.4.1, the approach itself looked applicable and meaningful. However, it was reasoned that this idea would not have exploited the full potential of the nonlinear mixed-effects modeling, and the outcoming process model describing the relationship between the length of stay and the covariates would not have helped to evaluate the impact of complex patient characteristics on the logistics of the ED system more than what a multiple regression approach would have done. Accordingly, this approach was discarded.

Concerning the **longitudinal model on day-wise time of arrival**, introduced in sub-section 2.4.2, the CC was used as the identifier onto which to group the entries, i.e., the observations associated with a patient with the same CC were interpreted by the employed software as coming from the same "individual". Due to this choice, however, the volume of information used to estimate the parameters was unbalanced among different chief complaints because of the high variability in terms of patients count by chief complaint (see sub-section 2.1.2). Therefore, the resulting model building would have very likely been skewed by producing strong assumptions on the chief complaints for which the count is high, whereas for several pathologies represented by small patient counts, e.g., "Abstinens", the estimated random effects would have very likely been too preponderant. In this case, it was reasoned that the technique in question could be meaningful in other similar situations, e.g., if the scope of the analysis would aim at focusing exclusively on the most represented sets of chief complaints. Given that this is not the research question of this thesis, the attempts with this technique were also discontinued.

Lastly, for what concerns the **longitudinal count data on hour-wise yearly time of arrival**, introduced in sub-section 2.4.3, the step that followed setting up the dataset for this approach, i.e., the design of the structural model, revealed itself to be particularly challenging in terms of formulation. Furthermore, further analysis of the data revealed that this approach would not have had any real meaningfulness for this thesis since, as mentioned in section 2.1.3, no significant crowding variability seems to exist throughout the

year in the emergency department of Akademiska sjukhuset. It was reasoned, however, that the technique in question still deserves mention in this thesis since it could be meaningful to use it for other related engineering issues, e.g., characterizing the variability between day hours and night hours in the logistics of an ED.

## 4.5. Future work

Due to the technical implementation challenges discussed in sub-section 4.3.2 and to the novelty of this approach, which required a careful and time-consuming exploration of the possible modeling techniques, a substantial part of the potential of the employed approach was exploited, but there would still be good margin for further studies. Therefore, to provide the reader with an advantageous starting point for further testing and application of the approach employed in this thesis, the focus in this section will be on the most important issues that should be addressed in future work.

First of all, a supplementary validation could be performed after having obtained sets of covariates like the ones shown in sub-section 3.2.3. Starting from the parameters and their distributions extracted from the real data according to the procedure followed in this thesis, the second validation would consist in generating a new population of patients based on these parameters for simulating their flow through the designed Markov Chains model and, consequently, calculating their length of stay. Then it would be possible to compare such simulated LOS with the one related to the real patients.

Afterward, the employed approach, which focused on patient flow modeling, could be used to inform operational research, with the advantage that the latter would allow the healthcare leaders to identify bottlenecks and answer "what-if" questions about real-world scenarios without having to face the costs and dangers of performing “trial and error” intervention on the real system. Among the techniques described in section 3.1, Discrete Event Simulation modeling or Agent-based Simulation modeling could be suitable choices for this purpose. To undertake the path of operational research, information regarding staffing, resources, and their utilization, would be needed to be able to properly design a simulation model. Since this kind of information was not included in the data, it was not possible to implement a DES or an ABS model in this thesis, and such implementation is thus highly recommended as potential future work.

Concerning the challenge in parallelizing the computation through Monolix GUI, which was discussed in sub-section 4.3.2, a future project could consist in designing such missing feature, or rather in re-writing the whole code from scratch. The latter approach, however, would be time-consuming due to the need to write a function for defining the NLME

approach and one for implementing the COSSAC algorithm, as well as to design the Markov Chains model with the parameters obtained through the application of NLME and COSSAC functions, and to finally optimize the code.

For what instead concerns the selection of the most significant covariates for each parameter and each transition, it would be meaningful to perform a sensitivity analysis to compare the results achieved with the COSSAC algorithm (see section 2.6) to what would be the output of more traditional selection techniques, such as "Lasso regression", which has already been applied for covariate selection in mixed-effects modeling [55], or machine learning feature selection techniques, e.g., "Boruta feature selection" [56]. Moreover, it could be possible to unfold some of the most represented ICD-10 codes from the grouping by macro diagnostic area used in this thesis and compare the potential new model with the previously validated one. This could allow for investigating some patient characteristics that are currently not distinguishable due to the grouping.

