FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Time-to-Event Prediction

Maria José Gomes Pedroto

Doctoral Program in Informatics Engineering

Supervisor: Prof. Dr. Alípio Mário Guedes Jorge

Co-Supervisor: Prof. Dr. João Pedro Carvalho Leal Mendes Moreira

August, 2024

Time-to-Event Prediction

Maria José Gomes Pedroto

Doctoral Program in Informatics Engineering

Dissertation submitted to Faculdade de Engenharia da Universidade do Porto to obtain the degree of Doctorate in Informatics Engineering

Approved in oral examination by the committee:

Chair: Doctor Carlos Miguel Ferraz Baquero-Moreno, University of Porto
External Examiner: Doctor Myra Spiliopoulou, University of Magdeburg, Germany
External Examiner: Doctor Manuel Filipe Vieira Torres dos Santos, University of Minho
Referee: Doctor Rui Carlos Camacho de Sousa Ferreira da Silva, University of Porto
Supervisor: Doctor Alípio Mário Guedes Jorge, University of Porto

August, 2024

Abstract

This work is centered on **modeling and predicting Time-to-Event (TTE) episodes**, and has two distinct purposes. The first purpose is to explore the usage of genealogical data for time-to-event prediction. Additionally, this work aims to aid medical professionals in improving the diagnosis and prognosis of patients afflicted with Hereditary Transthyretin Amyloidosis (ATTRv amyloidosis). This is a genetic disease with a strong historical background in the fishing villages of Póvoa do Varzim, in northern Portugal.

In order to explore the value of genealogical data for time-to-event prediction, this work has contributions in feature engineering, namely within the area of feature construction and feature selection. To this end, it explores and compares a summarizing approach focused on manually extracting meaningful features from genealogical trees, and which may require domain knowledge, with a more automated one using embeddings.

It contributes to model construction by creating a multivariate data-oriented approach that tracks a patient's risk of developing disease onset through different ages. In this case the implemented approach assumes as relevant differences in patient's age related characteristics. It also explores the impact of combining different age models of neighboring time windows in order to combine a set of base learners and construct a more reliable and less variance prone global model.

Finally, it contributes to model evaluation by addressing the implementation issues of a business-based approach to evaluate the expected return of changing clinical guidelines focused on defining how asymptomatic patients should be followed by interdisciplinary clinical teams. It also presents robust evaluation strategies that assist the multivariate data-oriented approach in selecting the best model. The evaluation strategies are also used to compare and evaluate the feature construction methods.

The application is focused on patients with ATTRv Amyloidosis.

To present and characterize the work done, this thesis is structured into four main sections. It begins with an introduction and a presentation of ATTRv Amyloidosis from a medically historic perspective. Then it presents the relevant background by dwelling into the connection of time-to-event prediction with feature engineering, model construction and model evaluation, as well as introducing key concepts of genealogical studies. After this, it presents its technical contributions, in the form of the main publications that constitute this work (one paper by chapter). It ends with an epilogue section which overviews the work performed, shares the main conclusions, and, finally, discusses the thesis from a technical and clinical perspective.

Keywords: Time-to-Event Data, Data Modeling, Regression Models, Decision Trees (DT), Elastic Net (EN), Lasso (LA), Linear Regression (LR), Random Forest (RF), Ridge (RI), Support Vector Regressor (SV), Extreme Gradient Boosting Trees (XG)

Resumo

Este trabalho foca-se em **modelar e prever episódios Time-to-Event (TTE)**, e tem dois propósitos. O primeiro é explorar o potencial de recorrer a dados genealógicos para previsão *time-to-event*. Adicionalmente, visa melhorar o diagnóstico feito pelos profissionais que trabalham com pacientes portadores de Hereditary Transthyretin Amyloidosis (ATTRv Amyloidosis). Esta é uma doença genética, hereditária, altamente debilitante, com um passado histórico que transporta para as aldeias piscatórias da Póvoa do Varzim, no norte de Portugal.

Para explorar o valor dos dados genealógicos para previsão *time-to-event*, este trabalho tem contribuições em *feature engineering*, nomeadamente na área de construção e seleção de *features*. Para este propósito, explora uma abordagem de sumarização focada na extração manual de características significativas de árvores genealógicas. De referir que esta abordagem pode necessitar de conhecimento de negócio específico. Posteriormente, compara esta abordagem com uma mais automatizada que usa *embeddings*.

Contribui para *model construction* criando uma abordagem multivariada, orientada a dados, para rastrear o risco de um paciente desenvolver o início de sintomas em diferentes idades. A abordagem implementada assume ser relevante a existência de diferenças à medida que os pacientes envelhecem. Ainda neste campo, explora o impacto da combinação de diferentes modelos focados em previsão de idades, usando janelas temporais próximas. Desta forma combina um conjunto de *base learners* e constrói uma solução com menor variabilidade.

Finalmente, contribui para *model evaluation*, abordando as questões de implementação de uma abordagem que indica o retorno expectável em alterar um conjunto de diretrizes clínicas, tomando partido de capacidades de previsão. Ainda neste campo, apresenta esquemas de avaliação robustos que suportam a abordagem multivariada orientada a dados na seleção do melhor modelo. Estes esquemas também são usados para avaliar e comparar as abordagens de construção de features delineadas.

A aplicação é focada em pacientes com ATTRv Amyloidosis.

Para caraterizar o trabalho realizado, esta tese está estruturada em quatro secções principais. Começa com uma introdução e uma apresentação de ATTRv Amyloidosis tendo por base uma perspetiva histórica. Posteriormente, apresenta o contexto relevante, relacionando a ligação entre previsão time-to-event com *feature engineering*, *model construction* e *model evaluation*. Ainda neste âmbito, apresenta conceitos-chave de estudos genealógicos. Em seguida, apresenta as principais contribuições, na forma das principais publicações que contituem este trabalho (um artigo por capítulo). Termina com um epílogo que resume o trabalho realizado, partilha as principais conclusões e, por fim, discute a tese numa perspetiva técnica e clínica.

Palavras-chave: Dados *Time-to-Event*, Modelação de Dados, Modelos de Regressão, Decision

Trees (DT), Elastic Net (EN), Lasso (LA), Linear Regression (LR), Random Forest (RF), Ridge (RI), Support Vector Regressor (SV), Extreme Gradient Boosting Trees (XG)

Acknowledgements

I express my gratitude to my thesis advisors, Prof. Dr. Alípio Mário Guedes Jorge and Prof. Dr. João Pedro Carvalho Leal Mendes Moreira, for the support and encouragement given over the years. Their knowledge has not only shaped this phase of my life but also fostered my academic growth. I am thankful for their combined guidance.

I am extremely grateful to Dra. Maria Teresa Pardal Monteiro Coelho, for her guidance, for sharing her decades-long experience, and for never giving up, despite never having officially been part of my advisory team. I extend my gratitude to the patients and dedicated team members of Unidade Corino de Andrade, without whose participation this thesis and the research it gave rise to would not have been possible. They are, without a doubt, a living testament to the commitment of the clinical professionals that operate the Portuguese *Serviço Nacional de Saúde* (SNS).

The journey of a PhD is full of paths that lead to unique experiences. I express my gratitude to Prof. Carlos Soares (FEUP/DEI), for the learning opportunities that my collaboration with PBS/PGBIA provided me. I also express my gratitude to Faculty of Engineering and Faculty of Sciences, in particular Prof. Bernardo Portela and Prof. Miguel Areias, without forgetting the students who crossed my path, namely Sara Silva. I would also like to thank Alexandra Ferreira (FCUP/DCC), Isabel Gonçalves (FCUP/DCC), Joana Dumas (INESC) and Sandra Reis (ex-FEUP/DEI), for all the help they have given me over the years.

I express my deepest gratitude to Alexandra Oliveira, for being included in the group of people who find my work interesting, as well as for all the help and encouragement given, even though I called her too close to the final whistle. Also to Shazia Tabassum for her kindness, encouragement and friendship, as well as for her help in reviewing one of my works. I would also like to thank Thiago Andrade for all the encouragement given, specially over the last few months. I also give special thanks to Tiago Costa, a valued colleague from ProDEI, for all the help and support he has given me over the years. Hope we can keep in touch.

I express my gratitude to my family (Pai, Irmãos, Patrícia) and friends (Angela Félix, Cláudia Moreiras, Cláudia Rodrigues, Eduarda Portela, Filipe Coelho, Gonçalo Pereira, Joana Fernandes, João Anselmo, João Bispo, Maria José Marques, Rui Sarmento, Ricardo Nobre), for their support, help, encouragement, patience and understanding, which have been a source of motivation and strength for me.

On a more personal note, I sincerely thank you Paulo. For your belief in me over the years. For your understanding and strength, besides your kindness and ethical standards, that have helped me during the ups and downs of this journey.

I close with thank you Mãe... you will always be in my heart.

Maria José Gomes Pedroto

"To improve is to change,
to be perfect is to change often."

Winston Churchill

"Francamente - porque pensam que eu escrevo?
Para incomodar o maior número de pessoas com o máximo de inteligência. "

Agustina Bessa-Luís

Contents

Appendices

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
AOO	Age-of-Onset
ATTR	Amyloid Transthyretin
ATTRv	Transthyretin-related amyloidosis variant
AVG	Average
BL	Baseline
CAPA	Cost Asymptomatic Prediction Approach
CDPT	Cost Delay Prediction Treatment
CV	Cross Validation
DT	Decision Tree
EN	Elastic Net
FAP	Familial Amyloid Polyneuropathy
FN	False Negative
FP	False Positive
GI	Gastrointestinal
GLM	Generalized Linear Model
LA	Least Absolute Shrinkage and Selection Operator
LR	Linear Regression
MAE	Mean Absolute Error
MAX	Maximum
MIN	Minimum
ML	Machine Learning
MSE	Mean Squared Error
NE	Network Embedding
NMR	Not Most Recent
OMR	Only Most Recent
PADO	Predicted Age of Disease Onset
RF	Random Forest
RI	Ridge
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
ROI	Return on Investment
SD	Standard Deviation
SOTA	State of the art
SV	Support Vector Machine
TP	True Positive
TN	True Negative
TTE	Time-to-Event
TTR-FAP	Transthyretin related Familial Amyloid Polyneuropathy
TTRVal30Met	methionine for valine in position 30 (of peptide chain in mature protein)
UCA	Unidade Corino de Andrade
XG	eXtreme Gradient Boosting

Chapter 1

Introduction

The concept of time has varied interpretations in different scientific fields. In physics, it represents a dimension measuring object movement, while in mathematics, it denotes changes in functions representing event duration. Also of note that philosophically, time can be seen as an illusion for addressing reality changes. In this work, time, in tandem with prediction, is important for studying the likely moment of future event occurrence.

To introduce the work performed, this chapter, explains why this theme is important, provides its context, outlines the goals, poses the most relevant research questions, and highlights the main contributions. It ends with the bibliographic note references.

1.1 Scope

The estimation of when an event is going to occur is important or even vital in different situations, and professionals across different domains are interested in modeling and predicting Time-to-Event (TTE) episodes [52]. These include **doctors** and **patients** aiming to pinpoint severe disease symptoms to enhance the quality of life, or reduce costs, as well as **engineers** who need to improve maintenance scheduling by predicting when the next failure of a specific component is most likely to occur.

Overall, the areas where Time-to-Event (TTE) modeling and prediction have had, currently have, and will, most likely, continue to have the greatest impact are **management** and **healthcare** [121]. In these domains, practitioners strive to identify predictors and create models that address business needs. This is achieved by continuously exploring business rules, finding suitable and concise representations of real case scenarios, as well by designing, comparing and evaluating different sets of prediction methodologies. These steps must be considered in conjunction with the holy grail that corresponds to quality of data, since, **garbage in, garbage out** [80]. Over time these can become never-ending tasks.

1.2 Context

This work focuses on the problem of modeling and predicting Time-to-Event (TTE) episodes [52], which present time-dependent variables, and whose domain applications have a tight connection with genealogical data sets. For its connection with modeling and prediction phases, within machine learning, this concept is explored in **feature engineering**, **model construction**, and **model evaluation**.

Time-to-Event (TTE) prediction deals with the study, development and exploitation of prediction models, to account for how much time will pass before an event or episode is experienced by a subject [52]. In this case, event or episode denotes occurrences that are of interest in medicine, epidemiology, demography, biology or sociology.

Genealogical data sets can be seen as relational or network connected entities, where the relationship between family members is maintained in *familial* or *kinship trees* [10]. These repositories of data serve as tools for genealogical-based studies, and can, for example, with the right manual or automatic feature extraction process, express tendencies of overall demographic evolution [38]. From a more philosophical point of view they serve as a way to know ones ancestors, as well as help to explore and understand genetic diseases [107]. As an example, consanguinity (marriage between close relatives) can increase the prevalence of autosomal recessive disorders in a population. This is due to the fact that close relatives are more likely to share genes, inherited from common ancestors. On this note, figure 1.1 depicts a close family where first degree cousins (depicted with labels 1 and 2), which share grandparents (depicted with labels 9 and 10), have 4 descendants (depicted with labels 3, 4, 5 and 6). These have a 25% chance to express genetic disorders passed down from their ancestors [13].

From a clinical standpoint, this work was born from the medical need to *predict the age-of-onset of patients diagnosed with ATTRv Amyloidosis*. This is a genetic hereditary disease, first diagnosed by Dr. Mário Corino de Andrade, a neurologist that worked in the "outpatients facility of Hospital Geral de Santo António" [8]. Since then, this disease has been the focus of multidisciplinary research teams working at the recently re-baptized Centro Hospitalar Universitário de Santo António (CHUdSA). These teams are known for their extensive work in enhancing the knowledge of this disease, while enduring day-to-day patient care.

When dealing with data, it is crucial to distinguish between descriptive, predictive, prescriptive, as well as diagnostic data analytics.

Descriptive analytics [93] is used to analyze past events and identify hypotheses worth testing, as well as highlight past trends. In **medicine**, it is relevant during retrospective studies. As an example, in the course of this work, an extensive longitudinal study was conducted to characterize profile changes of symptomatic ATTRVal30M patients followed and diagnosed at Centro Hospitalar Universitário de Santo António (CHUdSA), since the first patient was diagnosed in 1939 (see chapter 4).

Diagnostic analytics is a sub-stage of descriptive analytics. It consists on resourcing to data to figure out, for example, why a specific event occurred. Since it can enhance future recommen-

dations, without so much trial and error, it can impact prescriptive analytics. In medicine it can be seen as a task specifically developed to improve the explainability and understandability of sets of results.

Predictive analytics [93] answers to business questions by, for example, predicting the likelihood of specific outcomes, or by forecasting while resorting to historical data. In **medicine** this is relevant in the course of management tasks and it can, for example, estimate the medical costs a facility will incur over a fiscal year, by predicting the number of patients who will access the hospital emergencies.

Finally, **prescriptive analytics** [93] corresponds to the most advanced analytic task. It delivers actionable insights to decision managers and recommends what actions should be taken to optimize a desired outcome. In **medicine** it can be seen as an improvement that shows the impact of changes in diagnosis and follow-up procedures.

In the course of this work the last three phases correspond mostly to the cost and model evaluation tasks.

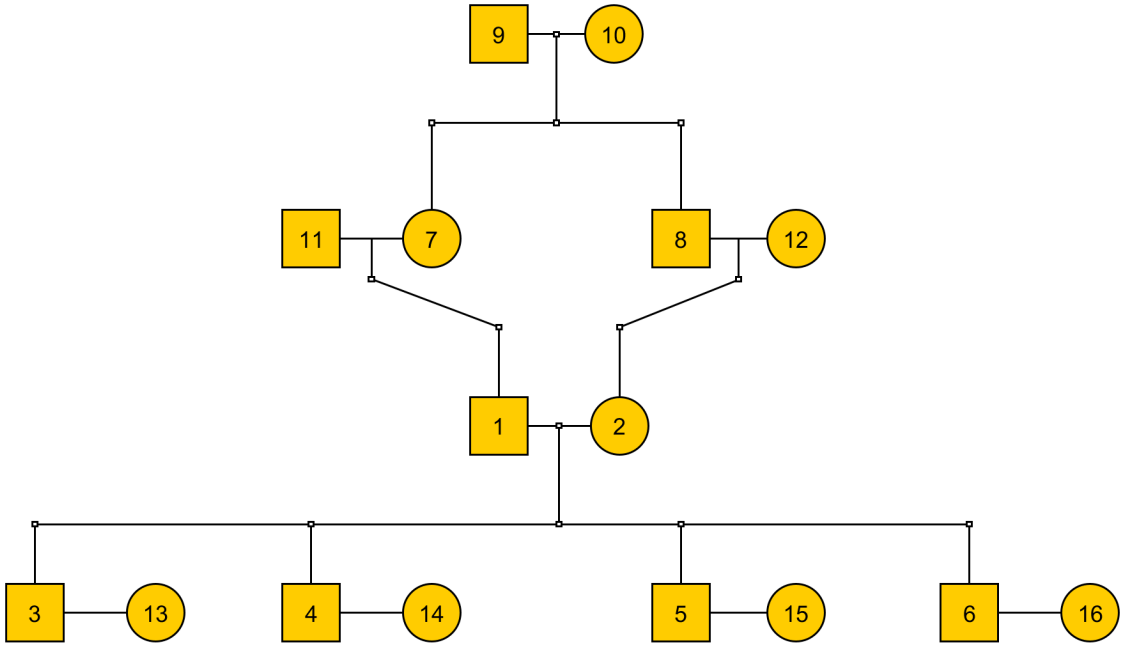


Figure 1.1: This image depicts a family tree. Squares depict males and circles depict females. One circle and one square connected by a strait line depict a marriage, while a line that drops from a marriage depicts a child.

1.3 Motivation and objectives

Medicine is a field in constant change. It evolves on a daily basis, in response to research, techniques, and expert guidelines. The application of discoveries that emerge from relevant research projects are leveraged by the need to prevent, diagnose and treat an ever growing range of medical diseases. This was seen in recent past with the SARS-CoV-2 (COVID-19) [47] pandemic.

Currently, this is the most well known example where predicting when events are going to occur showed to be of extreme importance, and capable to influence not only patient outcomes but society.

For reference purposes, the computational objectives that guided this work correspond to:

- develop approaches for feature construction (feature construction);
- build algorithms and methodologies for time-to-event machine learning (model construction and prediction);
- develop approaches to evaluate and compare different modeling and prediction approaches of TTE episodes, with a focus on improving their understanding (cost and model evaluation).

1.4 Research questions and main contributions

The main research questions dealt with were:

- **RQ1**: How can tightly coupled data be exploited to predict time-to-event episodes? In particular, how can this be achieved with genealogical data?
- **RQ2**: How to extract effective time-based feature sets for time-to-event episodes prediction?
- **RQ3**: How well is it possible to evaluate time-to-event episodes prediction occurrences with genealogical based data sets?
- **RQ4**: How is it possible to measure the impact of employing a data-oriented prediction approach in a time-to-event setting?

The contributions, which in academical terms consist of a list of publications referenced in 1.6, are aligned with three areas:

- **feature engineering (feature selection and feature construction)**, with the development and evaluation of approaches to represent patients evolution over time while, leveraging the most important feature sets ([85, 86]);
- **model construction**, with the development of a multivariate data-oriented approach to track a patient's risk of developing disease onset, over different ages, with the purpose of empowering medical professionals with the right tools to tailor treatments to patient needs, as well as exploring the impact of combining age models of relatively near time windows ([83, 84, 87]);
- **model evaluation**, with the implementation of a data-oriented approach to supply medical professionals with the cost assessment of following data oriented guidelines to real medical problems (to be submitted), as well as by studying and presenting a set of robust evaluation strategies that assist the multivariate data-oriented prediction methodology in selecting the best model, as well as in comparing the feature construction methods.

For reference purposes, the application is entirely focused on data sets of patients diagnosed with ATTRv Amyloidosis. Next, the thesis outline is presented.

1.5 Thesis outline

This thesis is organized in 10 chapters, organized in 4 thematic areas, outlined as follows.

Part I: Intro

Introduction

Chapter 1 delineates the scope and context, defines the problem, and sets forth the objectives and research questions.

Transthyretin-Related Familial Amyloid Polyneuropathy (ATTRv Amyloidosis)

In chapter 2, TTR-FAP is presented from a historical perspective. An overview of the profile of symptomatic patients from a time oriented manner is also presented, as well as main symptoms and diagnosis procedures.

