NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF MATHEMATICS

DULCE DA SILVA LOUREIRO

BSc in Mathematics

EXPLAINABLE MACHINE LEARNING FOR HEALTHCARE COST OPTIMIZATION

A TIME-DRIVEN APPROACH

MASTER IN MATHEMATICS AND APPLICATIONS

SPECIALIZATION IN DATA AND DECISION SCIENCES

NOVA University Lisbon

September, 2025

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF
MATHEMATICS

EXPLAINABLE MACHINE LEARNING FOR HEALTHCARE COST OPTIMIZATION

A TIME-DRIVEN APPROACH

DULCE DA SILVA LOUREIRO

BSc in Mathematics

Advisers: Maria Isabel Azevedo Rodrigues Gomes
Associate Professor, NOVA University Lisbon
Salomé Guedes Sequeira de Pádua Azevedo
Head of Digital Health, Value for Health CoLAB

Examination Committee

Chair: Manuel Valdemar Cabral Vieira
Assistant Professor, NOVA University Lisbon

Rapporteur: Tânia Rute Xavier de Matos Pinto Varela
Associate Professor, Instituto Superior Técnico, UL

Adviser: Maria Isabel Azevedo Rodrigues Gomes
Associate Professor, NOVA University Lisbon

MASTER IN MATHEMATICS AND APPLICATIONS
SPECIALIZATION IN DATA AND DECISION SCIENCES

NOVA University Lisbon
September, 2025

EXPLAINABLE MACHINE LEARNING FOR HEALTHCARE COST OPTIMIZATION
A TIME-DRIVEN APPROACH

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.3.7) [23].

Acknowledgements

This thesis marks the end of a five-year journey, one often accompanied by people reminding me of how challenging it must be to study Mathematics. The truth is, nothing feels that hard when you truly love it. For that reason, my first thanks goes to Mathematics itself, for being both my greatest passion and, at times, my biggest headache throughout these years.

Even though a dissertation is ultimately one's own work, this would not have been possible without the guidance and encouragement of my advisers. Thank you to Professor Isabel for all the ideas and for showing me how fun it can be to search for a solution, and to Dr. Salomé for pushing me out of my comfort zone and somehow making five-minute meetings last me an entire month. I also want to thank the VOH CoLAB team for all the help and support, for always having time for my questions, and for showing me what real teamwork looks like.

Moving to Lisbon for my master's was definitely a challenge, one I couldn't have overcome without the help of my family. Thank you to my mom and dad for always encouraging me to follow my dreams, and to my sister for making me miss even the times you annoyed me. To my grandma, for insisting I take as much food as I could fit in my bag, and to the rest of my family for always being my biggest cheerleaders.

I could never have made it through these academic years without my friends. Thank you to everyone who stood by me and listened to me explain my thesis, even if you didn't understand a word of it. A special thank you to my beloved GJ, for being my second home, and to my math girls, Megs and Inês, for pushing through this year with me and making sure I didn't procrastinate during our long hours in the library. Finally, to my safe harbor, my light, and my love, thank you, Zé, for being the voice I needed to hear and my strength when I had none.

Abstract

Efficient cost and resource allocation in healthcare is essential for the sustainability of hospital operations and patient-centered initiatives. However, this can be a complex issue due to the extensive scope of work and the difficulty in maintaining economic models designed for these evaluations. In recent years, machine learning (ML) has been increasingly adopted to support healthcare decision-making, but despite its predictive power, a major limitation remains: the lack of interpretability in many models, which hinders trust and usability by medical personnel.

In this study, we use the electronic medical records of 2800 cardiothoracic surgery patients of Santa Marta's Hospital and propose a novel approach that integrates ML with mathematical optimization to provide interpretable insights for healthcare cost analysis. Our methodology incorporates a set covering model within a clustering algorithm to identify representative patient cohorts, addressing the explainability gap in current ML approaches. Subsequently, we apply the Time-Driven Activity-Based Costing model to estimate the cost of each patient type by mapping clinical activities to time-based resource consumption.

The results show that patients can be grouped into meaningful cohorts that share clinical and resource-use characteristics, while still showing differences in costs. This allows for comparison of care pathways across patient types and provides cost estimates that capture both similarities and variations within the population.

To further demonstrate the adaptability of our framework, the clustering model is applied to another case study regarding Blockchain technology, highlighting its ability to extract interpretable patterns and optimize decision-making in domains beyond healthcare. By combining optimization techniques with interpretable machine learning, our approach provides a transparent framework for complex cost and resource analysis. This supports more informed decision-making, aligns with the growing demand for explainable ML, and enhances communication between technical and domain-specific stakeholders.

Keywords: Machine Learning, Optimization, Data-Driven Explainability, Cost Analysis, Time-Driven Activity-Based Costing

Resumo

No setor da saúde, é fundamental compreender de forma clara a alocação dos custos e dos recursos, de modo a garantir a viabilidade dos hospitais e dos projetos desenvolvidos para os seus doentes. No entanto, esta tarefa pode tornar-se uma questão complexa e desafiante, devido ao extenso volume de trabalho e à dificuldade em manter modelos económicos concebidos para este tipo de avaliações. Nos últimos anos, a aprendizagem automática tem sido cada vez mais adotada para apoiar a tomada de decisão no setor da saúde. Contudo, apesar do seu poder preditivo, uma limitação importante mantém-se: a falta de interpretabilidade de muitos modelos, o que compromete a confiança e a aplicabilidade por parte dos profissionais de saúde.

Neste trabalho, utilizamos os registos médicos eletrónicos de 2800 doentes submetidos a cirurgia cardiotorácica no Hospital de Santa Marta e propomos uma nova abordagem que integra aprendizagem automática com otimização para fornecer insights interpretáveis na análise de custos em saúde. A metodologia proposta baseia-se na utilização de conjuntos de coberturas incorporados num algoritmo de agrupamento para identificar coortes representativas de doentes, colmatando a lacuna de explicabilidade presente nas abordagens atuais de aprendizagem automática. De seguida, aplicamos o modelo de custeio Time-Driven Activity-Based Costing baseado em atividades e tempo para estimar o custo de cada tipologia de doente, mapeando atividades clínicas ao consumo de recursos com base no tempo.

Os resultados mostram que os doentes podem ser agrupados em grupos significativos que partilham características clínicas e de utilização de recursos, mantendo ainda diferenças nos custos. Isto permite a comparação dos percursos de cuidados entre diferentes tipos de doentes e fornece estimativas de custo que refletem tanto as semelhanças como as variações dentro da população.

Para evidenciar ainda mais a adaptabilidade da nossa metodologia, o modelo de agrupamento também é aplicado a outro caso de estudo relacionado com a tecnologia Blockchain, destacando sua capacidade de extrair padrões interpretáveis e otimizar o processo de tomada de decisão em domínios para além da área da saúde. Ao combinar técnicas de otimização com aprendizagem automática interpretável, a nossa abordagem fornece

uma estrutura transparente para a análise de custos em saúde, apoiando uma tomada de decisão mais informada, alinhando-se com a crescente exigência de explicabilidade em aprendizagem automática na área da saúde e facilitando uma melhor comunicação entre os intervenientes técnicos e clínicos.

Palavras-chave: Aprendizagem automática, Otimização, Explicabilidade baseada em dados, Análise de Custo, Time-Driven Activity-Based Costing

Contents

List of Figures

List of Tables

Acronyms

ABC Activity-Based Costing *(pp. 7, 16)*

CCR Cost Capacity Rate *(pp. 17, 18, 28, 39)*

PCA Principal Component Analysis *(pp. 5, 11)*

RCC Ratio of Cost to Charge *(p. 7)*
RVU Relative Value Unit *(p. 7)*

TDABC Time-Driven Activity-Based Costing *(pp. 2–4, 7–9, 16, 26, 32, 33, 39, 40, 45)*

1 Introduction

1.1 Context

In the context of public healthcare, the allocation and utilization of resources represent a significant factor in the analysis of costs [31]. When managing a large organization, limited resources depict difficult challenges that must be handled with great attention and care. This thesis considers this problem by proposing a method to evaluate and manage the costs and resources related to a patient subjected to cardiac surgery. The work is inserted into the CardioFollow.AI project, a research initiative aimed at designing and evaluating remote health interventions for patients subjected to cardiothoracic surgery. The intervention combines wearable sensors, a mobile application, and AI-driven alerts to support continuous monitoring and personalized follow-up.

Given this, it is necessary to comprehend and improve the management of the services provided, in a way to control and reduce the costs associated with those procedures. Although hospitals are not a typical business and profit is not the primary goal, accurate cost tracking is essential for long-term sustainability. Understanding the true cost of each service allows for better resource allocation and identifying areas for potential cost reduction [24].

Traditionally, healthcare institutions use a variety of top-down approaches that equally divide the assets across all departments. While straightforward, these methods usually result in under- or overestimation of the expense of certain practices [20, 24]. More recently, there has been an effort to implement cost analysis with a microcosting approach that is based on determining the price of each procedure and combining them to make a budget for each department. This new technique provides a more comprehensive insight into the distribution of expenditure and facilitates a more informed decision-making process when it comes to conducting the distribution of resources [20].

A key component of cost analysis in healthcare is the financial implications of each patient's hospitalization, which are shaped by the duration of their stay [29]. This is influenced by the sequence of events that occur from the moment of admission until the point of medical discharge. Therefore, it is important to outline the care pathway that

defines the trajectory of a patient enrolled in the CardioFollow.AI program.

As defined by Schrijvers (2012), the term "care pathway" is a “complex intervention for the mutual decision-making and organization of care processes for a well-defined group of patients during a well-defined period” [29]. In other words, a care pathway is a structured, multifaceted approach to healthcare delivery that coordinates decision-making and organizes care processes. It is designed for a specific patient group and implemented over a defined timeframe, aiming to improve treatment outcomes and resource utilization through collaborative planning and execution.

This concept, aligned with the microcosting approach, leads to an innovative cost model used to estimate each patient’s expenses based on their hospital journey called Time-Driven Activity-Based Costing (TDABC). TDABC is a technique widely used both in healthcare and other sectors such as in the scientific, manufacturing and food industries [21]. Specifically in care delivery, this method requires a detailed account of the patient’s journey within the hospital facilities, including their interactions with different services, as well as the monitoring that continues at home after discharge [14]. This can lead to an increased complexity in estimating the costs for each individual.

To combat this issue, this thesis combines TDABC with a clustering method to direct patients into meaningful model cohorts. This means that the algorithm is going to analyze the patients and, based on their common characteristics, create a representative patient for each cluster. Then it returns the forecasted expense of that patient’s treatment. This approach allows for a more accurate and actionable cost estimation, tailored to the specific needs of different patient groups.

By applying these tools, healthcare facilities can refine their financial management to enhance the allocation of resources and improve patient outcomes and satisfaction.

1.2 Objectives

The main goal of this work is to implement a cost estimation model using a novel framework that combines Time-Driven Activity-Based Costing with data-driven clustering of individuals based on their care pathways. This approach allows us to estimate the cost associated with each individual by assigning them to a representative cohort given their trajectory through a system or service. As a result, it supports a better understanding and management of resource allocation. The framework is also designed to be accessible to frontline professionals, enabling them to use it for real-time cost estimations. The specific objectives for this research include:

1. Aggregating the individuals into clusters with machine learning and optimization techniques that can be interpreted by medical professionals in a way that they can connect each patient to their cluster;

2. Implementing a TDABC model that can estimate the expense of a patient to the hospital given their care pathway;

3. Creating an easily interpretable model to facilitate the management of costs and the placement of funds and personnel in a hospital.

In the context of this thesis, we applied the methodology to data from a hospital setting, specifically in the context of cardiac surgery, and additionally to a Blockchain case study, highlighting its potential beyond healthcare.

2

Literature Review

This chapter is divided into two sections. Firstly, we explore the studies made with the objective of clustering individuals, both with black box models and by integrating machine learning and mathematical optimization. Then, we look into cost models, more specifically the Time-Driven Activity-Based Costing method, and how it performs compared to other frameworks.

2.1 Patient Clustering

Patient clustering is based on the idea of aggregating similar individuals with resembling paths. The objective of clustering algorithms in healthcare is to identify subpopulations with shared patterns, such as treatment response and healthcare utilization. By standardizing the trajectory followed by a specific group of individuals, patient clustering can personalize care delivery, estimate healthcare costs at the group level, and design targeted interventions that address the specific needs of each cluster. This approach is more efficient than managing patients individually, allowing for better resource allocation and more consistent, data-driven clinical decisions across similar cases.

2.1.1 Black Box Models

Traditionally, clustering is a machine learning approach to analyze patterns and behaviors of individuals, therefore, it is often described as a "black box" model. This means that the decisions made by the model are often given without any explanation, and the division of the individuals by the clusters is not interpretable [7]. In the healthcare context, this lack of transparency poses a significant barrier to model adoption [15]. Clinicians are less likely to trust or act on insights that they cannot interpret or validate through clinical reasoning. This constitutes a challenge in the grouping being endorsed by medical personnel, since the results are usually not intuitive for them.

Henao et al. [19] used electronic medical records with uncoded text to perform patient clustering. The methodology relies on the use of a probabilistic model called the Chinese Restaurant process, which consists of a distribution on the space of partitions of

$\mathbb{N}$. The framework functions as a restaurant, where customers (data points) choose tables (clusters) based on probabilities. The first instance starts in the initial cluster, and the following ones either create a new cluster with probability proportional to a parameter or join an existing cluster with probability proportional to the distance between the new point and the cluster. The results presented are relevant for the understanding of patient journeys, but the algorithm makes certain decisions for the cluster distribution that are not comprehensible at first glance. Later, the authors discovered that the model was not only grouping synonyms, but it was also putting pain locations together by medical condition. While this shows effectiveness in the model, it is not clear to doctors how these decisions are made, and it can constitute an obstacle to implementation.

Likewise, Simanjuntak et al. [30] resorted to patient clustering to study healthcare insurance benefits. They utilized insurance data to foresee patients' journeys based on their characteristics. The use of unsupervised learning, in this case, k-means clustering, makes the decision-making process not evident to the medical personnel. Furthermore, the application of Principal Component Analysis (PCA) for dimensionality reduction removes the intrinsic value of the treated variables, since it creates them from a linear combination of the input variables. Consequently, the reduced dataset has no practical meaning that can be interpreted within the context of the problem.

This specific case of patients in the CardioFollow project has been studied through machine learning algorithms by Laranja (2023). The models used were the k-means clustering, the BIRCH algorithm, the k-mode clustering, and the k-prototype clustering. All were implemented successfully, with good results, and positively rated by evaluation measures [22]. However, when the clusters were presented to the medical team, the results were met with skepticism, as they did not understand the model and did not trust its decisions.

In the context of machine learning models, there is a proven efficiency in identifying optimal solutions for complex problems within a multitude of fields [1, 13, 15]. However, in procedures where humans are heavily involved in the decision-making process, such as in healthcare, the need for explainable and interpretable solutions rises significantly. The lack of a concise thought process that can be interpreted and followed by physicians severely limits the applicability of these models in everyday hospitals. That is because the doctors do not have confidence in the model and its judgment [3].

In particular, when employing cluster analysis, it is crucial to facilitate interpretability by providing accurate and comprehensive explanations for the generated clusters [11].

Typically, there are two types of optimization approaches for clustering: those made in post-hoc settings and the ones that integrate optimization in the model from the beginning [17]. To our study, the latter are more relevant because they allow for understandable decisions and clear steps along the cluster implementation [3].

2.1.2 Machine Learning and Mathematical Optimization

The incorporation of mathematical optimization into machine learning models has recently been pursued with the objective of facilitating their interpretation and explanation. In light of these advances, there have been numerous efforts to integrate these two fields in the resolution of real-world problems. Some of the approaches seen in recent studies include the utilization of counterfactual explanations, which describe how input features should change to alter a model's prediction [12], observing the task as a data visualization problem [10], the use of set covers to do feature selection in data mining [25] and defining the clusters as a set of rules that each individual is required to adhere to [11].

When using group counterfactual explanations to better understand the outcomes of machine learning models [12], the focus shifts from a conventional approach of generating a single counterfactual for an individual instance to the generation of a group of counterfactuals for a group of instances. Therefore, we look for a collection of reasons for the possible alteration of the model's outcome and how they relate, rather than focusing on a single explanation. It is also relevant to observe how they relate to the initial entry and if they can apply to different instances. This technique is particularly pertinent due to its stakeholder-oriented and highly intuitive nature.

The problem might also be looked at from the perspective of data visualization [10]. The method relies on formulating the problem in a global optimization context where the objective function is a difference between two convex functions, individuals are represented as geometric objects whose volumes are proportional to a specified statistical value and whose positions reflect a dissimilarity measure. Ultimately, the objective is to have a visualization map that can be utilized by researchers to better interpret the data they are working with.

Furthermore, mathematical optimization and machine learning can be integrated through the Generalized Set Covering problem (gSC) [25]. Zhengyu Ma and Hong Seo Ryoo proposed a framework that focuses on feature selection, which is of vital importance for classification models in data mining. This perspective is based on looking at the situation as a combinatorial optimization problem where the goal is to minimize the cost associated with certain features, subject to a set of constraints. To solve the gSC problem, the authors present two procedures, a Surrogate and a Lagrangian relaxation. This research shows the importance of effective feature selection in data mining. By using optimization, the authors addressed the limitations of existing methods and proposed an enhanced technique that improves the performance of data mining algorithms.

Finally, the model that is presented in this work is based on the approach proposed by Emilio Carrizosa and Kseniia Kurishchenko [11], which enhances interpretability by providing clear rule-based explanations for cluster generation. The framework simultaneously allocates individuals to clusters while providing rules that characterize each cluster, which is achieved by formulating it as a multi-objective Mixed-Integer Linear Programming (MILP) problem. This is particularly important in the context of healthcare

because it turns the black box model into a set of rules that can be incorporated into a decision-making process that is accessible and understandable to medical personnel [15]. The model requires the input of two elements: the set of rules to be used and the dissimilarity matrix that compares each pair of individuals. Consequently, it is able to determine the allocation of each individual and each rule to the clusters.

2.2 Time-Driven Activity-Based Costing (TDABC)

Accurately measuring the cost of healthcare services at the patient level remains a significant challenge, as traditional costing methods often lack precision or require excessive resources [2].

Regarding cost models, there is a vast amount of research conducted across a multitude of contexts and fields [16, 26]. However, many conventional approaches, such as the Ratio of Cost to Charge (RCC) and Relative Value Unit (RVU) methods, have significant limitations [20].

The RCC is a top-down approach that divides the total expenses for the year by the total annual charges for the services provided. While this method allows for an overview of costs relative to charges, it does not grant an estimation of costs for a specified patient or procedure. As for the RVU, it is a bottom-up technique since it gives cost predictions for a service. Theoretically, the Relative Value Unit delivers accurate projections, but in reality, the method exhibits a lack of precision.

The Time-Driven Activity-Based Costing (TDABC) methodology, presented by Kaplan and Anderson as a bottom-up approach for cost analysis [21], emerges as an alternative that overcomes these challenges.

The TDABC model is a variation of the Activity-Based Costing (ABC) system. The ABC framework consists of identifying the activities required for a specified service or product and then estimating the costs associated with each activity to produce the price of the service or product itself.

A number of studies have been conducted to compare TDABC with other more traditional methods [2, 20, 21]. These studies conclude that TDABC is an easily comprehensible and implementable approach. This is due to the fact that this framework only requires the time required by the activity and the unit cost of supplying capacity, i.e., the cost associated with maintaining the service for a single time unit (for example, the cost of a surgeon per minute) [14].

When compared to the Activity-Based Costing system, which is considerably more time- and resource-consuming, TDABC performed better estimations of cost than ABC, and it was easier to identify the processes and procedures that have elevated costs or variability [2]. Moreover, adjustments to the model can be swiftly implemented, whereas the ABC model requires reestimating all the necessary parameters, thereby making model maintenance challenging [21].

By combining ease of implementation with accurate, patient-level cost estimates, TD-ABC aligns with the principles of value-based healthcare and effective resource allocation.

3

Methodology

The methodology is divided into five parts. Initially, a pre-processing and exploratory analysis of the data is conducted in order to evaluate which models and techniques are to be applied. Then, individuals are grouped based on similarities in their paths and profiles via clustering. Then, each one is assigned a group, i.e., a model cohort that serves as the representative for all members. For each of these groups, a care pathway that describes their interaction with the services is then matched. Subsequently, a cost analysis is performed for each type of trajectory using the model cohorts, which are defined by the centroids. The framework proposed in this thesis, described by Figure 3.1, offers significant potential for application in various contexts. Its primary aim is to simplify and enhance the cost estimation process in populations made up of individuals with different characteristics and needs. By using explainable clustering models, it becomes possible to define meaningful groups (in this case, of patients) and to estimate costs by assigning each group to a shared care trajectory under the Time-Driven Activity-Based Costing (TDABC) model.

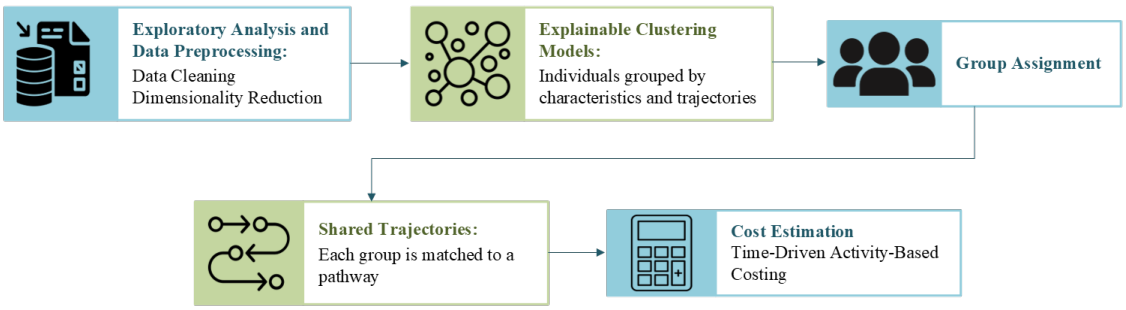


Figure 3.1: Proposed Framework Overview.

3.1 Exploratory Analysis

When working with real-world data, the facility requesting the analysis typically provides either structured or unstructured data. In healthcare settings such as hospitals, data is often unstructured, with high variability and numerous variable types, particularly text entries in clinical records [27]. This lack of standardization presents significant challenges and requires extensive preprocessing to prepare the dataset for analysis. In contrast, more structured environments, such as startups or businesses, usually rely on dedicated software systems for data collection and management. As a result, they tend to provide structured datasets that are easier to process and analyze. This subsection comprises some of the recommended techniques for dealing with unstructured and missing data.

The first obstacle that can be found is the inconsistency of the record constitution, which often leads to an extensive pre-processing of the data. This involves converting "Yes" and "No" into binary variables, applying direct rating techniques to ordinal variables, comparing categorical variables that appear to have the same information to confirm if there were any discrepancies, and standardizing text entries to be able to aggregate them later on.

The following step is to group the categorical text values so that they express significance, i.e., in order to reduce the number of outcomes of each variable, we aggregate them by similarity. To do this, the idea of concept hierarchies can be implemented [6]. Concept hierarchies are based on the principle that several variable outcomes can be generalized into a concept that presents the characteristics of every object within it, as shown in Figure 3.2.

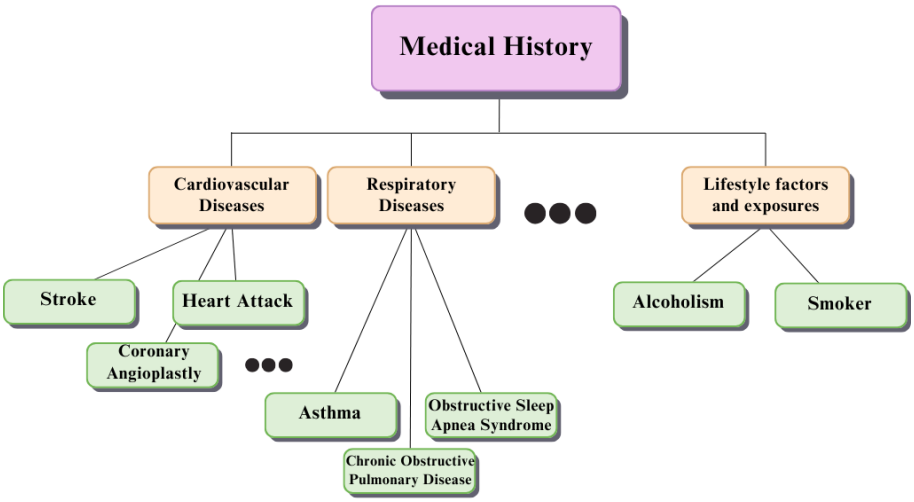


Figure 3.2: Example of a concept hierarchy for the variable Medical History.