Lastly, time-varying covariates could be introduced in the NLMEM formulation to account for the potential yearly variability within the system. Moreover, a covariate describing whether the arrival of the patients to the ED happens during the daytime or overnight could be added since several hospital wards do not admit patients during the night. The latter covariate could help further separate potential LOS variability deriving from the hospital's logistics from the LOS variability related to the patient characteristics. Additionally, since the data regarding patients entering the ED from June to August were excluded from the process model design, a new model could be generated just with the data coming from the three months in question. This would allow comparing such a scenario with the results produced with the data coming from the rest of the year. Alternatively to the latter possibility, a covariate describing whether the arrival of the patients to the ED happens during summer or the rest of the year could be added. Consequently, the effect of understaffing of the ED could be accounted for, and dataset D1 could thus be used almost in its entirety.

# 5 | Conclusions

In this thesis, mixed-effects modeling, an approach typically used in pharmacometrics, was applied to hospital medical records. Within the chosen approach, a Markov Chains model of patient flow that could capture and describe the impact of patient complex characteristics on the logistics of the emergency department was designed, tested, and validated. This was done with the purpose of establishing a bridge between logistical systems and the clinical insights of the hospital, which is particularly challenging due to the difficulty in dealing with high patient volumes and high clinical variability embedded into real clinical data. Accordingly, this work aimed at improving the understanding of how such data could be better exploited for healthcare modeling to potentially achieve a better organization of the hospitals in the future, and it managed to develop an approach for estimating covariate effects on parameters linked to the process description in the emergency department. Furthermore, due to how much the performance of the emergency department affects the functioning of the other hospital wards and, indirectly, healthcare systems and communities at large, the technique applied in this thesis, as well as the deriving model, were designed so that they could become the starting point for future operational research studies aiming at testing length of stay optimization procedures on the emergency department.

# Bibliography

[1] H. R. Rasouli, A. A. Esfahani, M. Nobakht, M. Eskandari, H. Goodarzi, and M. A. Farajzadeh, "Outcomes of Crowding in Emergency Departments; a Systematic Review," p. 10, 2019.

[2] S. P. LLC, *Hospital Acquired Infections*. StatPearls Publishing, 2022.

[3] S. Paling, J. Lambert, J. Clouting, J. González-Esquerré, and T. Auterson, "Waiting times in emergency departments: exploring the factors associated with longer patient waits for emergency care in England using routinely collected daily data," *Emergency Medicine Journal*, pp. emermed–2019–208 849, Sep. 2020. doi: 10.1136/emermed-2019-208849

[4] P. Bhattacharjee and P. K. Ray, "Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: A review and reflections," *Computers & Industrial Engineering*, vol. 78, pp. 299–312, Dec. 2014. doi: 10.1016/j.cie.2014.04.016

[5] M. M. Gunal, "A guide for building hospital simulation models," *Health Systems*, vol. 1, no. 1, pp. 17–25, Jun. 2012. doi: 10.1057/hs.2012.8

[6] J. I. Vázquez-Serrano, R. E. Peimbert-García, and L. E. Cárdenas-Barrón, "Discrete-Event Simulation Modeling in Healthcare: A Comprehensive Review," *International Journal of Environmental Research and Public Health*, vol. 18, no. 22, p. 12262, Nov. 2021. doi: 10.3390/ijerph182212262

[7] E. Hamrock, K. Paige, J. Parks, J. Scheulen, and S. Levin, "Discrete Event Simulation for Healthcare Organizations: A Tool for Decision Making:," *Journal of Healthcare Management*, vol. 58, no. 2, pp. 110–124, Mar. 2013. doi: 10.1097/00115514-201303000-00007

[8] J. W. Joseph and B. A. White, "Emergency Department Operations," *Emergency Medicine Clinics of North America*, vol. 38, no. 3, pp. 549–562, Aug. 2020. doi: 10.1016/j.emc.2020.04.005

[9] C. Zhang, K. P. Härenstam, S. Meijer, and A. S. Darwich, "Serious Gaming of Logistics

Management in Pediatric Emergency Medicine," *International Journal of Serious Games*, vol. 7, no. 1, pp. 47–77, Mar. 2020. doi: 10.17083/ijsg.v7i1.334

[10] K. A. A. Anantharaj, "Improving management of patient flow at Radiology Department using Simulation Models," p. 62.