Part II: Background

Time-to-Event prediction and genealogical studies

The 3*rd* chapter defines time-to-event prediction, and distinguishes it from survival analysis, since this is one of the most well known medical methods for analyzing distributional changes. Fundamentally, it presents the main areas of impact, highlights some of the main algorithms, while pointing a few of their main caveats. It then presents Time-to-Event prediction in feature engineering, model construction and model evaluation fields. For clarification purposes, initially, the focus is on approaches that represent entities which contain heterogeneous sets of data. It then proceeds to systematize the different methods considered state of the art (SOTA). Then, it presents key considerations of a modeling approach that highlights the utility of data-driven patient assessment. It also reviews regression and classification evaluation metrics, and highlighting both the advantages and disadvantages of each metric used. It closes by exploring a few key concerns of working with genealogical data sets, namely those regarding data quality issues.

Part III: Contributions

This section presents the main academic contributions by highlighting key concerns regarding each of the published and to be published papers. In this case, it is important to highlight that the order of the chapters does not correspond to the temporal order in which they were carried out. Thus, there are some relevant differences, such as the fact that the indicators for Q1, Q2, Q3 and Q4 time landmarks are only present in chapter 4, and currently it is not deemed relevant to expand their usage to the rest of the contribution chapters.

Study on the heterogeneity and variability of Portuguese families diagnosed with ATTRv Amyloidosis

Chapter 4 presents the results of an extensive longitudinal study which evaluates profile changes of the ATTRV30M patients diagnosed at Unidade Corino de Andrade, since the first patient was diagnosed until mid 2023. It also addresses the changes that occurred in the genealogical trees, by evaluating a set of genealogical-based indicators over different key time periods.

Preliminary attempt to answer the medical problem

Chapter 5 presents the preliminary results for modeling and selecting features for time-to-event prediction, with a focus on a data-oriented summarization (synopsis) approach.

What is the importance of feature selection?

Chapter 6 explores results of the data-oriented approach delineated to predict the AOO of TTR-FAP patients, while focusing on the importance of the different types of features. This way it highlights the ranks of the different sets of genealogical-based features considered.

Is important information being lost?

Chapter 7 presents results on a new methodology assembled to measure if the summarization approach leads to any loss of important information. In this case it resorts to feature construction from an embedding perspective and compares the predictive improvement in three types of models namely: data set A, data set B and full data set.

Is it possible to blend heterogeneous models?

Chapter 8 presents the results of a model construction improvement, focused on merging data from approximate time windows.

Clinical model for time-to-event prediction

Chapter 9 presents key computational concerns regarding the multivariate methodology for predicting the age-of-onset of ATTRv Amyloidosis. It also shares and discusses the results, both from a clinical as well as from a technical perspective. It increases the previously mentioned feature construction works by enhancing the different feature sets and exploring the clinical impact.

Finally, this chapter also presents the results of a data-oriented methodology to estimate the cost-benefit of changing clinical guidelines so as to take advantage of prediction capabilities. This can be used in other medical related problems to simulate the return on investment (ROI) [60] of following a set of guidelines for the follow-up of asymptomatic patients.

Of note that it this work considers that the social aspects of following patients are more pressing and relevant than monetary concerns, since the cheapest approach is simply not to follow patients that have not yet shown symptoms.

Part IV: Epilogue

Conclusions and future work

Chapter 10 discusses the results from a clinical as well as a technical perspective. It also leverages the main achievable conclusions and limitations of this work. It ends with a set of future work directions.

1.6 Bibliographic note

The thesis outcomes are available in the following publications:

Paper I: *Regression based Approaches for Predicting Age at Onset*, XXIV Jornadas de Classificação e Análise de Dados (JOCLAD 2017), **Maria Pedroto**, Alípio M Jorge, João Mendes-Moreira, Teresa Coelho, Unidade Corino de Andrade, Centro Hospitalar do Porto.

Paper II: *Predicting age of onset in TTR-FAP patients with genealogical features*, 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), **Maria Pedroto**, Alípio Jorge, João Mendes-Moreira, Teresa Coelho, IEEE.

Paper III: *Impact of genealogical features in transthyretin familial amyloid polyneuropathy age of onset prediction*, Practical Applications of Computational Biology and Bioinformatics, 12th International Conference, **Maria Pedroto**, Alípio Jorge, João Mendes-Moreira, Teresa Coelho, Springer International Publishing.

Paper IV: *Improving the Prediction of Age of Onset of TTR-FAP Patients Using Graph-Embedding Features*, Progress in Artificial Intelligence: 21st EPIA Conference on Artificial Intelligence, EPIA 2022, Lisbon, Portugal, August 31–September 2, 2022, Proceedings, **Maria Pedroto**, Alípio Jorge, João Mendes-Moreira, Teresa Coelho, Springer International Publishing.

Paper V: *Combining neighbor models to improve predictions of age of onset of ATTRv carriers*, Progress in Artificial Intelligence: 22nd EPIA Conference on Artificial Intelligence, EPIA 2023, Faial, Portugal, September 5- September 8, 2023, Proceedings, **Maria Pedroto**, Alípio Jorge, João Mendes-Moreira, Teresa Coelho, Springer International Publishing.

Paper VI: *Clinical model for Hereditary Transthyretin Amyloidosis Age of Onset Prediction*, **Maria Pedroto**, Teresa Coelho, Alípio Jorge, João Mendes-Moreira, Frontiers in Neurology.

Paper VII: *Heterogeneity in families with ATTRV30M Amyloidosis: A historical and longitudinal Portuguese case study impact for genetic counseling*, **Maria Pedroto**, Teresa Coelho, Joana Fernandes, Alexandra Oliveira, Alípio Jorge, João Mendes-Moreira, Amyloid, The Journal of Protein Folding Disorders.

The following paper is currently to be submitted:

Paper VIII: Cost Modeling for ATTRv Amyloidosis asymptomatic patient care.

Other works performed during the course of this thesis include:

Paper IX: *Comparison of Classification and Regression Approaches to Predict Age of Onset*, Doctoral Symposium Informatics Engineering 2017, **Maria Pedroto**.

Paper X: *Geovisualisation Tools for Reporting and Monitoring Transthyretin-Associated Familial Amyloid Polyneuropathy Disease*, Machine Learning and Principles and Practice and Knowledge Discovery in Databases, ECML PKDD 2022, PT I, Rúben Lôpo, Alípio Jorge, **Maria Pedroto**.

Chapter 2

Transthyretin-Related Familial Amyloid Polyneuropathy (ATTRv Amyloidosis)

The work conducted in this thesis addresses a pressing medical issue, that underlines the importance to **model and predict the age-of-onset of patients diagnosed with ATTRv Amyloidosis**. Breaking down the name, one has "**A**" for Amyloidosis, "**TTR**" for Transthyretin - the protein that transports the thyroid hormone thyroxine and retinol to the liver, and "**v**" for variant [15]. The event of interest - age-of-onset (AOO), represents the moment in time when first symptoms are felt.

Due to its heterogeneity, data scarcity and disease severity, predicting age-of-onset for ATTRv Amyloidosis is a challenging problem, both from an academic as well as from a medical perspective. From early times, this work was deemed clinically relevant due to published evidence [64, 102] showing the connection between the AOO of a patient and that of the ancestor. The studies focused on Portuguese families and presented different **descriptive** and statistical evaluations. Later, clinical researchers from other geographically dispersed cohorts conducted additional studies [4, 25, 37, 89], still from a **descriptive** and statistical standpoint.

On a follow-up as well as a more generic path, the study presented in chapter 4 and annex A points out, also from a **descriptive** and a statistical basis, the importance of family relations on AOO differences, between patients which share family relations, namely while considering **families**, **probands**, **parents** and **siblings**. One of its major findings consists on exhibiting clear statistical based evidences of a higher decrease in the AOO difference between **siblings**, than when considering other genealogical connections such as **parents**. Still, currently a few issues remain unclear, namely why it happens.

As such, this section follows to present the main characteristics of ATTRv Amyloidosis, while contextualizing as to its discovery and the symptoms that afflict its bearers. This is relevant since it

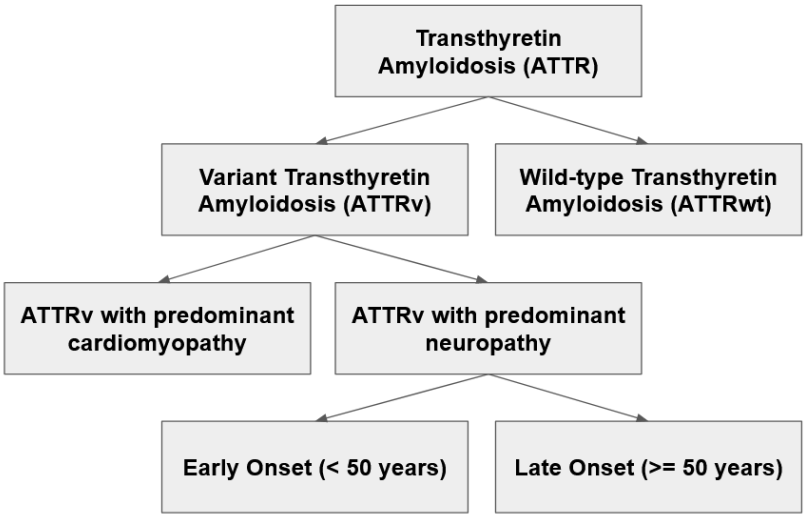


Figure 2.1: Types of Transthyretin Amyloidosis (ATTR)

adds context to the heterogeneity regarding age-related factors and its impact in rare disease management. It also refers a few of the current research lines that are followed by different research groups, with a focus on clinical practice.

2.1 ATTRv amyloidosis

ATTRv Amyloidosis was until recently known as Transthyretin Familial Amyloid Polyneuropathy (TTR-FAP) [15,67]. Characterized as a rare neurological hereditary disease, it is also classified as severe, progressive, and life-threatening.

On a more general perspective, Transthyretin Familial Amyloid Polyneuropathy (ATTRv Amyloidosis, where the *v* stands for variant), over time received different denominations, namely: Hereditary Amyloidosis, Familial Amyloid Polyneuropathy (TTR-FAP, FAP, PAF), ATTRv Amyloidosis, Transthyretin Amyloidotic Familiar Polyneuropathy, Portuguese Familial Amyloidosis, among others. Also referenced as an irreversible and rare neurological disease, it is worth mentioning that different types have been diagnosed over the years (for a full description please see figure 2.1).

From a historical perspective, ATTRv amyloidosis was first diagnosed in Portugal by Dr. Mário Corino da Costa de Andrade [8], in 1952. Over the following 42 years and until 1994, a total of 1233 cases were diagnosed at *Centro de Estudos de Paramiloidose* [81], the original name of the medical facility later denominated as *Unidade Corino de Andrade* (UCA) and which is highly valued by the knowledge it shares with other clinical facilities.

Currently, UCA, which acts under the umbrella of Centro Hospitalar Universitário de Santo António (CHUdSA), is one of the two medical facilities that in Portugal, diagnose and are responsible for the follow-up of patients and families afflicted with ATTRv Amyloidosis. It contains the oldest active data repository with information of formerly and currently active families of carriers

of this disease. For more information regarding the distribution of patients which currently are actively being followed in UCA please refer to chapter 4 and annex A. These contain an extensive longitudinal overview of the most current clinical and genealogical profile of Northern Portuguese carriers.

Thought initially to be restricted to Portugal, Sweden and Japan, over the last decades other *foci* of the disease emerged in France, Spain, Turkey, Germany and Austria [33], among other countries [7].

Depending on the stage at which a patient is, TTR-FAP can have a great impact on social, economic, familial, and working life quality, mainly due to the damage it causes to the sensory and autonomic nervous systems. In a more of a public policy perspective, it can also impact different social aspects such as main demographic indicators, specially in areas where the disease is not considered rare. From a clinical perspective and as an example, one of those symptoms, male impotence, directly influences the average dimension of families, namely when the carrier is male [72] even thought currently most patients undergo medically assisted pregnancies.

In Portugal, as well as Brazil and Sweden, the variant responsible by the largest percentage of active patient cases has the denomination of p.Val50Met [110], and was formerly known as ATTRVal30M [98]. Several studies have focused on analyzing which factors influence its phenotypic variations, and are responsible for the heterogeneity present on age-of-onset of ATTRVal30M bearers. It is worth noting that despite numerous attempts to identify the factors influencing this event, the answer remains unclear, due mainly to the wide dispersion of values observed in active patients, which in turn make it difficult to resort to classic clinical trials, where from a cohort of patients medical professionals select subset of patients with possibly a wide age range, to compare, for example, the distributional differences between symptomatic and control cases.

2.1.1 Amyloid and Amyloidosis

Amyloid is a medical term which references a substance that exists in the human body [20]. Over time, when this substance appears in abnormal deposits, it can lead to the development of a strain of degenerative diseases known as amyloidosis. Thus, amyloidosis occurs when proteins misfold and form insoluble fibrils that deposit in tissues belonging to different organs (e.g. heart, kidneys, liver, eyes) and interfere with their normal function [2].

For contextualization, in the case of proteins, "misfolding" [76] refers to the abnormal structure of a protein. Since the correct biological function of a protein depends on their correct and precise shape formation, when this does not occur, it can lead to a dysfunctional or even harmful clinical prognosis, which can lead to a degradation of tissues. Depending on the disease, this can occur due to environmental changes, or genetic mutations.

2.1.2 Impact in the human body

When amyloid sets in the human body it impacts different systems, such as the sensory and the autonomic nervous system [20].

The sensory system which is part of the central nervous system, is responsible for processing and transmitting information from different organs (eyes, ears, skin and tongue) to the brain. It ensures individuals are able to perceive and interpret different sensations such as touch, sound, light, taste and smell, besides distinguishing between hot and cold thermal variations [68].

The autonomic nervous system controls the body's involuntary functions, such as heartbeat and sweating. It is divided into two branches - the sympathetic nervous system, which prepares the body to respond to danger; and the parasympathetic nervous system, which helps the body to relax. The movements related to the autonomic nervous system are involuntary and reflexive, such as increased heart rate during exercise, sweating in response to heat, or digestion after a meal [114].

In clinical terms, a disease which affects the sensory as well as the autonomic nervous system can trigger a wide range of symptoms. In this case, it typically manifests as a severely progressive sensorimotor and autonomic polyneuropathy, together with cardiac dysfunction [33].

2.1.3 Symptoms and disease progression

TTR-FAP is an amyloid nerve based disease that affects patients' peripheral nerves, which control sensation and the movement of arms, as well as legs and autonomic nerves that control automatic body functions [20,27]. These symptoms define a set of categories that represent different severity degrees.

Worldwide, medical professionals agreed upon a Polyneuropathy Disability Score (PND) classification system to define the progression of patients diagnosed with TTR-FAP, and it corresponds to the following items [3]:

- PND I: Sensory disturbances in extremities and preserved walking capacity.
- PND II: Difficulty walking but no need for walking stick.
- PND IIIa: One stick or crutch required.
- PND IIIb: Two sticks or crutches needed.
- PND IV: Confined to a wheelchair or bed.

The main characteristic of this disease is that it is considered a complex disorder, whose symptoms can overlap with other conditions, especially in older ages. These factors contribute to the difficulty to pinpoint the start of symptoms onset, and why it is important to have a set of clinical guidelines, and not base the diagnosis solely on patient deposition. This is an even more pressing matter when the depositions are made by close-family relatives.

2.1.4 Characteristics of Portuguese subjects

In clinical terms and accordingly to several publications, TTR-FAP is defined by loss of superficial sensation, including thermal sensation, and a steady progression of the disease over the course of

10 to 15 years [63]. These characteristics can greatly impact patients' lives in all their dimensions, namely by presenting difficulties in social, familial, and work relationships.

In Portugal, TTR-FAP is characterized by a high penetrance rate. This defines the likelihood of a carrier of the disease to show its symptoms and in naive terms it states that the likelihood of a patient developing symptoms is near 100% when considering quite older ages. For references purposes, current distribution is depicted in figure 2.2. From it, it is clear the difference between male and female patients, besides the almost full penetrance, for patients over 80 years old. This factor was highly impacted, over time, by an increase in life expectancy.

In geographical terms, in Portugal, TTR-FAP is spread over both the coastal and interior country regions, and has one of the largest *foci* of the disease in the fishing villages of Póvoa de Varzim and Vila do Conde. Early epidemiological studies unraveled that *patient zero*, *i.e.* first patient to genetically acquire the disease, must have come from this region between 1500 and 1850 AC [72].

Currently, Portugal houses two national reference centers entrusted with diagnosing and overseeing patients with symptoms linked to ATTRv Amyloidosis. These facilities have teams of neurologists and professionals from other specialties (*e.g.* nurses, psychologists, psychiatrists, ophthalmologists, cardiologists, nephrologists, neurologists) with extensive practice in supporting these patients and addressing the different dimensions of this disease.

Within the northern reference center, Unidade Corino de Andrade (UCA), a total of 934 families are currently enlisted, but not all are focus of regular familial follow-up. This depends on the status of current and active family branches which study and classification can be addressed with careful retrospective studies.

In Portugal, the most prevalent mutation of ATTRv Amyloidosis corresponds to the exchange of the amino acid methionine for valine at protein position 50 (Val30Met following the traditional TTR classification; c.148G>A; p.Val50Met following the Human Gene Mutation Database) [33], while in the rest of the countries there is a wider genetic heterogeneity [81]. Of note that in clinical terms, this fact can induce a difference of diagnosis rates over different countries.

2.1.4.1 AOO patients profile

Accordingly to recent studies, age of disease onset (AOO) of ATTRv Amyloidosis carriers varies between the second and ninth decades of life, and presents variations across geographically distinct and distant *cohorts*. From a medical point of view, its correct determination is critical to determine when amyloidosis testing should be requisitioned by medical professionals [7].

The penetrance among positive patients carrying the same mutation varies between countries [98] and in individuals that belong to the same families. When focusing on the variation between early (when age of onset occurs before 50 years old) and late onset (when AOO occurs at, or after, 50 years old) there is a wide gender based heterogeneity.

[81] is one of the last published studies that focuses on the profile of ATTRv patients. There, authors perform a review across Europe. One of the fundamental objectives they had was to clarify the disease over the different geographical regions it appears. There, they indicate the patients in terms of early as well as late symptoms presentation, while also characterizing them in terms of

incidence and clinical presentation. One of the main clarification authors performed was regarding the definition of disease progression. Authors also expressed that it is not clear the best approach approach to define disease progression, specially in rare diseases, since they are highly dependent on medical experience which, can be quite narrow.

2.1.4.2 Progression

Notwithstanding the PND classification presented earlier, in Portugal clinical practice diverges from it and assesses the progression of the disease in three summary sets of stages that correspond to: Stage I, Stage II and Stage III [3]. This new category set corresponds to:

- sensory polyneuropathy, where a patient suffers from sensory impairment, which starts at the lower members and progresses to every member of the body;
- progressive walking disability, when patients start showing difficulties in walking and start using crutches or a second person to assist them;
- wheelchair bound or bedridden, when patients lose all their ability of walking and are confined to a wheelchair, or to a bed.

For ATTRv Amyloidosis, while in Stage I patients have some difficulty in walking but do not need assistance, in Stage II they requires assistance. This stage combines PND IIIa and PND IIIb, and is characterized by a motor disability progression in the lower limbs, by an overall weakness in hand muscles, and by a high level of impairment, although patients can still move with crutches or walking sticks. Lastly, in Stage III the patient is wheelchair or bed bound. This stage corresponds to PND IV [3, 26].

On a final note, the literature review shows that both symptoms and the terms that define disease progression can vary among carriers (due to different mutations). In some cases geographical dispersion can also influence a wide variability among positive individuals. Even so, in most cases, progression follows a progressive and debilitating pattern and can present in different forms, namely with a primary neuropathy or cardiomyopathy format. In the first case AOO tends to manifest early, while in the cardiomyopathy format it presents mainly as a late onset disease [32]. In general, symptoms tend to worsen over time, as the abnormal protein deposits accumulate and increase the damage to the affected tissues. Early stages may involve mild peripheral neuropathy, while later stages may present by significant disability, with patients experiencing severe neuropathy, cardiac complications, and autonomic dysfunction [26].