3.1.1 Missing data treatment

The dimensionality reduction methods often require the correlation matrix of the dataset, which, in return, requires the resolution of issues such as the presence of categorical variables and missing data. For variables with categorical outcomes, dummy variables can be created.

When handling missing data, it is important to consider the nature and meaning of each variable. Rather than relying solely on generic imputation strategies, one can adopt a context-aware approach that accounts for logical relationships within the data. In particular, missing values may result from the values of related variables. For example, a variable may be empty because another linked variable is set to zero, such as when the number of sessions is not recorded because the individual did not receive that type of service. In such cases, missing values can be safely replaced with zero, as this reflects the underlying data logic.

For variables that are not typically subject to significant changes, missing values can be addressed by selecting the most recent or the next available value for that individual. This approach assumes temporal stability and helps preserve the internal consistency of the data. Additionally, for response variables that are recorded across multiple time periods, mean imputation can be applied. This is done by filling in missing values using the average of the individuals' responses for that variable in the other available periods. This method promotes temporal coherence and ensures consistency over time.

To ensure meaningfulness in the data, variables with a consistently high rate of non-response, such as entire questionnaire sections, can be excluded from the dataset. Additionally, it is recommended to assess the percentage of missing data both per variable and per individual. Participants with minimal contribution to the dataset can be identified as "non-responsive" and excluded from the analysis. After removing such cases, if the remaining variables exhibit missing data rates below an acceptable threshold, mean imputation can be applied to handle residual missing values and improve dataset completeness.

3.1.2 Dimensionality Reduction

Following the treatment of the dataset, it may be necessary to reduce the dimensionality. To accomplish this objective, the R package `subselect`, developed by Cerdeira et al., is recommended for employment [28].

The choice of this method in place of Principal Component Analysis (PCA) is made in light of the objective of model interpretability. While commonly used in circumstances similar to these, PCA poses a difficulty since it performs linear combinations of the original variables [9]. As a consequence, it is challenging to interpret the modified dataset in terms of the sense of each variable. Additionally, PCA does not provide actual dimensionality reduction in the inputed data since it still requires the original dataset to perform.

In light of these considerations, this work utilizes the `subselect` package that incorporates three different search algorithms for larger datasets: `anneal`, `genetic`, and `improve`. For this work, the `improve` algorithm could be employed due to its explorative nature.

The `improve` algorithm searches for an optimal subset by iteratively replacing variables in the current set and comparing their evaluation criteria. This method allows the user to force the inclusion or exclusion of specific variables. This is particularly relevant to ensure flexibility in the method and allow the retention of variables considered essential by domain experts.

To determine the subset's quality, the package relies on the comparison of performance coefficients that measure the quality of the subset, the RM (Matrix Correlation) coefficient, the Generalized Coefficient of Determination (GCD) and the RV coefficient. After experimenting with all three, we decided that the GCD coefficient was the most appropriate for this work.

The Generalized Coefficient of Determination (GCD) measures the closeness of two subspaces: the subspace spanned by a subset of variables and the subspace spanned by a set of Principal Components from the full dataset. A higher GCD value indicates that the subset of variables is more closely aligned with the principal components, suggesting that it captures the essential structure of the data.

Before applying dimensionality reduction techniques, it is recommended to perform a correlation analysis to identify redundancy among variables. Highly correlated variables, based on established thresholds, such as those suggested by Asuero (2006) [4], can be identified for potential removal. This step helps to reduce multicollinearity and improves the effectiveness of subsequent variable selection methods.

Prior to applying the subselect method, it is essential to compute a valid correlation matrix. As a first step, columns with zero standard deviation should be removed, as they lack variability and do not contribute meaningful information. To ensure that the resulting correlation matrix is positive definite and suitable for further analysis, functions such as `cor.shrink` from the `corpor` package can be used. This process helps guarantee the validity of the matrix and supports reliable dimensionality reduction.

Our recommended approach to dimensionality reduction is to experiment with a wide range of subset variable sizes to evaluate the behavior of the algorithm. Comparing the performance coefficients for all possible sizes and selecting the one that represents a reasonable trade-off between dimensionality and information retention can be a good strategy. Analyzing the selected variables can reveal patterns, such as a preference for intervention-related variables or repeated measurements across time periods. Insights from this preliminary analysis can inform more targeted strategies, such as performing dimensionality reduction separately for each time period or comparing the selection outcomes with and without the inclusion of intervention-related variables.

It is also important to consider the computational implications of the final dataset. In high-dimensional settings, optimization models, especially those involving combinatorial components like Mixed-Integer Linear Programming (MILP), may become intractable

due to memory or processing limitations. When such issues arise, it may be necessary to further reduce the dataset's dimensionality.

Two possible strategies can be employed in this context. One approach is to identify and retain only the most frequently selected variables across different periods or subsets, as these are likely to hold consistent relevance. Alternatively, the dimensionality reduction process can be repeated with additional constraints, such as forcing the inclusion of essential variables (e.g., demographic or intervention-related features) while limiting the total number of selected features. These adjustments allow for the creation of a more compact and computationally feasible dataset while still preserving key information for analysis. The choice between these approaches should be made by comparing their performance coefficients.

3.1.3 Dataset Sampling

In cases where the full dataset presents computational challenges, particularly due to high dimensionality and memory issues, a sampling strategy can be employed to reduce the dataset size while preserving its analytical value. A representative subsample of manageable size should be selected to maintain sufficient diversity for modeling purposes. To ensure robustness and reproducibility, multiple independent random samples can be drawn using fixed random seeds. This approach balances computational feasibility with the need for reliable and interpretable results.

3.2 Clustering Method

The chosen model for this thesis was developed for clustering individuals by associating rules to instances and groups. The framework was designed by Emilio Carrizosa and Kseniia Kurishchenko [11]. To apply this model, we need to have the dissimilarity between every pair of individuals, δ_{ij}. Techniques such as the Euclidean and Gower distances are suitable candidates for consideration [18]. In addition, we need to construct η, a collection of N if-then rules. In some cases, the rule is going to direct the variable into a specific value (e.g., categorical variables). In other cases, a limit is going to be imposed on the variable's range (e.g., rules associated with numerical variables such as "Age"). The collection η should be divided into S groups where $\eta = \bigcup_{s=1}^{S} \eta_s$ and $\eta_s \cap \eta_{s'} = \emptyset$, if $s \neq s'$. Usually, each group η_s corresponds to a single variable, and the rules can, in some cases, direct the variable into a specific value (e.g., categorical variables, like "Sex=F" and "History=Tumours"). In other cases, a limit is going to be imposed on the variable's range (e.g., rules associated with numerical variables such as "Age"). The construction of these rules is explained in detail in Subsections 4.1.2.1 and 4.2.2.1. The explanations of a cluster are defined as the combination of rules using the AND operator so that it never exceeds the maximum ℓ. Also, only one rule out of each set η_s should be associated with one cluster, since these are mutually exclusive.

Given this, the clustering model is described as follows:

Indices and sets

$k \in \{1, \ldots, K\}$ for clusters
$i, j \in \{1, \ldots, I\} = \mathcal{I}$ for individuals
$s \in \{1, \ldots, S\}$ for groups of rules
$n \in \{1, \ldots, N\} = \eta = \bigcup_{s=1}^{S} \eta_s : \eta_s \cap \eta_{s'} = \emptyset$ for rules

Data

δ	Matrix of dissimilarities δ_{ij} between each pair of individuals i and j
b_{isn}	$= \begin{cases} 1, & \text{if individual } i \text{ is explained by rule } n \in \eta_s, \\ 0, & \text{otherwise.} \end{cases}$
$\theta_1 \geq 0$	Weight for true positive cases across the K clusters,
$\theta_2 \geq 0$	Weight for false positive cases across the K clusters,
ℓ	Maximum length of the clusters' explanations.

Decision variables

$$x_{ki} = \begin{cases} 1, & \text{if individual } i \text{ belong to cluster } k, \\ 0, & \text{otherwise.} \end{cases}$$

$$z_{ksn} = \begin{cases} 1, & \text{if rule } n \in \eta_s \text{ is chosen for cluster } k, \\ 0, & \text{otherwise.} \end{cases}$$

$$\alpha_i = \begin{cases} 1, & \text{if individual } i \text{ is a true positive case to the explanation assigned to its cluster,} \\ 0, & \text{otherwise.} \end{cases}$$

$$\beta_{ki} = \begin{cases} 1, & \text{if individual } i \text{ is outside cluster } k \text{ and is a false positive case to the explanation} \\ & \text{assigned to cluster } k, \\ 0, & \text{otherwise.} \end{cases}$$

Problem formulation

$$\min_{x,z,\alpha,\beta} \quad \sum_{k=1}^{K} \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \delta_{ij} x_{ki} x_{kj} \quad - \quad \theta_1 \sum_{i=1}^{I} \alpha_i \quad + \quad \theta_2 \sum_{k=1}^{K} \sum_{i=1}^{I} \beta_{ki} \tag{3.1}$$

s.t.

$$\sum_{k=1}^{K} x_{ki} = 1, \qquad i = 1, \ldots, I \tag{3.2}$$

$$\sum_{n \in \eta_S} z_{ksn} \leq 1, \qquad k = 1, \ldots, K \quad s = 1, \ldots, S \tag{3.3}$$

$$1 \leq \sum_{s=1}^{S} \sum_{s \in \eta_S} z_{ksn} \leq \ell, \qquad k = 1, \ldots, K \quad (3.4)$$

$$\alpha_i + x_{ki} + \sum_{n \in \eta_S} (1 - b_{isn}) z_{ksn} \leq 2, \qquad i = 1, \ldots, I \quad k = 1, \ldots, K \quad s = 1, \ldots, S \quad (3.5)$$

$$\beta_{ki} + x_{ki} + \sum_{s=1}^{S} \sum_{n \in \eta_S} (1 - b_{isn}) z_{ksn} \geq 1, \qquad i = 1, \ldots, I \quad k = 1, \ldots, K \quad (3.6)$$

$$x_{ki} \in \{0, 1\}, \qquad i = 1, \ldots, I \quad k = 1, \ldots, K \quad (3.7)$$

$$z_{ksn} \in \{0, 1\}, \qquad s = 1, \ldots, S \quad n \in \eta_S \quad k = 1, \ldots, K \quad (3.8)$$

$$\alpha_i \in [0, 1], \qquad i = 1, \ldots, I \quad (3.9)$$

$$\beta_{ki} \in [0, 1], \qquad i = 1, \ldots, I \quad k = 1, \ldots, K \quad (3.10)$$

The objective function, defined here by Equation 3.1, is divided into three terms. The first is to ensure intra-homogeneity as a result of minimizing the dissimilarity between each individual in each cluster by modifying the allocation of instances to groups. Then maximize the true positives connected to the decision variable α_i, and lastly minimize the false positives modeled to the variable β_{ki}. The last two terms are calculated with an associated weight that can vary to test the model's robustness.

The restrictions are designed to guarantee the model functions properly and the variables remain in domains that are meaningful and logical to the context. Consequently, restriction 3.2 assures that each individual belongs to only one cluster, restriction 3.3 ensures that for each cluster and each set of rules, one or no rule out of the group applies to the cluster. This constraint is reasonable because a set of rules is going to be constructed by limiting one or more variables. Restriction 3.4 verifies that the maximum length of the explanation for each cluster is respected. Restrictions 3.5 and 3.6 establish the good definition of α_i and β_{ki}, by ensuring that $\alpha_i = 0$ and $\beta_{ki} = 1$ are well-defined.

In the event of $\beta_{ki} = 1$, the individual can not belong to cluster k, therefore $x_{ki} = 0$. Then we have that for the set of rules associated with that cluster, $z_{ksn} = 1$, $b_{isn} = 1$, which means $\sum_{s=1}^{S} \sum_{n \in \eta_s} (1 - b_{isn}) z_{ksn} = 0$. Consequently $\beta_{ki} \geq 1$, along with the upper limit ensures that $\beta_{ki} = 1$.

Lastly, restrictions 3.7 and 3.8 are the domains of the variables x_{ki} and z_{ksn}. Likewise, restrictions 3.9 and 3.10 do the same for α_i and β_{ki}, and, without losing optimality, we can assume them to be continuous.

3.3 Time-Driven Activity-Based Costing (TDABC)

The cost analysis performed in this thesis uses the Time-Driven Activity-Based Costing model [21]. This framework originated from the Activity-Based Costing (ABC) model and is characterized by its simplicity and easy application. As explained in Chapter 2, the ABC model is extremely time- and resource-consuming since it requires employee surveys and interviews to be done periodically, making it impractical to maintain. To prevent this, TDABC is a more accurate and simpler approach that integrates time into cost estimation.

In order to implement a Time-Driven Activity-Based Costing model, two estimations are needed:

- **Unit cost of supplying capacity:** also called the capacity cost rate, this value measures the cost of providing an activity performed by a group of resources at their practical capacity.

- **Time required to perform an activity:** this measurement describes the number of unit times needed to complete a practice required for a product, service, and customer.

However, the application of TDABC can be heterogeneous, which influences the accuracy and limits the possibility of comparison between entities. Related to this, Da Silva Etges et al. proposed a framework for implementing this model, divided into eight steps [14]. This is the methodology that is going to be used in this work and is depicted by Figure 3.3.

Firstly, it is important to address the problem in question and identify the technologies that are going to be evaluated (Step 1). Knowing the purpose of the study helps to take a better approach, depending on what is being studied. Some perspectives are more relevant to external viewers, such as for evaluating methods. Others are refined to decision-making and internal management.

As mentioned in Chapter 1, the care pathway delineates the process of an individual in a specific institution. This concept relates to the care delivery value chain mentioned in Step 2. This information is relevant because it allows for the explicit definition of the individuals' trajectory and the depiction of the procedures that they interacted with.

Then it is time to estimate the expense of each main group and department (Steps 3 and 4 of the TDABC methodology). These costs are divided into two categories: direct and indirect. Materials, personnel, and equipment are part of the direct costs, along with electricity, rent, and others. Indirect costs include management and administrative logistics, like the human resources department and project coordination.

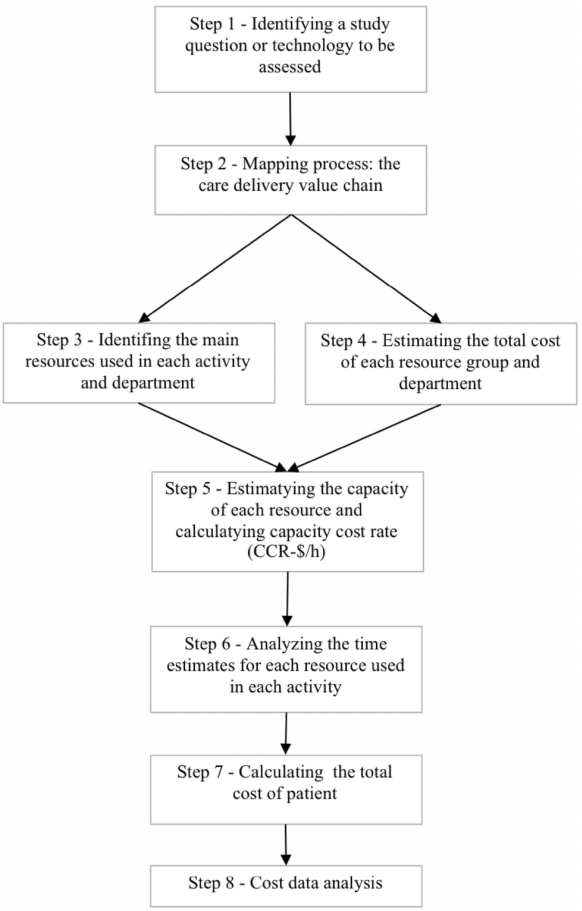


Figure 3.3: 8-step framework for implementing TDABC. Reproduced from [14].

Afterward, we calculate the Cost Capacity Rate (CCR) (Step 5) that represents the cost per unit time of performing a specific activity. This is calculated by dividing the expense of resources by the practical capacity.

$$CCR = \frac{\text{Cost of supplying capacity}}{\text{Practical capacity}}$$

Practical capacity refers to the amount of time when people and machines are working at full capability, which excludes employees' breaks, arrival and departure times, and other activities not related to meaningful work. For machines, maintenance, cleaning, and repair times should be accounted for. Theoretically, the practical capacity for an institution is estimated to be around 80% to 85% [21].

Next, we need to estimate the time each patient spends on an activity (Step 6). In order to obtain this information, we can use electronic medical records, such as timestamps indicating the entry and exit of patients for a department and hospital data systems. Another option is to do *in loco* observations to perform time predictions.

This way we can elaborate the total cost estimation of the individual, i.e., the expenditure of a participant from the beginning of their pathway to the end(Step 7). For that, we use Equation 3.11 that calculates the cost for each model cohort.

$$C_p = \sum_{act=1}^{A} \sum_{r=1}^{R} t_{act} CCR_r + y \tag{3.11}$$

C_p: Total cost of the model cohort p ($p = 1, \ldots, P$);
t_{act}: Time used by each resource for the activity act ($act = 1, \ldots, A$);
CCR_r: Cost Capacity Rate for resource r ($r = 1, \ldots, R$);
y: Other direct costs associated with the procedure (those that do not involve time)
R: number of resources;
A: number of activities in the pathway.

This way, we calculate the cost of each activity by first summing the time-dependent costs. These are obtained by multiplying the Cost Capacity Rate (CCR) of each required resource by the duration of the activity. Next, we add the time-independent costs, such as consumables and materials, which are represented here by y.

It is of interest to study the total cost of the individuals in the population. This total cost is calculated by multiplying the cost associated with each model cohort type by the number of participants that fit that model. This is given by Equation 3.12.

$$C_{total} = \sum_{p=1}^{P} C_p * N_p \tag{3.12}$$

C_p: Total cost of the model cohort p;
N_p: Number of patients that fit model cohort p;
P: number of model cohort types.

Consequently, we can discuss the results of the model and analyze them with the study question in mind (Step 8). Cost analysis should be performed by the use of tables and charts, and it is also relevant to identify departments with high unexpected expenditures and investigate the reasoning behind them. An evaluation of the resource allocation can be performed, and changes can be made to improve satisfaction and cost effectiveness.

4

Application of the Methodology

In this chapter, we provide a detailed description of how the proposed framework was applied to the case study of cardiac surgery patients from the CardioFollow.AI project. This application demonstrates the practical implementation of the methodology in a real-world healthcare context.

Notably, the same framework has also been applied in other initiatives by the Value for Health CoLAB, highlighting its adaptability and relevance across different domains. This chapter will also present one case study in a task related to a Blockchain project involving the Dotmoovs platform, which is a digital application that uses artificial intelligence and blockchain technology to gamify physical activity.

Together, these use cases highlight the flexibility and broad applicability of the proposed cost estimation approach, both within and beyond the healthcare sector.

4.1 Case Study: CardioFollow.AI

4.1.1 Exploratory Analysis

This thesis employs a dataset obtained from the electronic medical records of 2,789 cardiothoracic surgery patients of Hospital de Santa Marta. The ages of the individuals range from 18 to 88 years old, with a higher proportion of male subjects. The dataset was retrieved from the responses to a questionnaire divided into seven parts. The initial segment comprises the patient's personal information, while the later sections document the patient's well-being during the postoperative period. This period extends for one year, during which the patient is requested to respond to six questionnaires. These are to be completed at the following time points: preoperatively, three days post-surgery, thirty days post-surgery, three months post-surgery, six months post-surgery, and one year post-surgery.

Originally, there were 1,118 variables, but after the initial treatment described in Section 3.1, only 985 of them contained meaningful information (refer to Appendix A for a detailed description of the treated variables).

Next, we implemented the idea of concept hierarchies in order to ensure medical significance. To complete this step, it was essential to consult the clinical team involved in the CardioFollow.AI project. In this case, an interview was conducted with Dr. Jorge Pinheiro Santos, a Cardiac Surgery Consultant at Hospital de Santa Marta. During this meeting, the terminology was confirmed and standardized. Dr. Jorge provided assistance in finalizing several concept hierarchies and participated in a direct rating exercise to evaluate the importance of changes to certain variable outcomes.

In variables such as the ones regarding medical history and the type of medication taken by the patient, Dr. Jorge categorized the outcomes in general concepts, which was very important to extract meaning from those variables. For the type of medication, the questionnaire carried out by the nurses presented some issues because the tablets were divided by pharmacological class and were sometimes redundant. To overcome this, Dr. Jorge matched each outcome with the type of medication.

When looking over the variables regarding the medication adherence of each patient, Dr. Jorge referred us to the Morisky Medication Adherence Scale (MMAS) [8] that accurately represents the 6 variables in question. The objective of this technique is to assess how poorly patients adhere to their prescribed medication regimen by assigning a quantitative value based on their responses. Each question evaluates the degree of non-adherence by asking patients to match the frequency of their behavior (1 - Always, 2 - Almost always, 3 - Sometimes, 4 - Never). The score is then calculated by dividing the sum of said values by the number of questions. Consequently, higher scores correspond to patients demonstrating better medication adherence. The paper defines adherent patients as those with a score higher than the median, while non-adherent patients are defined as those with scores below the median.

4.1.1.1 Missing Data Treatment

To apply the selected dimensionality reduction method, we first needed to compute the correlation matrix of the dataset. This required addressing specific data preprocessing challenges, particularly the handling of categorical variables and missing data. Categorical variables were transformed into a numerical format by generating dummy variables. For example, features such as medical history and type of surgery were converted into binary ones.

As the first step to handle the missing data, we adopt a context-aware strategy that considers the meaning and logical structure of each variable, rather than relying solely on generic imputation methods. For instance, the variable representing the number of cigarettes smoked was missing for patients who did not smoke. In this case, the missing values were set to 0. For measurements such as weight and height, which are not expected to vary significantly over short periods, we selected the latest or next value registered for that patient. Similar to this approach, for the response variables present across all periods, the missing values were filled by inputting the average of the responses for the

corresponding variable in all other periods for that patient. This method is called mean imputation, and it was applied across all periods, which ensures consistency over time.

In the course of the analysis of the responses to the questionnaire, a consistent absence of participation in the Short Form Health Survey (SF-36) was observed. This section, which is designed to assess patients' health and well-being, comprises 36 items. The majority of patients provided no response to these questions, which led us to conclude that they provided no meaningful data and should therefore be removed entirely from the dataset.

Subsequently, an analysis was conducted to examine the percentage of missing data for each patient and each variable. This enabled the identification of a group of "non-responsive" patients, who contributed little to the overall dataset and were subsequently excluded. Following the removal of these patients, the remaining variables exhibited missing data rates of less than 20%. Consequently, we implemented a mean imputation technique to address the residual missing values, thereby ensuring the creation of a more comprehensive and reliable dataset for subsequent analysis.

4.1.1.2 Dimensionality Reduction

After the treatment of the dataset, a dimensionality reduction was performed using the R package `subselect`, developed by Cerdeira et al.[28].

Before implementing the subselect method, we studied the correlations between the variables in the treated dataset. This analysis showed that 105 variables presented high correlations to other variables as defined by Asuero (2006) [4]. Therefore, they were eliminated before the dimensionality reduction.

To apply the subselect method, we first needed to compute the correlation matrix. However, before this step, we identified and removed columns with a standard deviation of zero. Additionally, we used the `cor.shrink` from the `corpor` package to force the matrix to be positive definite. With this, our clean data had 595 variables (53% of the original variables) and 2706 patients (97% of the original patients).