[11] P. Yoon, I. Steiner, and G. Reinhardt, "Analysis of factors influencing length of stay in the emergency department," *CJEM*, vol. 5, no. 03, pp. 155–161, May 2003. doi: 10.1017/S1481803500006539

[12] H. Hajjarsaraei, B. Shirazi, and J. Rezaeian, "Scenario-based analysis of fast track strategy optimization on emergency department using integrated safety simulation," *Safety Science*, vol. 107, pp. 9–21, Aug. 2018. doi: 10.1016/j.ssci.2018.03.025

[13] J. C. Moskop, J. M. Geiderman, K. D. Marshall, J. McGreevy, A. R. Derse, K. Bookman, N. McGrath, and K. V. Iserson, "Another Look at the Persistent Moral Problem of Emergency Department Crowding," *Annals of Emergency Medicine*, vol. 74, no. 3, pp. 357–364, Sep. 2019. doi: 10.1016/j.annemergmed.2018.11.029

[14] B. Shirazi, "Fast track system optimization of emergency departments: Insights from a computer simulation study," *International Journal of Modeling, Simulation, and Scientific Computing*, vol. 07, no. 03, p. 1650015, Sep. 2016. doi: 10.1142/S179396231650015X

[15] K. W. McKinley, J. M. Chamberlain, Q. Doan, and D. Berkowitz, "Reducing Pediatric ED Length of Stay by Reducing Diagnostic Testing: A Discrete Event Simulation Model," *Pediatric Quality & Safety*, vol. 6, no. 2, p. e396, Mar. 2021. doi: 10.1097/pq9.0000000000000396

[16] P. L. Bonate, "Nonlinear Mixed Effects Models: Theory," in *Pharmacokinetic-Pharmacodynamic Modeling and Simulation*. Boston, MA: Springer US, 2011, pp. 233–301. ISBN 978-1-4419-9484-4 978-1-4419-9485-1

[17] M. Beaghen, "Canonical Variate Analysis with Longitudinal Data," 1997.

[18] I. Grigoryev, "AnyLogic in Three Days," p. 251, 2018.

[19] L. Standfield, T. Comans, and P. Scuffham, "MARKOV MODELING AND DISCRETE EVENT SIMULATION IN HEALTH CARE: A SYSTEMATIC COMPARISON," *International Journal of Technology Assessment in Health Care*, vol. 30, no. 2, pp. 165–172, Apr. 2014. doi: 10.1017/S0266462314000117

[20] J. Ladyman, J. Lambert, and K. Wiesner, "What is a complex system?" *Euro-*

*pean Journal for Philosophy of Science*, vol. 3, no. 1, pp. 33–67, Jan. 2013. doi: 10.1007/s13194-012-0056-8

[21] A. Goienetxea Uriarte, E. Ruiz Zúñiga, M. Urenda Moris, and A. H. Ng, "How can decision makers be supported in the improvement of an emergency department? A simulation, optimization and data mining approach," *Operations Research for Health Care*, vol. 15, pp. 102–122, Dec. 2017. doi: 10.1016/j.orhc.2017.10.003

[22] B. MIELCZAREK, "Review of modelling approaches for healthcare simulation," *OPERATIONS RESEARCH AND DECISIONS; ISSN 2081-8858*, 2016. doi: 10.5277/ORD160104 Medium: PDF Publisher: Wrocław University of Technology, Wrocław University of Economics, Polish Operational and Systems Research Society.

[23] R. Amalberti, Y. Auroy, D. Berwick, and P. Barach, "Five System Barriers to Achieving Ultrasafe Health Care," *Annals of Internal Medicine*, vol. 142, no. 9, p. 756, May 2005. doi: 10.7326/0003-4819-142-9-200505030-00012

[24] M. M. Wagner, W. R. Hogan, W. W. Chapman, and P. H. Gesteland, "Chief Complaints and ICD Codes," in *Handbook of Biosurveillance*. Elsevier, 2006, pp. 333–359. ISBN 978-0-12-369378-5

[25] A. Marazzi, F. Paccaud, C. Ruffieux, and C. Beguin, "Fitting the Distributions of Length of Stay by Parametric Models:," *Medical Care*, vol. 36, no. 6, pp. 915–927, Jun. 1998. doi: 10.1097/00005650-199806000-00014

[26] R. E. Weiss, *Modeling longitudinal data*, ser. Springer texts in statistics. New York ; London: Springer, 2005. ISBN 978-0-387-40271-0 OCLC: ocm61218216.