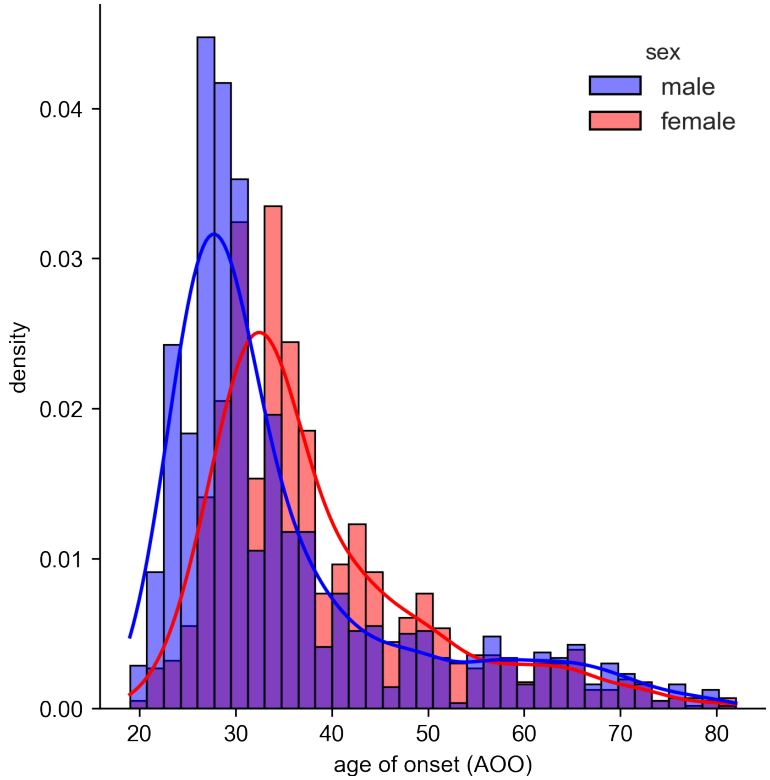
2.2 Summary

ATTRv Amyloidosis is a rare disease with the largest focus in Portugal. This assists local medical professionals to gain a high knowledge regarding disease progression, as well as its social and economic impact on both patients and families.

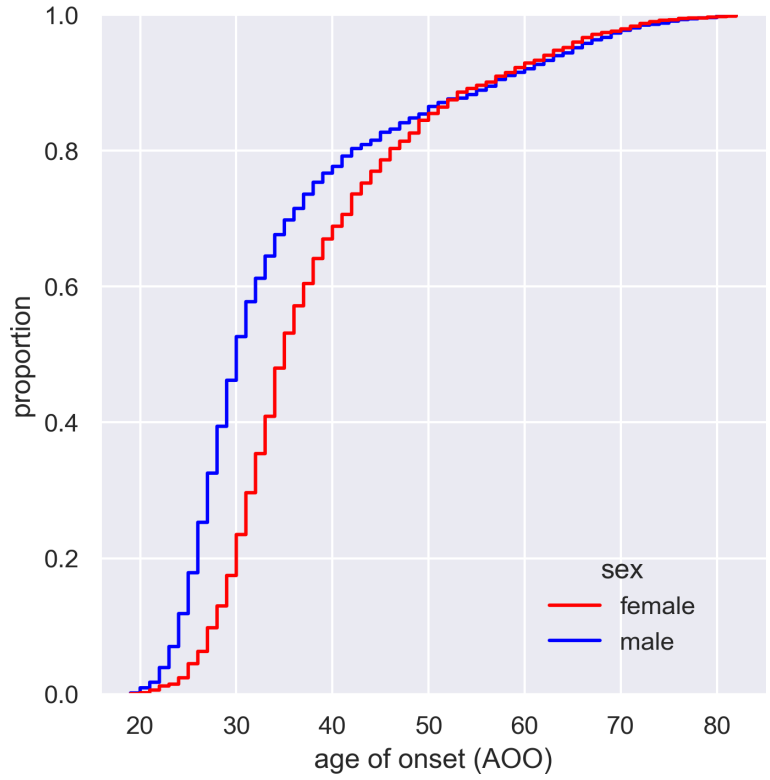
In the course of genealogical disease studies, a focus on family history holds considerable importance. As an example, the study presented in chapter 4 shows the results of an approach that evaluates the changes of age-of-onset (AOO) and other age-related clinical factors, within and among families affected by ATTRv amyloidosis. It reveals that currently Portugal faces a shift in current patient profile, with recent diagnosis showing an increase of male elderly cases, especially regarding probands. It also shows an overall reduction of time-to-diagnosis, and that families tend to have a long familial follow-up, since family members tend to be diagnosed over the course of several years. More importantly, results show the existence of families with both early and late-onset patients. The study also highlighted that symptomatic patients exhibit less variability towards siblings then other family members. These observations can have high impact in the course of family counseling sessions.

In clinical terms, the need to pinpoint the age-of-onset on cohorts of patients can improve medical resources, and clinical trials. This is not an easy task, especially when in the face of rare, highly heterogeneous and variable diseases, spread over different countries, and subject to different clinical guidelines, sometimes biased to local medical knowledge. From a computational-based technical perspective, the development of methodologies which resort to data sets with highly variable data completeness issues comes with sets of advantages and disadvantages. Notwithstanding, it is important to keep in mind that resorting to these data sets can be a good starting point. This can be the case for ATTRVal30M AOO prediction, since without it doctors need to continuously evaluate asymptomatic patients, at regular and close spanned time intervals, possibly over the course of several decades, thus straining complicated budget conditions, for health services. To address this issue, this thesis presents in chapters 7, 8 and 9, a set of models specially engineered with the purpose of predicting (or improving the prediction) of the AOO of ATTRv Amyloidosis.

On a final note, this chapter gave focus to disease progression since it is impacted by age-of-onset values, and its study can help to characterize the best age of onset prediction approach, which can also help in defining disease progression models.



(a) Distribution of TTR-FAP age of disease onset



(b) ECDF of TTR-FAP age of disease onset

Figure 2.2: Distribution of the age of onset of TTR-FAP patients diagnosed and followed at Unidade Corino de Andrade, since early records until 31st of July of 2023

Chapter 3

Time-to-Event prediction and genealogical studies

In recent decades, time-to-event prediction gained attention across various fields which lead to the development of different models for descriptive, predictive or prescriptive analyses [43]. This chapter discusses time-to-event prediction, relates it with feature engineering, model construction, as well as model evaluation topics, while explaining the flow of thought that lead to the different contributions that constitute this work. It also clarifies as to its relation with genealogical studies, by exploring a few key concerns related with working with genealogical data sets.

3.1 Time-to-Event prediction

The goal of Time-to-Event prediction is to understand and predict the timing of events. This can be achieved with relevant feature sets. It can also be addressed with different modeling approaches (classification, regression, time-series), in which case the **evaluation** schema holds importance.

In this work, for its importance in healthcare, the focus of model construction is on **regression algorithms**. This reinforces the need to work with **feature construction**, to accommodate with algorithms needs, as well as to delineate a careful **model evaluation** approach, which deals with heterogeneous data sets.

3.1.1 Definition

Time-to-event analysis is commonly used in clinical trials and medical research, where it is relevant to evaluate the time a patient took to experience a particular health outcome, or reach a specific endpoint. This is a wide term and for its intertwined relation with time-to-event prediction it is important to contextualize it.

So, when one diverges from the medical field, where the study of time-to-event occurrences analysis is known as **survival analysis** [112], one finds other terms referring to it, such as **reliability analysis** [117] in engineering, **duration models study** within economics, and **event history analysis** [6] within sociology, demography, psychology and political sciences [71]. For several decades [53] it was considered a statistical task for which researchers developed approaches capable to deal with censored data (*e.g.* instances for which the event of interest becomes incapable of being observed after a specific period of time, or when specific instances don't experience the event by the end of the follow up period). In these cases, the survival time is said to be censored, *i.e.* the observation period was cut off before the event occurred. Thus, the only available information is that it didn't happen by the end of the observation period [111], [5]. Traditional approaches simplify the problem while deleting censored records from the analysis [5].

When focusing on time-to-event episodes, and while formulating a list of examples, one has time-to-onset of first symptoms of a disease, time-to-death of a patient, time-to-divorce of a couple, time-to-unemployment of a professional, and time-to-failure of a mechanical device, as well known examples. These events can be characterized by both sets of static and dynamic characteristics that may need to be modeled over different periods of time. This is one of the interesting aspects of time-to-event data, which over time lead to the appearance of approaches, designed to cope with, for example, irregular trends, deal with problems, such as correlation among the predictors, deal with noisy records, or even deal with outliers or aberrant observations.

On a final note, time-to-event problems tend to start with a set of entities for which it is important to analyze their evolution, regarding a specific event or set of events. On the prediction side, it has the goal of modeling, predicting or prescribing one or several approaches for when the event occurs [116]. The study of these occurrences can be categorized between *one time only events*, *e.g.* when a patient will start to feel the symptoms of a disease, and *recurrent events* [1], *e.g.* when a patient with, for example, predisposition to develop cancer-based diseases, will develop the next one.

3.1.2 Time-to-event (TTE) prediction and survival analysis

Most literature reviews [24, 112] that focus on time-to-event emerge as an evolution of a statistical field known as survival analysis. For this work it is deemed important to differentiate between time-to-event prediction and survival analysis since, in some cases, they may appear to be the same.

Although they are often used interchangeably, depending on the context and field of application, time-to-event prediction and survival analysis can be seen as two different fields. In the context of this work survival analysis consists of a branch of statistics that deals with the analysis of time duration until one or more events happen. It is used to describe or explain the occurrence and timing of events.

[1] https://psiweb.org/docs/default-source/resources/psi-subgroups/scientific/2016/time-to-event-and-recurrent-event-endpoints/jrogers.pdf?sfvrsn=a86d2db_2

Statistical Methods	Non-Parametric	Kaplan-Meier		
		Nelson-Aalen		
		Life-Table		
	Semi-Parametric	Cox-Regression	Basic Cox-PH	
			Penalized Cox	Lasso-Cox
				Ridge-Cox
				EN-Cox
				OSCAR-Cox
			Time-Dependent Cox	
			Cox Boost	
	Parametric	Linear Regression	Tobit	
			Buckley James	
			Panelized Regression	Weighted Regression
				Structured Regularization
Machine Learning Models	Survival Trees			
	Bayesian Methods	Naive Bayes		
		Bayesian Network		
	Neural Network			
	Support Vector Machine			
	Ensemble	Random Survival Forests		
		Bagging Survival Trees		
	Advanced Machine Learning	Active Learning		
		Transfer Learning		
		Multi-Task Learning		
Related Topics	Early Prediction			
	Data Transformation	Uncensoring		
		Calibration		
	Complex Events	Competing Risks		
		Recurrent Events		

Table 3.1: Taxonomy of Time-to-Event (TTE) Analysis and Prediction Methods (adapted from [111])

On the other hand, time-to-event prediction is focused on predicting the amount of time until a specific event will occur. This can be accomplished with statistical or machine learning methods, with more or less complex approaches, which depend on the characteristics of the problem at hand, as well as the available data. So, in the context of this research, time-to-event prediction is considered a broader concept that involves predicting when an event will occur. In this case, the methods don't have any particular assumption. Although in [111] authors merge both concepts, we rely on their work for state of the art positioning, while considering these topics distinct.

Of note that for this work, due to the problem characteristics, the focus on ATTRv Amyloidosis, as well as for the restriction of resorting to genealogical data, the focus is exclusively on regression algorithms. In this case, since [111] states that methods fall under one of two categories, namely **Statistical** or **Machine Learning Approaches** (see table 3.1), this chapter follows for an explanation of the most important algorithms, while not focusing exclusively on the ones chosen for this work. The later ones can be found in sub-section 3.3.

3.1.3 Working with time dimension

Time is important when developing prediction models. By resorting to this dimension one copes with trend changes, and evaluates historical events to, for example, establish the risk or probability of future occurrences. For this work, for its relation with the healthcare ecosystem, it is important to distinguish between **prediction time** [94] and **prediction window** [90].

Prediction time answers when is the model going to generate a new prediction, while **prediction window** defines the time limit or period length for which the model is going to look for the answer. These concepts are affected if the interest is on developing descriptive, predictive, prescriptive or diagnostic approaches, and can transform the requirements of a given prediction problem.

3.1.4 Areas of impact, and their relation with the contributions

Time-to-event prediction is a specialized area of application of machine learning. From the point of view of this work, the most relevant machine learning stages where one needs to be aware of the most recent methods, tools and applications, although they are not limited to, consist of:

- **Data representation**: machine learning based approaches can handle the diverse types of data often encountered in time-to-event prediction. These data types can include continuous variables (e.g., patient age), categorical variables (e.g., disease type), and time-to-event data (e.g., time-to-death of a patient). In this case algorithms need to be able to work with mixed types of data. For more information refer to sub-section 3.2.

- **Feature engineering**: feature engineering plays a crucial role in time-to-event prediction since machine learning practitioners often need to create or transform features to capture relevant information about the event of interest, and its connection with potential predictors. For more information refer to sub-section 3.2.

- **Model selection**: machine learning offers a variety of models suitable for time-to-event prediction. Common choices include Cox proportional hazards models, Random Survival Forests, Support Vector Machines, and Deep Learning models like Recurrent Neural Networks (RNNs), or the new Long Short-Term Memory (LSTM) networks. Researchers can choose the most appropriate model based on the data set and problem characteristics. Specific algorithms can help to deal with problems such as multicolinearity or over fitting issues (e.g., regularization techniques). For more information refer to sub-section 3.3.

- **Ensemble methods**: ensemble methods, such as bagging or boosting, can be applied to time-to-event prediction models, to improve their robustness and predictive accuracy. For its possible complexity, it can be relevant to focus on simple approaches that can be easily understood by professionals with non-technical background. For more information refer to sub-section 3.3.

- **Evaluation metrics**: machine learning offers various metrics and approaches suitable to assess the performance of time-to-event prediction models. As an example, one approach to deal with multicolinearity is to assess its impact in the course of careful evaluation procedures. Another important requirement deals with results and how to make them explainable for non-technical users. For more information refer to sub-section 3.4.

The remainder of this chapter will consider these areas of impact, and how the contributions (publications) can be directly linked to them.

3.1.4.1 Approaches and methods

Currently, there is a multitude of approaches and algorithms for time-to-event prediction. An overview of the most relevant approaches is performed next. Of note that even thought they aren't relevant for the work presented here, they assist in sharing a clear picture of current state of time-to-event prediction in the healthcare ecosystem. This general overview also assists in further distinguishing survival analysis from time-to-event prediction. The study over different model construction algorithms continues in sub-section 3.3. There, the focus is on the algorithms employed in the methodologies delineated for this work.

Kaplan-Meier (KM) curves:

The Kaplan-Meier (KM) curve [92] is a non-parametric statistical method used to estimate the survival function from lifetime data. It is used in the course of descriptive analytic tasks to evaluate, for example, the time until death. In disease studies this event occurs after symptoms onset and it shows the cumulative survival probability of the event occurring after a specific moment.

One key point of KM curves consists in the fact that they can be very useful to understand if, for example, two related populations are distinguishable at any specific point in time, since its output triggers a good analytic view.

A few key disadvantages consists in how KM curves deal with missing values, estimation of the prognosis in the presence of competing risks, or how it assists the comparison of treatment

effects [16]. It is important to reference that it can be problematic to use KM curves when in the presence of covariates with time-dependent effects. This represents a very important limitation, since when dealing with rare diseases the amount of available subjects can be quite limited. Notwithstanding, a simple approach can be to subdivide the data set into distinct sub-groups.

Cox proportional hazards (CPH) regression:

The Cox Proportional Hazards (CPH) [109] regression model is a statistical method employed to analyze the effect of several variables on the time a specific event takes to happen. This corresponds, in its essence, to a multivariate regression model which employs the construction of a survival function to infer as to the probability of an event to happen in a specific moment in time.

The term *proportional hazards* indicates that the variables (features) [52] are constant over time. In general, this method is used to compare survival rates across different observational scenarios.

Random survival forests:

Random Survival Forests (RSF) [61] corresponds to a machine learning algorithm which extends Random Forests (RF) to handle TTE outcomes in the presence of right-censored data. It handles these cases by changing the splitting rule that allows for each tree to grow, while resorting to an ensemble of binary decision trees. These can be used to select the most important variables linked with the event of interest. It captures complex relationships between predictors and survival functions without requiring prior specification [58].

Deep learning (DL) approaches:

Deep learning for time-to-event modeling has shown promising results, especially when dealing with multivariate time series, since traditional methods can have difficulties to capture the complex nature of the underlying dynamical system, and the relationships between sequences of time values.

Although it has shown great promise in real applications [120], it can have issues when applied to problems where the amount of available data is small. For rare diseases, even in the cases data sets are compiled over extensive periods of time, this works as a disadvantage.

3.1.5 Difficulties

Time-to-event prediction poses several challenges that need to be addressed to obtain reliable predictions. One of them is related with missing data, since it can produce biased estimates with reduced statistical power. Also the existence of interactions between covariates can be considered an added difficulty. For the first case it is important to compare different approaches (i.e., have algorithms which deal with missing data), while in the second case it is important to work with algorithms capable to deal with such conditions (i.e., Lasso [108], Ridge or Elastic Net [123]).

On top of these, one of the hardest characteristics deals with how it relates with generalizability issues, since models developed on one population may not generalize well on other populations or settings. As such it is essential to assess the external validity of the model and to validate it on an independent data set, which can be hard in some projects. In this work the external data set schema is considered in chapter 9.

Addressing these and other challenges requires careful consideration on what are the most appropriate approaches, and how they work with feature engineering, model construction and model evaluation. These are taken into consideration from chapters 4 to 9. Important to reference that for its relation with medicine, the focus of this thesis is on regression based algorithms and approaches for each of these areas. These will be referenced next.

3.2 Feature Engineering and TTE

In time-to-event prediction, accurate and informative **feature engineering** is important to produce effective models. This deals with transforming raw data to create new, stable inputs for data modeling and prediction. It can also refer to the construction of new features [35].

In this work it is relevant to subdivide feature engineering into **feature construction** [74] and **feature learning** [14]. While with the first the focus is on the creation of new features from raw data, the later deals with encountering new representations of data.

In this sub-section a set of algorithms for feature selection are also presented [97]. These consist on general purpose algorithms that can also work in the model construction phase. The selected algorithms were: Elastic Net [123] (EN), Lasso (L1-regularization, also known as Least Absolute Shrinkage and Selection Operator) (LA) [108] and Ridge (L2-regularization) (RI) [55]. A set of other algorithms are presented in the next sub-section, namely Linear Regression (LR), Decision Tree (DT), Random Forest (RF), and Support Vector Machines (SV), despite focused on the prediction phase. In this case, EN, LA and RI are refereed in feature selection for their capabilities in working with multicolinearity issues. For a complete study on regression methods please see [54].

Refocusing on **feature construction** [35, 74], it involves identifying and transforming the available predictors to create new features that are more informative for a prediction task. Its importance is emphasized by the number and nature of available predictors, which can vary depending on the application. Of note that the number and size of available predictors can also make it challenging to develop accurate, highly performing, and interpretable models. A few of the most common feature transformation techniques are explored in table 3.2.

As the field continues to advance, it is likely that new and innovative methods for feature construction will play an increasingly important role in improving the predictive performance of these models. One such case is related with the development of deep learning based approaches such as **feature learning** [14] methods.

Transformation	Operation	Description
Non-Numeric to Numeric	Count	Count the number of times each value appears
	One-hot Encoding	Process to transform categorical variables in order they can be better processed by ML algorithms
	"Hash Trick"	Process to transform a string to a numeric equivalent (*e.g.* apply a hexadecimal function)
	Leave-one-out Encoding	Similar to one-hot encoding but discards current row's target in order to reduce the effect of outliers.
Single Variable Transformations	Mathematical Transformations	Apply different sets of mathematical transformations (*e.g.* x^2, $sqrt(x)$, $log(x)$) If the variables are positive and skewed (*e.g.* their distribution has a long right or left tail) apply $log(x)$.
	Bucketing/ Binning (Data Discretization)	When converting numeric values into ranges (*e.g.* instead of using the age of a set of patients it can be important to replace the value with interval values). Use in presence of noisy data. Algorithms: equal-width (distance) partitioning; equal-depth (frequency) partitioning.
	Scaling	Apply equivalent scaling to variables (*e.g.* a data set shouldn't have a variable in kilometres and another variable in centimetres)
	Standardization	Process of transformation the data set to have zero mean and unit variance using for example the equation: $x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$
	Normalization	Process of converting the values in a column into a normalised range (*e.g.* converting heights into cm values for records between 100 cm to 220 cm). In this case the process can consider 0 as the mean value of all the records with a standard deviation of 1.
	BoxCox	Process of using the power transformations developed by statisticians George E.P.Box and David Cox in 1964 to identify an appropriate exponent.
Two-Variable Combinations	Add	Sum similar-scaled variables
	Subtract	When in the presence of time-relative data sets a transformation which can be applied in the difference to a single time point
	Scaling	
Model Based Approaches	PCA	PCA can be applied in order to compact the information in a data set. ML algorithms should use the results of the PCA in order to perform their predictive tasks.
	Factor Analysis	Factor analysis is a method aimed at reducing a large number of variables into a smaller set of factors. It is valuable for extracting the maximum common variance from all variables and is part of the general linear model (GLM).
	Clustering	Clustering is an unsupervised machine learning technique which groups unlabeled examples, based on their similarity to each other.