Our first attempt at a dimensionality reduction was using the entire dataset, and we experimented with subset sizes ranging from 1 to 100. The GCD coefficients for this trial can be seen in Figure 4.1. From this analysis, we observed that the algorithm was selecting mainly the variables regarding interventions (educational reinforcement, anticipation of the consultation, etc) and sequential variables, i.e., the same measurement at different periods. These results guided us to experiment implementing the dimensionality reduction for each period of the questionnaire separately, and also to evaluate the behaviour of the algorithm with and without interventions.

Our approach to dimensionality reduction by periods was to divide the dataset into seven groups: patient characteristics, pre-operative questionnaire, 3-day questionnaire, 30-day questionnaire, 3-month questionnaire, 6-month questionnaire, and 1-year questionnaire. For each of those groups, we applied the algorithm across all possible subset sizes and selected the best size for each one by comparing the GCD coefficients. For instance,

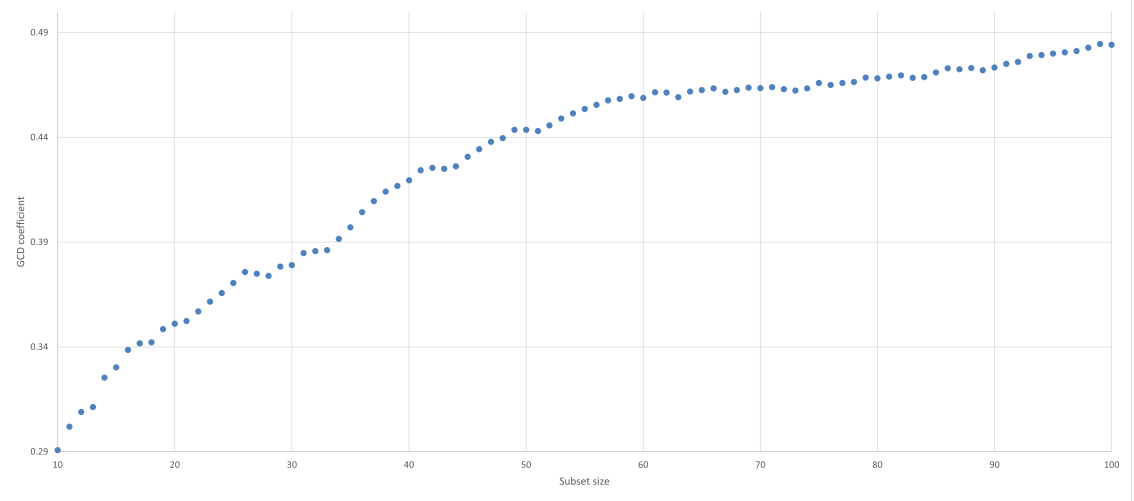


Figure 4.1: Generalized Coefficient of Determination (GCD) for different subset sizes when employing the dimensionality reduction for the whole dataset.

the data regarding the 3-day questionnaire had 124 variables, so we executed the `improve` command from the `subselect` package with kmin = 1 and kmax = 123. Afterwards, we observed the GCD coefficients (Figure 4.2) and selected the subset of size 8, as it exhibited a local maximum in the GCD graph and represented a reasonable trade-off between dimensionality and information retention. We applied this strategy to the data with and without interventions. The reduced data with interventions had several of the intervention variables, which led us to conclude that these were, as predicted, important to the dataset. It also chose the same variables for several periods, like Medication Adherence, Presence of Pain, and Consumption of Two Vegetables.

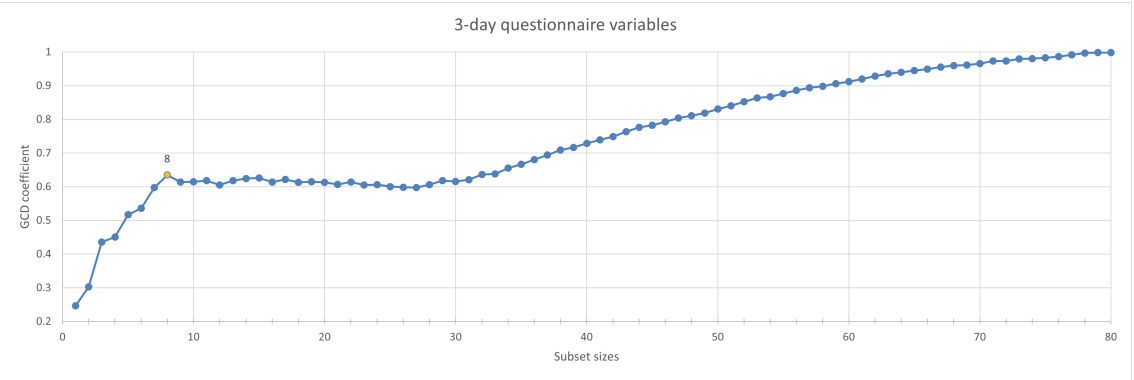


Figure 4.2: GCD coefficients for the 3-day subsets.

The dimensionality reduction for the data with no interventions selected four variables from the pre-operative period, eight from the 3-day questionnaire, twelve from the 30-day questionnaire, thirteen from the 3-month questionnaire, nine from the 6-month questionnaire, and eight from the 1-year questionnaire. The selection process for all seven periods can be found in Appendix B.

Ideally, we would keep all variables regarding patient characteristics like age, sex,

medical history, and medications, which added twenty more variables for a total of ninety-four variables. This would be our ideal dataset that represents the original with a GCD coefficient of 0.4286. However, the clustering model encountered issues when running on such a large volume of data, mainly because of the computational demands imposed by both the dataset's size and the complexity of the optimization problem. The model chosen was a Mixed-Integer Linear Programming (MILP) problem, which is known to be computationally intensive, especially as the number of variables and constraints increases. In our case, the model required the construction of a dissimilarity matrix and a rule compliance matrix. Given that each variable was associated with two rules, this resulted in a total of $94 \times 2 = 188$ rules per patient, leading to over 508,000 rule entries across the entire dataset. Furthermore, the dissimilarity matrix alone involved computing pairwise distances between all patients, resulting in over 3.6 million entries. This combination of high-dimensional data and complex constraints caused the model to exceed memory and processing limits, often failing to complete execution.

To address this issue, we attempted to reduce the dataset's size a second time. Initially, we examined the most frequently occurring variables resulting from the prior dimensionality reduction. When disregarding the period component, the variables "Medication Adherence", "Attempted to Quit Smoking", "Nurse Appointment", "Consumption of 2 Vegetables", "Presence of Pain", and "Blood Pressure Monitoring" appeared for more than three periods. To reduce the dimensionality of the dataset, we retained only those variables that consistently appeared across multiple periods, along with the intervention-related variables. This approach resulted in a subset of 30 variables and a GCD coefficient of 0.2348.

The second strategy consisted of reapplying the `improve` algorithm a second time. This time, we forced the inclusion of the 5 variables regarding interventions and the variables for age and sex. The remaining patient-specific variables were included in the dimensionality reduction. As in the previous analysis, we evaluated the results across all possible subset sizes and selected the one that offered a good balance between dimensionality and GCD coefficient. This method yielded a more compact subset of 21 variables and a GCD coefficient of 0.2823. Since the subset for the second technique is better and it was computacionally bearable for the model we advanced with those variables, which can be found in Table 4.1.

4.1.1.3 Dataset Sampling

While working with the reduced subset of variables and all of the 2706 patients, we still encountered significant computational limitations, particularly memory-related issues that caused the model to crash during execution. The high dimensionality and volume of the data exceeded the available system resources, making it impractical to proceed with the full dataset. To address this, we adopted a sampling strategy aimed at reducing the dataset size while preserving its representativeness and analytical integrity. We selected

Table 4.1: Variable names for the reduced dataset.

Age	Sex	Discharge Destination	AVR Surgery
Statin Medication	Anticoagulant Medication	Presence of Preoperative Pain	Preoperative Consumption of Two Vegetables
Blood Pressure Monitoring at 30 Days	Scar Hydration at 3 Months	Medication Adherence at 3 Months	Attempted to Quit Smoking at 1 Year
Lack of Knowledge for Blood Pressure at 1 Year	Blood Glucose Monitoring at 1 Year	Lack of Means for Blood Glucose at 1 Year	Intervention: Education Reinforcement
Intervention: Other Professional	Intervention: Consultation	Intervention: Treatment	Intervention: Emergency Services

a sample size of 100 individuals, as it provided sufficient diversity to capture distinct model patients while remaining computationally feasible within an acceptable processing time. To ensure robustness and consistency in the results, five independent samples of this size were drawn randomly from the full dataset, each using a fixed seed to guarantee reproducibility.

4.1.2 Clustering Method

Now we implement the clustering model described in Section 3.2. Firstly, we computed the dissimilarity matrices for each sample using `Python`. For each dataset, numerical features were normalized with the `StandardScaler`, and then the Euclidean distance matrix was calculated using the `pdist` function from `scipy`.

4.1.2.1 Rule Construction

To employ the clustering model, it was then necessary to construct the set of N rules for each group, i.e., each variable from the dataset listed in Table 4.1. To do so, each variable got two mutually exclusive rules that represented the compliance of that variable to a value or an interval. Each treatment can be described as detailed in Table 4.2. This enabled us to calculate the rule compliance matrix necessary for the model that represented whether each individual obeyed the rules.

Table 4.2: Rule definition

Variable	**Rule**	**Definition**
Age	Aged 65 and over Aged under 65	Age ≥ 65 Age < 65
Sex	Male Female	
Discharge Destination	Home Other facilities	Destination after surgery
AVR Surgery	Yes No	Underwent Aortic Valve Replacement
Statin Medication	Yes No	Prescribed statins
Anticoagulant Medication	Yes No	Prescribed anticoagulants
Preoperative Pain	Yes No	Reported pain before surgery
Preoperative Vegetable Intake	Yes No	Ate 2+ vegetables daily before surgery
BP Monitoring at 30 Days	Always Not always	Reported monitoring blood pressure daily
Scar Hydration at 3 Months	Always Not always	Reported consistently hydrating scar
Medication Adherence at 3 Months	Adherent Non-adherent	Score above or below median [8]
Quit Smoking Attempt at 1 Year	Tried Did not try	Attempted to quit smoking
BP Lack of Knowledge at 1 Year	Yes No	Lack of knowledge to monitor blood pressure
BS Monitoring at 1 Year	Not never Never	Reported monitoring blood sugar at least sometimes
BS Means Barrier at 1 Year	Yes No	Lack of resources to monitor blood sugar
Educational Reinforcement	Yes No	Nurse provided educational reinforcement
Other Professional	Yes No	Nurse referred to another healthcare professional
Consultation	Yes No	Anticipated in-person consult
Treatment	Yes No	Surgeon changed treatment plan
Emergency Service	Yes No	Referred to emergency care

4.1.3 Time-Driven Activity-Based Costing

In this section, we follow the 8-step methodology proposed by Da Silva Etges et al. for implementing the TDABC model.

The first step is to identify the study question and what technologies are to be assessed. In the specific case of this thesis, the study question is geared toward improving resource allocation and performing cost analysis as a means to support more efficient and informed healthcare management. This way, we can evaluate the trajectory of a patient who undergoes cardiothoracic surgery and the associated costs.

Then it is time to define the care pathway for each patient in the CardioFollow.AI program. In this phase, we developed an activity map that describes the processes the patient went through and how they are connected, as shown in Figure 4.3. This involved collecting data in the hospital from patients and medical personnel.

The patient starts by being administratively registered, followed by an appointment with the surgeon and nurse. After that, the patient has an appointment with the anesthetist and, on a later date, closer to the surgery, proceeds to administrative admission. Once admitted, the surgeon and the nurse visit the patient for continuous ward monitoring a total of four times. Then the patient undergoes pre-operative tests and preparation before surgery. Surgery is followed by continuous monitoring, first in the ICU and, usually before the one-day mark, the patient is transferred to the ward with visits from the surgeon, nurses, and other professionals. Upon stabilization, a discharge appointment with the surgeon and nurse is scheduled, followed by administrative discharge. The patient then goes home and is expected to follow recovery instructions. Post-discharge care includes a series of nurse call appointments to monitor recovery. Based on patient feedback, the nurse may choose to intervene by reinforcing education, referring the patient to another professional, or requesting the surgeon's intervention. If the surgeon is involved, they may adjust treatment, anticipate an in-person consultation, or direct the patient to emergency care. These follow-up phone calls are scheduled for the 3rd and 30th day, the 3rd and 6th month, and one year after discharge. Additionally, on the 12th day post-discharge, there is an in-person hospital appointment.

To complete Steps 3 and 4 of the TDABC methodology, it is necessary to estimate the costs associated with each department. The data from Santa Marta's Hospital was retrieved previously to this work by the Value for Health CoLAB team and divided into four categories: Human Resources (HR), Spaces, Equipment and Technology, and Consumables. Since the data was collected during a previous study, future work should focus on updating the cost estimates to improve the accuracy and relevance of the findings.

For Human Resources, the team conducted interviews with the hospital staff and identified the positions of the necessary people for each activity. The costs associated with this department were calculated by analyzing the salaries of each role involved.

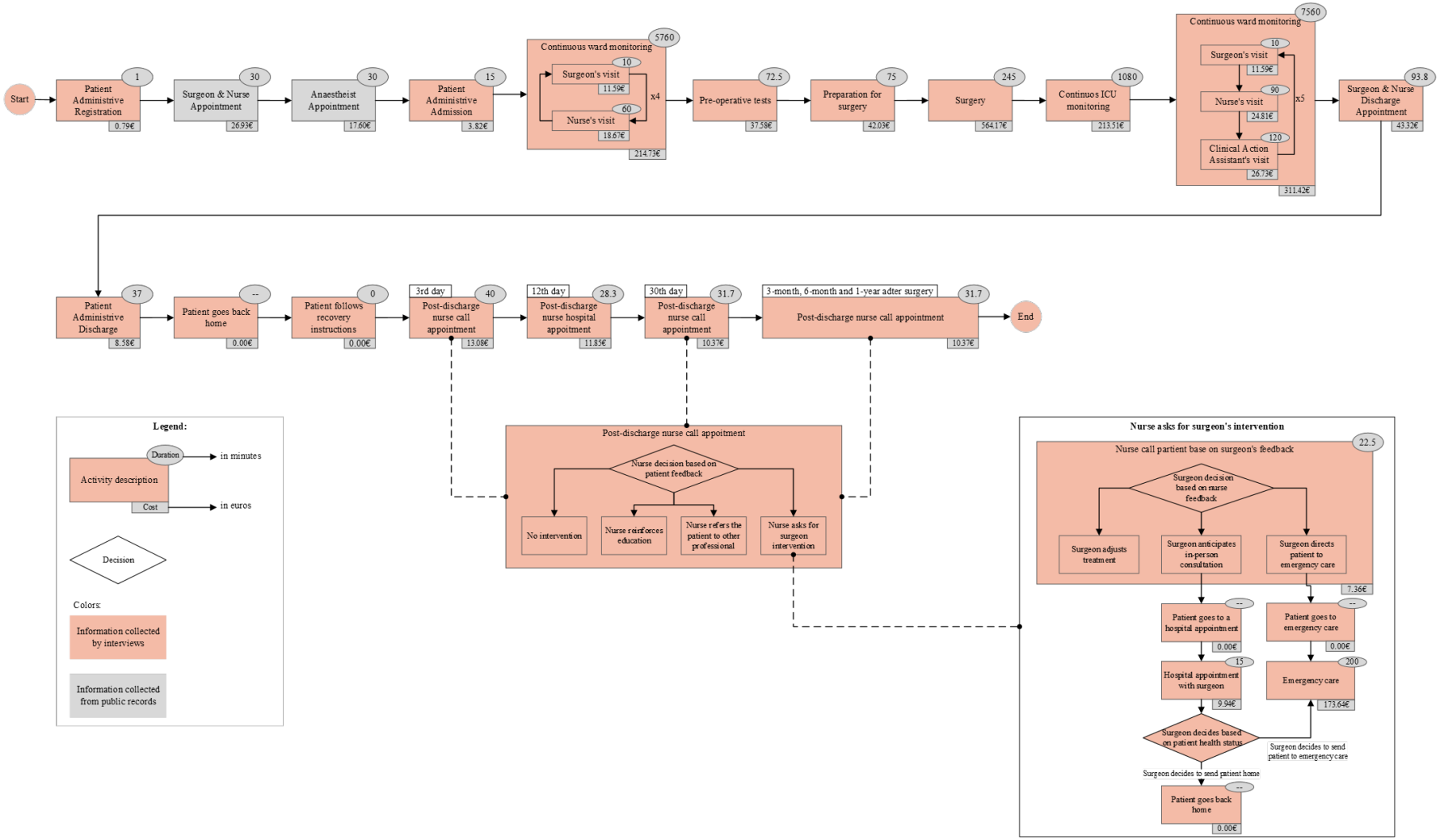


Figure 4.3: Care pathway for CardioFollow.AI patients.

The Spaces section required a different approach, as individual rooms did not have predefined costs. Instead, the team first identified the spaces needed for each activity. Then, they estimated a cost per square meter and calculated the total expense for each room by multiplying this rate by the room's area in square meters, as shown in Table 4.3.

Table 4.3: Annual Cost and Area of Clinical Use Spaces.

Space	**Clinical Use Area (m^2)**	**Cost per Year (Thousands of €)**
Nurse Room	36	25.54
Reception	22	15.61
Call Center Room	9	6.39
Work Room	27	19.16
Consultation Office	15	10.64
Operating Room	110	78.05
Surgical Room	60	42.57
Recovery Room	24	17.03
Intensive Care Unit	90	63.86
Intermediate Care Unit	90	63.86
Ward Room	5	3.55
Cost per square meter: 709.55 €		

For Equipment, Technology, and Consumables, the team identified the resources required for each activity and gathered the corresponding cost data. In Equipment and Technology, time was incorporated into the cost estimation through the application of depreciation rates and an annual maintenance percentage, allowing for the calculation of each item's yearly cost.

It is important to note that the cost estimation for emergency care was based on the available literature. Future work should consider variations across hospitals and incorporate hospital-specific data to enhance the accuracy and contextual relevance of the analysis.

Next, we needed to calculate the Cost Capacity Rate (CCR), which depended on the estimation of the practical capacity of each resource. As defined in Section 3.3, practical capacity refers to the time when people and machines are working at full capability. For the purpose of this thesis, HR practical capacity meant we removed weekends, holidays, vacations, and sick days. For certain roles, it made sense to account for days used for conferences and workshops. Depending on the position, we considered between 7 and 8 hours of work per day, along with an additional 1 to 2 hours daily allocated for breaks, training, and other non-patient-related activities. We also considered an 80% effective time, which resulted in the following times for each position:

Table 4.4: Work Days and Practical Capacity (in Hours) by Role in the Human Resources Department.

HR	Available Work Days per Year	Annual Practical Capacity
Nurse	221	1,237.6
Technical Assistant	221	1,237.6
Doctor	214	1,369.6
Operational Assistant	221	1,237.6

For the Spaces component, availability was assumed to be 24 hours per day and 365 days per year. However, a 20% reduction in available time was applied to account for activities such as cleaning, maintenance, and other non-operational uses. The same strategy was applied to the Equipment and Technology.

For Step 6, it is necessary to get the time that each patient needs for a specific activity. In order to obtain this information, the team used a mixture of electronic medical records (e.g., timestamps indicating the entry and exit of patients for a department), hospital data systems, and *in loco* observations. These sources enabled the team to make informed estimations of time allocation for each activity.

Finally, we can estimate the total cost of patient care, i.e., the expenditure incurred from the moment of admission to the point of medical discharge. To compute this, we apply Equation 3.11, described in Section 3.3, to calculate the cost for each patient in the CardioFollow program.

The cost of each activity is calculated as explained before, by first summing the time-dependent costs, which include Human Resources, Spaces, Equipment, and Technology. Then, we add the time-independent costs, such as Consumables.

It is of interest to study the total cost of the patients in the CardioFollow.AI program for the hospital. This total cost is calculated by multiplying the cost associated with each patient model type by the number of patients that fit that model, as shown in Equation 3.12.

4.2 Case Study: Blockchain

To further illustrate the adaptability of the proposed framework beyond the healthcare sector, we also applied it to a use case within the Blockchain and digital technology domain. This case involved the Dotmoovs platform, a mobile application that uses artificial intelligence and blockchain to gamify physical activity through skill-based challenges and tokenized incentives. Dotmoovs enables users of various athletic levels to participate in skill-based challenges, receive real-time feedback, and earn token-based rewards. The goal of applying this methodology to this use case was also to explore how the clustering method performs in a very different setting, particularly one driven by user engagement.

The task focused on developing a tool that can analyze social media data, identify the communities involved in these conversations, and uncover the topics being discussed in

the context of digital sports. By using Natural Language Processing (NLP) and clustering techniques to identify recurring themes, hashtags, and influential users, the tool empowers marketers with valuable insights for more informed decision-making. These insights will support the refinement of digital sports marketing strategies, help monitor campaign performance, and uncover new opportunities for targeted engagement and community growth.

4.2.1 Exploratory Analysis

The data used in this application of the methodology was collected through web scraping directly from the Twitter platform. As a result, the dataset is considered structured, requiring minimal preprocessing compared to unstructured data sources such as the previous case study.

The dataset was obtained by searching for the keyword "dotmoovs" in the "Recent" tab on the Twitter website, and extracting the corresponding tweet data and performance metrics. A total of 1067 tweets were collected, with the following variables: Username, Tweet content, Replies, Retweets, Likes, and Views.

Minor preprocessing steps were required to standardize the dataset. Specifically, the Views variable, originally stored as a character string representing values in thousands, was converted to a numeric format. Additionally, missing values across all variables were replaced with zeros to ensure consistency.

To construct the rules required by the clustering model, which are used to define communities within the Dotmoovs population, it was necessary to develop strategies for transforming the extracted variables into meaningful indicators. These indicators aimed to distinguish between positive vs. negative communities and influential vs. non-influential communities. We used sentiment analysis to identify the tone of tweets and influence ranking metrics to assess user impact.

4.2.1.1 Sentiment Analysis

To extract valuable information from the tweet's text, we applied Sentiment Analysis, a Natural Language Processing (NLP) technique that helps identify and quantify the emotional tone portrayed in text. In this context, each tweet is assigned three sentiment scores, positive, neutral, and negative, which collectively represent the overall emotion behind the message.

For this task, we used a model based on the BERT (Bidirectional Encoder Representations from Transformers) language model, an open-source machine learning framework. Unlike traditional models that analyze text in a single direction, BERT interprets text by understanding the context around each word, thereby facilitating the understanding of how each word is related to the ones that precede and follow it.

The selection of the model was made with consideration for the specific needs of the problem as well as the context. For these reasons, we opted for a pre-trained BERT-based

model that has been fine-tuned specifically for sentiment analysis in the Twitter context [5].

For our dataset, the model was applied to the Cleaned Tweet variable, which is derived from the original Tweet variable after removing URLs, punctuation, and converting all text to lowercase.

4.2.1.2 Influence Ranking

Our next challenge was to define influence when looking at the performance metrics that characterized each tweet. This will later help us to draft rules that can define which tweets are influential and those that have less impact in the community.

Instead of directly comparing the performance metrics to determine the influential tweets, we used the relative order of the tweets' ranks as a distance measure. By doing this, we define the influence distribution of each of the performance metrics. This is done by sorting the tweets by the performance metric whose distribution we are attempting to model and then matching a rank so that number 1 is allocated to the most influential tweet, and as the rank increases, the influence of the tweet decreases. Therefore, the lower the rank, the more influential the tweet is. For the instances with the same value for a performance metric, their rank is the average of the ranks amongst them.

For our dataset, this technique was applied to the variables Replies, Retweets, Likes, and Views, converting each into its respective rank. The influence thresholds for each metric, defined as the top 10%, are presented in Table 4.5. For instance, any tweet with a rank below 128 in the Views variable is classified as influential. This ranking approach provides a systematic way to identify tweets that exhibit higher engagement across different metrics.

Table 4.5: Influence threshold for each metric.