[27] C. J. Pinheiro and M. B. Bates, *Mixed-Effects Models in S and S-PLUS*, ser. Statistics and Computing. New York: Springer-Verlag, 2000. ISBN 978-0-387-98957-0. [Online]. Available: http://link.springer.com/10.1007/b98882

[28] Mathworks, "Nonlinear Mixed-Effects Modeling - MATLAB & Simulink - MathWorks United Kingdom." [Online]. Available: https://uk.mathworks.com/help/simbio/ug/what-is-nonlinear-mixed-effects-modeling.html

[29] Lixoft, "Monolix Homepage." [Online]. Available: https://lixoft.com/products/monolix/

[30] Lixoft, "Population parameter using SAEM algorithm," 2017. [Online]. Available: https://monolix.lixoft.com/tasks/population-parameter-estimation-using-saem/

[31] Lixoft, "Probability distribution of the individual parameters in Monolix." [Online]. Available: https://monolix.lixoft.com/data-and-models/individualdistribution/

[32] Lixoft, "Check initial estimates and auto init in Monolix." [Online]. Available: https://monolix.lixoft.com/tasks/check-initial-estimates-and-auto-init/

[33] Lixoft, "Initial estimate for Monolix." [Online]. Available: https://monolix.lixoft.com/tasks/initialization/

[34] Lixoft, "Model for individual covariates using Monolix." [Online]. Available: https://monolix.lixoft.com/data-and-models/covariate/

[35] Lixoft, "Algorithms convergence assessment." [Online]. Available: https://monolix.lixoft.com/tasks/algorithms-convergence-assessment/

[36] Lixoft, "Bayesian estimation using Monolix." [Online]. Available: https://monolix.lixoft.com/data-and-models/bayesianestimation/

[37] Lixoft, "Conditional distribution calculation using Monolix," 2017. [Online]. Available: https://monolix.lixoft.com/tasks/conditional-distribution/

[38] Lixoft, "EBEs (conditional mode) calculation for Monolix." [Online]. Available: https://monolix.lixoft.com/tasks/ebes/

[39] Lixoft, "Tasks and results." [Online]. Available: https://monolix.lixoft.com/tasks/

[40] J. A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, Jan. 1965. doi: 10.1093/comjnl/7.4.308

[41] Lixoft, "Standard error using the Fisher Information Matrix." [Online]. Available: https://monolix.lixoft.com/tasks/standard-error-using-the-fisher-information-matrix/

[42] Lixoft, "Statistical tests for model building on Monolix." [Online]. Available: https://monolix.lixoft.com/tasks/tests/

[43] Lixoft, "Log Likelihood estimation using Monolix." [Online]. Available: https://monolix.lixoft.com/tasks/log-likelihood-estimation/

[44] Lixoft, "Time to event data models in Monolix." [Online]. Available: https://monolix.lixoft.com/data-and-models/ttedata/

[45] Lixoft, "Categorical data modeling using Monolix." [Online]. Available: https://monolix.lixoft.com/data-and-models/categoricaldata/

[46] A. Leijon and G. E. Henter, "Pattern Recognition," p. 283, 2015.

[47] G. Ayral, J. Si Abdallah, C. Magnard, and J. Chauvin, "A novel method based on unbiased correlations tests for covariate selection in nonlinear mixed effects models: The COSSAC approach," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 10, no. 4, pp. 318–329, Apr. 2021. doi: 10.1002/psp4.12612

[48] Lixoft, "Automatic covariate model building." [Online]. Available: https://monolix.lixoft.com/model-building/automatic-covariate-model-building/

[49] K. B. Ahsan, M. R. Alam, D. G. Morel, and M. A. Karim, "Emergency department resource optimisation for improved performance: a review," *Journal of Industrial Engineering International*, vol. 15, no. S1, pp. 253–266, Dec. 2019. doi: 10.1007/s40092-019-00335-x