Table 3.2: High level overview of feature transformation approaches (adapted from [35])

These approaches can be seen as an improvement over manual feature construction methods since they allow the machine to learn representations and use this knowledge in the mechanics of prediction models. Nonetheless, it is worth noting that by losing user input they tend to suffer from a loss of interpretability which undermines their acceptance by non-technical users.

3.2.1 Feature Construction

As stated, feature construction deals with manual or automated approaches to generate new, more informative features from raw data records [74]. Overall, for time-to-event prediction, there has been some progress in recent times, namely regarding the appearance of new approaches to deal with network and sequence data. In either cases, algorithms vary from simple transformations of available predictors, to more complex approaches focused on, for example, deep learning, or incorporating mathematical algorithms. Of note that both can in turn be integrated with more complex model construction necessities by streamlining the feature construction and model constructions processes. Next, these items are taken into consideration and used for assessing feature construction for genealogical tree data sets.

3.2.1.1 From genealogical trees to feature construction

Overall, this work focuses on the use of genealogical data in time-to-event prediction scenarios. In such cases data sets are characterized by complex relationships between the observations which can make it challenging to construct informative features.

In simple terms genealogical data describes familial relationships between individuals who are related to one another through a common ancestor. These relationships can be seen as connectors, and can provide valuable information about groups of events, such as the risk of diseases, as well as key genetic or environmental factors that are shared across family members ([38]). In such cases, a simple abstraction approach is to see these data sets as graph-based data, where the individuals correspond to **nodes** and the relationships to **connectors** (e.g., edges), that in turn assist in the task of informing about the interactions that connect entities. In this case the challenge of constructing informative features relies in identifying the key network characteristics that are associated with the outcome of interest. Afterwards, one can resort to classic machine learning algorithms to construct models to answer to business needs, in the form of predictions. This was the approach followed in this work.

To navigate from **genealogical trees** to **feature construction** it is relevant to survey the scientific literature while focusing on medical and computer science articles.

While the literature is a bit reduced, one of the most interesting works resorts to genealogical data sets to study different social aspects of a group of individuals. So, in [38], authors identify a set of 21 features to reveal lifespan patterns in human population. To extract these features authors considered genealogical trees as networks of nodes and resorted to statistical measurements to characterize lifespan variations over time. Later, they constructed a classification-based model

to identify the main characteristics of individuals who outlived 50, which would also outlive 80 years, thus focusing on a specific subset of individuals.

When diverging from social sciences and entering into scientific networks, the type of features found are focused on graph, vertex and edge based statistics. In [113], authors analyzed the evolution of the number of *entering* and *exiting* nodes to identify writing trends, while evaluating their scientific impact over time. Their objective was to evaluate the degree of importance of scientific papers over time while their approach was to model the scientific papers as networks of data.

When inferring as to the usage and application of genealogical trees in the medical sciences to assess events, it is clear that researchers resort to these data structures to store data to be later target of manual analysis. Their focus deals with extracting information regarding the ancestor that transmitted the disease and little more. For reference purposes and as an early example, please check [17]. In this work authors take a data set from patients diagnosed with Huntington disease and use as features year of onset, year of birth, gender of transmitting ancestor and gender of offspring to delineate a linear model that infers as to the age of onset of a patient diagnosed with Huntington disease. In this case they relate a subject's AOO with the age-of-onset of the transmitting ancestor and age at the moment of the affected individual's birth. They conclude as to the independence of the gender of affected child with transmitting ancestor. Another example can be seen in [96], where the purpose is to identify key factors that influence age of symptoms onset of *Autosomal Dominant Alzheimer Disease* (ADAD). In this case, authors assess the contribution of parental age of onset, age of onset by mutation type, family contribution, besides APOE genotype and patient gender. In this case, they identify a correspondence of the variance of the age of onset with the type of mutation.

From the above works it is clear the focus on establishing relations between a patient or subject age-of-onset with that of the transmitting ancestor, neglecting other familial relations such as brothers, sisters or cousins. One reason can be the need to evaluate how different levels of information can impact the results. As such, the need to take several different types of familial connections into account suggests it may be important to summarize genealogical trees. This is evaluated next and can be seen in chapter 5.

Notwithstanding, from a clinical perspective it is noteworthy that the effect of environmental factors should be evaluated, since over time the evolution of different habits can impact ones health. For the need to collect different types of information this will not be subject of study in this thesis.

3.2.1.2 Summarizing genealogical trees

From the previous literature review an idea emerged: the importance to summarize genealogical trees. This lead to a key question, namely ***what type of information is relevant when inferring general hypothesis regarding hereditary diseases***?

To allow for a computational based answer it is important to evaluate what frameworks exist that can help to indicate a more general path for the summarization (synopsis) of genealogical trees. So, the next stage of this research lead to the works of [46] and [45]. From these it was

possible to compile a list of metrics that leverage genealogical based characteristics. The selected characteristics consist of:

- number of women or male relatives that have a specific strain of a disease;
- number of women or male relatives present in a genealogy;
- percentage of genealogical data completeness;
- number of children of a couple;
- generation number (class) of an individual;
- ratio of family members that are or were affected by an illness, taking into account their familiar relationships.

These served as initial requirements for the information to be gathered from multi-rooted pedigrees and which ended up differentiating the genealogical trees. Chapters 5 and 9 present the work that followed this survey as they list the most relevant manual features extracted from the genealogical trees studied.

After this assessment another key question emerged. It consisted on ***if it was possible to define a rank for different sets of features and what approaches could be used to assess their predictive value***? The background issues relevant to answer it are referenced next and its answer is on chapter 6.

3.2.2 Feature Selection

To answer the previous question the path lead to feature selection approaches. Variable selection approaches consist of algorithms and methodologies that identify which predictive variables should be retained in a model and a possible strategy consists on penalize the features accordingly to their degree of importance.

3.2.2.1 Penalized or shrinkage regression methods

When working on regression problems it is relevant to delineate approaches capable to deal with different sets of problems, compare the results and focus on the best approach. One example of a common problem consists on dealing with the possibility of over fitting. In simple terms this consists on developing models that don't generalize well, and it happens when the model developed becomes considerably attached to the training data set. Another possible problem consists in how to deal with multicollinearity issues. Of note that multicollinearity consists on having a predictor variable, or feature, highly correlated with other features which can generate models biased to specific situations. Other examples of possible problems consist on: working with a large number of variables, or working with a low ratio of instances per number of variables. To deal with a few of the most common data set problems one can resort to **shrinkage methods** [54]. In this work the

focus relies on Elastic Net (EN), Lasso (LA) and Ridge (RI) algorithms. Although resorting to penalized linear models constrains the sizes of the coefficients, it brings the need for a tuning phase to identify the optimal subset of relevant features. This introduces additional complexity to the knowledge data discovery pipeline and increases the overall computational cost of the modeling strategy [97], which can lead to the need of special evaluation strategies.

Lasso (L1-regularization) is designed for feature selection in sparse feature spaces. Since it reduces the number of variables upon which the given solution is dependent, it is considered useful in some contexts. This means it is good for variable selection and bad for group selection when the predictive variables are strongly correlated.

Ridge regression (L2-regularization) addresses some of the problems of Linear Regression by imposing a penalty on the size of coefficients. It is good for multicollinearity, and group feature selection, while having difficulties in variable selection cases.

Finally, **Elastic Net** is a model that is capable of dealing with multicollinearity issues that need variable selection (combines the strengths of Lasso and Ridge algorithms) [123].

After the summarizing works and the feature selection study, another important question emerged, namely ***if important information was being lost while resorting to such summarizing methods***. In this case the path lead to study **representation** and **feature learning** approaches and their results are depicted in chapter 7 while the associated background will be introduced next. As for the impact of feature selection by resorting to shrinkage methods, it is explored in chapter 6.

3.2.3 Representation Learning

Feature construction deals with the creation of new features based on raw data sets. This can be achieved in (semi) automatic approaches such as Principal Component Analysis (PCA) [59] which transforms a data set into a new set of features represented by their principal components. Both genealogical trees and representation learning involve the mapping of complex structures into a more computationally understandable format. In this case, just as a genealogical tree simplifies the understanding of family relationships, representation learning simplifies the understanding of complex data.

Another interesting aspect related with genealogical trees is that individuals (e.g., nodes) and relationships (e.g., edges) can appear with a time component, since they appear and disappear accordingly with event occurrences. In these cases the features (or reduction methods) need to evolve accordingly. In this work the chosen approach was to resort to embeddings to bypass temporal restrictions. This topic is introduced next from an academic perspective, while from a contribution point of view it is explored in chapter 7.

3.2.3.1 Evaluating possible loss of relevant information

Machine learning (ML) systems express complex real-world scenarios so as to transform inputs into descriptive, predictive or prescriptive answers. One of its principal phases corresponds to the

construction, selection, extraction, and engineering of features [74]. This consists on the application of sets of constructive operators over existing data records, to learn or generate new sets of predictors [74]. One of the most interesting approaches deals with predicting while resorting to network data. Here problems acquire a different dimension since prediction over nodes or edges, sometimes with directionally concerns, requires approaches to handle time-related characteristics [50]. One field of research that deals with different representation learning aspects in these complex scenarios is network embedding (NE).

Network embedding (NE) is a research field that focuses on the representation of complex systems in low dimensional spaces. Since its creation, researchers have dealt with the conceptualization, development, and application of different solutions, namely: analysis of social networks; language modeling, for example, by studying co-occurrence of words; and networks of communication [48]. NE algorithms focus on resourcing to the underlying structure present in network data, its vertex content, as well as any other available secondary information [119], while providing answers to intricate questions. Of note that several "new" approaches are developed on top of graph convolutional networks (GCN). This is the case in the works of [105] and [65]. In the first reference, authors propose a general framework that learns embeddings of emails and users to improve a classification task over state-of-the-art approaches. In the second case, authors propose a node embedding approach to disentangle the information shared by the structure of a network and the characteristics presented in its nodes, with a multitask graph convolutional network variational autoencoder. For this, different dimensions of the final embeddings are dedicated to encode feature information, network structure, and also shared feature-network information.

Embeddings correspond to numerical vectors that represent discrete and relational objects, into a fixed set of numerical features. In the early years of its usage, embeddings were useful as a form of dimensionality reduction [22]. These *arrays* of numbers, which lack human meaning, are obtained at a feature construction phase. Further transformation approaches, focused, for example, on retrieving distance or location patterns, help data-mining algorithms in solving specific problems.

In general terms, given a set of independent and identically distributed (i.i.d.) data points, graph embedding algorithms construct an affinity graph, and embed them in a new lower-dimensional space. With this, the algorithm transforms itself into a new model of lower dimensionality while maintaining the main characteristics of the network, since connected vertices are kept closer to each other in the new embedding space [119]. Current research trends focus on one of two paths: (i) learning the best representation expressing multipurpose scenarios when combining graphs, their sub-components, or sub-parts [75]; and node embedding approaches, where the focus is on representing nodes or their connections [44] to express meaningful patterns. In the first case, current works extend the **graph2vec** [75] algorithm, while in the latter most significant contributions are based on **node2vec** [50] or **edge2vec** [44] algorithms.

A **node2vec** model is constructed by learning a mapping of nodes in a low-dimensional space of features that maximizes the likelihood of preserving the actual neighborhood of a node [50].

A **graph2vec** model, on the other hand, extends document embedding neural networks to learn

representations of entire graphs. It forces structurally similar graphs to be closer to each other in the embedding space [75].

An **edge2vec** model extends information regarding connections between individuals to leverage them on the representation onto a new feature space [44]. With the latest embedding innovations, it became possible to learn task-agnostic representations of complex networks and to use these features as inputs in different prediction algorithms. Chapter 7 assesses the impact of summarizing genealogical trees through manual feature extraction, by exploring a new feature construction methodology based on embeddings, and comparing both.

3.2.4 Discussion

The previous background reveals key aspects to retain.

Current studies examining how familial relationships influence event prediction often simplify the problem and focus solely on a subject's ancestors, neglecting other familial relationships (siblings). This fact suggests that evaluating different familial relationships could be beneficial. Additionally, researchers resort to genealogical trees in a simplistic manner. Particularly, they do not assess the impact of time and use it mostly for descriptive or visual assessment in genealogical studies. This suggests that developing methods that are able to condense the information present in genealogical trees, and integrate it into prediction models can represent a valuable application contribution. This corresponds to the manual feature construction approach developed.

Another key consideration consists on developing a method to evaluate the potential loss of relevant information, triggered by manual feature construction over genealogical trees. In this case the focus is on representation learning approaches and how they can be used to capture information from genealogical trees to represent individuals. One possible path is to convert the graphs (e.g., genealogical trees) into low-dimensional feature vectors. It's worth noting that in genealogical data, each individual is represented as a node in a graph, with edges connecting individuals who share a familial relationship. This suggests leveraging current graph-embedding approaches to create genealogical-based embeddings. After constructing the embeddings, the next step is to evaluate their effectiveness as input features for traditional time-to-event prediction models. This corresponds to the automatic feature construction approach developed and is connected with the model evaluation contributions.

Lastly, and since feature construction and representation learning can lead to over ambiguous sets of features, it can be important to assess the most important sets of covariates (or feature sets). In this case another research path can be to resort to feature selection with penalized or shrinkage methods to assist in evaluating each approach. In this case it can be also relevant to resort to PCA, with the loss of explainability it entails, since features loose actual meaning, to further enhance this path.

3.3 Model Construction and TTE

Understanding, comparing and evaluating various approaches is paramount for a successfully completion of a prediction task. Some of the key considerations involved in constructing time-to-event prediction models, which include the choice of an appropriate algorithm, which modeling approach to use, and what are the main considerations for developing a proper evaluation for model performance, are explored next. The focus is on Regression [53] (see chapters 5 to 7 for application results), Ensemble Learning [70] (see chapter 8 for application results) as well as Classification [42] tasks (see chapter 9 for application results). The two last with considerations regarding the need for an easy adaptation from regression models.

3.3.1 Regression

Regression algorithms model the relationship between one or more independent variables, also known as **predictors**, **covariates** or **regressors**, and a dependent variable, also known as the **response** or **target** variable [53]. When working under the constraint that the response variable has a continuous value, its goal is to estimate the effect of the independent variables on the dependent response. Depending on the case, it is possible to resort to mathematical functions to model such relationships.

In more objective and mathematical terms, regression modeling consists on approximating a function f from a set of input variables X to a continuous target variable y, which can later be used to estimate the relationship between the input variables, to the output target. There are several relevant algorithms considered state of the art [77] and a few of them are explored next.

3.3.1.1 Algorithms

For this work, the best typification of regression methods corresponds to **Multivariate Linear Regression**, **Tree-Based methods**, and **Support Vector Machine** Algorithms [30].

When doweling over **Multivariate Linear Regression Algorithms**, one focuses on Linear Regression (LR), Lasso, referenced as L1-regularization or as Least Absolute Shrinkage or Selection Operator (LA) [108], Ridge, referenced as L2-regularization (RI) [55], and Elastic Net, reference also as L1 and L2 regularization (EN) [122].

For **Tree-Based methods** one can work with Decision Tree (DT) [18], Random Forest (RF) [19], and XGBoost (XG) [23].

For **Support Vector Machine** algorithms the focus is on Support Vector Regressor (SVR) [100].

To deal with some of the most common problems in regression problems such as overfitting, multicolinearity, and low ratio of instances per variables, in [54], authors suggest using Lasso (LA). Despite LA was explored previously, namely in the feature selection phase, the main characteristics of Lasso (LA), Ridge (RI) and Elastic Net (EN) will be presented next, from a prediction perspective, so as to allow for their comparison with the other regression algorithms.

Lasso Regression (LA) helps to select features in sparse feature spaces. It is useful to reduce the number of variables upon which the solution is dependent, making it a good choice for variable selection, but a bad choice for group selection when there are predictive variables that are strongly correlated [122].

It is noteworthy to reference that *Ridge regression* (RI) also addresses some of the problems of Linear Regression by imposing a penalty on the size of the coefficients. Considered to be good to deal with multicollinearity and group feature selection, it can have adverse issues when working towards single variable selection [122].

Elastic Net (EN) was born by the need to merge LA and RI. As such, it is a model capable of dealing with multicollinearity issues that need variable selection [122].

As for *Decision Tree Regressors* (DT) [42], their focus is on predicting quantitative values by partitioning the predictor variables into non-overlapping regions. In this case partitioning continues until all of the instance records in each leaf-node have the same target value, or until some of predefined size limits are reached. If the latter happens, then the predicted value needs to be measured. How? By calculating the average the target values of all the records that share the same leaf. This corresponds to a simple and well regarded approach, with the advantage of producing very explainable results.

Random Forest (RF) [42, 70] generally improves the performance of a modeling approach, by resorting to large numbers of different sets of *Decision Trees* (DT). Of note that this is considered an ensemble based algorithm which captures properties of a data set while leveraging the amount of available data. It works by training different decision trees on, for example, a different subset of the training data, as well as randomly selecting a subset of the features at each split node based on the split criterion. In this case the prediction results are obtained by introducing a new record to all trees and returning the average values encountered.

XGBoost (XG) [23] is one of the most used approaches to answer to a wide range of predictive questions. It stands for Extreme Gradient Boosting and corresponds to an implementation that conforms to Gradient Boosting, Stochastic Gradient Boosting or Regularized Gradient Boosting. It has high computational and performance speed, is capable to automatically handle missing data values. From an engineering perspective its most important characteristisc relates with the way it parallelizes the tree construction procedure, while still being updatable when new data is available, since it can greatly impact the re-training phase of a model.

Finally, *Support-Vector Machine* (SV) is an algorithm that uses a hyperplane, or a set of hyperplanes, and focuses on classification, regression or outlier detection tasks. In the special case of Support Vector Regressor, one defines a limit (ε) to find a function $f(x)$ that has at most ε deviation from the obtained targets (y_i). In other words, this is an algorithm that, based on a threshold (ε), fits a line to the training data by minimizing a cost function [100].

3.3.1.2 Answering time dependent questions

In the course of a regression based modeling approach, time-oriented responses are crucial, specially when dealing with time-changing covariates. They assist in understanding and predicting

the behavior of a variable over time, while taking into account the temporal dependencies that exist in the data [53]. This is relevant to deal with the independence assumption.

Important to reference that several approaches can be used to deal with answering time dependent questions.

In a time-oriented regression model, the response variable is typically a function of time, and the predictors can include lagged values of the response variable, known as autoregressive terms. Another approach is to transform the covariates and have independent models to deal with time changes. This is considered a more understandable approach for non-technical users and serves as the foundation for this work's approach.

In this work the main approach developed to deal with time changing effects for time-to-event prediction questions deals with transforming the covariates to conform to time changes. This substantiates to having models focused on the characteristics of patients at specific ages and it is relevant since the patients evolve over time thus increasing the risk of developing symptoms. Of note that in these cases the validation phase corresponds to an important step in the prediction pipeline, due to its need to evaluate the underlying data-generating process and to account for independence over the covariates.

3.3.2 Ensemble Learning

In regression, base learners forecast a value given a feature vector [53]. This feature vector corresponds to a set of independent inputs. Important to note that different regression algorithms operating over the same problem and while working with the same data sets can generate different responses, specially when working with data sets that contain outliers, are unbalanced, and/or contain missing information. In such cases ensemble methods [70] can be used to obtain better predictive performance, as well as to deal with problems of traditional regression or classification approaches.

In general terms, the basic idea of ensemble learning is to create a group (or ensemble) of models that work together to produce a more accurate prediction than any single model could achieve alone and it can be constructed in different ways. One of the most simple model construction approaches consists of averaging a group of predictions (regression setting) generated by multiple models, sometimes instantiated by the same algorithm but with different parameters. Another approach is to resort to a voting mechanism (classification setting) to select the most common prediction.

On a wider spectrum, ensemble methods represent generalization approaches that combine methods created with the joint purpose of predicting a specific outcome [70]. Since the strategies often consist of leveraging statistical concepts, such as mean or median functions, or concatenating sequential results, it can be done in stages. In the first stage, i.e. **model construction**, one can build different models with different training subsets or different algorithms. In the second stage, i.e. **model selection**, the aim is to reduce the number of models to combine. The third stage, which corresponds to the **integration** phase, combines the results of the chosen models. For this thesis, and for ensemble learning strategy engineered, the focus is on this third phase [70].

Of note that overall, the results obtained with ensemble models reduce the generalization error of individual learners. Standard strategies incorporate **bagging**, **boosting** and **stacking** approaches [70].

While **bagging** corresponds to the aggregation of multiple models trained and instantiated in parallel and **boosting** instantiates a new model sequentially to the previous one, **stacking** combines a set of different estimators (first-layer) using a meta-model that, depending on the type of model, combines them to maximize the robustness of the final model.