Metric	**Replies**	**Retweets**	**Likes**	**Views**
10% Influence	105	107	108.5	128

4.2.1.3 Dataset Sampling

While working with the full set of 1067 tweets and all associated variables, we faced significant computational challenges, particularly, the clustering model required an impractically long processing time due to the volume and dimensionality of the data. To address this, we employed a stratified sampling approach based on k-means clustering to reduce the dataset size while preserving its structural diversity and representativeness. By first grouping tweets into clusters with k-means, we ensured that samples were drawn proportionally from each cluster, maintaining the heterogeneity of the original data. This approach allowed us to select a manageable subset of tweets that captured key community patterns and user behaviors, while significantly reducing processing time.

4.2.2 Clustering Method

We now proceed to implement the clustering model described in Section 3.2.

Firstly, we needed to compute the dissimilarity matrix. For this, we used the `daisy` function from the `cluster` package in R.

4.2.2.1 Rule Construction

To employ the clustering model in the context of this dataset, it was necessary to construct a set of N rules for each group, corresponding to the variables listed before. For each variable, two mutually exclusive rules were defined to indicate whether a tweet belonged to a specific category, such as positive vs. negative sentiment or influential vs. non-influential, based on predefined thresholds or classifications derived from the data. Each treatment, i.e., rule assignment, can be described as follows:

Table 4.6: Rule definition.

Variable	**Rule**	**Definition**
Rank Replies	Non-influencers	Rank Replies ≥ 105
	Influencers	Rank Replies < 105
Rank Retweets	Non-influencers	Rank Retweets ≥ 107
	Influencers	Rank Retweets < 107
Rank Likes	Non-influencers	Rank Likes ≥ 108.5
	Influencers	Rank Likes < 108.5
Rank Views	Non-influencers	Rank Views ≥ 128
	Influencers	Rank Views < 128
Negative Score	Negative	Negative Score ≥ 0.65
	Not negative	Negative Score < 0.65
Positive Score	Positive	Positive Score ≥ 0.9
	Not positive	Positive Score < 0.9

This allowed us to construct the rule compliance matrix required by the model, indicating whether each individual satisfied the defined rules.

4.2.3 Time-Driven Activity-Based Costing

For this case study, a cost analysis based on the e TDABC methodology was not conducted, as it was not part of the scope of the task. However, applying such an analysis could offer valuable insights, for example, by estimating the expenses associated with each community in the context of marketing campaigns and their utilization of company resources.

5

Discussion

This chapter is divided into two parts. First, we examine the CardioFollow.AI use case, analyzing both the clustering results and the Time-Driven Activity-Based Costing analysis. This is followed by the Blockchain use case, which focuses solely on clustering.

5.1 Case Study: CardioFollow.AI

In this section, we begin by exploring the clustering results for the CardioFollow.AI case study, followed by the presentation of the Time-Driven Activity-Based Costing results.

5.1.1 Clustering Results

The clustering model described in Section 3.2 was run using Gurobi with Python via the GurobiPy interface on a PC with an Intel® Ultra 7 155H processor, 32 GB RAM, and a 1 TB SSD, running Windows 11 Pro. For both models, the number of rules per cluster was limited to four ($\ell = 4$).

As stated previously, we begin by examining the results for cardiothoracic surgery patients. For each of the five samples of 100 patients, we explored clustering configurations from two to five clusters, yielding four solutions per sample. Across all samples and configurations, 70 model patients were identified.

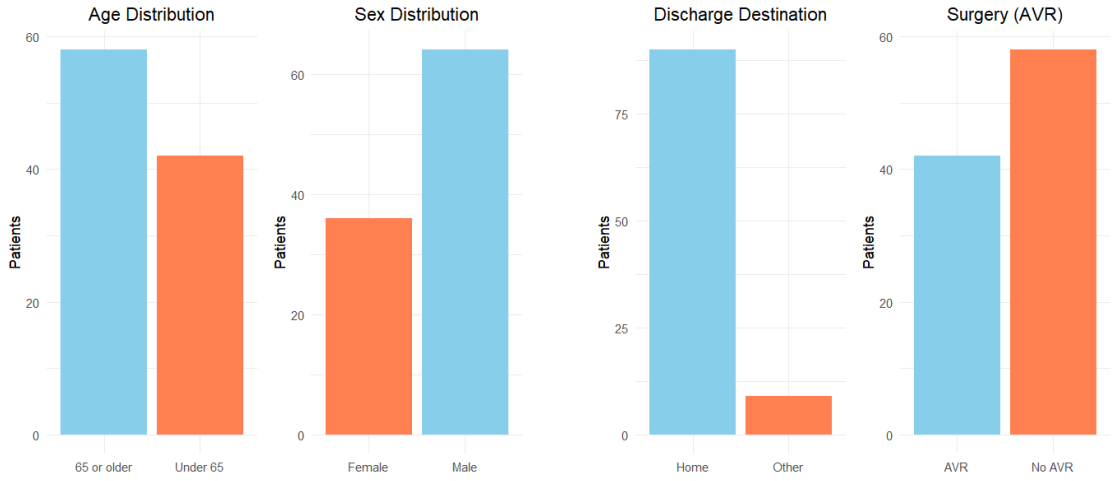
As shown in Figure 4.3, only "Emergency Service," "Treatment," and "Consult" interventions changed patient pathways. Each patient could experience one, two, or all three interventions, resulting in eight unique care pathways.

For simplicity, we present results from sample 1 with a four-model-patient clustering solution, as findings were similar across other samples and configurations.

Before presenting the clustering results, it is important to provide a brief descriptive analysis of sample 1 to better understand the patients' general characteristics and behaviors. This is also important to understand some of the clustering model's choices.

The sample consists of 100 patients, with a slightly higher proportion of older adults: 58 individuals were aged 65 or older at the time of surgery, while 42 were younger than

65. Regarding sex distribution, the sample is predominantly male, with 64 men and 36 women (Figure 5.1).



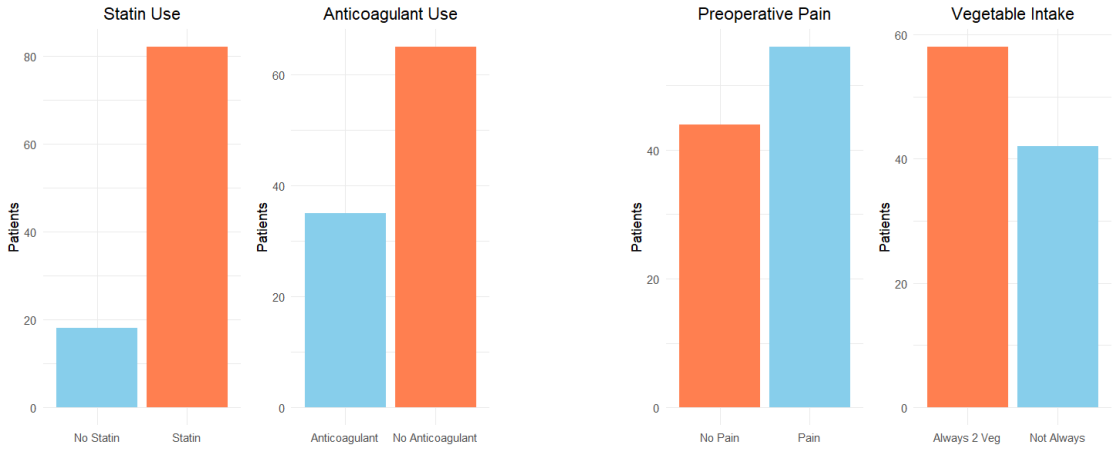
(a) Age and sex distribution of patients in sample 1.

(b) Destination and AVR Surgery distribution of patients in sample 1.

Figure 5.1: Overview of demographics and surgery distributions for sample 1.

In terms of discharge destination, the majority of patients were discharged to their homes, while the remaining 9 were sent to other destinations such as other hospitals and nursing homes. Concerning surgical procedures, 42 patients underwent Aortic Valve Replacement (AVR), and 58 did not (Figure 5.1).

Regarding medication, most patients were prescribed statins (82), while only 18 did not report statin use. The use of anticoagulants was less frequent, with 35 patients reporting use and 65 not taking them (Figure 5.2).



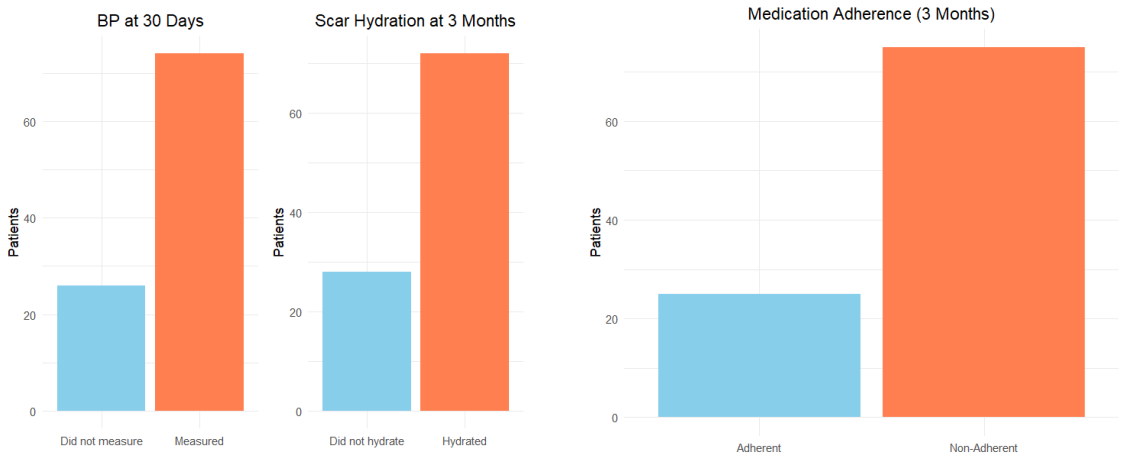
(a) Medication prescription distribution of patients in sample 1.

(b) Preoperative responses distribution of patients in sample 1.

Figure 5.2: Overview of medications prescriptions and preoperative responses distributions for sample 1.

When analyzing pre-operative conditions and behaviors, 56 patients reported experiencing pain before surgery, while 44 did not. Regarding dietary habits, 58 patients stated they consistently consumed at least two vegetables per day, whereas 42 did not report the same level of consistency (Figure 5.2).

Post-operative monitoring indicators show that 74 patients monitored blood pressure levels consistently after 30 days of discharge, while 26 did not. At three months, 72 patients reported consistently hydrating their surgical scars, whereas 28 did not. However, medication adherence was relatively low: only 25 patients were classified as adherent, with adherence values above the median (3.975), while 75 were considered non-adherent (Figure 5.3).



(a) Postoperative responses distribution of patients in sample 1.

(b) Medication adherence distribution of patients in sample 1.

Figure 5.3: Overview of postoperative responses and medication adherence distributions for sample 1.

Long-term health behaviors reveal several areas of concern. Only 3 patients reported having attempted to quit smoking by one year post-discharge. When it comes to chronic condition monitoring, both knowledge and access appear to be limited. Specifically, 2 patients indicated a lack of awareness regarding the importance of monitoring blood pressure, while the remaining 98 understood its significance. Regarding blood glucose monitoring, 50 patients reported never having measured their blood sugar levels. Furthermore, 35 patients stated they lacked the necessary equipment to perform the measurements, whereas 65 had access to the required tools (Figure 5.4).

Finally, follow-up interventions primarily involved educational reinforcement, which was received by 98 patients. Additionally, 30 patients were referred to another health professional, 20 were referred for a medical consultation, 2 patients each received changes for specific treatment, and 2 were referred to emergency services (Figure 5.5).

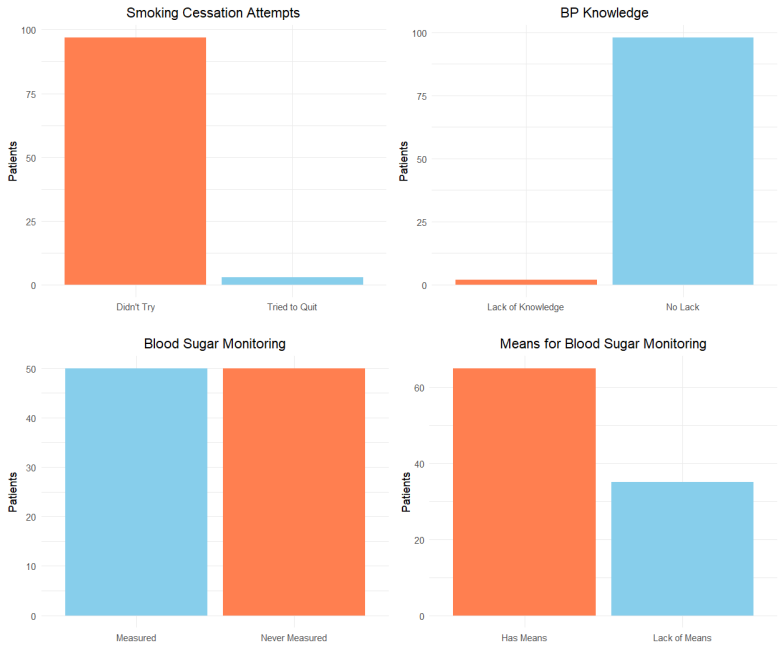


Figure 5.4: Overview of long term behaviours distributions for sample 1.

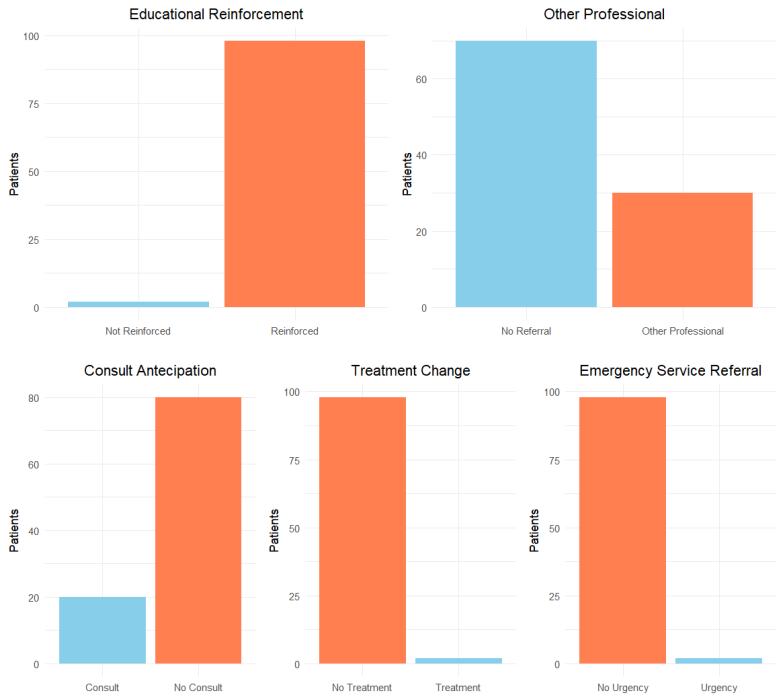


Figure 5.5: Overview of interventions distributions for sample 1.

We now proceed to analyze the results for sample 1, focusing on the scenario in which the model was configured to generate four clusters. This scenario was selected in consultation with Dr. Salomé Azevedo, PhD, Head of Digital Health at VOH CoLAB and a recognized specialist in clinical contexts.

The allocation of a set of rules to the clusters can be seen in Figure 5.6.This distribution forms the basis for defining the representative model patient within each cluster.

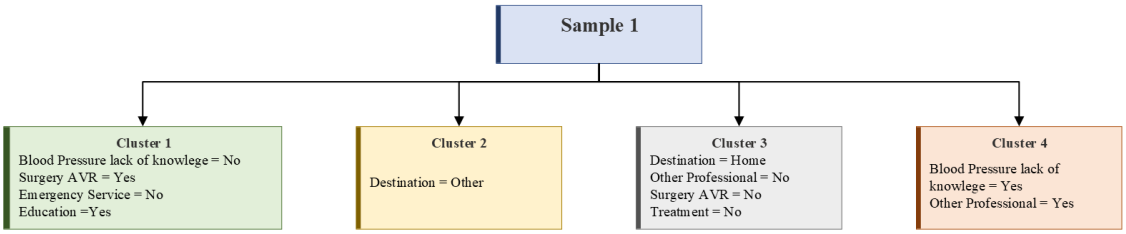


Figure 5.6: Clustering results for sample 1.

Cluster 1 is characterized by patients who underwent Aortic Valve Replacement (AVR) surgery, did not make use of emergency services, and, despite already being aware of the importance of blood pressure monitoring, still received educational reinforcement. This cluster can be interpreted as representing our standard AVR patients, i.e., individuals who follow clinical expectations, do not place significant demand on hospital resources, and are distinctly marked by the specific surgical procedure they underwent.

Cluster 2 comprises patients who were not discharged to their homes. Instead, they were transferred to other hospitals, nursing homes, or similar facilities. As a result, the hospital is no longer able to track their subsequent healthcare utilization or associated costs, making them less visible in follow-up or cost-monitoring systems.

Cluster 3, in contrast to Cluster 1, consists of patients who did not undergo Aortic Valve Replacement (AVR) surgery. These individuals also had minimal intervention, particularly with respect to changes in treatment or early consultations. Unlike Cluster 2, patients in this group were discharged to their homes. Overall, these represent typical non-AVR patients, whose low level of follow-up intervention suggests they are likely to generate minimal additional costs for the hospital.

Cluster 4 is characterized by patients who were referred to other healthcare professionals, such as primary care providers, dietitians, or diabetes specialists. These patients also demonstrated a lack of knowledge regarding blood pressure monitoring, which may reflect limited health literacy and suggest a need for closer clinical follow-up. As a result, individuals in this group are likely to require more support and resources, potentially leading to higher healthcare costs for both the hospital and the broader healthcare system.

We now turn to an analysis of the actual behavior associated with each of the four clusters, in order to better understand the rationale behind the model's earlier decisions.

Cluster 1 predominantly consists of older adults (75%), with a slight male majority (57%). Nearly all individuals in this group underwent aortic valve replacement (AVR) surgery (89%) and used statins (89%). All patients were discharged home. Short-term

indicators were favorable, with 71% reporting consistent blood pressure monitoring at 30 days and 86% showing proper scar hydration at 3 months. In terms of long-term behavior, only 25% were classified as medication adherent, and 4% attempted to quit smoking. All patients received educational reinforcement, but no additional follow-up, such as consultations, treatments, or emergency services, was reported. These observed behaviors align with the rule-based definition of Cluster 1, characterized by AVR surgery, no emergency service usage, and receipt of educational reinforcement despite prior knowledge of the importance of monitoring blood pressure.

Cluster 2 shares a similar demographic profile with Cluster 1, composed mainly of older adults (78%) and a higher proportion of men (67%). However, discharge destinations differed, as only 56% of individuals in this group returned home, indicating a notable proportion were discharged to other facilities such as nursing homes or hospitals. AVR surgery was less common (39%), and statin use was slightly lower (72%), with anticoagulant use remaining limited (22%). Post-operative outcomes included moderate levels of pain and blood pressure control, and scar hydration at 3 months was reported by 56% of patients. Medication adherence reached 39%, and 6% attempted to quit smoking. Most patients (89%) received educational reinforcement, and a subset also engaged with other healthcare professionals or emergency services. These behaviors are consistent with the defining rules of Cluster 2, which characterizes patients based on not being discharged home.

Cluster 3 stands apart due to its younger demographic, 68% of patients were under 65, and the complete absence of AVR surgery. Statins use remained relatively high at 75%. This group exhibited relatively high levels of positive lifestyle behaviors, including regular vegetable intake (82%) and strong blood pressure monitoring at 30 days (86%). However, they also had the lowest rate of medication adherence (18%) and the highest proportion of individuals reporting a lack of means to monitor blood sugar levels (54%), indicating limited access to necessary equipment. Follow-up was limited to educational reinforcement, with no reports of additional consultations, treatments, or emergency service use. These patterns are consistent with the cluster definition, where patients typically had little to no intervention, were discharged home, and did not undergo AVR surgery.

Cluster 4 presents a more balanced age distribution, with 54% of patients aged 65 or older, and the highest proportion of males (73%). AVR surgery was moderately represented (38%), and statin use was high (88%). This group reported the highest levels of post-operative pain (69%), while maintaining moderate rates of vegetable intake (54%), blood pressure control at 30 days (69%), and scar hydration at 3 months (69%). Medication adherence remained low (23%). Unlike the other clusters, all patients in Cluster 4 received both educational reinforcement and referrals to other healthcare professionals. Additionally, 73% had at least one consultation after discharge. No treatments or emergency service usage were recorded, indicating that consistent outpatient follow-up may have contributed to preventing acute care utilization. These behaviors match the cluster

definition: patients were referred to other professionals and showed limited knowledge of blood pressure monitoring, indicating a need for closer follow-up.

This cluster analysis enabled the definition of distinct model patients, characterized by their unique trajectories and clinical features. These profiles provide a solid foundation for subsequent cost analyses, allowing a more precise evaluation of healthcare resource utilization tailored to the specific needs and outcomes of each patient group.

5.1.2 Time-Driven Activity-Based Costing Results

In this section, we present a detailed explanation of the Time-Driven Activity-Based Costing (TDABC) analysis performed for the previously selected scenario. We also include additional examples that provide relevant insights into potential interventions and alternative trajectories.

The selected example involves no interventions across all model patient types. Consequently, all patient cohorts follow an identical pathway, as illustrated in Figure 4.3. For each model patient, we report the minimum, average, and maximum cost estimates (Table 5.1), calculated using Equation 3.11 and the CCR tables.

Table 5.1: Cost summary for sample 1 with 4 clusters in euros.

Patient ID	Total Minimum Cost (€)	Total Average Cost (€)	Total Maximum Cost (€)
1	994.13	1020.82	1049.92
2	994.13	1020.82	1049.92
3	994.13	1020.82	1049.92
4	994.13	1020.82	1049.92

As expected, the cost estimates for all patient cohorts in this case are identical. However, this may not fully reflect real-world situations.

Cluster 1 is characterized by the presence of Aortic Valve Replacement (AVR) surgery and educational interventions. In the TDABC analysis, surgery costs are not specified at the procedure level, meaning that the actual cost of AVR surgery could be higher or lower than assumed. Furthermore, educational reinforcement is delivered during post-discharge calls, which can extend call duration and potentially increase costs. These factors are not fully captured in the current model, which may place the estimates toward the higher end of the range while still not fully reflecting the true cost of care for this cluster.

Cluster 2 includes patients who are not discharged home but are instead transferred to other facilities, such as hospitals, care centers, or nursing homes. In the TDABC analysis, these patients were assigned the same estimated costs as the rest of the other model patients. However, this assumption may not be accurate, as once patients leave the hospital's care, additional costs incurred at the receiving facility are not captured in the model. As a result, the current estimates may underestimate the true cost of their care pathway.

Cluster 3 represents the average patient profile, typically not requiring major interventions. In the TDABC analysis, these patients incur minimal additional costs for the hospital, as their care pathway closely follows the standard process without significant deviations.

Cluster 4 is characterized by high utilization of hospital resources. However, the associated costs are not fully captured in this analysis, as they involve a wide range of professionals, such as dietitians, primary care providers, and others, each with different and unmeasured expenses. Additionally, patients in this cluster often exhibit low health literacy, which may necessitate further interventions that incur additional costs not accounted for in the current TDABC model.

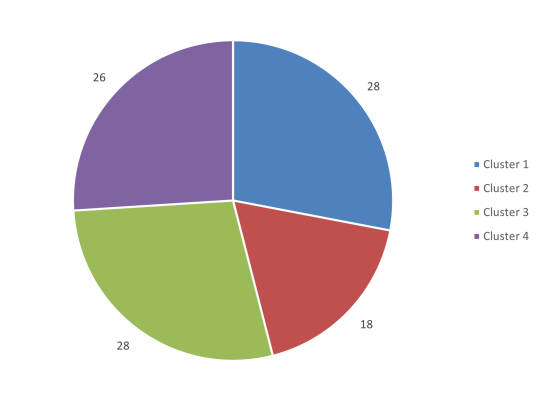
The following table presents the costs for other model patients relevant to this analysis. Unlike the previous example, these patients follow different care pathways and therefore make use of different resources. Model patient 1 underwent consultation anticipation, model patient 52 utilized emergency services, and model patient 37 experienced a treatment change.

Table 5.2: Cost summary for patients of interest in euros.