[50] S. Adeyemi, T. Chaussalet, and E. Demir, "Nonproportional random effects modelling of a neonatal unit operational patient pathways," *Statistical Methods & Applications*, vol. 20, no. 4, pp. 507–518, Nov. 2011. doi: 10.1007/s10260-011-0174-z

[51] C. Zhang, W.-W. Wang, M.-M. Pan, and Z.-C. Gu, "Simulation of anticoagulation in atrial fibrillation patients with rivaroxaban—from trial to target population," *Reviews in Cardiovascular Medicine*, vol. 22, no. 3, p. 1019, 2021. doi: 10.31083/j.rcm2203111

[52] M. Tracy, M. Cerdá, and K. M. Keyes, "Agent-Based Modeling in Public Health: Current Applications and Future Directions," *Annual Review of Public Health*, vol. 39, no. 1, pp. 77–94, Apr. 2018. doi: 10.1146/annurev-publhealth-040617-014317

[53] S. Adeyemi and T. J. Chaussalet, "Models for Extracting Information on Patient Pathways," in *Intelligent Patient Management*, S. McClean, P. Millard, E. El-Darzi, and C. Nugent, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 171–182. ISBN 978-3-642-00179-6

[54] K. Katsaliaki and N. Mustafee, "Applications of Simulation within the Healthcare Context," p. 44.

[55] J. Ribbing, J. Nyberg, O. Caster, and E. N. Jonsson, "The lasso—a novel method for predictive covariate model building in nonlinear mixed effects models," *Journal of Pharmacokinetics and Pharmacodynamics*, vol. 34, no. 4, pp. 485–517, Aug. 2007. doi: 10.1007/s10928-007-9057-1

[56] M. B. Kursa and W. R. Rudnicki, "Feature Selection with the Boruta Package," *Journal of Statistical Software*, vol. 36, no. 11, 2010. doi: 10.18637/jss.v036.i11

# A | Datasets

## A.1. Explanation of the variables

This section presents and describes the variables composing the two acquired datasets (D1 and D2). Whereas "***contact_id***" and "***person_id***" are respectively a case identifier and a patient identifier, as it is described in sub-section 2.1.1, "***municipality***" refers to the municipality of residence of each patient, and no explanation for the variables "***sex***" and "***age***" is needed.

"***ARRIVAL_DATE***" in dataset D1 and "***Contact_Start***" in dataset D2 have the same meaning, which consists of the date and time at which each patient entered the ED. Similarly, both "***DISCHARGE_DATE***" in dataset D1 and "***Contact_End***" in dataset D2 consist of the date and time at which each patient left the ED. Contextually, "***arrival_method***" describes whether a patient reached the ED by ambulance or by other means of transportation, and "***reason_for_discharge***" consists of the explanation of the reason because of which each of them left the ED. The list of these possible reasons is shown in table 2.3.

"***first_doctor_contact_date***" refers to the date and time at which each patient was visited by a doctor or dentist, or medical student for the first time after having entered the ED, where the variable "***contact_type***" describes which of these three kinds of healthcare providers performed such a first visit for each patient. "***last_doctor_contact_date***" refers to the date and time at which each patient was visited by a healthcare provider among the kinds mentioned above for the last time before leaving ED in one of the possible ways that are shown in table 2.3. Through Python programming, it was possible to determine that for 36 462 patients (73,02% of the total) the variable "*last_doctor_contact_date*" assumes the same value that the variable "*first_doctor_contact_date*" shows, i.e., these patients were visited by a doctor, or dentist, or medicine student only once before leaving the ED.