For its current research focus, ensemble learning benefits from diversity. In this case, by resorting to multiple models that have different strengths and weaknesses, the ensemble can cover a wider range of possible solutions and reduce the risk of over fitting, where the model performs well on the training data but poorly on new data. In this case diversity is achieved by using different algorithms, different subsets of the data, or by varying the hyper parameters of the models.

On a final note, ensemble learning has become increasingly popular since it can improve the accuracy and robustness of predictive models, and constitutes an ever growing field of application. Although there are approaches specially directed for ensemble learning, such as Random Forest (RF) [19] and Extreme Gradient Boost (XG) [39, 40], sometimes they are referenced as base learners since they act as already implemented tool sets. In this case and for this work, to delineate a strategy and since they were already used in the early attempts to answer the medical problem, it is relevant to consider them as base learners and further aggregate their predictive results. Practical results are presented in chapter 8.

3.3.2.1 Integration with regression

As stated previously, there are several approaches for an ensemble based predictive answer to a specific problem. One of the most well known cases is represented by the Random Forest (RF) since it combines several DTs, by merging individual results in a group answer. On a follow-up note, this can be seen as the base of the approach developed and presented in chapter 8. This is the case since by combining the results of individual weak learners it is possible to develop more stable and less variance prone results. Of note that this is still considered as a powerful technique although its impact is not always clear, since it can be easily grasped by non-technical individuals and produce more acceptable results.

It is noteworthy that in the case when several independent models are used to generate an ensemble predictive response, the objective is to reduce variance and bias, while improving final results. This exploits the capabilities of individual algorithms, each capturing different aspects of the data, and leveraging the strengths of each individual model, while mitigating their individual weaknesses.

3.3.3 Classification

Classification [42] consists in learning a target function f that maps each input set of attributes X_i with $i \in [1,2,...N[$, to one of the predefined target labels y_k. If $k = 2$ this is denominated as a

binary classification task, and if $k > 2$ it is known as a multi-classification problem. As such, in a more formal notation, and by considering a set of input data points, it can be represented as

$$(X_1, y_1), (X_2, y_2), ...(X_n, y_n) \tag{3.1}$$

where X_i corresponds to the feature vector and y_i corresponds to the target variable. In this case a prediction model learns the function f that with the input space X maps the target y. This target function, also known as a classification model, is useful both for **descriptive** and **predictive** modeling.

When focusing on the descriptive setting, the target function is used to summarize input features. As such, it is worth to indicate that in this work, there is special focus into which models can be applied to Time-to-Event prediction, with and without explicit feature engineering approaches.

3.3.3.1 Algorithms

Currently there are several state of the art algorithms applicable for classification based predictive questions. Most approaches rely on **decision trees** (DT), **random forests** (RF), and **support vector classifiers** (SVMs) [118].

In this case, both **SVM** [28] and **RF** [19] are popular algorithms used in a variety of applications. In the first case the algorithm blends linear modeling with instance-based learning, by selecting a small number of critical boundary samples from each category, thus building linear discriminate functions that separate each category. When no linear separation is possible the algorithm resorts to a kernel function to automatically inject the training samples into a higher-dimensional space. Although not easily understandable and loosing on key features such as interpretability, SVM is considered as one of the most reliable algorithms to tackle classification questions. On the other hand **RF** [42] is an ensemble learning method that, in the classification scenario, constructs a multitude of **DT** at training time, and that outputs the class that corresponds to the mode of the individual trees. RF corrects DT disadvantage of being prone to over fitting over the training set, and is widely recognized for its interpretability and highly accurate performance.

As for **DT** [118], when operating alone, the main difference when considering its regression counterpart relates with the cost function. In this case the cost considers the Gini Index metric, also known as Gini Impurity, which assesses the likelihood of an incorrect classification when a randomly selected data point is assigned to a target, based on the distribution of classes.

When focusing on **XG** [23], this represents an improvement over DT, since it integrates a process named boosting to improve performance. It also integrates pre-processing methods to handle categorical data, namely with binning and one-hot encoding procedures.

As for the **EN**, **LA** and **RI**, main characteristics state that they can handle data with multicolinearity issues. In binary classification these can be integrated with logistic regression, thus generating as output a probability referencing each class.

Lastly, regarding **LR** and binary classification it isn't suitable on its own as it needs to transform the results.

3.3.3.2 Integration with regression

Regression modeling deals with answering problems where the target variable has a continuous value (for example, predicting a person's weight or the price of a house). When it is relevant to model a discrete label or category, such as if a tumor is malignant or benign, classification comes into play [118]. Although one can consider it as a completely independent prediction problem, transforming a regression-based response into a classification prediction is considered as an interesting practice, when the output or target variable of interest is categorical (discrete). The transformation process involves defining a threshold to separate the continuous output into distinct classes. For example, if one has a regression model that predicts a person's risk of developing a certain disease based on various features (weight, age, gender), one might decide that any predicted value above 0.8 classifies the person as "at risk", while any result below that threshold classifies the person as "not at risk". The choice of the threshold can impact the performance of the classification model and must be chosen carefully, both with business as well as data fundamentals. Of note that one of the most common problems deals with having to deal with unbalanced issues. This can be dwelt with by evaluating with precision and recall metrics, or by applying smote to the training set [91].

3.3.4 Discussion

Overall, independently of the solution (classification, regression, ensemble, time-series, outlier detection) one of the main issues of model construction deals with the definition of the right amount of data needed. For models with small data, researchers can resort to components such as bootstrapping to increase the available background knowledge.

Regression is an important field of research in machine learning prediction, and in the health-care setting this is an even more pressing matter. In regression based modeling main tasks deal with the choice of the algorithm, the best feature set, as well as the diagnosis of the most relevant questions. Nonetheless it is important to keep in mind that the most relevant characteristics of regression algorithms focus on their: ability to deal with missing values; sensibility to outliers; dependence on specific characteristics of the data, such as normality or independence of the observations; and model interpretability. From the previously referenced algorithms, the ones that seem most promising for dealing with genealogical data sets correspond to: EN (seeing it combines both RI and LA advantages); DT (for their explanatory capabilities); RF (for its generalization capabilities, specially when compared with DT); and XG (for being a very scalable approach with good predictive capabilities).

Worth mentioning that after addressing a problem with a regression based setting it can be interesting to explore other approaches.

Ensemble learning is a powerful tool for improving both interpretability and predictive performance. In simple terms it explores approaches which combine multiple models, thus improving

weaker learning results. In addition, since multiple ML models are involved, this enhances reliability, improves stability, and ensures accurate predictions. In time it can lead to a better acceptance rate of prediction approaches from non-machine learning field experts. Another important characteristic of ensemble methods relates with the increase of time and resources needed during training time. For example, training 500 trees in a RF consumes more time and resources than with a single DT. So, it is relevant to keep in mind memory constraints when developing data oriented prediction answers.

On a final note, another approach can be to explore classification based answers. This area of application can assist in making the results more easily understood by non-technical users. As such, it can trigger the generation of new knowledge. In other cases and when integrated with regression based modeling it can assist in evaluating the relevance of sets of predictive questions. In this work this corresponds to the intended use case.

3.4 Assessing the utility of TTE models

In the healthcare setting, assessing the utility of a prediction model is essential to evaluate its performance, as well as to determine its clinical applicability. In this case it is deemed important that researchers employ a multitude of assessment approaches to measure their utility and value.

By providing a comprehensive overview of different assessment approaches, this section helps to understand the impact of choosing a method over the other. The focus relies on the metrics used considering the **classification**, **regression** or **ensemble** prediction results. A few key considerations over the best evaluation strategies are also presented.

Finally, this subsection also introduces a cost modeling approach that measures the utility of resorting to prediction based approaches to define the follow-up of patients positively diagnosed with a clinical condition.

3.4.1 Metrics based evaluation

The usefulness of any forecasting model often needs to be measured against specific standards. Overall, the choice of the best performance metric depends on the application and the goals of the analysis.

3.4.1.1 Regression based metrics

Over time, researchers developed and analyzed a diverse set of metrics to gauge different perspectives of prediction models. These provide actionable insights that go beyond a simple comparison of predicted and actual values.

Mean Absolute Error (MAE)

The mean absolute error (MAE) [42] corresponds to a statistical estimation of the absolute error of a model. For a given set of results, where y_{real} represents the observed results of a set of N records and y_{pred} represents the results obtained by a model, MAE is given by:

$$MAE = \frac{\sum_{i=1}^{N} |y_{real_i} - y_{pred_i}|}{N} \tag{3.2}$$

This is an easy to understand measure, as it is expressed in the same units as the predictive variables. It allows for non-technical and non-data expert researchers to easily understand the implications of the results. It has the downside of being less sensitive to very large errors, as it does not magnify the importance of bigger errors in the difference assessment.

Mean Squared Error (MSE)

On the other hand, mean squared error (MSE) [42], which is also known as mean squared deviation (MSD), corresponds to an estimator that measures the average squared difference between the estimated and the real values. Is is always non-negative, with values closer to zero representing better results. The equation is given as:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_{real_i} - y_{pred_i})^2 \tag{3.3}$$

The major drawback is that the results are expressed in a metric that differs from the target value, which complicates interpretation.

Root Mean Squared Error (RMSE)

In a regression model, the root mean squared error (RMSE) [42] represents the average difference between the predicted and the actual values. Mathematically, it represents the standard deviation of the residuals, and it indicates the how well the data fits the model (*e.g.* how equal the observed and predicted points are). Its formula is represented by:

$$RMSE = \sqrt{MSE} \tag{3.4}$$

It has as main disadvantage the fact of being hard to interpret, due to the *sqrt* of the results. It is also quite sensitive to outliers, which can distort the interpretation of the results.

R Squared

The coefficient of determination [53], also known as R^2 or r^2 provides a measure of how well a model incorporates the variance present on the data, and is defined as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{3.5}$$

This is a metric useful for descriptive scenarios since it answers if a model is adequate or not. In regression predictive settings it can lead to erroneous assumptions. As such its usage depends if one is conducting a predictive or descriptive study.

3.4.1.2 Classification based metrics

The landscape of classification models demands for an understanding as to their efficiency. To unravel the performance of these models, a group of classification-based metrics has emerged, offering a multifaceted evaluation of the predictive value of a set of results. Next a list of the pros and cons of a few of the most important metrics is presented.

For reference purposes, in this context, **TN** corresponds to True Negative, **TP** corresponds to True Positive, **FP** corresponds to False Positive and **FN** corresponds to False Negative.

Precision

Precision [42] refers to the ability of the model to correctly identify positive data points.

$$Precision = \frac{TP}{FP+TP} \tag{3.6}$$

Accuracy

Accuracy [42] informs as to the rate of the values that were classified correctly. It is calculated as the sum of all correct classifications divided by all the values.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{3.7}$$

Its usefulness ends when the problem is imbalanced, as the results can be misleading.

Recall

Recall [42] is also known as sensitivity and measures the percentage of correctly classified positives among the entire sum of actual positives.

$$Recall = \frac{TP}{TP+FN} \tag{3.8}$$

It is helpful when in the presence of an imbalanced data set, or when the goal is to identify a few critical cases among a larger sample (outlier or fraud detection).

Specificity

Specificity [42] also known as true negative rate measures how well a model can identify true negatives.

$$Specificity = \frac{TN}{TN+FP} \tag{3.9}$$

F1-Score

F1-Score [42] combines Precision and Recall into a single metric.

$$F1-Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.10}$$

3.4.2 Evaluation strategies

Evaluating the performance of machine learning models is critical in the course of model development life-cycle. The choice of an appropriate evaluation schema can impact the reliability and generalization assumptions of a model, and conversely lead to its acceptance or rejection. Next, a few of the most important considerations over the evaluation strategies used in this work are shared.

3.4.2.1 Holdout

When resorting to holdout [62], a data set is divided into two distinct disjoint sets: a **training set** used for model training, and a **holdout** or **testing** set, set aside and reserved for assessing the performance of the model.

Over the model construction phase, the training of the model is exclusively done on the training set, and its effectiveness is then evaluated with the holdout set. This is a quick schema that evaluates model performance in a rapid manner. It is used during initial assessments and considered suitable when the number of training and testing samples is high. In this case, it is relevant to ensure that the split reflects the characteristics of the overall data set, and, if need be, it can be combined with additional techniques like stratified or temporal based sampling, to maintain class distribution or record sequence.

3.4.2.2 Train, Test, Validation

The **train-test-validation** [42] schema is a strategy that involves dividing a data set into three distinct subsets: the **training** set used to train the model, the **test** set used for example for tuning the parameters of the model, and the **validation** set used to assess the predictive accuracy and reliability of the model.

While the training set is used for training the model, thus allowing it to learn patterns and relationships within the data, the test set is employed during model development to, for example, fine-tune a set of hyper parameters of the proposed algorithm, besides assisting in specific decisions about the model's architecture. Of note that this subset acts as an additional safeguard against over fitting, by providing a set of data records not used in the training process.

The final set named as the validation set remains unseen during the training and tuning phases and serves as an independent data set to assess the model's performance on newly, previously unseen examples.

This schema offers a balanced approach, ensuring that the model is trained on a sufficiently large data set, evaluated on independent data, and fine-tuned with a separate set, thus contributing to the development of more robust and reliable machine learning models.

Its main disadvantage deals with the need to have a large-enough data set, since it is important to subdivide them on three bulks of data. But since the need to have a large-enough data set is one of the most important facts in model development it has to be considered as a normal handicap in most cases.

3.4.2.3 Cross validation

Cross validation (CV) [62] is a technique employed to compare different modeling approaches by assessing the predictive performance of each model. This is a well-known robust procedure that starts by dividing the data set into a *k* number of slots. Afterwards, interactively, it takes the records of *k-1* of these slots and trains them in a specific predictive modeling pipeline. It then tests the results on the unused slot and measures the predictive performance of this instance of the model. This procedure is performed a total of *k* times, with the final result measures being the average values of all the *k* runs [42].

One of the important aspects when dealing with cross validation procedures is the definition of the evaluation metrics for the different runs of the model since it is relevant to check if over the different experiments the results have relatively low standard deviation (SD). If it is high enough, it shows possible problems in the data set or in the different phases before model training (e.g., feature construction, data transformations, pre-processing). Another problem can be related with the existence of drift.

Of note that cross validation can be applied to perform an informed model selection and evaluation. Related with this, an in-depth analysis of the work of [9] was performed. In this case the authors assessed different types of cross validation procedures to be used according to, mainly, the volume of the data set and the problem approach (*i.e.* classification versus regression vs time oriented data). These different approaches vary accordingly with the initial problems, the modeling approach, the amount of available data, and the algorithms used for model selection.

To nest or not to nest

Nested cross validation [21] corresponds to an advanced statistical-based technique that addresses the potential bias introduced by standard cross validation during hyper-parameter and model selection phases.

In traditional cross validation, a data set is split into multiple folds for training and validation phases. This process is then repeated for hyper parameter tuning. One of its main issues is that it can lead to over fitting of the validation set, as the same data is used for both hyper-parameter selection and performance evaluation.

Nested cross-validation mitigates over fitting by introducing an outer loop that performs the standard cross-validation for model evaluation and an inner loop that has the special purpose of focusing on hyper parameter tuning.

By interactively partitioning the data set into training and test sets in both the inner and outer loops, nested cross-validation provides a robust framework for unbiased model evaluation. This methodology is particularly valuable when working with limited data or when the choice of hyper parameters significantly influences model performance. This nested structure allows for a more accurate assessment of a model's generalization performance, ensuring that the hyper parameters selected during the inner loop are tested on completely unseen data in the outer loop, providing a more realistic estimate of model performance on new, unseen data.

3.4.3 Cost based assessment

Cost estimation is relevant in healthcare, as it influences accessibility, quality, and financial sustainability of healthcare services. The greatest impact occurs during budgeting planning activities since accuracy can ultimately influence the treatments available to patients afflicted by highly debilitating and impacting diseases.

In clinical terms, and to allow for an earliest as possible treatment time, asymptomatic patients known to be carriers of a disease should be monitored for symptoms [103]. In the case of ATTRv Amyloidosis [15], due to the highly variable penetrance across different worldwide cohorts, as well as wide range of symptoms, it is not easy, even though current literature assesses the importance of symptoms onset inference methods.

Presently, patient cost evaluations focus on annual healthcare costs estimations, associated with management of patients, according to disease stage ([56], [57]). These stages occur after the symptoms onset [3].

Accordingly to the current literature, if left untreated, patients suffering of ATTRv Amyloidosis have a decreased mean life-expectancy of 12 years, after symptoms onset, with yearly mean healthcare costs of 125.645€ per patient ([56], [57]). This suggests a follow-up approach focused on prediction based results can greatly impact overall budgets in healthcare facilities. As such, there is a need to focus on what approaches are possible to assist management professionals in estimating the return of investment (ROI) of resorting to a prediction based approach to improve the follow-up of asymptomatic patients. Chapter 9 focuses on this topic, while also presenting an in-depth overview of the best prediction based AOO results.

3.4.4 Discussion

Metrics-based evaluation provides a quantitative assessment of a model's performance. While different metrics have different purposes they tend to provide an indication of the goodness-of-fit of a model, and therefore a measure of how well unseen samples are likely to be predicted by a specific model. This can be impacted by the chosen evaluation schema.

Model-based evaluation methods such as cross validation or holdout procedures are employed to evaluate machine learning model results, based on data characteristics. These methods provide a robust measure of the model's performance, by partitioning the data set into different sets of records and evaluating results on unseen records. Depending on the case, it can be relevant to follow data sequence characteristics as well as to evaluate other types of strategies such as time-oriented data split or sliding window to name a few.

Lastly, business-oriented evaluation focuses on the impact of model predictions on business metrics. This type of assessment is crucial in understanding the work's effectiveness on real cases, as well as its potential return on investment. It often involves domain-specific metrics or domain-expert evaluations, and requires a deep understanding of the business context.

Next section presents an overview of key characteristics of genealogical data sets and the best approaches to deal with them.

3.5 Genealogical studies

3.5.1 Introduction

Event prediction is considered relevant in healthcare since it can impact patient care, help to optimize resource allocation, as well as enhance the overall efficiency of healthcare systems. Some of the key medical areas where event prediction plays a crucial role include **disease diagnosis and progression**, where it assists in predicting the progression of existing conditions. These results can assist in developing personalized treatment approaches.

In recent years, the fusion of computational-based research with the field of genealogy has led to an evolution in the ability to trace and understand human ancestry. This synergy breathed new life into the age-old pursuit of unraveling one's roots.

Presently, **genealogy** is considered as a research field that deals with the extraction of knowledge from families of individuals, while **genealogists** compile and use different historical records to explore *kinship* or *pedigree* hierarchies between different sets of people. These networks can then be used to understand the impact of demographic, sociological, as well as epidemiological events [38].

In Portugal, a few of the most well known genealogical studies leaned over the study of typically Portuguese diseases such as *Transthyretin Familial Amyloid Polyneuropathy* (TTR-FAP), which initial research findings were conducted by Prof. Corino de Andrade and his team [102], and *Machado-Joseph* which is a disease originated from Azores and which early findings were lead by Prof. Dra. Paula Coutinho [29], also under the tuterage of Prof. Corino de Andrade.

Internationally, other studies dwelt with *Cystic Fibrosis* (Sag.-Lac-St.Jean - Canadá) and *Myotonic Dystrophy* (Saguenay – Canadá) [72].

An introductory overview of concepts related with genealogical studies is presented next. It serves as an introduction to relevant aspects that one needs to have in mind when developing computational based procedures that act on genealogical trees.

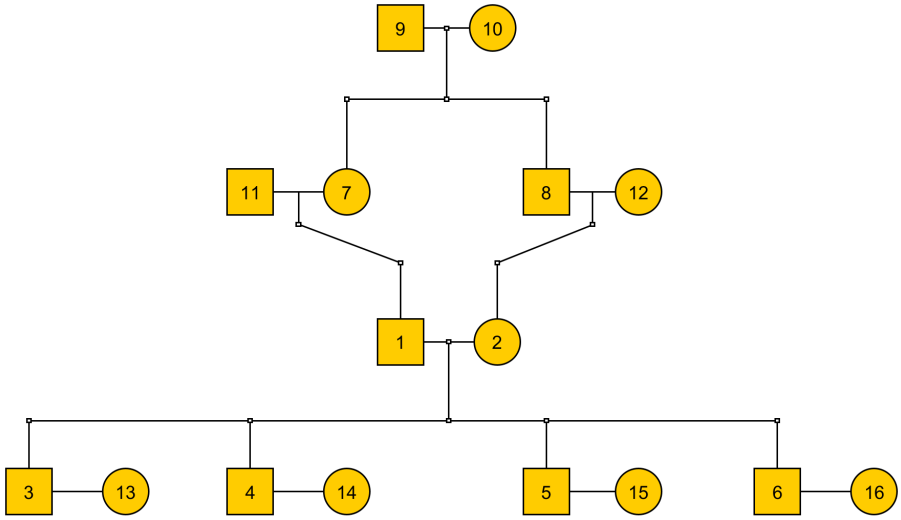
3.5.1.1 Family or kinship trees

A family or kinship tree, also known as genealogical or *pedigree* chart consists on a visual representation of families, independent of their size. These visual representations, which can be directly transformed into a directed graph, represent an abstraction of two types of objects: vertices or nodes (which represent individuals) and edges (which represent family ties between individuals).