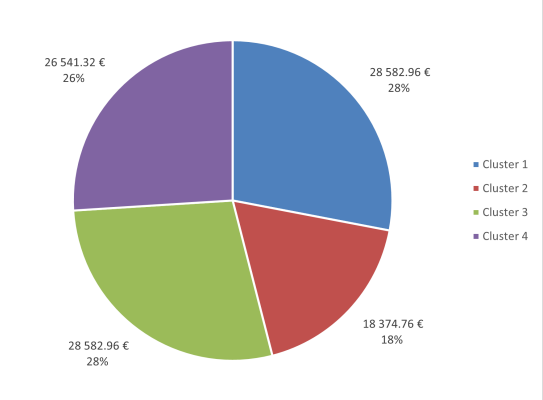
Patient ID	Total Min Cost (€)	Total Average Cost (€)	Total Maximum Cost (€)
1	1007.87	1038.11	1070.78
37	1000.67	1028.18	1058.10
52	1131.33	1201.82	1274.72

It is important to note that, as mentioned earlier, the emergency service cost estimates were derived from available literature rather than hospital-specific data, which may result in discrepancies in the findings. Additionally, consultation anticipation and treatment changes can affect the duration of calls, thereby impacting the associated costs.

As discussed in Chapter 3, it is also relevant to examine the total cost associated with patients in each sample. Continuing with sample 1, we maintain the four-cluster configuration. While each patient theoretically incurs the same hospital expense, the clusters differ in size, as shown in Figure 5.7a. By multiplying the average cost per patient by the size of each cluster, we obtain the estimated total costs presented in Figure 5.7b, resulting in a total of 102,081.99 € for the entire sample.



(a) Cluster sizes for the 4-cluster configuration of sample 1.



(b) Cost estimations for each cluster.

Figure 5.7: Overview of cluster sizes and corresponding cost estimations for sample 1.

5.2 Case Study: Blockchain

In this section, we examine the Blockchain use case. We begin by characterizing the sample, followed by an analysis of the clustering results.

The sample consists of 100 tweets as described in Subsection 4.2.1. In terms of influence, approximately 10% of tweets rank as influencers across key interaction types, including replies, retweets, likes, and views. Specifically, 11 tweets fall below the threshold rank of 105 for replies, indicating higher engagement, while 10 tweets are similarly classified as influencers by retweet rank (< 107), 11 by likes rank (< 108.5), and 10 by views rank (< 128).

The sentiment analysis results indicate that the vast majority of tweets demonstrate a neutral or positive sentiment. Only 2 tweets have a high Negative Score (≥ 0.65), reflecting a generally positive or neutral tone in the sample. Specifically, 20 tweets reach or surpass the Positive Score threshold of 0.9, representing a meaningful subset of highly positive content.

This stratified sample accurately represents the overall dataset extracted, preserving the essential characteristics of the full data. The distribution of influencer ranks across replies, retweets, likes, and views, as well as the balance of positive, neutral, and negative sentiment scores, closely mirrors that of the entire dataset. This ensures that the sample provides a reliable basis for analysis while reflecting the broader patterns of engagement and sentiment present in the full collection of tweets.

We will now analyze the results for the stratified sample, focusing on the scenario with three clusters. This configuration was chosen because it produced the most insightful outcomes for community studies and demonstrated a favorable objective value compared to other clustering solutions. The results can be seen in Figure 5.8.

Cluster 1 is characterized by users with relatively low influence, as indicated by higher ranks in replies (rank_replies ≥ 105), retweets (rank_retweets ≥ 107), and views

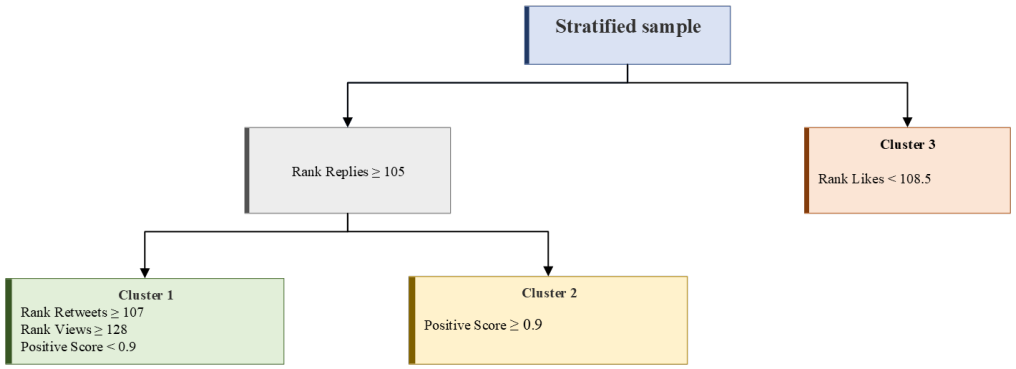


Figure 5.8: Clustering results for the stratified sample.

(rank_views ≥ 128). In addition, their sentiment scores were below 0.9, leading to a classification of Not positive. This group can be interpreted as low-impact, low-positivity participants.

Cluster 2 comprises users who also displayed low influence in replies (rank_replies ≥ 105), similar to Cluster 1, but with a Positive Score of at least 0.9, resulting in a Positive classification. These users can be seen as low-impact but constructive participants, contributing positively to discussions despite limited influence.

Cluster 3, in contrast, is characterized by more influential users in terms of likes (rank_likes < 108.5), leading to an Influencer classification. Despite having a stronger influence compared to the other clusters, their higher engagement in likes may not necessarily extend across other interaction types, suggesting a focused but impactful role within the network.

We will now compare this classification with the actual behavior observed in each cluster to gain a deeper understanding of the model's decisions.

Cluster 1 predominantly consists of tweets with low engagement, since none of the tweets in this cluster are classified as influencers across replies, retweets, likes, or views. This confirms a consistently low interaction profile. Sentiment analysis reveals that 100% of these tweets fall below the positive sentiment threshold, with only 2.4% showing a high negative score. This behavioral profile aligns closely with the rule-based definition of Cluster 1, characterized by high rank values indicating non-influencer status and predominantly non-positive sentiment.

In Cluster 2, similarly to Cluster 1, none of the tweets are classified as influencers across replies, retweets, likes, or views, indicating a consistently non-influential profile. However, this cluster differs notably in sentiment because 50% of the tweets exhibit positive sentiment, while none display negative sentiment. This mix of low engagement but higher positive sentiment aligns with the rule-based definition of Cluster 2, characterized by non-influencer ranks paired with predominantly positive sentiment scores.

Cluster 3 is distinguished by a notably higher level of engagement. Approximately 48% of tweets in this cluster are classified as influencers in replies, with similar proportions for retweets, likes, and views, indicating a strong presence of influential users. Despite this

higher engagement, only 9% of tweets exhibit positive sentiment, while the vast majority (91%) are non-positive, including a single tweet marked as negative. This cluster aligns with the rule-based definition of Cluster 3, characterized by a mix of influencer activity and predominantly neutral or non-positive sentiment.

This analysis enabled a deeper exploration of these communities, providing valuable insights into their key topics and priorities. As a result, it facilitated the development of more targeted and effective marketing strategies.

5.3 Framework Implementation and Strategic Recommendations

The framework proposed in this thesis is adaptable for implementation across a wide range of contexts. In this section, we discuss the practical implications of applying this model and identify scenarios in which its use is most appropriate. This discussion is centered on the CardioFollow.AI case study, which has been a primary focus throughout the development of this project.

Regarding implementation, this model is best suited for higher organizational levels, providing insights that are particularly relevant for strategic decision-making. Operations leads, hospital management, and administrators are likely to find the model's findings more valuable in guiding resource allocation, policy development, and long-term planning than frontline doctors and clinicians focused on day-to-day patient care. However, as noted earlier, it remains essential that the model is accessible and understandable to all personnel, regardless of their technical expertise. This is especially important because many hospital administrators have clinical backgrounds themselves, and for the model to be effectively adopted, its recommendations must align with their clinical judgment and experience. Ensuring this alignment promotes trust and facilitates better integration of the model's outputs into clinical workflows and organizational processes. Also, making sure everyone understands the model helps different teams work better together, which can lead to better patient care and smoother hospital operations.

The recommendations mainly concern the hospital's internal information system and the way data is recorded and managed. The primary issue with the provided dataset was the large volume of unstructured data and the lack of clarity regarding what was being measured. To address this, we suggest restructuring the questionnaires completed by the nursing team to be electronically recorded with drop-down options for answers, making the data more structured and easier to analyze. At the same time, we recognize the need for detailed notes, which can still be entered through a free-text field to complement the predefined choices. Although this free-text information may not be directly used by the model, it remains crucial for effective communication among healthcare professionals and teams.

It is also important to maintain detailed records of the costs and times associated with procedures and activities. This could be achieved by continuously collecting and updating the relevant data, as done in this study. Additionally, the hospital could integrate

electronic medical records to capture specific time stamps linked to individual patients. Such integration would improve cost estimations, for example, by providing accurate durations for each phone call, as mentioned earlier.

Although the variables selected to characterize the model patients may not be immediately intuitive to doctors and clinicians, they were identified as the most significant factors representing patient profiles within the dataset. To strengthen the model's relevance and acceptance, it would be valuable to engage with medical personnel through discussions or workshops. This could help determine whether their clinical experience aligns with the data-driven findings or offers additional insights. For example, clinicians might provide context around why certain variables are important or suggest alternative factors worth considering. Such collaboration could also improve communication between data analysts and healthcare professionals, ultimately enhancing the model's practical application.

6

Conclusion

In conclusion, this thesis aimed to develop a cost estimation model that combined Time-Driven Activity-Based Costing with data-driven clustering of patients based on their care pathways. By grouping individuals into representative cohorts and associating each patient with a model trajectory, this framework enables accurate and interpretable cost estimates while reducing the burden of detailed individual analyses. This approach strengthens efficiency in healthcare cost analysis, supports more informed resource allocation, and provides hospital administrators and decision-makers with a practical tool to guide organizational planning and strategic management.

This work builds on the substantial body of existing research reviewed in Chapter 2, particularly studies on healthcare costing and interpretable machine learning. While prior models have either lacked interpretability or proved too resource-consuming for routine use, the approach presented here fills this gap by integrating optimization-based clustering with TDABC, offering both transparency and operational relevance at the organizational level.

The results of applying this model to the CardioFollow.AI case study demonstrate its ability to capture cost variations across patient clusters, reflecting differences in procedures, discharge destinations, and resource utilization. Although some limitations exist due to data availability and variability, the model offers valuable cost approximations that can inform hospital management and operational planning.

Beyond the specific case study explored here, the proposed framework is adaptable and can be applied across a variety of healthcare contexts. Its flexibility allows healthcare organizations to tailor the model to different specialties, patient populations, and data availability scenarios, making it a valuable tool for strategic decision-making and resource management in diverse clinical settings. Ultimately, this work contributes a practical approach that supports improved financial planning and operational efficiency in healthcare systems.

For future research, expanding the model to incorporate richer clinical data and real-time cost tracking could further enhance its accuracy and adaptability. Additionally, integrating patient outcomes alongside cost metrics would support a more comprehensive

value-based assessment. On the implementation side, ongoing collaboration with clinical teams and hospital administrators will be essential to refine usability and ensure alignment with organizational workflows, complemented by training and communication strategies to facilitate adoption. Ultimately, this thesis aimed to develop an interpretable and practical framework for cost estimation at the organizational level, and its contribution also lies in advancing the emerging field of interpretable clustering methods for healthcare cost analysis. As interest in explainable machine learning continues to grow, future research in this area has the potential to provide decision-makers with increasingly transparent, data-driven tools to support sustainable and value-based healthcare management.

Bibliography

[1] K.-M. M. Aigner, F. Liers, M. Goerigk, M. Hartisch, A. Miehlich, F. Liers, A. Miehlich, M. Goerigk, M. Hartisch, and A. Miehlich. "A Framework for Data-Driven Explainability in Mathematical Optimization". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 19. 2024, pp. 20912–20920. DOI: 10.1609/aaai.v38i19.30081. URL: https://www.researchgate.net/publication/373164098 (cit. on p. 5).

[2] S. Akhavan, L. Ward, and K. J. Bozic. "Time-driven Activity-based Costing More Accurately Reflects Costs in Arthroplasty Surgery". In: *Clinical Orthopaedics and Related Research* 474.1 (2016-01), pp. 8–15. ISSN: 15281132. DOI: 10.1007/s11999-015-4214-0 (cit. on p. 7).

[3] N. Alangari, M. El Bachir Menai, H. Mathkour, and I. Almosallam. "Exploring Evaluation Methods for Interpretable Machine Learning: A Survey". In: *Information (Switzerland)* 14.8 (2023), pp. 1–29. ISSN: 20782489. DOI: 10.3390/info14080469 (cit. on p. 5).

[4] A. G. Asuero, A. Sayago, and A. G. González. "The correlation coefficient: An overview". In: *Critical Reviews in Analytical Chemistry* 36 (1 2006), pp. 41–59. ISSN: 10408347. DOI: 10.1080/10408340500526766 (cit. on p. 12).

[5] F. Barbieri, J. Camacho-Collados, L. E. Anke, and L. Neves. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2020, pp. 1644–1650. DOI: 10.18653/v1/2020.findings-emnlp.148 (cit. on p. 31).

[6] S. Behzadi, N. S. Müller, C. Plant, and C. Böhm. "Clustering of mixed-type data considering concept hierarchies: problem specification and algorithm". In: *International Journal of Data Science and Analytics* 10.3 (2020-09), pp. 233–248. ISSN: 23644168. DOI: 10.1007/s41060-020-00216-2 (cit. on p. 10).

[7] D. Bertsimas, A. Orfanoudaki, and H. Wiberg. "Interpretable clustering: an optimization approach". In: *Machine Learning* 110 (1 2021), pp. 89–138. DOI: 10.1007/s10994-020-05896-2. URL: https://doi.org/10.1007/s10994-020-05896-2 (cit. on p. 4).

[8] A. C. Cabral. "Cross-cultural adaptation and validation of a European Portuguese version of the 8-item Morisky medication adherence scale". In: *Revista Portuguesa de Cardiologia* 37 (4 2018). DOI: 10.1016/j.repce.2017.09.017 (cit. on pp. 20, 25).

[9] J. Cadima, J. O. Cerdeira, and M. Minhoto. "Computational aspects of algorithms for variable selection in the context of principal components". In: *Computational Statistics and Data Analysis* 47.2 SPEC. ISS. (2004-09), pp. 225–236. ISSN: 01679473. DOI: 10.1016/j.csda.2003.11.001 (cit. on p. 11).

[10] E. Carrizosa, V. Guerrero, and D. Romero Morales. "Visualizing data as objects by DC (difference of convex) optimization". In: *Mathematical Programming* 169.1 (2018-05), pp. 119–140. ISSN: 14364646. DOI: 10.1007/s10107-017-1156-1 (cit. on p. 6).

[11] E. Carrizosa, K. Kurishchenko, A. Marín, and D. Romero Morales. "On clustering and interpreting with rules by means of mathematical optimization". In: *Computers and Operations Research* 154.October 2022 (2023-06), p. 106180. ISSN: 03050548. DOI: 10.1016/j.cor.2023.106180. URL: https://doi.org/10.1016/j.cor.2023.106180 (cit. on pp. 5, 6, 13).

[12] E. Carrizosa, J. Ramírez-Ayerbe, and D. Romero Morales. "Mathematical optimization modelling for group counterfactual explanations". In: *European Journal of Operational Research* 319.2 (2024-12), pp. 399–412. ISSN: 03772217. DOI: 10.1016/j.ejor.2024.01.002 (cit. on p. 6).

[13] F. Clautiaux and I. Ljubić. *Last fifty years of integer linear programming: A focus on recent practical advances*. 2024. DOI: 10.1016/j.ejor.2024.11.018 (cit. on p. 5).

[14] A. P. B. da Silva Etges, L. N. Cruz, R. K. Notti, J. L. Neyeloff, R. P. Schlatter, C. C. Astigarraga, M. Falavigna, and C. A. Polanczyk. "An 8-step framework for implementing time-driven activity-based costing in healthcare studies". In: *European Journal of Health Economics* 20.8 (2019-11), pp. 1133–1145. ISSN: 16187601. DOI: 10.1007/s10198-019-01085-8 (cit. on pp. 2, 7, 16, 17).

[15] S. Deshmukh, B. K. Behera, P. Mulay, E. A. Ahmed, S. Al-Kuwari, P. Tiwari, and A. Farouk. "Explainable quantum clustering method to model medical data". In: *Knowledge-Based Systems* 267 (2023-05). ISSN: 09507051. DOI: 10.1016/j.knosys.2023.110413 (cit. on pp. 4, 5, 7).

[16] M. M. Garrido, P. Deb, J. F. Burgess, and J. D. Penrod. "Choosing Models for Health Care Cost Analyses: Issues of Nonlinearity and Endogeneity". In: *Health Services Research* 47.6 (2012-12), pp. 2377–2397. ISSN: 00179124. DOI: 10.1111/j.1475-6773.2012.01414.x (cit. on p. 7).

[17] M. Goerigk, M. Hartisch, S. Merten, and K. Sharma. "Feature-Based Interpretable Surrogates for Optimization". In: (2024-09). arXiv: 2409.01869. URL: http://arxiv.org/abs/2409.01869 (cit. on p. 5).

[18] J. C. Gower. "A General Coefficient of Similarity and Some of Its Properties Published by : International Biometric Society Stable URL : http://www.jstor.org/stable/2528823". In: *International Biometric Society* 27.4 (1971), pp. 857–871 (cit. on p. 13).

[19] R. Henao, J. Murray, G. Ginsburg, L. Carin, and J. E. Lucas. "Patient clustering with uncoded text in electronic medical records." In: *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2013.3 (2013), pp. 592–599. ISSN: 1942597X (cit. on p. 4).

[20] W. P. Hennrikus, P. M. Waters, D. S. Bae, S. S. Virk, and A. S. Shah. "Inside the Value Revolution at Children's Hospital Boston: Time-Driven Activity-Based Costing in Orthopaedic Surgery". In: *The Harvard Orthopaedic Journal* 14.December (2012). URL: http://www.orthojournalhms.org (cit. on pp. 1, 7).

[21] R. S. Kaplan and S. R. Anderson. *Time-Driven Activity-Based Costing*. Tech. rep. 2003 (cit. on pp. 2, 7, 16, 17).

[22] M. Laranja. *Estratificação de pacientes para o desenho de intervenções de alteração de comportamento utilizando técnicas de apredizagem automática não supervisionada e de explicabilidade*. 2023 (cit. on p. 5).

[23] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on p. i).

[24] I. D. Lutilsky and L. Juroš. "Business Processes in the Function of Cost Management in Healthcare Institutions". In: *Eurasian Studies in Business and Economics*. Vol. 5. Springer Science and Business Media B.V., 2017, pp. 571–589. DOI: 10.1007/978-3-319-46319-3_35 (cit. on p. 1).

[25] Z. Ma and H. S. Ryoo. "General Set Covering for Feature Selection in Data Mining". In: *Management Science and Financial Engineering* 18.2 (2012-11), pp. 13–17. ISSN: 2287-2043. DOI: 10.7737/msfe.2012.18.2.013 (cit. on p. 6).

[26] A. S. Malehi, F. Pourmotahari, and K. A. Angali. "Statistical models for the analysis of skewed healthcare cost data: a simulation study". In: *Health Economics Review* 5.1 (2015-05). ISSN: 21911991. DOI: 10.1186/s13561-015-0045-7 (cit. on p. 7).

[27] K. Noor, L. Roguski, X. Bai, A. Handy, R. Klapaukh, A. Folarin, L. Romao, J. Matteson, N. Lea, L. Zhu, F. W. Asselbergs, W. K. Wong, A. Shah, and R. J. Dobson. "Deployment of a Free-Text Analytics Platform at a UK National Health Service Research Hospital: CogStack at University College London Hospitals". In: *JMIR Medical Informatics* 10 (8 2022-08), e38122. ISSN: 2291-9694. DOI: 10.2196/38122 (cit. on p. 10).

[28] J. C. Orestes, P. D. Silva, J. Cadima, and M. Minhoto. "The subselect Package". In: (2008), pp. 1–35 (cit. on pp. 11, 21).

[29] G. Schrijvers, A. van Hoorn, and N. Huiskes. "The care pathway: Concepts and theories: An introduction". In: *International Journal of Integrated Care* 12.SPECIAL EDITION I (2012). ISSN: 15684156. DOI: 10.5334/ijic.812 (cit. on pp. 1, 2).

[30] H. T. A. Simanjuntak, H. M. Simanungkalit, E. R. Tampubolon, and T. D. Natalia. "Patients Clustering on BPJS Health Insurance Data Using Partition Clustering Algorithm". In: *2023 IEEE International Conference of Computer Science and Information Technology: The Role of Artificial Intelligence Technology in Human and Computer Interactions in the Industrial Era 5.0, ICOSNIKOM 2023*. IEEE, 2023, pp. 1–8. ISBN: 9798350360752. DOI: 10.1109/ICoSNIKOM60230.2023.10364367 (cit. on p. 5).

[31] M. Sutton, S. Garfield-Birkbeck, G. Martin, R. Meacock, S. Morris, M. Sculpher, A. Street, S. I. Watson, and R. J. Lilford. "Economic analysis of service and delivery interventions in health care". In: *Health Services and Delivery Research* 6.5 (2018-02), pp. 1–16. ISSN: 2050-4349. DOI: 10.3310/hsdr06050 (cit. on p. 1).