"***priority***" is supposed to refer to whether patients had life-threatening conditions that granted them an overriding over other patients. However, feedback coming from experts of

Akademiska sjukhuset revealed that the hospital did not have a clear and unified protocol for when the healthcare providers had to write a patient in the registry as a priority patient. It was thus advised by the experts against making use of such parameter for modeling the process. "***triage***" refers to the triage category to which each patient was assigned according to Akademiska sjukhuset's 2019 triage system. An analysis of the possible values that this variable could take during 2019 and a consultation of experts from the hospital revealed the existence of six levels: "RED" for the patients in need of immediate and continuous monitoring, "ORANGE" for those who needed monitoring or surveillance every 20 minutes and to be seen by a doctor within 30 minutes to 60 minutes, "YELLOW" for those who needed monitoring or surveillance every hour and to be seen by a doctor within two hours, "GREEN" for those who needed monitoring or surveillance every two hours to four hours and to be seen by a doctor within four hours, "BLUE" for those who needed monitoring or surveillance every four hours and to be seen by a doctor within four hours, "WHITE" for those whose triage surveillance process was over. However, the same experts from the hospital referred that it was common during 2019 not to have a systematic way of reporting the updates of the triage color over time, and thus they advised against making use of the parameter in question for modeling the process. "***triage_level***" is the numerical coding of the variable "*triage*", where 1 is assigned to the "RED" codes, 2 is assigned to the "ORANGE" ones, 3 to the "YELLOW" ones, 4 to the "GREEN" ones, 5 to the "BLUE" ones and 6 to the "WHITE" ones. However, by computing the count of patients assigned to each of the levels, it was found that to no entry in dataset D1 the "BLUE" code seems to be assigned. However, a consultation with an expert from Akademiska sjukhuset revealed that the protocol for all the patients with a "BLUE" code was to re-record the triage level when their need for monitoring ceased, i.e., when the newly assigned code would have been "WHITE".

"***MA_unit***" refers to a label assigned to each patient during the first contact with a doctor. This label indicates the type of medical team whose operators' skills were thought to be the most appropriate to treat a specific patient, whereas "***team_care_contact***" refers to the actual medical team to which each patient was assigned, e.g., Acute care team ("akutteam") or Medical team 1 ("Med team 1").

"***cause_of_visit***" refers to the "chief complaint" that was assigned to each of the patients. An explanation of what is a chief complaint is provided in sub-section 2.1.2, whereas the distribution of this variable among the patients can be seen in figure 2.2. Contextually, "***main_diagnosis***" refers to the formal classification given to each patient according to the "ICD-10" classification of their main diagnosis. An explanation of what is an ICD-10 code is provided in sub-section 2.1.2, whereas a grouping of the patients by macro

diagnostic areas thanks to a simplification of the ICD-10 codes to their first letter can be seen in table 2.1.

"***LoS_hours***" in dataset D1 and "***duration_contact***" in dataset D2 have the same meaning, which consists in the value of the metric "Length of stay", whose meaning is explained in sub-section 1.1.1, expressed in hours.

"***RegistrationDate***" indicates on which date and at which time a doctor requested the medical imaging session. "***ExaminationDate***" indicates on which date and at which time the medical imaging session was performed. "***ReportDate***" indicates on which date and at which time the outcome of the performed medical imaging session was written in the hospital's electronic record.

"***RadiologyInvestigation***" describes which type of medical imaging was performed and in which body district. The entries for which "*RadiologyInvestigation*" only mentions a body district consist of sessions of x-ray imaging; for the other entries, the type of performed medical imaging is also specified.

"***Contact_Status***" consists of a categorical variable meant for declaring whether the corresponding medical imaging session was actually performed or not. By analyzing the whole dataset D2, it was possible to observe that all the listed sessions were performed. "***Radiology_Status***" indicates whether the corresponding entry refers to a partial decision or to a final decision over the outcome of a performed medical imaging session. "***Contact_Type***" from dataset D2, which has a different meaning than the variable "*contact_type*" from dataset D1, explains the type of the corresponding medical contact for imaging purposes. The count of entries by "*Contact_Type*" was performed, and it resulted in 53 408 entries being labeled as "reception visit" ("Mottagningsbesök"), 110 as "activity" ("Aktivitet"), and 34 as "consultation" ("Konsultation").