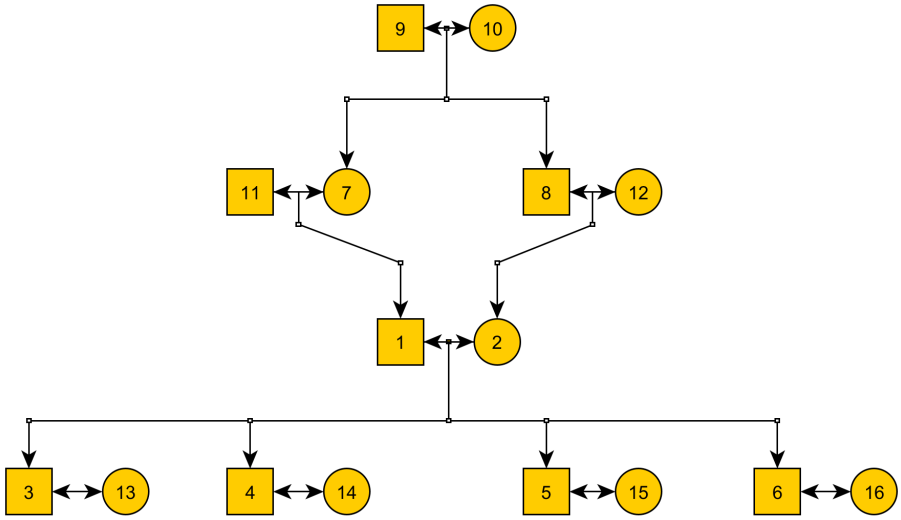
In the cases where there is a type of relation (father or mother to a son or daughter) the representation considers an edge with a start and an end node. In the cases there isn't a blood connection, for example when the relation is for marital concerns, then it can be considered as an undirected or bi-connected relation [10].

Figure 3.1 shows a family of individuals represented as a kinship. In this case figure 3.1.a represents the kinship structure, while a graph-based format is represented in figure 3.1.b. This

abstraction can help to identify possible metrics that can enhance the study of genealogical based diseases.



(a) Kinship Structure



(b) Graph or Network Structure

Figure 3.1: Transforming Family Trees in Graph based Structures

3.5.1.2 Dealing with genealogical data sets

When dealing with genealogical data sets it is important to have two main concerns in mind. In one hand it is important to evaluate its data quality standards, since data sources compiled over extensive periods of time can lead to inaccurate sources of data. On the other hand it is important to continuously process these data records since they can, in the mean time, suffer different types of corrections. As such a first step when resorting to these lakes of data records deals with researching

different types of problems that can occur, so as to easily identify sources of noise, corruptions or incorrect data records [36, 79].

A few of the most common and important sources of data quality issues in genealogical data sets consists of:

- structural differences in data records: historical records may be incomplete or have missing information such as birth dates, names, or locations.
- errors: transcription errors, typos, and data entry mistakes can introduce inaccuracies into genealogical databases. Thus cross-referencing and validating information from multiple sources must be taken into account.
- data quality fluctuations over time: individuals may be recorded under different names or aliases, especially when dealing with international or multicultural genealogies. In such cases it is relevant to be aware that variations in spelling or naming conventions, since these can lead to confusion.
- date discrepancies: discrepancies in dates, such as different birth or marriage dates for the same individual in different records, can complicate the construction of kinship trees.
- geographic changes: changes which can incur in borders or regarding the name of regions or streets can make it challenging to locate ancestral locations accurately.
- inconsistent formats: historical records may use different date formats, calendars, and languages, making data interpretation and standardization challenging.
- privacy concerns: in modern genealogy, privacy considerations are essential, especially when dealing with living individuals. Thus it is important to be mindful of data protection laws and ethical guidelines.

Table 3.3 completes the previous list of problems, since it shows the most well known occurrences that can impact any data extraction process, specially when dealing with historical unstructured data records compiled over long periods of time.

Condition	Description	Example
Proprietary File Format	Occurs when a file is designed to work only in a specific application/ system	-
Wrong Variables Type	Occurs when a variables has the wrong data type. This can generate problems when working with specific business intelligence integration processes as validation processes are more difficult to define (*e.g.* define an customer identifier as text.)	-
Tabular Data Design	Occurs when variables form both rows and columns and/ or column headers are values and not variable names.	-
Illegal Values	Occurs when variables have a value that is out of range	Onset Date = '31-12-2100'
Violation of Values logical dependencies	Occurs when the logical dependency of some values are broken.	Patient (age = 35, birth date = '19-09-1970')
Wrong Data Type or Syntax Violation	Occurs when a value does not respect the data entry constraints.	Patient (age = '19-09-1970')
Missing Data	Occurs when data is encoded with a dummy variable or not present.	Patient (age = NA)
Incorrect Data	Occurs when data contains valid values that do not correspond to reality.	Patient (age = 53) but actually his age is 35
Misspelled	Occurs when data is misspelled	Patient (country = 'Ingland')
Ambiguous Data	Occurs when data values have different interpretation.	Patient (Hospital = 'Maria') (*e.g.* it can be Hospital Santa Maria or Hospital Sta. Maria Pia)
Outdated Temporal Data	Occurs when data changed in a period in time (*e.g.* name of a street or postal code)	
Duplicates	Occurs when the same data appears mode than once.	Patient (id = 1, name = 'Elsa Manuela'); Patient (id = 2, name = 'Elsa Manuela')
Outliers	Occurs when a variable has an outlier which may indicate an error.	Patient (age = 99)
Missing Variables	Occurs when an important variable is missing for analysis (*e.g.* cardiovascular exams missing from a cardiovascular clinical trial).	-
Lack of Information Diversity	Occurs when a variable has few or under-represented unique values (*e.g.* all observations of country in a Portuguese Clinical Trial are set to Portugal, or the id column has a unique value for each record)	-

Table 3.3: Data quality conditions adapted from [79]

To deal with different data quality issues, researchers need to adopt practices that measure, evaluate and correct data records. A few of the most common ones include:

- implement mechanisms to cross-reference different sources of information.
- keep detailed information regarding the different sources of information, such as people, dates and file formats.
- standardize all possible data records such as names, dates and locations, in order to set high levels of consistency in the databases.
- use genealogy software as they offer tools for source citation and data validation.

More information regarding this issue can be obtained in [79] but it is important to reference that when first dealing with the genealogical trees from Unidade Corino de Andrade these and other problems were encountered. Although chapter 5 contains a list of the most relevant problems encountered during data cleaning phases this process had to be continuously refined so as to deal with the appearance of new problems.

3.6 Summary

From its inception, this work had two distinct purposes. While on one hand it was deemed important to explore the value of genealogical data sets, in the other hand it was considered relevant to focus on ATTRv Amyloidosis. These purposes, from a technical standpoint, suggest it is more relevant to focus the development on data driven assisted methodologies which can be rapidly transformed, depending on requirements changes.

So, from a technical standpoint, both statistical and machine learning-based approaches have sets of strengths and weaknesses, and can be used in different situations, depending on the features, target and research questions. While on one hand, statistical models may be preferred when the goal is to address specific hypotheses, on the other hand, machine learning approaches may be preferred when the focus is on making accurate predictions, or when the data is complex, suffers from high-dimensionality, or when it has conditions which are more suitable for data-driven personalized methodologies.

Also noteworthy that a careful data exploring procedure can assist on understanding possible future data drifts which can lead, from a business perspective, to additional difficulties in turning the results actionable. Of note that in the case of rare diseases this is an even more pressing concern since the data sets tend to be compiled over extensive periods of time, sometimes dispersed over different clinical units, with different guidelines.

On a final note, in this work, it is important to reference that when developing a data-driven approach for medical problems it is considered relevant to enforce a proper evaluation schema. As such in the application studies different types of evaluation strategies were taken into account.

Chapter 4

Study on the heterogeneity and variability of Portuguese families diagnosed with ATTRv Amyloidosis

Hereditary transthyretin amyloidosis, clinical named ATTRv amyloidosis, is an inherited disease with significant variability, where the study of family history holds importance in baseline evaluations and clinical follow-up. This study aims to evaluate the changes of age-of-onset and other age-related clinical factors.

Overall, this descriptive work is relevant since it gives a clear picture of the heterogeneity that surrounds carriers and their families, and expresses how their shared profile changed since the first patient was diagnosed. The initial format of this study is incorporated in this thesis since it was published early this year in Amyloid, The Journal of Protein Folding Disorders (for more information please check the publication in [82]). This is a medical journal which consistently publishes articles on amyloid based diseases.

On a quick summary and preview of the main conclusions achieved, this study reveals that there has been a consistent increase in elderly cases over recent decades, primarily among male proband patients. When focusing on the analysis of familial data, it shows that symptomatic patients exhibit less variability toward siblings, especially when compared to other family members, namely the transmitting ancestors' age-of-onset. It is also noteworthy that both the clinical as well as the genealogy-based indicators studied are characterized by high standard deviation (SD) values, which emphasizes the heterogeneity present in ATTRv Amyloidosis over the different decades of its presence in Northern Portugal.

4.1 Introduction

Genetic studies delve into the evolution of genes, heredity, and how certain traits are passed from parents to offspring [107].

Most genetic diseases are rare, thus affecting a small number of people, while others are more common and may be prevalent in specific populations or ethnic groups. Depending on the duration for which these groups are monitored by disease specialists, there is a need to document family history over time.

From a medical perspective, the notion of family history encompasses an individual's health and disease records, as well as their familial connections, both living and diseased. This historical survey helps to determine a patient's risk of developing specific clinical conditions. Its study is important as it can shed a light on the hereditary, as well as the biological or environmental factors that affect a patient's prognosis.

4.2 Objectives

This study aims to evaluate the changes of age-of-onset (AOO) and other age-related clinical factors, within and among families affected by ATTRv Amyloidosis.

4.3 Patients and methods

As analysis of the information in 934 genealogical trees was performed, encompassing family, close and extended relatives, especially siblings, ancestors and descendants. The clinical variability of the patients cohort was also assessed, specifically in relation to a key familial figure, known as the proband. This represents the initial patient diagnosed in a family and prompts further familial investigation. The results are presented alongside a comparison of current cases with historical records.

4.3.1 Subjects

At present, Portugal has two national reference centers entrusted with diagnosing and overseeing patients with symptoms linked to ATTRv Amyloidosis. Within the northern reference center, Unidade Corino de Andrade (UCA), which operates under the umbrella of Centro Hospitalar Universitário de Santo António, a total of 934 families are enlisted, with 4611 symptomatic and asymptomatic patients (49.89% male, 50.12% female). From these, 873 families (93.5%) are listed as ATTRV30M families. This is the variant with the highest expression in Portugal families.

4.3.2 Materials and Methods

To ensure a comprehensive evaluation of the entire patient cohort, and a distinct understanding of the characteristics of the most recent patients, the last, most well organized version of the data

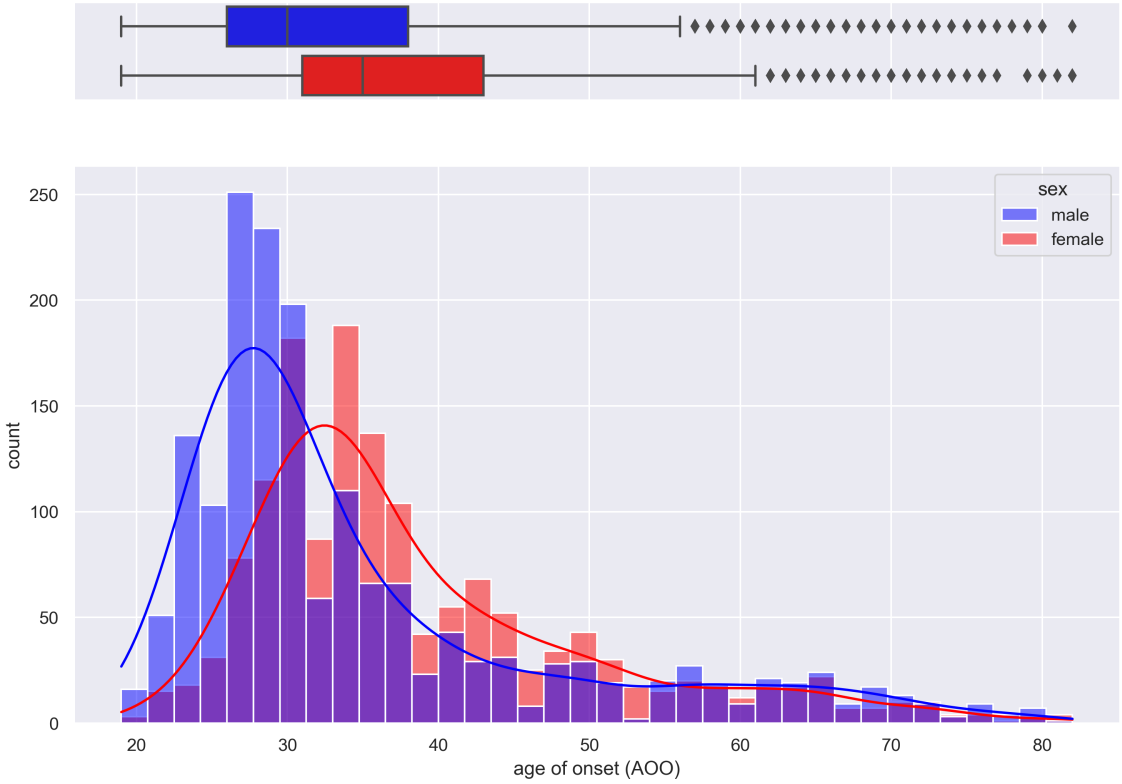


Figure 4.1: Age-of-onset (AOO) of all observed carriers. Blue shows the male symptomatic carriers AOO, while red shows the female symptomatic carriers AOO distribution. There is a difference close to five years for the peak of each distribution, with male patients having a risk of presenting symptoms earlier. Mann-Whitney has a p-value result below 0.05 which shows there is a significant difference in AOO between genders.

sets of clinical and genealogical information was divided into clusters (groups of symptomatic carriers), based on their year of diagnosis.

'Year of diagnosis' refers to the year when the primary medical professional overseeing a patient's clinical process accurately documents that the patient's onset already commenced.

After plotting the distribution of the age-of-onset of symptomatic patients (see figure 4.1), and performing several experiments, the chosen procedure was to divide the cohort into four equal sized distinct groups of patients. This way it is possible to account for the increase and decrease in the 1960s and 2010s decades, as well as the appearance of two distinct peaks of patients' diagnosis in between. The groups, henceforth referred to as Q1, Q2, Q3, and Q4 are illustrated in figure 4.2.

Each group (Q1, Q2, Q3, and Q4) comprises approximately 25% of the complete set of positively diagnosed patients, who have valid age of onset information. This includes valid year of birth and year of onset. This is crucial, as patients with unclear year of birth or onset cannot have their AOO accurately determined. With the stratification of the entire cohort, it is possible to compare both between and within each group. This assists in assessing the potential differences in the diagnostic procedure and evaluate the impact of advancements in disease research over time.

To align with recognized clinical categories, and within each category labeled as Q1, Q2, Q3,

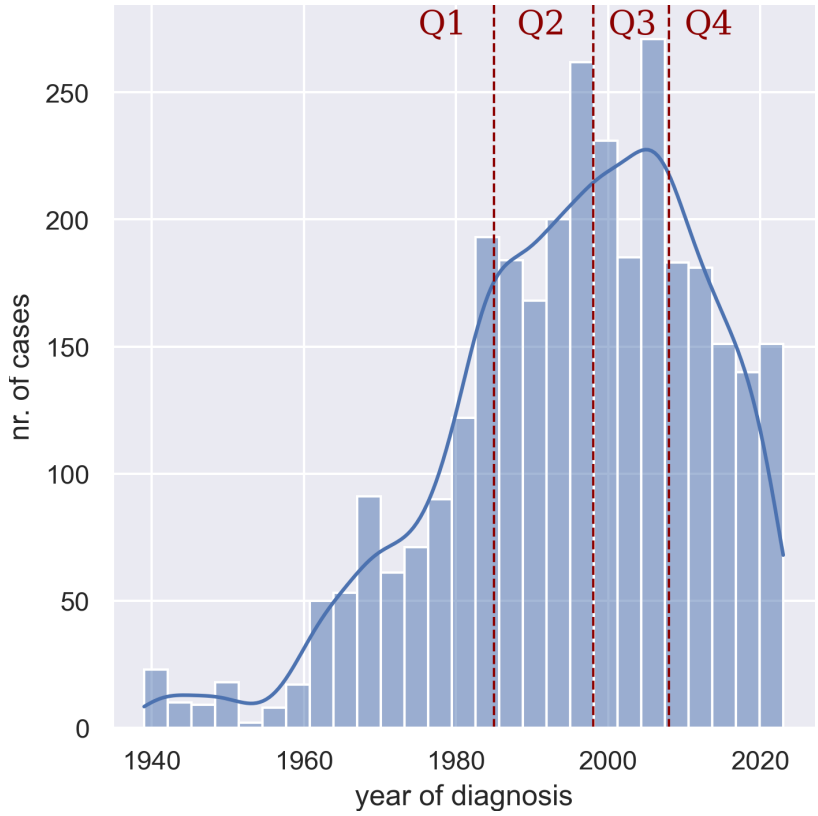


Figure 4.2: Distribution of positive carriers diagnosis showing the 25% (until 1985), 50% (until 1998), and 75% (until 2008) percentiles. While Q1 represents the patients diagnosed between 1939 and 1985 (46 years), Q2 represents the patients diagnosed between 1985 and 1998 (13 years), Q3 represents the patients diagnosed between 1998 and 2008 (10 years), and, Q4 represents the patients diagnosed after 2008 (15 years).

Table 4.1: Statistics of clinical indicators of symptomatic male patients. Bold values indicate smaller variability. Only significant differences obtained by the Mann-Whitney between each Q's and Q4 are present.

		early onset				***late onset***			
		Q1	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**
n		425	368	331	268	26	45	45	99
age.onset	average	29.4	30.1	30.1	32.2	55.8	59.1	60.2	64.2
	stdev	6.1	5.6	5.8	6.7	5.1	6.5	6.9	7.4
	coef.var	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1
	median	28	29	29	31	55	57	60	64
	range	[19,49]	[20,49]	[19,49]	[19,49]	[50,68]	[50,78]	[50,82]	[50,80]
	statistic	42654.5	40678.0	36686.0		479.0	1354.5	1524.0	
	p - value	< 0.001*	**0.0002***	**0.0003***		< 0.001*	**0.0002***	**0.002***	
age.diag	average	33	33.1	31.5	33.4	60.8	62.9	63.3	66.9
	stdev	6.5	6.4	6.4	7.1	5.8	6.8	7.4	7.8
	coef.var	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1
	median	31	32	30	32	60	63	62	68
	range	[19,54]	[22,60]	[20,57]	[20,54]	[52,74]	[52,79]	[50,84]	[51,82]
	statistic			37281.0		701.5	1552.0	1613.0	
	p-value			**0.0008***		0.0004	0.004	0.008	
time.diag	average	3.5	3	1.4	**1.3**	4.9	3.8	3.1	**2.7**
	stdev	2.8	2.6	1.7	1.5	2.8	2.5	2.7	2
	coef.var	0.8	0.9	1.2	1.2	0.6	0.7	0.9	0.7
	median	3	2	1	1	4.5	3	2	2
	range	[0,18]	[0,17]	[0,13]	[0,11]	[1,10]	[0,11]	[0,10]	[0,10]
	statistic	90330.5	72420.0			1897.5	2839.0		
	p-value	< 0.001*	< 0.001*			0.0002	0.007		
age.death	average	**40.8**	43.7	43.4	42.8	**67**	67.1	70.7	73.8
	stdev	7	8	8.6	7.1	6.1	5.6	5.8	6.3
	coef.var	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1
	median	40	43	43	42	66	67	71	75
	range	[24,78]	[26,71]	[24,65]	[30,59]	[56,81]	[56,81]	[57,85]	[57,85]
	statistic					212.0	318.0	573.5	
	p-value					0.0001	< 0.001	0.009	
time.death	average	11.3	13.2	11.2	**8.8**	10.6	8.5	10	**7.1**
	stdev	4.7	6.4	5.8	4.5	4.2	3.7	3.7	2.6
	coef.var	0.4	0.5	0.5	0.5	0.4	0.4	0.4	0.4
	median	11	12	11	9	9	8	10	7
	range	[1,55]	[1,32]	[2,27]	[1,20]	[6,21]	[3,18]	[3,20]	[0,13]
	statistic	5748.5	4734.5			763.0		1278.5	
	p-value	0.004	0.0004			0.0005		0.0001	

and Q4, the patient set was further divided into early and late onset. These are clinical categories that determine whether a patient is classified as having early onset, that is, if they exhibit symptoms before the age of 50, or late onset, which indicates if symptoms occur at the age of 50 or older. These sub-groups were then examined to check for differences in age-of-onset, age-at-diagnosis, time-to-diagnosis, age-at-death and time-to-death (refer to tables 4.1 and 4.2).