A

Appendix A: Dataset Codebook

Variable	L	N	Class
PROCESS_NUMBER	A	1	numeric
AGE_AT_SURGERY	B	2	numeric
SEX	C	3	binary
DIAGNOSES	D	4	character
HISTORY	E	5	character
SURGERIES	F	6	character
SURGICAL_COMPLICATIONS	G	7	binary
SURGICAL_DEATH	H	8	binary
DESTINATION_AFTER_SURGERY	I	9	character
CAUSE_DEATH_CARDIAC_SURGERY	AKO	977	binary

PRE_OP_QUESTIONNAIRE	L	N	Class
FATIGUE_PO	J	10	ordinal
ACTIVITY_AUTONOMY_PO	K	11	binary
ACTIVITY_INCREASE_PO	L	12	binary
ACTIVITY_REST_PO	M	13	binary
ACTIVITY_OTHER_PO	N	14	binary
WORK_SITUATION_PO	O	15	character
PAIN_PRESENCE_PO	P	16	binary
PAIN_LOCATION_PO	Q	17	character
PAIN_SCALE_PO	R	18	ordinal
PRECORDIAL_PAIN_PO	S	19	binary
MUSCULOSKELETAL_PAIN_PO	T	20	binary
PAIN_TYPE_OTHER_PO	U	21	binary
ANALGESIC_THERAPY_PO	V	22	binary
COMFORT_POSITION_PO	W	23	binary
MASSAGE_PO	X	24	binary
PAIN_ACTION_OTHER_PO	Y	25	binary
ABSENT_PO	Z	26	binary
ANALGESIC_INSTRUCTIONS_PO	AA	27	binary
ANALGESIC_INCENTIVE_PO	AB	28	binary
PAIN_RELIEF_TEACHING_PO	AC	29	binary
PAIN_INTERVENTION_OTHER_PO	AD	30	binary
MORE_FIVE_MEALS_PO	AE	31	ordinal
TWO_FRUITS_PO	AF	32	ordinal
TWO_VEGETABLES_PO	AG	33	ordinal
VARIED_DIET_PO	AH	34	ordinal
REDUCE_SALT_PO	AI	35	ordinal
DRINK_WATER_PO	AJ	36	ordinal
DRINK_WINE_BEER_PO	AK	37	ordinal
WEIGHT_PO	AL	38	numeric
HEIGHT_PO	AM	39	numeric
DIET_INSTRUCTIONS_PO	AN	40	binary
HYDRATION_INSTRUCTIONS_PO	AO	41	binary
ALCOHOL_INSTRUCTIONS_PO	AP	42	binary
DIET_INTERVENTION_OTHER_PO	AQ	43	binary
SMOKING_HABITS_PO	AR	44	binary
NUMBER_CIGARETTES_PO	AS	45	numeric
SMOKING_CESSATION_PO	AT	46	binary
CONSULTATION_CESSATION_PO	AU	47	binary
MEDICAL_CESSATION_PO	AV	48	binary
CESSATION_ALONE_PO	AW	49	binary
VASODILATOR_PO	AX	50	binary
VASODILATOR_TYPE_PO	AY	51	character
ANTI_ARRHYTHMIC_PO	AZ	52	binary
ANTI_ARRHYTHMIC_TYPE_PO	BA	53	character
ANTI_HYPERTENSIVE_PO	BB	54	binary
ANTI_HYPERTENSIVE_TYPE_PO	BC	55	character
ANTIPLATELETS_PO	BD	56	binary
ANTIPLATELETS_TYPE_PO	BE	57	character
ANTICOAGULANTS_PO	BF	58	binary
ANTICOAGULANTS_TYPE_PO	BG	59	character
ANTI_DYSLEPIDEMIC_PO	BH	60	binary
ANTI_DYSLEPIDEMIC_TYPE_PO	BI	61	character
OTHER_THERAPY_PO	BJ	62	binary
OTHER_THERAPY_PO	BK	63	character
SPOUSE_CAREGIVER_PO	BL	64	binary
CHILD_CAREGIVER_PO	BM	65	binary
OTHER_CAREGIVER_PO	BN	66	binary
FAMILY_SUPPORT_PO	BO	67	binary
FORMAL_CAREGIVER_PO	BP	68	binary
INFORMAL_CAREGIVER_PO	BQ	69	binary
REFERRAL_PO	BR	70	binary
SOCIAL_REFERRAL_PO	BS	71	binary
DIETETICS_REFERRAL_PO	BT	72	binary
SMOKER_REFERRAL_PO	BU	73	binary
FAMILY_REFERRAL_PO	BV	74	binary
OTHER_REFERRAL_PO	BW	75	binary

SF36PO_HEALTH_STATUS_PO	BX	76	ordinal
SF36PO_ONE_YEAR_COMPARISON_PO	BY	77	ordinal
SF36PO_VIGOROUS_ACTIVITY_PO	BZ	78	ordinal
SF36PO_MODERATE_ACTIVITY_PO	CA	79	ordinal
SF36PO_SHOPPING_ACTIVITY_PO	CB	80	ordinal
SF36PO_MULTIPLE_FLIGHTS_ACTIVITY_PO	CC	81	ordinal
SF36PO_ONE_FLIGHT_ACTIVITY_PO	CD	82	ordinal
SF36PO_BENDING_ACTIVITY_PO	CE	83	ordinal
SF36PO_MORE_THAN_1KM_ACTIVITY_PO	CF	84	ordinal
SF36PO_HUNDREDS_METERS_ACTIVITY_PO	CG	85	ordinal
SF36PO_100_METERS_ACTIVITY_PO	CH	86	ordinal
SF36PO_BATHING_ACTIVITY_PO	CI	87	ordinal
SF36PO_F_WORKING_TIME_PO	CJ	88	ordinal
SF36PO_PHYS_WORKED_LESS_PO	CK	89	ordinal
SF36PO_PHYS_LIMITED_WORK_PO	CL	90	ordinal
SF36PO_PHYS_WORK_DIFFICULTY_PO	CM	91	ordinal
SF36PO_AND_WORKING_TIME_PO	CN	92	ordinal
SF36PO_EMO_WORKED_LESS_PO	CO	93	ordinal
SF36PO_EMO_CAREFULLY_PO	CP	94	ordinal
SF36PO_PHYS_EMO_INTERFERENCE_PO	CQ	95	ordinal
SF36PO_HAD_PAIN_PO	CR	96	ordinal
SF36PO_PAIN_WORK_PO	CS	97	ordinal
SF36PO_VITALITY_PO	CT	98	ordinal
SF36PO_NERVOUS_PO	CU	99	ordinal
SF36PO_DEPRESSED_PO	CV	100	ordinal
SF36PO_CALM_PO	CW	101	ordinal
SF36PO_ENERGY_PO	CX	102	ordinal
SF36PO_SAD_PO	CY	103	ordinal
SF36PO_EXHAUSTED_PO	CZ	104	ordinal
SF36PO_HAPPY_PO	DA	105	ordinal
SF36PO_TIRED_PO	DB	106	ordinal
SF36PO_SOCIAL_ACTIVITY_PO	DC	107	ordinal
SF36PO_GOT_SICK_PO	DD	108	ordinal
SF36PO_HEALTHY_PO	DE	109	ordinal
SF36PO_HEALTH_WORSENING_PO	DF	110	ordinal
SF36PO_OPTIMAL_HEALTH_PO	DG	111	ordinal
LIVES_ALONE_PO	AKP	978	binary

DS_3_DAYS_QUESTIONNAIRE	L	N	Class
WALKS_3days	DH	112	character
WALK_FATIGUE_3days	DI	113	binary
WALK_MOBILITY_3days	DJ	114	binary
WALK_TRIED_3days	DK	115	binary
WALK_IMPORTANT_3days	DL	116	binary
WALK_OTHER_3days	DM	117	binary
STAIRS_3days	DN	118	character
STAIRS_FATIGUE_3days	DO	119	binary
STAIRS_MOBILITY_3days	DP	120	binary
STAIRS_TRIED_3days	DQ	121	binary
STAIRS_IMPORTANT_3days	DR	122	binary
STAIRS_OTHER_3days	DS	123	binary
PARTIAL_DEPENDENCY_3days	DT	124	binary
FATIGUE_3days	DU	125	ordinal
ACTIVITY_AUTONOMY_3days	DV	126	binary
ACTIVITY_INCREASE_3days	DW	127	binary
ACTIVITY_REST_3days	DX	128	binary
ACTIVITY_DRIVING_3days	DY	129	binary
ACTIVITY_OTHER_3days	DZ	130	binary
PAIN_PRESENCE_3days	EA	131	binary
PAIN_LOCATION_3days	EB	132	character
PAIN_SCALE_3days	EC	133	ordinal
PRECORDIAL_PAIN_3days	ED	134	binary
MUSCULOSKELETAL_PAIN_3days	EE	135	binary
INCISIONAL_PAIN_3days	EF	136	binary
PAIN_TYPE_OTHER_3days	EG	137	binary
ANALGESIC_THERAPY_3days	EH	138	binary
COMFORT_POSITION_3days	EI	139	binary
MASSAGE_3days	EJ	140	binary
PAIN_ACTION_OTHER_3days	EK	141	binary
ABSENT_3days	EL	142	binary
ANALGESIC_INSTRUCTIONS_3days	EM	143	binary
ANALGESIC_INCENTIVE_3days	EN	144	binary
PAIN_RELIEF_TEACHING_3days	EO	145	binary
PAIN_INTERVENTION_OTHER_3days	EP	146	binary
PHYSICAL_EMOTIONAL_STATE_3days	EQ	147	binary
SLEEP_PATTERN_CHANGE_3days	ER	148	binary
CONSTIPATION_3days	ES	149	binary
PHYSICAL_EMOTIONAL_OTHER_3days	ET	150	binary
SLEEP_INTERVENTION_GENERAL_3days	EU	151	binary
INTESTINAL_INTERVENTION_GENERAL_3days	EV	152	binary
GENERAL_INTERVENTION_OTHER_3days	EW	153	binary
SUTURE_CLEANING_3days	EX	154	ordinal
NO_SUTURE_CLEANING_KNOWLEDGE_3days	EY	155	binary
NO_SUTURE_CLEANING_BELIEF_3days	EZ	156	binary
NO_SUTURE_CLEANING_OTHER_3days	FA	157	binary
WOUND_EVOLUTION_3days	FB	158	ordinal
HEALED_3days	FC	159	binary
HEALING_PROCESS_3days	FD	160	binary
WOUND_INFLAMMATION_3days	FE	161	binary
WOUND_INFECTION_3days	FF	162	binary
WOUND_DEHISCENCE_3days	FG	163	binary
WOUND_EXUDATE_3days	FH	164	binary
WOUND_OTHER_3days	FI	165	binary
WOUND_HYDRATION_3days	FJ	166	ordinal
ECONOMIC_HYDRATION_3days	FK	167	binary
BELIEF_HYDRATION_3days	FL	168	binary
KNOWLEDGE_HYDRATION_3days	FM	169	binary
NON_HEALED_WOUND_HYDRATION_3days	FN	170	binary
HYDRATION_OTHER_3days	FO	171	binary
SUN_PROTECTION_WOUND_3days	FP	172	ordinal
SUN_PROTECTION_BELIEF_3days	FQ	173	ordinal
SUN_PROTECTION_KNOWLEDGE_3days	FR	174	binary
SUN_PROTECTION_OTHER_3days	FS	175	binary
LIMB_ELEVATION_3days	FT	176	ordinal
ELEVATION_BELIEF_3days	FU	177	binary
ELEVATION_KNOWLEDGE_3days	FV	178	binary
ELEVATION_NO_EDEMA_3days	FW	179	binary
ELEVATION_OTHER_3days	FX	180	binary
SUTURE_INTERVENTION_INSTRUCTION_3days	FY	181	binary
SUTURE_INTERVENTION_TEACHING_3days	FZ	182	binary
SUTURE_INTERVENTION_HYDRATION_3days	GA	183	binary
SUTURE_INTERVENTION_SUN_PROTECTION_3days	GB	184	binary
SUTURE_INTERVENTION_ELEVATION_3days	GC	185	binary
SUTURE_INTERVENTION_CROSS_3days	GD	186	binary
SUTURE_INTERVENTION_OTHER_3days	GE	187	binary
MORE_FIVE_MEALS_3days	GF	188	ordinal
TWO_FRUITS_3days	GG	189	ordinal

DS_30_DAYS_QUESTIONNAIRE	L	N	Class
WALKS_30days	KD	290	character
WALK_FATIGUE_30days	KE	291	binary
WALK_MOBILITY_30days	KF	292	binary
WALK_TRIED_30days	KG	293	binary
WALK_IMPORTANT_30days	KH	294	binary
WALK_OTHER_30days	KI	295	binary
STAIRS_30days	KJ	296	character
STAIRS_FATIGUE_30days	KK	297	binary
STAIRS_MOBILITY_30days	KL	298	binary
STAIRS_TRIED_30days	KM	299	binary
STAIRS_IMPORTANT_30days	KN	300	binary
STAIRS_OTHER_30days	KO	301	binary
DAILY_ACTIVITY_30days	KP	302	character
PARTIAL_DEPENDENCY_30days	KQ	303	binary
FATIGUE_30days	KR	304	ordinal
ACTIVITY_AUTONOMY_30days	KS	305	binary
ACTIVITY_INCREASE_30days	KT	306	binary
ACTIVITY_REST_30days	KU	307	binary
ACTIVITY_DRIVING_30days	KV	308	binary
ACTIVITY_OTHER_30days	KW	309	binary
WORK_SITUATION_30days	KX	310	character
PAIN_PRESENCE_30days	KY	311	binary
PAIN_LOCATION_30days	KZ	312	character
PAIN_SCALE_30days	LA	313	ordinal
PRECORDIAL_PAIN_30days	LB	314	binary
MUSCULOSKELETAL_PAIN_30days	LC	315	binary
PAIN_TYPE_OTHER_30days	LD	316	binary
ANALGESIC_THERAPY_30days	LE	317	binary
COMFORT_POSITION_30days	LF	318	binary
MASSAGE_30days	LG	319	binary
PAIN_ACTION_OTHER_30days	LH	320	binary
ABSENT_30days	LI	321	binary
ANALGESIC_INSTRUCTIONS_30days	LJ	322	binary
ANALGESIC_INCENTIVE_30days	LK	323	binary
PAIN_RELIEF_TEACHING_30days	LL	324	binary
PAIN_INTERVENTION_OTHER_30days	LM	325	binary
SUTURE_CLEANING_30days	LN	326	ordinal
NO_SUTURE_CLEANING_KNOWLEDGE_30days	LO	327	binary
NO_SUTURE_CLEANING_BELIEF_30days	LP	328	binary
NO_SUTURE_CLEANING_OTHER_30days	LQ	329	binary
WOUND_EVOLUTION_30days	LR	330	ordinal
HEALED_30days	LS	331	binary
HEALING_PROCESS_30days	LT	332	binary
WOUND_INFLAMMATION_30days	LU	333	binary
WOUND_INFECTION_30days	LV	334	binary
WOUND_DEHISCENCE_30days	LW	335	binary
WOUND_EXUDATE_30days	LX	336	binary
WOUND_OTHER_30days	LY	337	binary
WOUND_HYDRATION_30days	LZ	338	ordinal
ECONOMIC_HYDRATION_30days	MA	339	binary
BELIEF_HYDRATION_30days	MB	340	binary
KNOWLEDGE_HYDRATION_30days	MC	341	binary
NON_HEALED_WOUND_HYDRATION_30days	MD	342	binary
HYDRATION_OTHER_30days	ME	343	binary
SUN_PROTECTION_WOUND_30days	MF	344	ordinal
SUN_PROTECTION_BELIEF_30days	MG	345	ordinal
SUN_PROTECTION_KNOWLEDGE_30days	MH	346	binary
SUN_PROTECTION_OTHER_30days	MI	347	binary
LIMB_ELEVATION_30days	MJ	348	ordinal
ELEVATION_BELIEF_30days	MK	349	binary
ELEVATION_KNOWLEDGE_30days	ML	350	binary
ELEVATION_NO_EDEMA_30days	MM	351	binary
ELEVATION_OTHER_30days	MN	352	binary
SUTURE_INTERVENTION_INSTRUCTION_30days	MO	353	binary
SUTURE_INTERVENTION_TEACHING_30days	MP	354	binary
COLLOID_INTERVENTION_30days	MQ	355	binary
SUTURE_INTERVENTION_HYDRATION_30days	MR	356	binary
SUTURE_INTERVENTION_SUN_PROTECTION_30days	MS	357	binary
SUTURE_INTERVENTION_ELEVATION_30days	MT	358	binary
SUTURE_INTERVENTION_CROSS_30days	MU	359	binary
SUTURE_INTERVENTION_OTHER_30days	MV	360	binary
MORE_FIVE_MEALS_30days	MW	361	ordinal
TWO_FRUITS_30days	MX	362	ordinal
TWO_VEGETABLES_30days	MY	363	ordinal
VARIED_DIET_30days	MZ	364	ordinal
REDUCE_SALT_30days	NA	365	ordinal
DRINK_WATER_30days	NB	366	ordinal
DRINK_WINE_BEER_30days	NC	367	ordinal

TWO_VEGETABLES_3days	GH	190	ordinal
VARIED_DIET_3days	GI	191	ordinal
REDUCE_SALT_3days	GJ	192	ordinal
DRINK_WATER_3days	GK	193	ordinal
DRINK_WINE_BEER_3days	GL	194	ordinal
WEIGHT_3days	GM	195	numeric
HEIGHT_3days	GN	196	numeric
DIET_INSTRUCTIONS_3days	GO	197	binary
HYDRATION_INSTRUCTIONS_3days	GP	198	binary
ALCOHOL_INSTRUCTIONS_3days	GQ	199	binary
DIET_INTERVENTION_OTHER_3days	GR	200	binary
VASODILATOR_3days	GS	201	binary
VASODILATOR_TYPE_3days	GT	202	character
ANTI_ARRHYTHMIC_3days	GU	203	binary
ANTI_ARRHYTHMIC_TYPE_3days	GV	204	character
ANTI_HYPERTENSIVE_3days	GW	205	binary
ANTI_HYPERTENSIVE_TYPE_3days	GX	206	character
ANTIPLATELETS_3days	GY	207	binary
ANTIPLATELETS_TYPE_3days	GZ	208	character
ANTICOAGULANTS_3days	HA	209	binary
ANTICOAGULANTS_TYPE_3days	HB	210	character
ANTI_DYSLEPIDEMIC_3days	HC	211	binary
ANTI_DYSLEPIDEMIC_TYPE_3days	HD	212	character
OTHER_THERAPY_3days	HE	213	binary
OTHER_THERAPY_3days	HF	214	character
THERAPY_FORGOT_3days	HG	215	ordinal
THERAPY_NEGLECT_3days	HH	216	ordinal
THERAPY_IMPROVED_3days	HI	217	ordinal
THERAPY_WORSENED_3days	HJ	218	ordinal
THERAPY_TOOK_MORE_3days	HK	219	binary
THERAPY_INTERRUPTED_3days	HL	220	ordinal
THERAPY_CHANGED_OTHER_3days	HM	221	ordinal
THERAPY_CHANGED_ROUTINE_3days	HN	222	ordinal
THERAPY_CHANGED_COSTS_3days	HO	223	ordinal
THERAPY_CHANGED_EFFECTS_3days	HP	224	binary
THERAPY_CHANGED_MANY_3days	HQ	225	binary
THERAPY_CHANGED_USEFUL_3days	HR	226	binary
THERAPY_CHANGED_REMINDER_3days	HS	227	binary
THERAPY_CHANGED_PRESCRIPTION_3days	HT	228	ordinal
THERAPY_CHANGED_LONG_3days	HU	229	ordinal
THERAPY_CHANGED_CONFIDENCE_3days	HV	230	ordinal
THERAPY_CHANGED_WANTED_3days	HW	231	ordinal
THERAPY_CHANGED_REASON_3days	HX	232	binary
ANTICOAGULANT_SCHEDULE_3days	HY	233	binary
ANTICOAGULANT_DOSE_3days	HZ	234	binary
ANTICOAGULANT_INR_3days	IA	235	binary
ANTICOAGULANT_GENERAL_3days	IB	236	binary
ANTICOAGULANT_INSTRUCTIONS_3days	IC	237	binary
ANTICOAGULANT_TEACHING_3days	ID	238	binary
THERAPY_PRESCRIPTION_TEACHING_3days	IE	239	binary
PRESCRIPTION_NAME_TEACHING_3days	IF	240	binary
PRESCRIPTION_ACTION_TEACHING_3days	IG	241	binary
PRESCRIPTION_POSOLOGY_TEACHING_3days	IH	242	binary
PRESCRIPTION_EFFECTS_TEACHING_3days	II	243	binary
PRESCRIPTION_OTHER_TEACHING_3days	IJ	244	binary
THERAPY_BENEFITS_TEACHING_3days	IK	245	binary
THERAPY_INTERVENTION_OTHER_3days	IL	246	binary
BLOOD_PRESSURE_3days	IM	247	ordinal
BLOOD_PRESSURE_MEASURES_3days	IN	248	binary
BLOOD_PRESSURE_KNOWLEDGE_3days	IO	249	binary
BLOOD_PRESSURE_BELIEF_3days	IP	250	binary
BLOOD_PRESSURE_OTHER_3days	IQ	251	binary
HEART_RATE_3days	IR	252	ordinal
HEART_RATE_MEASURES_3days	IS	253	binary
HEART_RATE_KNOWLEDGE_3days	IT	254	binary
HEART_RATE_BELIEF_3days	IU	255	binary
HEART_RATE_OTHER_3days	IV	256	binary
TEMPERATURE_3days	IW	257	ordinal
TEMPERATURE_MEASURES_3days	IX	258	binary
TEMPERATURE_KNOWLEDGE_3days	IY	259	binary
TEMPERATURE_BELIEF_3days	IZ	260	binary
TEMPERATURE_OTHER_3days	JA	261	binary
BLOOD_SUGAR_3days	JB	262	ordinal
BLOOD_SUGAR_NOT_APPLICABLE_3days	JC	263	binary
BLOOD_SUGAR_MEASURES_3days	JD	264	binary
BLOOD_SUGAR_KNOWLEDGE_3days	JE	265	binary
BLOOD_SUGAR_BELIEF_3days	JF	266	binary
BLOOD_SUGAR_OTHER_3days	JG	267	binary
MONITORING_TEACHING_3days	JH	268	binary

WEIGHT_30days	ND	368	numeric
HEIGHT_30days	NE	369	numeric
DIET_INSTRUCTIONS_30days	NF	370	binary
HYDRATION_INSTRUCTIONS_30days	NG	371	binary
ALCOHOL_INSTRUCTIONS_30days	NH	372	binary
DIET_INTERVENTION_OTHER_30days	NI	373	binary
SMOKING_HABITS_30days	NJ	374	binary
NUMBER_CIGARETTES_30days	NK	375	numeric
SMOKING_CESSATION_30days	NL	376	binary
CONSULTATION_CESSATION_30days	NM	377	binary
MEDICAL_CESSATION_30days	NN	378	ordinal
CESSATION_ALONE_30days	NO	379	binary
VASODILATOR_30days	NP	380	binary
VASODILATOR_TYPE_30days	NQ	381	character
ANTI_ARRHYTHMIC_30days	NR	382	binary
ANTI_ARRHYTHMIC_TYPE_30days	NS	383	character
ANTI_HYPERTENSIVE_30days	NT	384	binary
ANTI_HYPERTENSIVE_TYPE_30days	NU	385	character
ANTIPLATELETS_30days	NV	386	binary
ANTIPLATELETS_TYPE_30days	NW	387	character
ANTICOAGULANTS_30days	NX	388	binary
ANTICOAGULANTS_TYPE_30days	NY	389	character
ANTI_DYSLEPIDEMIC_30days	NZ	390	binary
ANTI_DYSLEPIDEMIC_TYPE_30days	OA	391	character
OTHER_THERAPY_30days	OB	392	binary
OTHER_THERAPY_30days	OC	393	character
THERAPY_FORGOT_30days	OD	394	ordinal
THERAPY_NEGLECT_30days	OE	395	ordinal
THERAPY_IMPROVED_30days	OF	396	binary
THERAPY_WORSENED_30days	OG	397	ordinal
THERAPY_TOOK_MORE_30days	OH	398	ordinal
THERAPY_INTERRUPTED_30days	OI	399	ordinal
THERAPY_CHANGED_OTHER_30days	OJ	400	ordinal
THERAPY_CHANGED_ROUTINE_30days	OK	401	binary
THERAPY_CHANGED_COSTS_30days	OL	402	ordinal
THERAPY_CHANGED_EFFECTS_30days	OM	403	binary
THERAPY_CHANGED_MANY_30days	ON	404	binary
THERAPY_CHANGED_USEFUL_30days	OO	405	binary
THERAPY_CHANGED_REMINDER_30days	OP	406	binary
THERAPY_CHANGED_PRESCRIPTION_30days	OQ	407	binary
THERAPY_CHANGED_LONG_30days	OR	408	ordinal
THERAPY_CHANGED_CONFIDENCE_30days	OS	409	ordinal
THERAPY_CHANGED_WANTED_30days	OT	410	ordinal
THERAPY_CHANGED_REASON_30days	OU	411	binary
ANTICOAGULANT_SCHEDULE_30days	OV	412	binary
ANTICOAGULANT_DOSE_30days	OW	413	binary
ANTICOAGULANT_INR_30days	OX	414	binary
ANTICOAGULANT_GENERAL_30days	OY	415	binary
ANTICOAGULANT_INSTRUCTIONS_30days	OZ	416	binary
ANTICOAGULANT_TEACHING_30days	PA	417	binary
THERAPY_PRESCRIPTION_TEACHING_30days	PB	418	binary
PRESCRIPTION_NAME_TEACHING_30days	PC	419	binary
PRESCRIPTION_ACTION_TEACHING_30days	PD	420	binary
PRESCRIPTION_POSOLOGY_TEACHING_30days	PE	421	binary
PRESCRIPTION_EFFECTS_TEACHING_30days	PF	422	binary
PRESCRIPTION_OTHER_TEACHING_30days	PG	423	binary
THERAPY_BENEFITS_TEACHING_30days	PH	424	binary
THERAPY_INTERVENTION_OTHER_30days	PI	425	binary
BLOOD_PRESSURE_30days	PJ	426	ordinal
BLOOD_PRESSURE_MEASURES_30days	PK	427	binary
BLOOD_PRESSURE_KNOWLEDGE_30days	PL	428	binary
BLOOD_PRESSURE_BELIEF_30days	PM	429	binary
BLOOD_PRESSURE_OTHER_30days	PN	430	binary
HEART_RATE_30days	PO	431	ordinal
HEART_RATE_MEASURES_30days	PP	432	binary
HEART_RATE_KNOWLEDGE_30days	PQ	433	binary
HEART_RATE_BELIEF_30days	PR	434	binary
HEART_RATE_OTHER_30days	PS	435	binary
TEMPERATURE_30days	PT	436	ordinal
TEMPERATURE_MEASURES_30days	PU	437	binary
TEMPERATURE_KNOWLEDGE_30days	PV	438	binary
TEMPERATURE_BELIEF_30days	PW	439	binary
TEMPERATURE_OTHER_30days	PX	440	binary
BLOOD_SUGAR_30days	PY	441	ordinal
BLOOD_SUGAR_NOT_APPLICABLE_30days	PZ	442	binary
BLOOD_SUGAR_MEASURES_30days	QA	443	binary
BLOOD_SUGAR_KNOWLEDGE_30days	QB	444	binary
BLOOD_SUGAR_BELIEF_30days	QC	445	binary
BLOOD_SUGAR_OTHER_30days	QD	446	binary