# B | Monolix code

```
DESCRIPTION: categorical data
Markovian dependence
continuous time (estimation of transition rates)

[LONGITUDINAL]
input = {p,q13, q23, q34, q35, q36}

DEFINITION:
State = {type = categorical, categories = {1,2,3,4,5,6}, dependence = Markov
  P(State_1=1)=p, P(State_1=3)=0, P(State_1=4)=0, P(State_1=5)=0, P(State_1=6)=0
  transitionRate(1,2) = 0, transitionRate(2,1) = 0, transitionRate(1,3) = q13
  transitionRate(3,1) = 0, transitionRate(1,4) = 0, transitionRate(4,1) = 0
  transitionRate(1,5) = 0, transitionRate(5,1) = 0, transitionRate(1,6) = 0
  transitionRate(6,1) = 0, transitionRate(2,3) = q23, transitionRate(3,2) = 0
  transitionRate(2,4) = 0, transitionRate(4,2) = 0, transitionRate(2,5) = 0
  transitionRate(5,2) = 0, transitionRate(2,6) = 0, transitionRate(6,2) = 0
  transitionRate(3,4) = q34, transitionRate(4,3) = 0, transitionRate(3,5) = q35
  transitionRate(5,3) = 0, transitionRate(3,6) = q36, transitionRate(6,3) = 0
  transitionRate(4,5) = 0, transitionRate(5,4) = 0, transitionRate(4,6) = 0
  transitionRate(6,4) = 0, transitionRate(5,6) = 0, transitionRate(6,5) = 0
}

OUTPUT:
output=State
```

Figure B.1: Code for the 6-states CTMC structural model.

# List of Figures

# List of Tables

# Acknowledgements

I would like to thank all the people who helped me, supported me, and lived with me part of this challenging academic path.

In the first place, thanks to my supervisors, Adam Darwich, Luca Marzano, and Jayanth Raghothama, to my external supervisor, Enrico Gianluca Caiani, and to my KTH examiner, Sebastiaan Meijer, for their outstanding support and guidance. Thanks also to Harsha Krishna for his priceless help with handling the server and to Maksims Kornevs, my degree project group supervisor, for his feedback and support throughout the writing process of this thesis.

Vorrei poi ringraziare la mia famiglia, che mi ha sempre supportato, non soltanto per quanto riguarda la mia istruzione, ed ha sempre creduto in me nonostante tutti gli ostacoli che si sono presentati sul mio cammino. Sono i familiari migliori che si possano desiderare, e sarò loro sempre grato per tutto ciò che hanno fatto per me.

Grazie a mio padre, Nicola, per avermi sempre esortato a puntare in alto, e per aver sempre fatto tutto ciò che era in suo potere per supportarmi ed aiutarmi ad inseguire i miei sogni. Per menzionare giusto un esempio, non dimenticherò mai le volte in cui mi ha portato a cena fuori perché aveva capito che avessi bisogno di parlare. In generale, senza di lui non avrei mai potuto scegliere di studiare al Politecnico di Milano e poi in Svezia al KTH, e probabilmente oggi la mia vita sarebbe stata molto diversa. Di conseguenza, ora non avrei di fronte a me un così ampio orizzonte per la mia vita futura e per la mia carriera.

Grazie a mia madre, Paola, per essersi sempre presa ottima cura di me, anche a 3000Km di distanza, e per aver fatto sempre del proprio meglio per assicurarsi la mia felicità. È ed è sempre stata lì per me, specialmente quanto ne avevo più bisogno, e nulla è mai riuscito ad impedirle di supportarmi in tutti i modi possibili. Per menzionare giusto un esempio, non dimenticherò mai tutte le ore che ha trascorso a preparare ed impacchettare cibi prelibati per poi spedirmeli e farmi sentire a casa pur vivendo lontano.

Grazie a mia sorella, Roberta, alla quale sono profondamente legato. Grazie alla sua dolcezza, al suo affetto, ed alla sua capacità di tirarmi su di morale. Sono davvero fortunato

ad avere una sorella come lei, che si dimostra sempre autentica, amorevole, fedele ai propri ideali, e felice di trascorrere del tempo con me.

Special thanks to my true love, Rebecka, with whom I desire to spend my entire existence. Thanks to her for having supported me in my darkest and most difficult times and for having introduced one kind of happiness into my life that is truly difficult to express with words. Thanks to her for making me feel loved and at home here in Sweden, despite its differences from my homeland, and for everything she does for me and shares with me every day. I still struggle to believe how lucky I am to have her in my life. Finally, thanks to her also for having taught me how to use LaTeX properly. Without her, this thesis would have looked much uglier and less professional than it does.

Grazie ai miei nonni Carla, Lina ed Umberto, ed un ringraziamento speciale a nonno Carmine, che purtroppo non è potuto vivere abbastanza a lungo da vedermi laureato ma continua a vivere nel mio cuore. Grazie a lui per avermi supportato e per essere stato la persona più orgogliosa sulla faccia della terra ogniqualvolta ho superato uno dei miei esami.