To assess genealogical variability, the focus went to the number of patients represented in the trees, during each of the time periods corresponding to the Q1, Q2, Q3, and Q4 clusters. This allowed for the categorization of the genealogical trees into **new families** or **old families**, within each individual group, as well as to quantify their changes over time.

Table 4.2: Statistics of clinical indicators of symptomatic female patients. Bold values indicate smaller variability. Only significant differences obtained by the Mann-Whitney between each Q's and Q4 are present.

		early onset				***late onset***			
		Q1	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**
n		290	321	316	271	16	51	62	75
age.onset	average	32.7	34.9	34.6	36	57.3	59.2	59.3	61.6
	stdev	5.7	6.1	6.3	6.3	6.1	7.5	7.1	8.3
	coef.var	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1
	median	32	34	34	35	55	57	58	60
	range	[22,49]	[21,49]	[21,49]	[20,49]	[51,70]	[50,76]	[50,79]	[50,82]
	statistic	27101.5	39070.0	37308.0					
	p-value	< 0.001	0.032	0.007					
age.diag	average	36.4	38	**36.1**	37.5	**61.3**	62.5	63	63.6
	stdev	6.1	7.3	6.9	6.8	6.4	7	7.5	8.3
	coef.var	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1
	median	35	37	35	36	59.5	61	63	64
	range	[24,56]	[22,62]	[22,59]	[21,60]	[52,75]	[51,77]	[51,83]	[51,84]
	statistic	35461.0		37603.5					
	p-value	0.045		0.011					
time.diag	average	3.6	3.1	**1.5**	1.6	3.9	3.4	3.6	**2**
	stdev	2.9	3.2	2	2	2.7	3	3.6	1.7
	coef.var	0.8	1	1.3	1.3	0.7	0.9	1	0.9
	median	3	2	1	1	3	3	3	2
	range	[0,15]	[0,20]	[0,16]	[0,14]	[1,11]	[0,14]	[0,16]	[0,8]
	statistic	58743.5	58429.5			888.5	2410.5	2848.0	
	p-value	< 0.001	< 0.001			0.002	0.012	0.022	
age.death	average	**45.5**	48.8	47.6	47.6	**65.4**	70	71.3	73.3
	stdev	6.7	7.8	9	10	6.7	7.2	6.9	8.5
	coef.var	0.1	0.2	0.2	0.2	0.1	0.1	0.1	0.1
	median	45	48	48	47	64	69.5	71	72
	range	[28,81]	[30,73]	[27,68]	[29,68]	[56,79]	[58,87]	[56,91]	[58,90]
	statistic					58.5			
	p-value					0.017			
time.death	average	12.7	13.6	10.4	**9.6**	**8.1**	10.9	11.4	9.7
	stdev	4.9	5.7	6.3	5	2.8	5	4.9	3.4
	coef.var	0.4	0.4	0.6	0.5	0.3	0.5	0.4	0.4
	median	12	13	9	9	8.5	10	11.5	9
	range	[3,35]	[1,32]	[2,29]	[2,20]	[4,13]	[4,27]	[0,21]	[4,15]
	statistic	4529.0	4915.5						
	p-value	0.002	0.0004						

On average, the genealogical trees clearly identify between 78.02% and 99.77%. These values depend on the category of each patient, which rests on the type of clinical patient, onset status, as well as living status.

For this work, a family is classified as a **new family** if the proband (i.e., the first diagnosed patient) receives the diagnosis within the time period under consideration. Conversely, a family is categorized as an **old family** if the proband was diagnosed in a preceding time period, relative to the current period for which the current measures are being compiled (i.e., Q1, Q2 or Q3). In each individual assessment, only families with known diagnosed symptomatic patients were included. Each of these categories were further differentiated between **early onset**, **late onset**, and **mixed onset** families, and these can be considered as the active families at that given period.

To evaluate genealogical variability in both new and older families, four genealogical-based metrics were identified and calculated (please refer to tables from A.1 to A.4). These were chosen to capture the variability in age-of-onset within and between families, and consist of **average largest difference of age of disease onset**, **average largest age of disease onset difference between siblings**, **largest age of onset difference between parents and children**, and **largest age of disease onset difference between probands and patients**. For old families, these metrics were assessed for both **most recent cases** (OMR), which only considers records that share the current time period, and **not most recent cases** (NMR), which also considers patients diagnosed in previous time periods.

As per analysis results and to comprehensively evaluate genealogical variability, this chapter includes an analysis of the complete age-of-onset variability within pairs of patients who share a family connection, or that have proband-patient, parent-child, or sibling relationships. These assessments were conducted across both the full variability spectrum (multiple records per family) and the maximum variability spectrum (one record per family). For these results, please refer to figures 4.6 and 4.7.

4.3.3 Descriptive and statistical analysis

The results are reported with a set of statistical metrics. These are average, standard deviation (stdev), coefficient of variation (coef.var), median and range ([minimum, maximum]). The coefficient of variation is a statistical metric that measures variability changes across data sets. When it shows changes across data or time frames it means that the relative variability of the data is not consistent.

Disparities were evaluated using a Mann-Whitney U test. This consists of a non-parametric test used to identify significant differences between groups. To account for overall changes, the evolution in diagnosed patients was analysed with Generalized Linear Models (GLM). This aided in the comparison of the clinical indicators with the Q's and gender variables (see table 4.3).

All tests were two-sided, and a p-value of $<= 0.05$ was deemed statistically significant. All statistical analyses were performed using Python [version 3.9.7].

4.3.3.1 Comparing distributions

To compare the distributions of the different groups, including those categorised as Q1, Q2, Q3, and Q4, density-based distribution plots were employed. These are valuable when the purpose is to compare distribution shapes across different data sets or variables, regardless of their sample sizes. Since the area under each histogram bar is normalised, the raw values are available in tabular format, as needed.

4.4 Results

The upcoming results encompass an analysis of the distribution of symptomatic carriers, to show heterogeneity among patients, and a study of the familial-based indicators.

4.4.1 Diagnosed patients variability

Since 1939, UCA has diagnosed and provided follow-up care for a total of 4611 subjects classified as positive for the Val30Met mutation. However, due to the variable nature of symptom onset, fluctuations in patient follow-up caused by the natural progression of the disease, or the simple unavailability of current patient data, a total of 407 patients (230 females and 177 males) could not be definitively classified between symptomatic and asymptomatic and will not be included in this study. Also, 237 patients (115 females and 122 males) lacked matched birth dates and/or onset dates. This leaves 3967 individuals (1966 females and 2001 males), among which there are 1494 symptomatic females and 1712 symptomatic males.

4.4.1.1 Symptomatic patients

The average age of diagnosis for early-onset cases is 32.3 years (SD=6.6), while for late-onset cases, it corresponds to 64.5 years (SD=2.7). Across Q1 to Q4, there was a 0.75 increase in early-onset cases, while late-onset cases exhibited an average increase of four times that (4.2 years). This metric differs from the age of symptom onset, as it is determined within the medical context during patients' clinical follow-up. As such, it aids in identifying key factors for day-to-day clinical practice.

As for gender-specific variations (refer to tables 4.1 and 4.2) results show that these have an irregular coefficient of variation values in the age-of-diagnosis, time-of-diagnosis and time-of-death, both for early and late-onset female symptomatic patients. This suggests it is important to further evaluate these subgroups.

When considering the AOO among early-onset female patients, from Q1 to Q4, there has been an average increase of 3.3 years. Among male patients, the growth was less pronounced, with an increase of 2.8 years. Notably, the greatest change was observed in late-onset cases. Among female patients, the AOO increased by 4.3 years, while among males, this value nearly doubled, reaching 8.4 years.

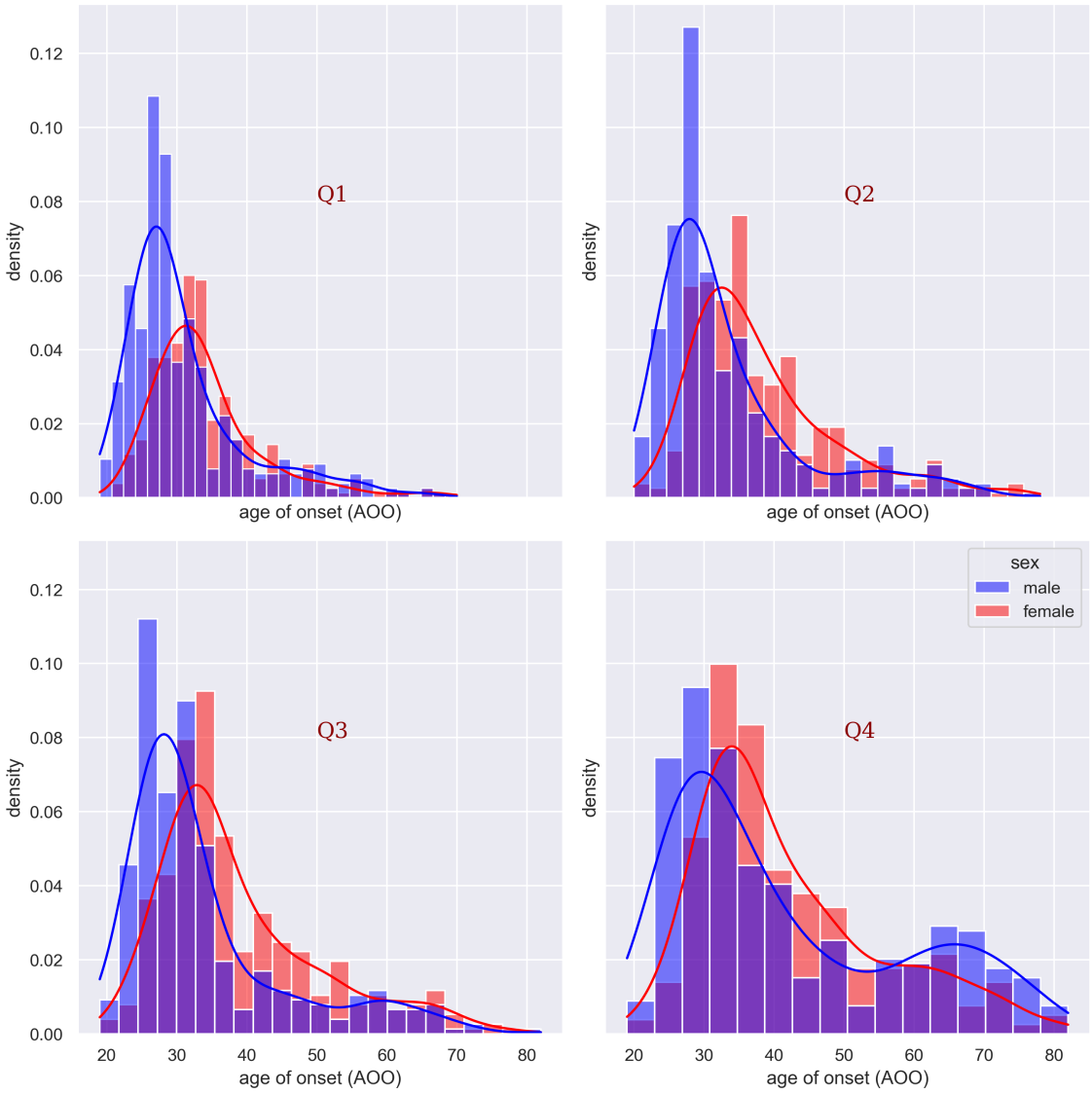


Figure 4.3: Distribution of age-of-onset of symptomatic carriers for the 25% (1985), 50% (1998), 75% (2008) and 100% (2023) year landmarks. The distributions show an inversion in the proportion of early as well as male onset patients, with an increase of female early-onset patients, as well as male late-onset patients, in Q4.

Time-to-diagnosis (difference between age-of-onset and age-of-diagnosis) exhibited a general decrease, with the most distinct difference observed in late-onset female cases (1.9 years). Interestingly, the other differences between Q1 to Q4 for early (2.5 years) and late male (2.2 years), as well as early-onset female patients (2 years) showed similar trends.

As for the age-of-death, results show an increase in average values both for early as well as late, male and female patients. A closer look at the time-of-death (difference between the age-of-onset and age-of-death) shows an increase for late-onset female patients (1.6 years). As for male patients, they show a decrease of 2.5 years for early-onset and 3.5 years for late-onset male patients. In the case of early-onset female patients results show a decrease of 3.1 years since Q1 average results.

Moreover, results show that, in every variable, Q4 is statistically significantly different from any previous time fold.

In the early-onset group of male patients, the age-of-onset in Q1, Q2, and Q3 has a smaller median than the group of patients in Q4 (28, 29, 29, 31 respectively) and the statistical test revealed that these groups are statistically different from Q4 ($p \approx < 0.001$, $p \approx 0.0002$ and $p \approx 0.0003$ respectively). Moreover, these patterns are also presented on the late-set groups where Q1 as a median of 55, Q2 of 57, Q3 of 60 and Q4 of 64. Also, Q1, Q2 and Q3 are statistically different from Q4 ($p \approx 0.001$, $p \approx 0.0002$ and $p \approx 0.0002$, respectively).

In terms of age-of-diagnosis and for the early-onset group, when compared to Q4, Q3 had a smaller median (32 for Q4 and 30 for Q3). Q3 is statistically different from Q4 ($p \approx 0.0008$). On the late-onset groups, the latter groups present lower medians than Q4 (60 for Q1, 63 for Q2, 62 for Q3 and 68 for Q4) and there are statistically differences between Q1, Q2 and Q3 to Q4 ($p \approx 0.0004$, $p \approx 0.004$ and $p \approx 0.008$ respectively). The time-of-diagnosis had a higher median in Q1 and Q2 when compared to Q4 (3, 2 and 1 respectively) and statistical tests revealed that the latter groups are statistically different from Q4 ($p \approx < 0.001$ for both) for the group of early-onset. Similar findings on the late-onset groups with a more expressive values of medians of 4.5 for Q1, 3 for Q2 and 2 for Q4. Statistical test found that the results of Q1, Q2 are statistically different from Q4 ($p \approx 0.0002$ and $p \approx 0.007$ respectively). The age-of-death, and on the late-onset group, the groups Q1, Q2 and Q3 are statistically different from Q4 ($p \approx 0.0001$, $p < 0.001$ and $p \approx 0.009$ respectively), moreover the median values are lower in the older times (66, 67 and 71 respectively).

The time-of-death, on the early-onset group, is higher on the Q1 and Q2 when compared to the Q4 group (11, 12, 9 respectively). These groups (Q1 and Q2) are statistically different from Q4 ($p \approx 0.004$ and $p \approx 0.0004$ respectively). In the late onset group, the older groups represented by Q1 and Q3 have higher values than the Q4 (9, 10, 7). These groups (Q1 and Q3) are statistically different from group Q4 ($p \approx 0.0005$ and $p \approx 0.0001$ respectively) (see table 4.1).

In the early-onset group of female patients, the age of onset in Q1, Q2, and Q3 has a smaller median than the group of patients in Q4 (32, 34, 34, 35 respectively. Also noteworthy that the statistical tests revealed that these groups are statistically different from Q4 ($p \approx < 0.001$, $p \approx 0.032$ and $p \approx 0.07$ respectively). In terms of age-of-diagnosis, when compared to Q4, Q1 and Q3 had

a smaller median (35 for Q1 and Q3 and 36 for Q4). Q1 and Q3 are statistically different from Q4 (p≈0.045 and p ≈ 0.011 respectively). The time-of-diagnosis had a higher median in Q1 and Q2 when compared to Q4 (3, 2 and 1 respectively) and statistical tests also revealed that the latter groups are statistically different from Q4 (p≈ 0.001 for both). Similar findings are also present in the time-to-death variable, where Q1 and Q2 have higher median values than Q4 (12,13 and 9 respectively) and later groups are statistically different from Q4 (p≈0.002,p≈0.0004).

In the late onset group, time of diagnosis have higher median values on the older groups Q1, Q2 and Q3 when comparing to Q4 (3, 3, 3, 2). In this case, statistical tests also revealed that the latter groups are statistically different from Q4 (p≈ 0.002, p≈ 0.012 and p≈ 0.022). The age of death was lower in the Q1 when compared to the Q4 group. These two groups are statistically different ($p \approx 0.017$), see table 4.2).

There were no statistical differences between all the other Q's and Q4 for the other variable and on the different groups.

The results presented in table 4.3, which correspond to the generalized linear model results, provide additional support for the descriptive assessments. Specifically, they indicate that in early-onset patients, women tend to experience symptoms at a later stage of life, when compared to men. More precisely, women experienced the first symptoms approximately 4 years later than men ($p < 0.01$) and when receiving the diagnosis ($p < 0.01$) in the overall time groups. Additionally, patients diagnosed in earlier time periods tend to have significantly lower ages of onset (e.g., β =-2.70 and $p < 0.001$ for Q1, β = -1.55 and $p < 0.001$ for Q2, and β = -1.71 and $p < 0.001$ for Q3) and lower ages of diagnosis (e.g. β = -1.62 and $p < 0.001$ for Q3), when compared to the more recent time period Q4.

The time to diagnosis, in the early onset group, has been decreasing consistently throughout the years. When compared with Q4, the time to diagnosis was approximately 2 years later ($p < 0.001$) in Q1 and 1.6 years ($p < 0.001$) in Q2. This pattern is also expressed in the late-onset group, where time to diagnosis was at least 2 years higher in Q1 ($p < 0.001$) and 1 year higher in Q2 and Q3 ($p \approx 0.001$, $p \approx 0.007$ respectively).

For late-onset patients, results indicate that a male patient in Q1, Q2 and Q3 tends to develop symptoms significantly earlier than males in the reference group Q4. The ones in Q1 experienced approximately 9 years earlier ($p < 0.001$), the ones in Q2 approximately 6 years ($p < 0.001$) and the ones in Q3, 5 years earlier ($p < 0.001$). The women in Q1 also developed symptoms significantly 3 years earlier than women in Q4 ($p < 0.013$).

Regarding the age of diagnosis, results show that a male patient in Q1, Q2 and Q3 tends to develop symptoms significantly earlier than males in the reference group Q4. The ones in Q1 experienced approximately 7 years earlier ($p < 0.001$), and the ones in Q2 and Q3 were approximately 5 years ($p < 0.001$ in both groups). The women in Q3 also developed symptoms significantly 5 years earlier than women in Q4 ($p < 0.044$).