MONITORING_INSTRUCTIONS_3days	JI	269	binary
MONITORING_OTHER_3days	JJ	270	binary
SPOUSE_CAREGIVER_3days	JK	271	binary
CHILD_CAREGIVER_3days	JL	272	binary
OTHER_CAREGIVER_3days	JM	273	binary
FAMILY_SUPPORT_3days	JN	274	binary
FORMAL_CAREGIVER_3days	JO	275	binary
INFORMAL_CAREGIVER_3days	JP	276	binary
SURGEON_APPOINTMENT_3days	JQ	277	binary
CARDIOLOGIST_APPOINTMENT_3days	JR	278	binary
DOCTOR_APPOINTMENT_3days	JS	279	binary
NURSE_APPOINTMENT_3days	JT	280	binary
OTHER_APPOINTMENT_3days	JU	281	binary
REFERRAL_3days	JV	282	binary
SOCIAL_REFERRAL_3days	JW	283	binary
CARDIOLOGY_REFERRAL_3days	JX	284	binary
SURGERY_REFERRAL_3days	JY	285	binary
DIETETICS_REFERRAL_3days	JZ	286	ordinal
SMOKER_REFERRAL_3days	KA	287	binary
FAMILY_REFERRAL_3days	KB	288	binary
OTHER_REFERRAL_3days	KC	289	binary
LIVES_ALONE_3days	AKQ	979	binary
INDEPENDENT_DAILY_ACTIVITY_3days	AKR	980	binary
TOTAL_DEPENDENCY_3days	AKT	982	binary
DEPENDENT_TASKS_3days	AKU	983	character

MONITORING_TEACHING_30days	QE	447	binary
MONITORING_INSTRUCTIONS_30days	QF	448	binary
MONITORING_OTHER_30days	QG	449	binary
SURGEON_APPOINTMENT_30days	QH	450	binary
CARDIOLOGIST_APPOINTMENT_30days	QI	451	binary
DOCTOR_APPOINTMENT_30days	QJ	452	binary
NURSE_APPOINTMENT_30days	QK	453	binary
OTHER_APPOINTMENT_30days	QL	454	binary
REFERRAL_30days	QM	455	binary
SOCIAL_REFERRAL_30days	QN	456	binary
CARDIOLOGY_REFERRAL_30days	QO	457	binary
SURGERY_REFERRAL_30days	QP	458	binary
DIETETICS_REFERRAL_30days	QQ	459	ordinal
SMOKER_REFERRAL_30days	QR	460	binary
FAMILY_REFERRAL_30days	QS	461	binary
OTHER_REFERRAL_30days	QT	462	binary
INDEPENDENT_DAILY_ACTIVITY_30days	AKS	981	binary
TOTAL_DEPENDENCY_30days	AKV	984	binary
DEPENDENT_TASKS_30days	AKW	985	character

DS_3_MONTHS_QUESTIONNAIRE	L	N	Class
FATIGUE_3months	QU	463	ordinal
ACTIVITY_AUTONOMY_3months	QV	464	binary
ACTIVITY_INCREASE_3months	QW	465	binary
ACTIVITY_REST_3months	QX	466	binary
ACTIVITY_OTHER_3months	QY	467	binary
WORK_SITUATION_3months	QZ	468	character
PAIN_PRESENCE_3months	RA	469	binary
PAIN_LOCATION_3months	RB	470	character
PAIN_SCALE_3months	RC	471	ordinal
PRECORDIAL_PAIN_3months	RD	472	binary
MUSCULOSKELETAL_PAIN_3months	RE	473	binary
PAIN_TYPE_OTHER_3months	RF	474	binary
ANALGESIC_THERAPY_3months	RG	475	binary
COMFORT_POSITION_3months	RH	476	binary
MASSAGE_3months	RI	477	binary
PAIN_ACTION_OTHER_3months	RJ	478	binary
ABSENT_3months	RK	479	binary
ANALGESIC_INSTRUCTIONS_3months	RL	480	binary
ANALGESIC_INCENTIVE_3months	RM	481	binary
PAIN_RELIEF_TEACHING_3months	RN	482	binary
PAIN_INTERVENTION_OTHER_3months	RO	483	binary
WOUND_EVOLUTION_3months	RP	484	ordinal
HEALED_3months	RQ	485	binary
HEALING_PROCESS_3months	RR	486	binary
WOUND_INFLAMMATION_3months	RS	487	binary
WOUND_INFECTION_3months	RT	488	binary
WOUND_DEHISCENCE_3months	RU	489	binary
WOUND_EXUDATE_3months	RV	490	binary
WOUND_OTHER_3months	RW	491	binary
WOUND_HYDRATION_3months	RX	492	ordinal
ECONOMIC_HYDRATION_3months	RY	493	binary
BELIEF_HYDRATION_3months	RZ	494	binary
KNOWLEDGE_HYDRATION_3months	SA	495	binary
NON_HEALED_WOUND_HYDRATION_3months	SB	496	binary
HYDRATION_OTHER_3months	SC	497	binary
SUN_PROTECTION_WOUND_3months	SD	498	ordinal
SUN_PROTECTION_BELIEF_3months	SE	499	binary
SUN_PROTECTION_KNOWLEDGE_3months	SF	500	binary
SUN_PROTECTION_OTHER_3months	SG	501	binary
LIMB_ELEVATION_3months	SH	502	ordinal
ELEVATION_BELIEF_3months	SI	503	binary
ELEVATION_KNOWLEDGE_3months	SJ	504	binary
ELEVATION_NO_EDEMA_3months	SK	505	binary
ELEVATION_OTHER_3months	SL	506	binary
SUTURE_INTERVENTION_INSTRUCTION_3months	SM	507	binary
SUTURE_INTERVENTION_TEACHING_3months	SN	508	binary
SUTURE_INTERVENTION_HYDRATION_3months	SO	509	binary
SUTURE_INTERVENTION_SUN_PROTECTION_3months	SP	510	binary
SUTURE_INTERVENTION_ELEVATION_3months	SQ	511	binary
SUTURE_INTERVENTION_CROSS_3months	SR	512	binary
SUTURE_INTERVENTION_OTHER_3months	SS	513	binary
MORE_FIVE_MEALS_3months	ST	514	ordinal
TWO_FRUITS_3months	SU	515	ordinal
TWO_VEGETABLES_3months	SV	516	ordinal
VARIED_DIET_3months	SW	517	ordinal
REDUCE_SALT_3months	SX	518	ordinal
DRINK_WATER_3months	SY	519	ordinal
DRINK_WINE_BEER_3months	SZ	520	ordinal
WEIGHT_3months	TA	521	numeric
HEIGHT_3months	TB	522	numeric
DIET_INSTRUCTIONS_3months	TC	523	binary
HYDRATION_INSTRUCTIONS_3months	TD	524	binary
ALCOHOL_INSTRUCTIONS_3months	TE	525	binary
DIET_INTERVENTION_OTHER_3months	TF	526	binary
SMOKING_HABITS_3months	TG	527	binary
NUMBER_CIGARETTES_3months	TH	528	numeric
SMOKING_CESSATION_3months	TI	529	binary
CONSULTATION_CESSATION_3months	TJ	530	binary
MEDICAL_CESSATION_3months	TK	531	binary
CESSATION_ALONE_3months	TL	532	binary
VASODILATOR_3months	TM	533	binary
VASODILATOR_TYPE_3months	TN	534	character
ANTI_ARRHYTHMIC_3months	TO	535	binary
ANTI_ARRHYTHMIC_TYPE_3months	TP	536	character
ANTI_HYPERTENSIVE_3months	TQ	537	binary
ANTI_HYPERTENSIVE_TYPE_3months	TR	538	character
ANTIPLATELETS_3months	TS	539	binary
ANTIPLATELETS_TYPE_3months	TT	540	character
ANTICOAGULANTS_3months	TU	541	binary

DS_6_MONTHS_QUESTIONNAIRE	L	N	Class
FATIGUE_6months	XW	647	ordinal
ACTIVITY_AUTONOMY_6months	XX	648	binary
ACTIVITY_INCREASE_6months	XY	649	binary
ACTIVITY_REST_6months	XZ	650	binary
ACTIVITY_OTHER_6months	YA	651	binary
WORK_SITUATION_6months	YB	652	character
PAIN_PRESENCE_6months	YC	653	binary
PAIN_LOCATION_6months	YD	654	character
PAIN_SCALE_6months	YE	655	ordinal
PRECORDIAL_PAIN_6months	YF	656	binary
MUSCULOSKELETAL_PAIN_6months	YG	657	binary
PAIN_TYPE_OTHER_6months	YH	658	binary
ANALGESIC_THERAPY_6months	YI	659	binary
COMFORT_POSITION_6months	YJ	660	binary
MASSAGE_6months	YK	661	binary
PAIN_ACTION_OTHER_6months	YL	662	binary
ABSENT_6months	YM	663	binary
ANALGESIC_INSTRUCTIONS_6months	YN	664	binary
ANALGESIC_INCENTIVE_6months	YO	665	binary
PAIN_RELIEF_TEACHING_6months	YP	666	binary
PAIN_INTERVENTION_OTHER_6months	YQ	667	binary
SUN_PROTECTION_WOUND_6months	YR	668	ordinal
SUN_PROTECTION_BELIEF_6months	YS	669	binary
SUN_PROTECTION_KNOWLEDGE_6months	YT	670	binary
SUN_PROTECTION_OTHER_6months	YU	671	binary
SUTURE_INTERVENTION_INSTRUCTION_6months	YV	672	binary
SUTURE_INTERVENTION_TEACHING_6months	YW	673	binary
SUTURE_INTERVENTION_HYDRATION_6months	YX	674	binary
SUTURE_INTERVENTION_SUN_PROTECTION_6months	YY	675	binary
SUTURE_INTERVENTION_ELEVATION_6months	YZ	676	binary
SUTURE_INTERVENTION_CROSS_6months	ZA	677	binary
SUTURE_INTERVENTION_OTHER_6months	ZB	678	binary
MORE_FIVE_MEALS_6months	ZC	679	ordinal
TWO_FRUITS_6months	ZD	680	ordinal
TWO_VEGETABLES_6months	ZE	681	ordinal
VARIED_DIET_6months	ZF	682	ordinal
REDUCE_SALT_6months	ZG	683	ordinal
DRINK_WATER_6months	ZH	684	ordinal
DRINK_WINE_BEER_6months	ZI	685	ordinal
WEIGHT_6months	ZJ	686	numeric
HEIGHT_6months	ZK	687	numeric
DIET_INSTRUCTIONS_6months	ZL	688	binary
HYDRATION_INSTRUCTIONS_6months	ZM	689	binary
ALCOHOL_INSTRUCTIONS_6months	ZN	690	binary
DIET_INTERVENTION_OTHER_6months	ZO	691	binary
SMOKING_HABITS_6months	ZP	692	binary
NUMBER_CIGARETTES_6months	ZQ	693	numeric
SMOKING_CESSATION_6months	ZR	694	binary
CONSULTATION_CESSATION_6months	ZS	695	binary
MEDICAL_CESSATION_6months	ZT	696	ordinal
CESSATION_ALONE_6months	ZU	697	binary
VASODILATOR_6months	ZV	698	binary
VASODILATOR_TYPE_6months	ZW	699	character
ANTI_ARRHYTHMIC_6months	ZX	700	binary
ANTI_ARRHYTHMIC_TYPE_6months	ZY	701	character
ANTI_HYPERTENSIVE_6months	ZZ	702	binary
ANTI_HYPERTENSIVE_TYPE_6months	AAA	703	character
ANTIPLATELETS_6months	AAB	704	binary
ANTIPLATELETS_TYPE_6months	AAC	705	character
ANTICOAGULANTS_6months	AAD	706	binary
ANTICOAGULANTS_TYPE_6months	AAE	707	character
ANTI_DYSLEPIDEMIC_6months	AAF	708	binary
ANTI_DYSLEPIDEMIC_TYPE_6months	AAG	709	character
OTHER_THERAPY_6months	AAH	710	binary
OTHER_THERAPY_6months	AAI	711	character
THERAPY_FORGOT_6months	AAJ	712	ordinal
THERAPY_NEGLECT_6months	AAK	713	ordinal
THERAPY_IMPROVED_6months	AAL	714	ordinal
THERAPY_WORSENED_6months	AAM	715	ordinal
THERAPY_TOOK_MORE_6months	AAN	716	ordinal
THERAPY_INTERRUPTED_6months	AAO	717	ordinal
THERAPY_CHANGED_OTHER_6months	AAP	718	ordinal
THERAPY_CHANGED_ROUTINE_6months	AAQ	719	binary
THERAPY_CHANGED_COSTS_6months	AAR	720	binary
THERAPY_CHANGED_EFFECTS_6months	AAS	721	binary
THERAPY_CHANGED_MANY_6months	AAT	722	ordinal
THERAPY_CHANGED_USEFUL_6months	AAU	723	ordinal
THERAPY_CHANGED_REMINDER_6months	AAV	724	binary
THERAPY_CHANGED_PRESCRIPTION_6months	AAW	725	binary

ANTICOAGULANTS_TYPE_3months	TV	542	character
ANTI_DYSLEPIDEMIC_3months	TW	543	binary
ANTI_DYSLEPIDEMIC_TYPE_3months	TX	544	character
OTHER_THERAPY_3months	TY	545	binary
OTHER_THERAPY_3months	TZ	546	character
THERAPY_FORGOT_3months	UA	547	ordinal
THERAPY_NEGLECT_3months	UB	548	ordinal
THERAPY_IMPROVED_3months	UC	549	ordinal
THERAPY_WORSENED_3months	UD	550	ordinal
THERAPY_TOOK_MORE_3months	UE	551	ordinal
THERAPY_INTERRUPTED_3months	UF	552	ordinal
THERAPY_CHANGED_OTHER_3months	UG	553	ordinal
THERAPY_CHANGED_ROUTINE_3months	UH	554	binary
THERAPY_CHANGED_COSTS_3months	UI	555	ordinal
THERAPY_CHANGED_EFFECTS_3months	UJ	556	binary
THERAPY_CHANGED_MANY_3months	UK	557	ordinal
THERAPY_CHANGED_USEFUL_3months	UL	558	binary
THERAPY_CHANGED_REMINDER_3months	UM	559	binary
THERAPY_CHANGED_PRESCRIPTION_3months	UN	560	binary
THERAPY_CHANGED_LONG_3months	UO	561	ordinal
THERAPY_CHANGED_CONFIDENCE_3months	UP	562	ordinal
THERAPY_CHANGED_WANTED_3months	UQ	563	binary
THERAPY_CHANGED_REASON_3months	UR	564	binary
ANTICOAGULANT_SCHEDULE_3months	US	565	binary
ANTICOAGULANT_DOSE_3months	UT	566	binary
ANTICOAGULANT_INR_3months	UU	567	binary
ANTICOAGULANT_GENERAL_3months	UV	568	binary
ANTICOAGULANT_INSTRUCTIONS_3months	UW	569	binary
ANTICOAGULANT_TEACHING_3months	UX	570	binary
THERAPY_PRESCRIPTION_TEACHING_3months	UY	571	binary
PRESCRIPTION_NAME_TEACHING_3months	UZ	572	binary
PRESCRIPTION_ACTION_TEACHING_3months	VA	573	binary
PRESCRIPTION_POSOLOGY_TEACHING_3months	VB	574	binary
PRESCRIPTION_EFFECTS_TEACHING_3months	VC	575	binary
PRESCRIPTION_OTHER_TEACHING_3months	VD	576	binary
THERAPY_BENEFITS_TEACHING_3months	VE	577	binary
THERAPY_INTERVENTION_OTHER_3months	VF	578	binary
BLOOD_PRESSURE_3months	VG	579	ordinal
BLOOD_PRESSURE_MEASURES_3months	VH	580	binary
BLOOD_PRESSURE_KNOWLEDGE_3months	VI	581	binary
BLOOD_PRESSURE_BELIEF_3months	VJ	582	binary
BLOOD_PRESSURE_OTHER_3months	VK	583	binary
HEART_RATE_3months	VL	584	ordinal
HEART_RATE_MEASURES_3months	VM	585	binary
HEART_RATE_KNOWLEDGE_3months	VN	586	binary
HEART_RATE_BELIEF_3months	VO	587	binary
HEART_RATE_OTHER_3months	VP	588	binary
BLOOD_SUGAR_3months	VQ	589	ordinal
BLOOD_SUGAR_NOT_APPLICABLE_3months	VR	590	binary
BLOOD_SUGAR_MEASURES_3months	VS	591	binary
BLOOD_SUGAR_KNOWLEDGE_3months	VT	592	binary
BLOOD_SUGAR_BELIEF_3months	VU	593	binary
BLOOD_SUGAR_OTHER_3months	VV	594	binary
MONITORING_TEACHING_3months	VW	595	binary
MONITORING_INSTRUCTIONS_3months	VX	596	binary
MONITORING_OTHER_3months	VY	597	binary
SURGEON_APPOINTMENT_3months	VZ	598	binary
CARDIOLOGIST_APPOINTMENT_3months	WA	599	binary
DOCTOR_APPOINTMENT_3months	WB	600	binary
NURSE_APPOINTMENT_3months	WC	601	binary
OTHER_APPOINTMENT_3months	WD	602	binary
REFERRAL_3months	WE	603	binary
SOCIAL_REFERRAL_3months	WF	604	binary
CARDIOLOGY_REFERRAL_3months	WG	605	binary
SURGERY_REFERRAL_3months	WH	606	binary
DIETETICS_REFERRAL_3months	WI	607	ordinal
SMOKER_REFERRAL_3months	WJ	608	binary
FAMILY_REFERRAL_3months	WK	609	binary
OTHER_REFERRAL_3months	WL	610	binary
SF36F3M_HEALTH_STATUS_3months	WM	611	ordinal
SF36F3M_ONE_YEAR_COMPARISON_3months	WN	612	ordinal
SF36F3M_VIGOROUS_ACTIVITY_3months	WO	613	ordinal
SF36F3M_MODERATE_ACTIVITY_3months	WP	614	ordinal
SF36F3M_SHOPPING_ACTIVITY_3months	WQ	615	ordinal
SF36F3M_MULTIPLE_FLIGHTS_ACTIVITY_3months	WR	616	ordinal
SF36F3M_ONE_FLIGHT_ACTIVITY_3months	WS	617	ordinal
SF36F3M_BENDING_ACTIVITY_3months	WT	618	ordinal
SF36F3M_MORE_THAN_1KM_ACTIVITY_3months	WU	619	ordinal
SF36F3M_HUNDREDS_METERS_ACTIVITY_3months	WV	620	ordinal
SF36F3M_100_METERS_ACTIVITY_3months	WW	621	ordinal

THERAPY_CHANGED_LONG_6months	AAX	726	ordinal
THERAPY_CHANGED_CONFIDENCE_6months	AAY	727	binary
THERAPY_CHANGED_WANTED_6months	AAZ	728	binary
THERAPY_CHANGED_REASON_6months	ABA	729	binary
ANTICOAGULANT_SCHEDULE_6months	ABB	730	binary
ANTICOAGULANT_DOSE_6months	ABC	731	binary
ANTICOAGULANT_INR_6months	ABD	732	binary
ANTICOAGULANT_GENERAL_6months	ABE	733	binary
ANTICOAGULANT_INSTRUCTIONS_6months	ABF	734	binary
ANTICOAGULANT_TEACHING_6months	ABG	735	binary
THERAPY_PRESCRIPTION_TEACHING_6months	ABH	736	binary
PRESCRIPTION_NAME_TEACHING_6months	ABI	737	binary
PRESCRIPTION_ACTION_TEACHING_6months	ABJ	738	binary
PRESCRIPTION_POSOLOGY_TEACHING_6months	ABK	739	binary
PRESCRIPTION_EFFECTS_TEACHING_6months	ABL	740	binary
PRESCRIPTION_OTHER_TEACHING_6months	ABM	741	binary
THERAPY_BENEFITS_TEACHING_6months	ABN	742	binary
THERAPY_INTERVENTION_OTHER_6months	ABO	743	binary
BLOOD_PRESSURE_6months	ABP	744	ordinal
BLOOD_PRESSURE_MEASURES_6months	ABQ	745	binary
BLOOD_PRESSURE_KNOWLEDGE_6months	ABR	746	binary
BLOOD_PRESSURE_BELIEF_6months	ABS	747	binary
BLOOD_PRESSURE_OTHER_6months	ABT	748	binary
HEART_RATE_6months	ABU	749	ordinal
HEART_RATE_MEASURES_6months	ABV	750	binary
HEART_RATE_KNOWLEDGE_6months	ABW	751	binary
HEART_RATE_BELIEF_6months	ABX	752	binary
HEART_RATE_OTHER_6months	ABY	753	binary
BLOOD_SUGAR_6months	ABZ	754	ordinal
BLOOD_SUGAR_NOT_APPLICABLE_6months	ACA	755	binary
BLOOD_SUGAR_MEASURES_6months	ACB	756	binary
BLOOD_SUGAR_KNOWLEDGE_6months	ACC	757	binary
BLOOD_SUGAR_BELIEF_6months	ACD	758	binary
BLOOD_SUGAR_OTHER_6months	ACE	759	binary
MONITORING_TEACHING_6months	ACF	760	binary
MONITORING_INSTRUCTIONS_6months	ACG	761	binary
MONITORING_OTHER_6months	ACH	762	binary
SURGEON_APPOINTMENT_6months	ACI	763	binary
CARDIOLOGIST_APPOINTMENT_6months	ACJ	764	binary
DOCTOR_APPOINTMENT_6months	ACK	765	binary
NURSE_APPOINTMENT_6months	ACL	766	binary
OTHER_APPOINTMENT_6months	ACM	767	binary
REFERRAL_6months	ACN	768	binary
SOCIAL_REFERRAL_6months	ACO	769	binary
CARDIOLOGY_REFERRAL_6months	ACP	770	binary
SURGERY_REFERRAL_6months	ACQ	771	binary
DIETETICS_REFERRAL_6months	ACR	772	binary
SMOKER_REFERRAL_6months	ACS	773	binary
FAMILY_REFERRAL_6months	ACT	774	binary
OTHER_REFERRAL_6months	ACU	775	binary
SF36F6M_HEALTH_STATUS_6months	ACV	776	ordinal
SF36F6M_ONE_YEAR_COMPARISON_6months	ACW	777	ordinal
SF36F6M_VIGOROUS_ACTIVITY_6months	ACX	778	ordinal
SF36F6M_MODERATE_ACTIVITY_6months	ACY	779	ordinal
SF36F6M_SHOPPING_ACTIVITY_6months	ACZ	780	ordinal
SF36F6M_MULTIPLE_FLIGHTS_ACTIVITY_6months	ADA	781	ordinal
SF36F6M_ONE_FLIGHT_ACTIVITY_6months	ADB	782	ordinal
SF36F6M_BENDING_ACTIVITY_6months	ADC	783	ordinal
SF36F6M_ACTIV_MOREKM_6months	ADD	784	ordinal
SF36F6M_HUNDREDS_METERS_ACTIVITY_6months	ADE	785	ordinal
SF36F6M_100_METERS_ACTIVITY_6months	ADF	786	ordinal
SF36F6M_BATHING_ACTIVITY_6months	ADG	787	binary
SF36F6M_F_WORKING_TIME_6months	ADH	788	ordinal
SF36F6M_PHYS_WORKED_LESS_6months	ADI	789	ordinal
SF36F6M_PHYS_LIMITED_WORK_6months	ADJ	790	ordinal
SF36F6M_PHYS_WORK_DIFFICULTY_6months	ADK	791	ordinal
SF36F6M_AND_WORKING_TIME_6months	ADL	792	ordinal
SF36F6M_EMO_WORKED_LESS_6months	ADM	793	ordinal
SF36F6M_EMO_CAREFULLY_6months	ADN	794	ordinal
SF36F6M_PHYS_EMO_INTERFERENCE_6months	ADO	795	ordinal
SF36F6M_HAD_PAIN_6months	ADP	796	ordinal
SF36F6M_PAIN_WORK_6months	ADQ	797	ordinal
SF36F6M_VITALITY_6months	ADR	798	ordinal
SF36F6M_NERVOUS_6months	ADS	799	ordinal
SF36F6M_DEPRESSED_6months	ADT	800	ordinal
SF36F6M_CALM_6months	ADU	801	ordinal
SF36F6M_ENERGY_6months	ADV	802	ordinal
SF36F6M_SAD_6months	ADW	803	ordinal
SF36F6M_EXHAUSTED_6months	ADX	804	ordinal
SF36F6M_HAPPY_6months	ADY	805	ordinal