Among the grandparents who deserve to be thanked, I also consider Rebecka's grandmother, Katalin, as my own. Thanks to Großmama Katalin for having "adopted" me as her new grandson, for always supporting me, and for having helped me "survive" my last academic year. Thanks to her for always making me feel at home with her.

Grazie anche alle mie zie Annamaria, Damla, Rkia, e Tina, ai miei zii Alessandro, Antonio, Ferdinando, Sandro, e Sergio, ed ai miei cugini Amir, Banchialem, Carmine, Carol, Giovanni, Noemi, e Seble. Non ho parole per descrivere quanto io sia contento di averli, e non dimenticherò mai tutte le cose che han fatto per me.

I want to thank my dear friends Dario, Rolando, Francesco, Lorenzo, Daniel, Fabio, Sofia, Andy Bashforth, and all the friends from the game. Thanks to all of them for having always been there for me. There are so many memories that I share with them and that I would like to cite, but I would have to write a new thesis to fit them all. I could not desire better friends than them, and sometimes I wonder how they could stand my extraordinary long absence due to my university deadlines. Moreover, I still wonder how Rolando survived my phone calls at 3 a.m. during our first year of university.

Special thanks go then to Alessandra. She is not just a friend but rather a dear sister. Thanks to her for having endured knowing me for almost six years, it must have been tough! No matter the difference in our life choices, even when I was studying in Sweden and she was doing so in the United States, she was always there for me and never let me

down. She always gave me the strength I needed not to give up during my years at Polimi, no matter how tough those were. I also still wonder how we could successfully prepare together for the second part of the exam in biomechanics (and some other courses) in just a few days...

Grazie a zio Andrea e zia Michela, zio Daniele e zia Mariarosaria, e Nino Stabile e Tiziana, per essersi sempre presi cura di me come fossi loro figlio. In particolare, grazie a zio Andrea per le notti trascorse con me a studiare, e grazie a Nino per aver salvato la mia sessione d'esame prestandomi il suo portatile lo scorso anno.

Un ringraziamento speciale va a zio Michele e zia Enza. Grazie per il loro affetto, per il loro supporto, e per essere entrati a far parte della famiglia. Grazie a zio Michele per essere stato il primo a raggiungermi in ospedale quando sono stato ricoverato nel 2018, nonostante il suo terrore per gli ospedali.

Grazie a Domenico Coseglia e Teresa Vacchiano per tutto ciò che hanno fatto per me in questi anni, e grazie a Stefania Visconti per avermi supportato in modo straordinario verso la fine del mio percorso liceale. Grazie al professore Francesco Palo per avermi fatto capire l'importanza di imparare l'inglese e per avermi dato solide basi in inglese accademico, e grazie a tutti gli altri straordinari docenti che mi hanno aiutato a diventare chi sono oggi.

Grazie a Lazzaro Lenza, Annamaria Esposito, Luciano Santimone, Donato Scotillo, Giancarlo Giolitto, Elisa Di Iaconi, e Matteo De Caro, per l'inestimabile aiuto e supporto che mi hanno fornito in questi ultimi anni.

Grazie anche a tutti gli altri amici di famiglia, i quali son diventati negli anni anche cari amici miei. Ogniqualvolta trascorro del tempo al sud, la loro piacevole e ristoratrice compagnia mi fa sempre moltissimo piacere. Sono troppi per poterli elencare uno ad uno, ma il mio apprezzamento e la mia gratitudine nei loro confronti provengono dal profondo del mio cuore.

Thanks also to my roommate Nicolas, or "The Frenchman", the best and friendliest roommate I have ever had. If I lived comfortably in Sweden for two years and completed my studies, it is also thanks to him and his kindness.

Grazie a Davide, il mio carissimo barista. Se sono riuscito a sopravvivere alla Laurea Triennale ed al primo anno di Laurea Magistrale è anche grazie al suo "bicchiere con otto espressi" ed ai suoi santini. Grazie a lui per essersi preso così tanta cura di me quando vivevo a Milano.

Finally, thanks to all the people I may have forgotten to mention! I hope they will not be too mad at me for not being included.