SF36F3M_BATHING_ACTIVITY_3months	WX	622	ordinal
SF36F3M_F_WORKING_TIME_3months	WY	623	ordinal
SF36F3M_PHYS_WORKED_LESS_3months	WZ	624	ordinal
SF36F3M_PHYS_LIMITED_WORK_3months	XA	625	ordinal
SF36F3M_PHYS_WORK_DIFFICULTY_3months	XB	626	ordinal
SF36F3M_AND_WORKING_TIME_3months	XC	627	ordinal
SF36F3M_EMO_WORKED_LESS_3months	XD	628	ordinal
SF36F3M_EMO_CAREFULLY_3months	XE	629	ordinal
SF36F3M_PHYS_EMO_INTERFERENCE_3months	XF	630	ordinal
SF36F3M_HAD_PAIN_3months	XG	631	ordinal
SF36F3M_PAIN_WORK_3months	XH	632	ordinal
SF36F3M_VITALITY_3months	XI	633	ordinal
SF36F3M_NERVOUS_3months	XJ	634	ordinal
SF36F3M_DEPRESSED_3months	XK	635	ordinal
SF36F3M_CALM_3months	XL	636	ordinal
SF36F3M_ENERGY_3months	XM	637	ordinal
SF36F3M_SAD_3months	XN	638	ordinal
SF36F3M_EXHAUSTED_3months	XO	639	ordinal
SF36F3M_HAPPY_3months	XP	640	ordinal
SF36F3M_TIRED_3months	XQ	641	ordinal
SF36F3M_SOCIAL_ACTIVITY_3months	XR	642	ordinal
SF36F3M_GOT_SICK_3months	XS	643	ordinal
SF36F3M_HEALTHY_3months	XT	644	ordinal
SF36F3M_HEALTH_WORSENING_3months	XU	645	ordinal
SF36F3M_OPTIMAL_HEALTH_3months	XV	646	ordinal

SF36F6M_TIRED_6months	ADZ	806	ordinal
SF36F6M_SOCIAL_ACTIVITY_6months	AEA	807	ordinal
SF36F6M_GOT_SICK_6months	AEB	808	ordinal
SF36F6M_HEALTHY_6months	AEC	809	ordinal
SF36F6M_HEALTH_WORSENING_6months	AED	810	ordinal
SF36F6M_OPTIMAL_HEALTH_6months	AEE	811	ordinal

DS_1_YEAR_QUESTIONNAIRE	L	N	Class
FATIGUE_1year	AEF	812	ordinal
ACTIVITY_AUTONOMY_1year	AEG	813	binary
ACTIVITY_INCREASE_1year	AEH	814	binary
ACTIVITY_REST_1year	AEI	815	binary
ACTIVITY_OTHER_1year	AEJ	816	binary
WORK_SITUATION_1year	AEK	817	character
PAIN_PRESENCE_1year	AEL	818	binary
PAIN_LOCATION_1year	AEM	819	character
PAIN_SCALE_1year	AEN	820	ordinal
PRECORDIAL_PAIN_1year	AEO	821	binary
MUSCULOSKELETAL_PAIN_1year	AEP	822	binary
PAIN_TYPE_OTHER_1year	AEQ	823	binary
ANALGESIC_THERAPY_1year	AER	824	binary
COMFORT_POSITION_1year	AES	825	binary
MASSAGE_1year	AET	826	binary
PAIN_ACTION_OTHER_1year	AEU	827	binary
ABSENT_1year	AEV	828	binary
ANALGESIC_INSTRUCTIONS_1year	AEW	829	binary
ANALGESIC_INCENTIVE_1year	AEX	830	binary
PAIN_RELIEF_TEACHING_1year	AEY	831	binary
PAIN_INTERVENTION_OTHER_1year	AEZ	832	binary
SUN_PROTECTION_WOUND_1year	AFA	833	ordinal
SUN_PROTECTION_BELIEF_1year	AFB	834	binary
SUN_PROTECTION_KNOWLEDGE_1year	AFC	835	binary
SUN_PROTECTION_OTHER_1year	AFD	836	binary
SUTURE_INTERVENTION_INSTRUCTION_1year	AFE	837	binary
SUTURE_INTERVENTION_TEACHING_1year	AFF	838	binary
SUTURE_INTERVENTION_HYDRATION_1year	AFG	839	binary
SUTURE_INTERVENTION_SUN_PROTECTION_1year	AFH	840	binary
SUTURE_INTERVENTION_ELEVATION_1year	AFI	841	binary
SUTURE_INTERVENTION_CROSS_1year	AFJ	842	binary
SUTURE_INTERVENTION_OTHER_1year	AFK	843	binary
MORE_FIVE_MEALS_1year	AFL	844	ordinal
TWO_FRUITS_1year	AFM	845	ordinal
TWO_VEGETABLES_1year	AFN	846	ordinal
VARIED_DIET_1year	AFO	847	ordinal
REDUCE_SALT_1year	AFP	848	ordinal
DRINK_WATER_1year	AFQ	849	ordinal
DRINK_WINE_BEER_1year	AFR	850	ordinal
WEIGHT_1year	AFS	851	numeric
HEIGHT_1year	AFT	852	numeric
DIET_INSTRUCTIONS_1year	AFU	853	binary
HYDRATION_INSTRUCTIONS_1year	AFV	854	binary
ALCOHOL_INSTRUCTIONS_1year	AFW	855	binary
DIET_INTERVENTION_OTHER_1year	AFX	856	binary
SMOKING_HABITS_1year	AFY	857	binary
NUMBER_CIGARETTES_1year	AFZ	858	numeric
SMOKING_CESSATION_1year	AGA	859	binary
CONSULTATION_CESSATION_1year	AGB	860	binary

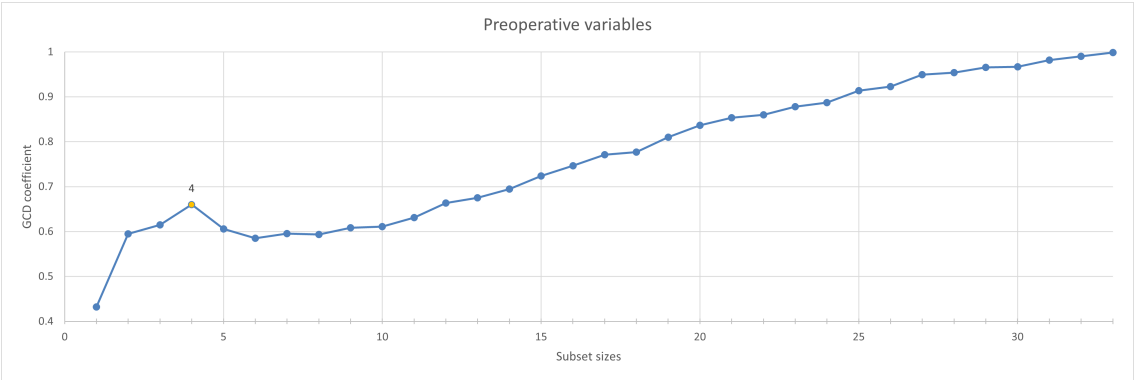
MEDICAL_CESSATION_1year	AGC	861	binary
CESSATION_ALONE_1year	AGD	862	binary
VASODILATOR_1year	AGE	863	binary
VASODILATOR_TYPE_1year	AGF	864	character
ANTI_ARRHYTHMIC_1year	AGG	865	binary
ANTI_ARRHYTHMIC_TYPE_1year	AGH	866	character
ANTI_HYPERTENSIVE_1year	AGI	867	binary
ANTI_HYPERTENSIVE_TYPE_1year	AGJ	868	character
ANTIPLATELETS_1year	AGK	869	binary
ANTIPLATELETS_TYPE_1year	AGL	870	character
ANTICOAGULANTS_1year	AGM	871	binary
ANTICOAGULANTS_TYPE_1year	AGN	872	character
ANTI_DYSLEPIDEMIC_1year	AGO	873	binary
ANTI_DYSLEPIDEMIC_TYPE_1year	AGP	874	character
OTHER_THERAPY_1year	AGQ	875	binary
OTHER_THERAPY_1year	AGR	876	character
THERAPY_FORGOT_1year	AGS	877	ordinal
THERAPY_NEGLECT_1year	AGT	878	ordinal
THERAPY_IMPROVED_1year	AGU	879	ordinal
THERAPY_WORSENED_1year	AGV	880	ordinal
THERAPY_TOOK_MORE_1year	AGW	881	ordinal
THERAPY_INTERRUPTED_1year	AGX	882	binary
THERAPY_CHANGED_OTHER_1year	AGY	883	ordinal
THERAPY_CHANGED_ROUTINE_1year	AGZ	884	ordinal
THERAPY_CHANGED_COSTS_1year	AHA	885	ordinal
THERAPY_CHANGED_EFFECTS_1year	AHB	886	ordinal
THERAPY_CHANGED_MANY_1year	AHC	887	binary
THERAPY_CHANGED_USEFUL_1year	AHD	888	ordinal
THERAPY_CHANGED_REMINDER_1year	AHE	889	binary
THERAPY_CHANGED_PRESCRIPTION_1year	AHF	890	ordinal
THERAPY_CHANGED_LONG_1year	AHG	891	ordinal
THERAPY_CHANGED_CONFIDENCE_1year	AHH	892	ordinal
THERAPY_CHANGED_WANTED_1year	AHI	893	binary
THERAPY_CHANGED_REASON_1year	AHJ	894	binary
ANTICOAGULANT_SCHEDULE_1year	AHK	895	binary
ANTICOAGULANT_DOSE_1year	AHL	896	binary
ANTICOAGULANT_INR_1year	AHM	897	binary
ANTICOAGULANT_GENERAL_1year	AHN	898	binary
ANTICOAGULANT_INSTRUCTIONS_1year	AHO	899	binary
ANTICOAGULANT_TEACHING_1year	AHP	900	binary
THERAPY_PRESCRIPTION_TEACHING_1year	AHQ	901	binary
PRESCRIPTION_NAME_TEACHING_1year	AHR	902	binary
PRESCRIPTION_ACTION_TEACHING_1year	AHS	903	binary
PRESCRIPTION_POSOLOGY_TEACHING_1year	AHT	904	binary
PRESCRIPTION_EFFECTS_TEACHING_1year	AHU	905	binary
PRESCRIPTION_OTHER_TEACHING_1year	AHV	906	binary
THERAPY_BENEFITS_TEACHING_1year	AHW	907	binary
THERAPY_INTERVENTION_OTHER_1year	AHX	908	binary
BLOOD_PRESSURE_1year	AHY	909	ordinal
BLOOD_PRESSURE_MEASURES_1year	AHZ	910	binary

BLOOD_PRESSURE_KNOWLEDGE_1year	AIA	911	binary
BLOOD_PRESSURE_BELIEF_1year	AIB	912	binary
BLOOD_PRESSURE_OTHER_1year	AIC	913	binary
HEART_RATE_1year	AID	914	ordinal
HEART_RATE_MEASURES_1year	AIE	915	binary
HEART_RATE_KNOWLEDGE_1year	AIF	916	binary
HEART_RATE_BELIEF_1year	AIG	917	binary
HEART_RATE_OTHER_1year	AIH	918	binary
BLOOD_SUGAR_1year	AII	919	ordinal
BLOOD_SUGAR_NOT_APPLICABLE_1year	AIJ	920	binary
BLOOD_SUGAR_MEASURES_1year	AIK	921	binary
BLOOD_SUGAR_KNOWLEDGE_1year	AIL	922	binary
BLOOD_SUGAR_BELIEF_1year	AIM	923	binary
BLOOD_SUGAR_OTHER_1year	AIN	924	binary
MONITORING_TEACHING_1year	AIO	925	binary
MONITORING_INSTRUCTIONS_1year	AIP	926	binary
MONITORING_OTHER_1year	AIQ	927	binary
SURGEON_APPOINTMENT_1year	AIR	928	binary
CARDIOLOGIST_APPOINTMENT_1year	AIS	929	binary
DOCTOR_APPOINTMENT_1year	AIT	930	binary
NURSE_APPOINTMENT_1year	AIU	931	binary
OTHER_APPOINTMENT_1year	AIV	932	binary
REFERRAL_1year	AIW	933	binary
SOCIAL_REFERRAL_1year	AIX	934	binary
CARDIOLOGY_REFERRAL_1year	AIY	935	binary
SURGERY_REFERRAL_1year	AIZ	936	binary
DIETETICS_REFERRAL_1year	AJA	937	ordinal
SMOKER_REFERRAL_1year	AJB	938	binary
FAMILY_REFERRAL_1year	AJC	939	binary
OTHER_REFERRAL_1year	AJD	940	binary
SF36F6M_HEALTH_STATUS_1year	AJE	941	ordinal
SF36F6M_ONE_YEAR_COMPARISON_1year	AJF	942	ordinal
SF36F6M_VIGOROUS_ACTIVITY_1year	AJG	943	ordinal
SF36F6M_MODERATE_ACTIVITY_1year	AJH	944	ordinal
SF36F6M_SHOPPING_ACTIVITY_1year	AJI	945	ordinal
SF36F6M_MULTIPLE_FLIGHTS_ACTIVITY_1year	AJJ	946	ordinal
SF36F6M_ONE_FLIGHT_ACTIVITY_1year	AJK	947	ordinal
SF36F6M_BENDING_ACTIVITY_1year	AJL	948	ordinal
SF36F6M_ACTIV_MOREKM_1year	AJM	949	ordinal
SF36F6M_HUNDREDS_METERS_ACTIVITY_1year	AJN	950	ordinal
SF36F6M_100_METERS_ACTIVITY_1year	AJO	951	ordinal
SF36F6M_BATHING_ACTIVITY_1year	AJP	952	ordinal
SF36F6M_F_WORKING_TIME_1year	AJQ	953	ordinal
SF36F6M_PHYS_WORKED_LESS_1year	AJR	954	ordinal
SF36F6M_PHYS_LIMITED_WORK_1year	AJS	955	ordinal
SF36F6M_PHYS_WORK_DIFFICULTY_1year	AJT	956	ordinal
SF36F6M_AND_WORKING_TIME_1year	AJU	957	ordinal
SF36F6M_EMO_WORKED_LESS_1year	AJV	958	ordinal
SF36F6M_EMO_CAREFULLY_1year	AJW	959	ordinal
SF36F6M_PHYS_EMO_INTERFERENCE_1year	AJX	960	ordinal

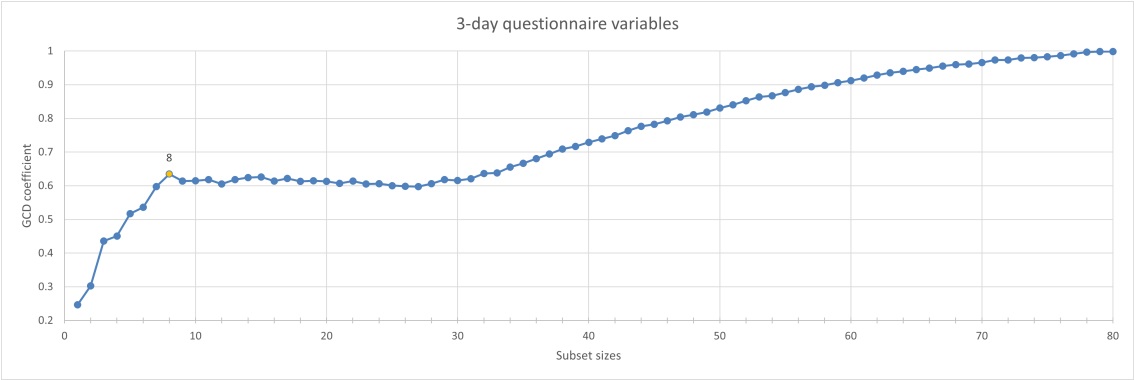
SF36F6M_HAD_PAIN_1year	AJY	961	ordinal
SF36F6M_PAIN_WORK_1year	AJZ	962	ordinal
SF36F6M_VITALITY_1year	AKA	963	ordinal
SF36F6M_NERVOUS_1year	AKB	964	ordinal
SF36F6M_DEPRESSED_1year	AKC	965	ordinal
SF36F6M_CALM_1year	AKD	966	ordinal
SF36F6M_ENERGY_1year	AKE	967	ordinal
SF36F6M_SAD_1year	AKF	968	ordinal
SF36F6M_EXHAUSTED_1year	AKG	969	ordinal
SF36F6M_HAPPY_1year	AKH	970	ordinal
SF36F6M_TIRED_1year	AKI	971	ordinal
SF36F6M_SOCIAL_ACTIVITY_1year	AKJ	972	ordinal
SF36F6M_GOT_SICK_1year	AKK	973	ordinal
SF36F6M_HEALTHY_1year	AKL	974	ordinal
SF36F6M_HEALTH_WORSENING_1year	AKM	975	ordinal
SF36F6M_OPTIMAL_HEALTH_1year	AKN	976	ordinal

B

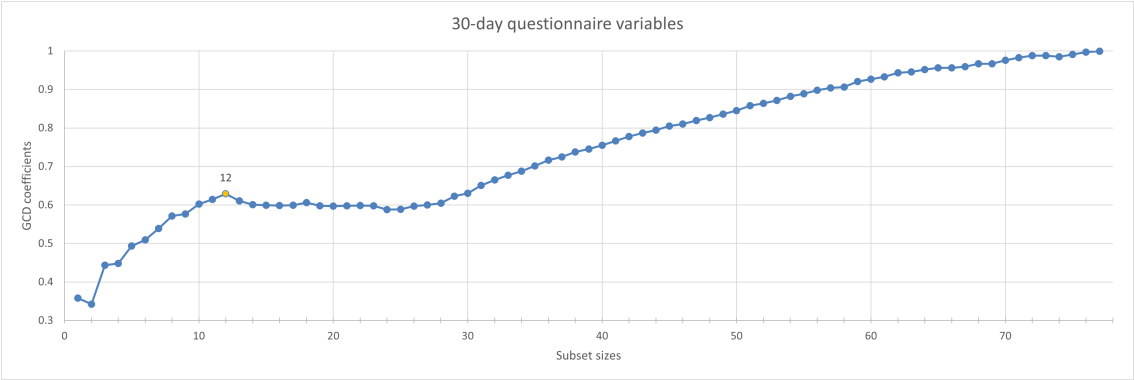
Appendix B: GCD coefficients for different follow-up periods



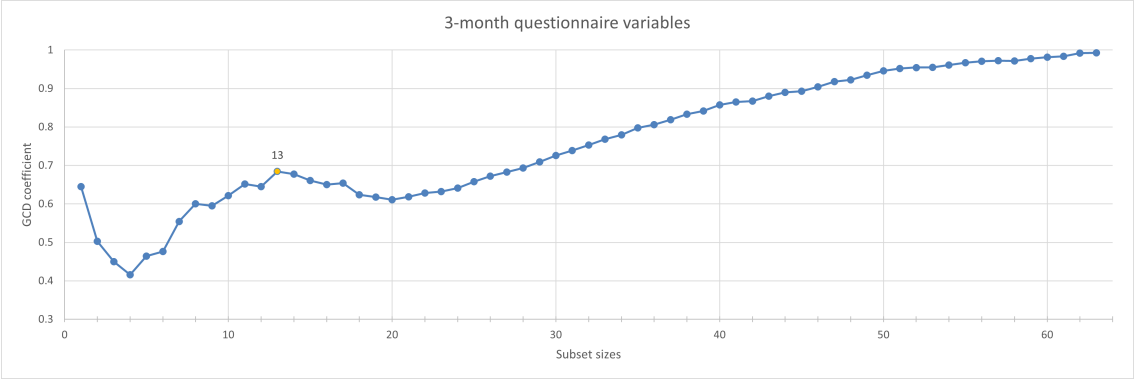
(a) GCD coefficients for preoperative subsets.



(b) GCD coefficients for the 3-day subsets.

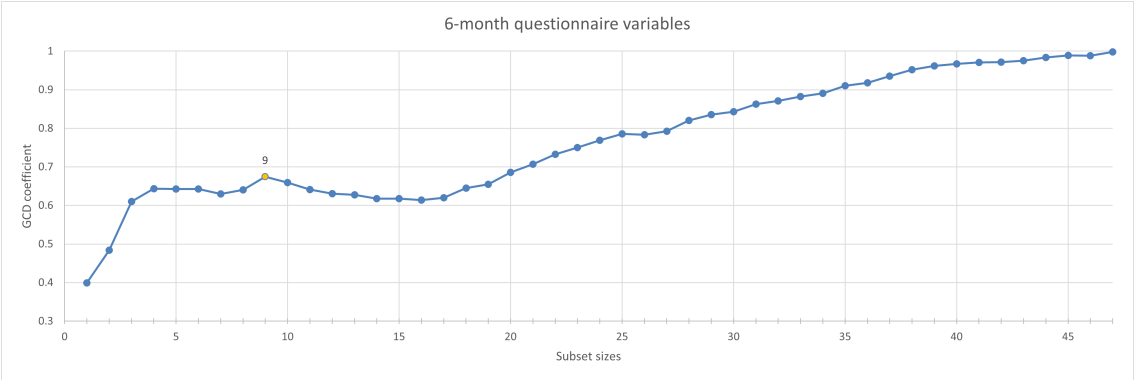


(c) GCD coefficients for the 30-day subsets.

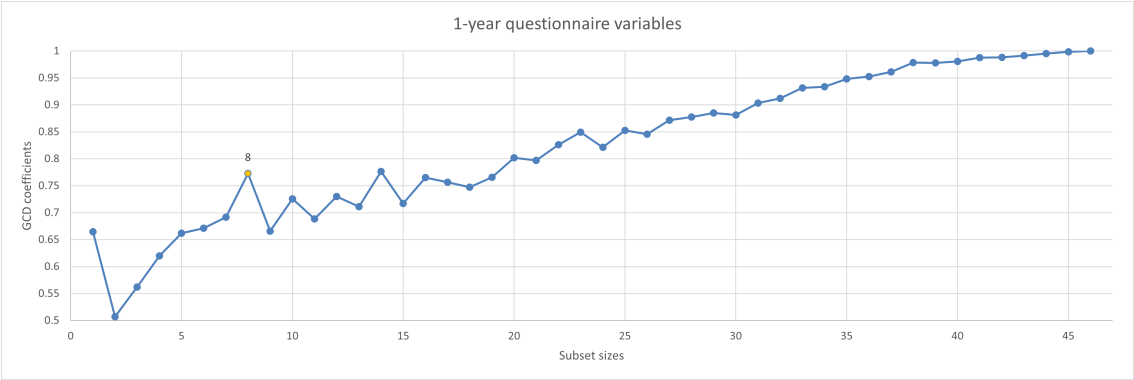


(d) GCD coefficients for the 3-month subsets.

Figure B.1: GCD coefficients for the different follow-up periods.



(e) GCD coefficients for the 6-month subsets.



(f) GCD coefficients for the 1-year subsets.

Figure B.1: GCD coefficients for the different follow-up periods.