All the models in the early-onset group had a R^2 ranging from 10% until 15% and in the late-onset group ranging from 6% until 13%. This shows that the age onset, the age of diagnosis and

Table 4.3: Generalized Linear Models (GLM) for age.onset, age.diag and time.diag to evaluate iterations over time groups (Q1, Q2, Q3, Q4). We considered Q4 and male as the reference classes.

model	Variables	**coef**	**std.err**	**z**	**p-value**	**[0.025**	**0.975]**	R^2
early.onset and age.onset	Intercept	32.0366	0.281	114.118	< 0.001	31.486	32.587	0.13
	Q[T.Q1]	-2.7016	0.335	-8.06	< 0.001	-3.359	-2.045	
	Q[T.Q2]	-1.5469	0.346	-4.468	< 0.001	-2.225	-0.868	
	Q[T.Q3]	-1.7108	0.351	-4.879	< 0.001	-2.398	-1.023	
	sex[T.women]	4.03	0.237	17.035	< 0.001	3.566	4.494	
early.onset and age.diag	Intercept	33.2869	0.308	108.14	< 0.001	32.684	33.89	0.10
	Q[T.Q1]	-0.4884	0.374	-1.305	0.192	-1.222	0.245	
	Q[T.Q2]	0.1387	0.379	0.366	0.714	-0.603	0.881	
	Q[T.Q3]	-1.6266	0.384	-4.241	< 0.001	-2.378	-0.875	
	sex[T.women]	4.2828	0.262	16.327	< 0.001	3.769	4.797	
early.onset and time.diag	Intercept	1.3756	0.102	13.505	< 0.001	1.176	1.575	0.15
	Q[T.Q1]	2.2652	0.136	16.606	< 0.001	1.998	2.533	
	Q[T.Q2]	1.6777	0.138	12.138	< 0.001	1.407	1.949	
	Q[T.Q3]	0.0824	0.140	0.588	0.556	-0.192	0.357	
late.onset and age.onset	Intercept	65.416	0.676	96.733	< 0.001	64.091	66.741	0.13
	Q[T.Q1]	-8.8755	1.415	-6.272	< 0.001	-11.649	-6.102	
	Q[T.Q2]	-6.2827	1.314	-4.78	< 0.001	-8.859	-3.706	
	Q[T.Q3]	-5.291	1.284	-4.121	< 0.001	-7.807	-2.775	
	sex[T.women]	-3.1239	1.049	-2.979	0.003	-5.179	-1.069	
	Q[T.Q1]:sex[T.women]	5.5433	2.221	2.496	0.013	1.191	9.896	
	Q[T.Q2]:sex[T.women]	3.3175	1.863	1.781	0.075	-0.333	6.968	
	Q[T.Q3]:sex[T.women]	2.7176	1.784	1.523	0.128	-0.780	6.215	
late.onset and age.diag	Intercept	68.104	0.688	98.968	< 0.001	66.755	69.453	0.09
	Q[T.Q1]	-7.3897	1.609	-4.594	< 0.001	-10.543	-4.237	
	Q[T.Q2]	-5.1707	1.338	-3.866	< 0.001	-7.792	-2.549	
	Q[T.Q3]	-4.9582	1.306	-3.795	< 0.001	-7.519	-2.398	
	sex[T.women]	-3.5197	1.067	-3.299	0.001	-5.611	-1.428	
	Q[T.Q1]:sex[T.women]	4.6943	2.558	1.835	0.066	-0.318	9.707	
	Q[T.Q2]:sex[T.women]	3.2595	1.895	1.720	0.085	-0.455	6.974	
	Q[T.Q3]:sex[T.women]	3.6551	1.816	2.013	0.044	0.096	7.214	
late.onset and time.diag	Intercept	2.5234	0.177	14.285	< 0.001	2.177	2.87	0.06
	Q[T.Q1]	2.0636	0.42	4.914	< 0.001	1.24	2.887	
	Q[T.Q2]	1.0333	0.316	3.267	0.001	0.413	1.653	
	Q[T.Q3]	0.807	0.301	2.678	0.007	0.216	1.398	

the time of diagnosis can only be partially explained by the Q groups and gender. The lower values of R^2 in the late-onset groups can be explained by the lower sample size.

4.4.2 Genealogical variability

Turning the attention to the families, out of the 934 profiles, 873 of them include patients with positive diagnoses of Val30M mutation at Unidade Corino de Andrade, Centro Hospitalar Universitário de Santo António.

As previously mentioned, for this study, it is deemed important to distinguish the genealogical variability between the follow-up of **new families** and **old families**. This to study differences of overall diagnosis and follow-up accordingly to patient historical knowledge. So, among these, 7.45% (65) do not have symptomatic patients, 17.41% (152) encompass both early and late onset patients, 58.65% (512) consist of early onset patients, while 16.49% exclusively comprise late onset patients (refer to tables from A.1 to A.4). In total, the family trees account for 4292 subjects who have been positively diagnosed, with 111 patients appearing in more than one tree. Of these, 3313 are symptomatic patients with valid age of onset, leaving 209 patients with missing data to determine the age of onset.

With respect to new families, it is important to note that the total number has decreased by 237 from Q1 to Q4, signifying a reduction of nearly 65%. Conversely, concerning old families and their progression from Q2 to Q4, the total number of records increased by 86. This marks a change of nearly 42%.

4.4.2.1 Symptomatic patients

Across recent decades, in percentage terms, there has been a rise in new families, primarily driven by the late onset category. These have expanded by almost 50% (refer to tables A.1 and A.2).

In old families, there has been a decrease in the number of patients. This, coupled with the consistent count of diagnosed families, results in a broader range of patients per family in the most recent cases, particularly those falling within group Q4. Notably, there has also been an increase in mixed-onset families, contributing to a heightened variability between familial patient clusters. Additionally, changes have emerged in the distribution of age of onset (AOO) between new and old families, and across each data cluster (refer to figure 4.4). Particularly noticeable is the reduced variation in old families between male and female patients, as the peaks have largely converged in Q3-Q4. In contrast, among new families, the latest landmarks have seen late male onset carriers surpassing early male onset carriers.

The distribution of age of onset for symptomatic probands diagnosed within each data group is depicted in figure 4.5. One of the key characteristics observed is the relationship between the age and type of probands diagnosed. In Q1, early onset patients significantly drive familial follow-up. However, this profile undergoes a shift, particularly around Q3 and Q4, with probands becoming predominantly characterized by middle (after 40 years old) or late onset cases for male patients (equal or greater to 50 years old).

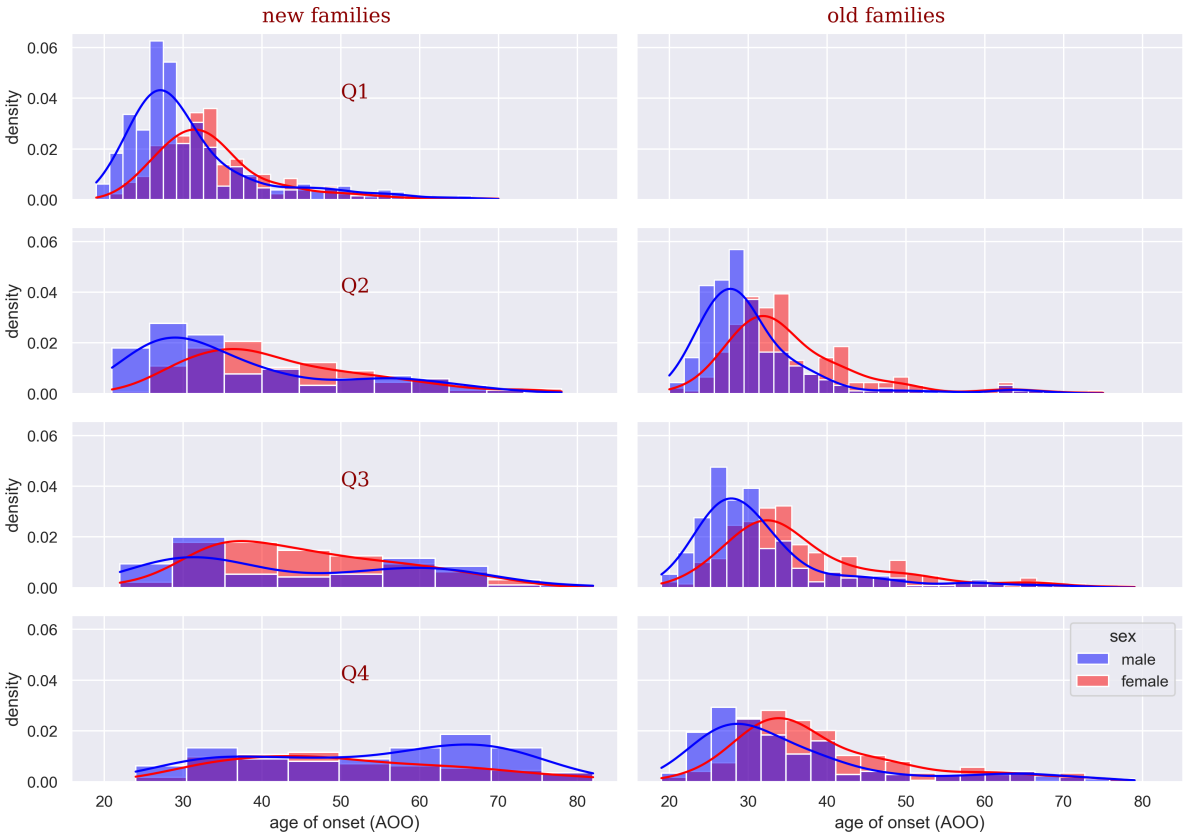


Figure 4.4: Distribution of symptomatic carriers with distinction between newly diagnosed families and older diagnosed families (with historical data). On the right we have older families (patient with family records diagnosed in previously landmarks) while on the left we have newly diagnosed families. Most relevant differences appeared over the last two landmarks with late male onset carriers surpassing early male onset carriers in the newly diagnosed families. Also the difference between male and female carriers has been decreasing over the older diagnosed families.

The relationship between parents' and their children's age of onset has been the subject of numerous studies in recent decades (see [64] and [27]). This focus arises from the presence of an anticipation factor in the AOO difference between parents and children, which varies according to the gender of the involved symptomatic patients. Given that this is a familial genealogical disease, and considering potential genealogical factors that may influence the distribution of AOO differences among siblings and between parents and children, the attention turned to both the average and distribution of the variability difference exhibited within each family (refer to figures 4.6 and 4.7, and tables from A.1 to A.4). Although distribution of the variability of genealogical indicators (across familial, proband, siblings and parents-based relations) should be focused on the maximum or average variability, in order to remove possible bias, in cases when a subset of families has growing number of symptomatic patients, the results of the full variability were also assessed, to present a full picture of onset changes.

So, the genealogical family indicator shows a relatively low average AOO difference in new families, although these values increase in mix-onset families as well as late-onset families (refer

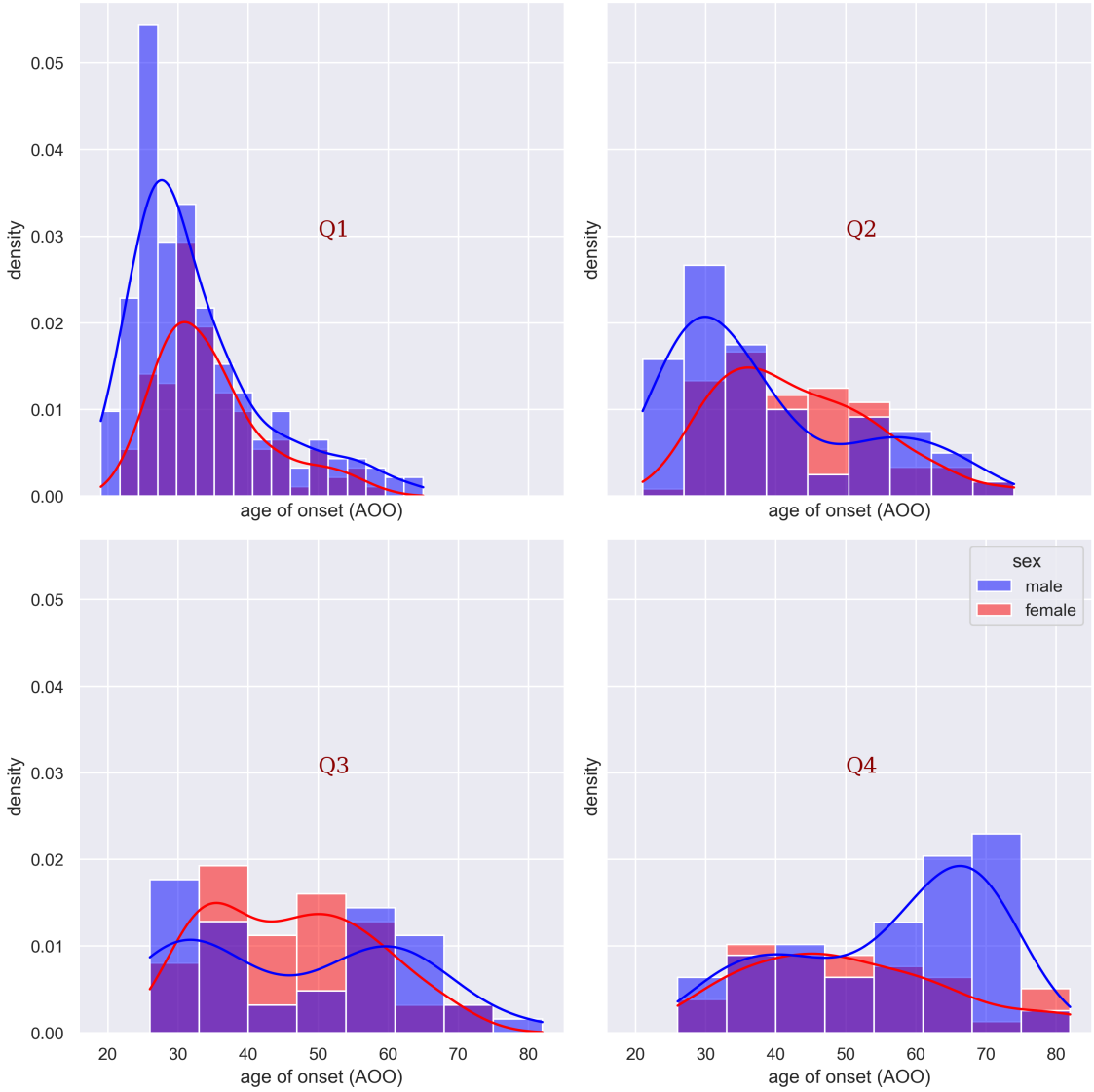


Figure 4.5: Distribution of the age of onset for symptomatic probands for the 25% (1985), 50% (1998), 75% (2008) and 100% (2023) year landmarks. It is clear the change in the type of probands diagnosed. If in Q1 we have clearly early onset patients this shifts especially around Q3 and Q4 with probands being clearly defined by middle (after 40 years old) or late onset cases for the male patients.

to tables A.1 and A.2). In old families, there is a concentration on early onset families, with Only Most Recent (OMR) values showing a smaller average value when compared to Not Most Recent (NMR) in most cases, due to a decrease in the number of records present in each group. The change between NMR and OMR in unique family values shows that there is a tendency for patients of the same family to have dispersal diagnosis years, that is, results show that usually, patients aren't diagnosed in familial concentrations.

As for proband-patients genealogical indicator, in general terms, there is a decrease in newly diagnosed families and an increase in old diagnosed families. Also, in old diagnosed families there is a tendency for only having values in NMR cases. This shows that in these families there

is a wide discrepancy in the diagnosis of probands and other family relatives which suggests that families tend to have a long clinical follow-up.

Regarding parent-child relations, most of the cases are present in old families, more exactly in not most recent cases (NMR) (see tables A.3 and A.4). This suggests that there is a considerable difference in the year of diagnosis between parents and their children over the different clusters (i.e., Q1, Q2, Q3 and Q4). In new families, the average difference in disease onset between parents and children has been increasing, as well as the raw number of families over the different groups of data has been decreasing. This can be due to a significant difference in the time of diagnosis between parents and children.

As for sibling relationships between symptomatic patients, the average difference of AOO has been decreasing in early onset families over the considered groups, while in late as well as mix-onset families it has been increasing. It's worth noting that 68% of the cases exhibit a difference in the age of onset below 10 years. This suggests a high aggregation of values in terms of sibling relations.

As for the statistical tests and when considering the extended family evolution for newly diagnosed families, the value of the largest AOO difference of Q1 is statistically different from the values in Q4 ($p \approx 0.038$). It is important to notice that the median value of the largest AOO difference on Q1 is 8 while the same metric on Q4 is 4. No other statistical differences were found (table A.1).

In the case of the extended family evolution for old diagnosed families, female patients in Q2 had the largest AOO difference proband-patient significantly different from female patients in Q4 ($p < 0.001$). The median value for this characteristic in Q2 is 9, and 12 in Q4. Also, female patients in Q2 are statistically different from females in Q4 in terms of the largest AOO difference siblings ($p \approx 0.008$). In this case, the median in Q2 is 7 and reduces to 5 in Q4 (table A.3).

In this study generalized linear models are used to analyze the relationship between the age-of-onset of a patient and that of their previously symptomatic familial relations. This analysis provides technical justification for the empirical results presented in figures 4.6 and 4.7 and identifies trends among the different families being the target of clinical follow-up. Since each patient can have a highly variable number of relations under the patients, probands, parents and siblings categories, and to ascertain possible relation bias, the patients were sampled into groups of independent records. For each relation, there were created as many groups as the highest number of relations of a patient. In the end, the groups of models assist on reaching average insights into the different variables. Of note, that it was considered only viable if there were as many as 10 independent patient records for each coefficient. In the end there were 14 models for patients, 1 model for probands, 1 model for parent relations, and 3 models for sibling relations.

So, the **family relation** models (refer to table 4.4) indicate that women tend to develop symptoms later than other male patients. Specifically, they show that women, in general, tend to experience symptoms, on average, 5.35 years after male patients and, on average, 1.48 years before their male relatives. Additionally, results reveal that their age-of-disease onset occurs, on average, 0.21 years later than that of their relatives. As for the **proband relation**, it indicates that women

develop symptoms, on average, 4.84 years after other male patients, while women probands tend to experience symptoms, on average, 1.50 years before male probands. Additionally, it shows that patients, on average, tend to develop symptoms 0.49 years after their proband relations. When considering **parent relation**, results indicate that women, on average, tend to experience symptoms 4.45 years after males, and women parents tend to develop symptoms 2.72 years before male parents. In this case, it also shows that a patient tends to experience symptoms, on average, 0.39 years after their parents. The model that characterizes patients that have **siblings** as their relations shows that women tend to develop symptoms, on average, 3.48 years after males, and 2.26 years before their male siblings. These final results also indicate that a patient, on average, typically shows symptoms 0.75 years after their siblings' relations.

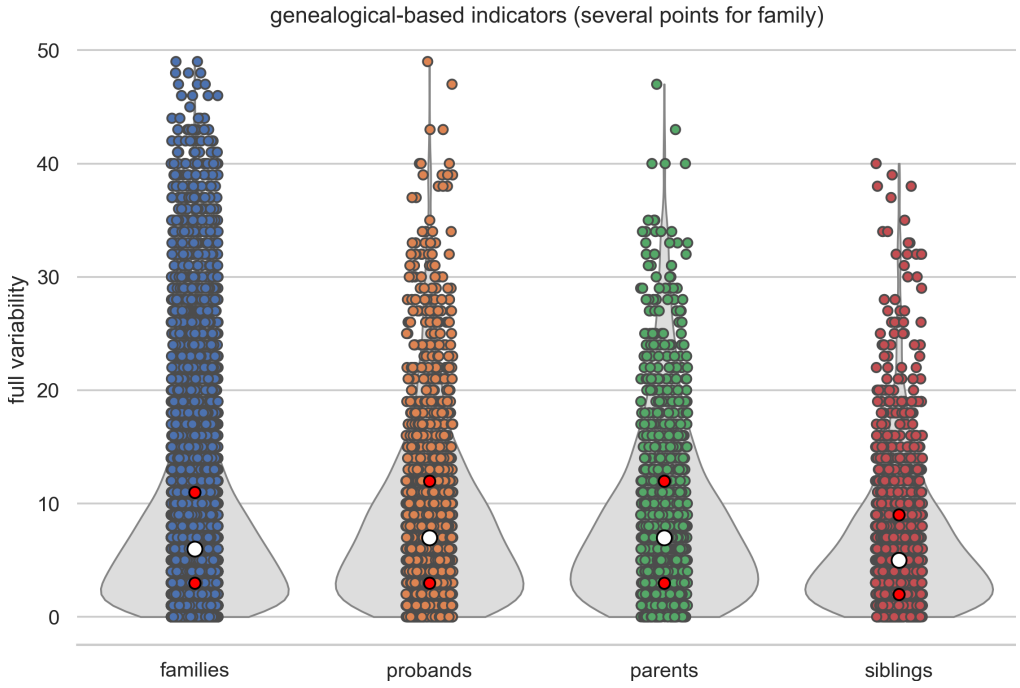


Figure 4.6: Violin plots which depict the distribution of the overall age of onset variability across familial, proband, siblings, and parents-based relations (several record per family). The central white dot represents the 50% quantile, while the upper and lower red dots signify the 75% and 25% quantiles, respectively. These plots, along with the quantile positions, highlight that siblings exhibit the least overall variability within the studied ATTRVal30M families, even though there are less instances then by taking the family based comparison.

All results have an average R^2 ranging from 13% (family relation) to 66% (sibling relation), indicating an increasing explanatory power of the selected variables for the variability in the age-of-onset, when considering a rank in the familial relations. This rank starts in the siblings relationship (1^{st}), proceeds to the parents (2^{nd}), crosses over to the probands (3^{rd}), and ends with the family relationship (4^{th}).

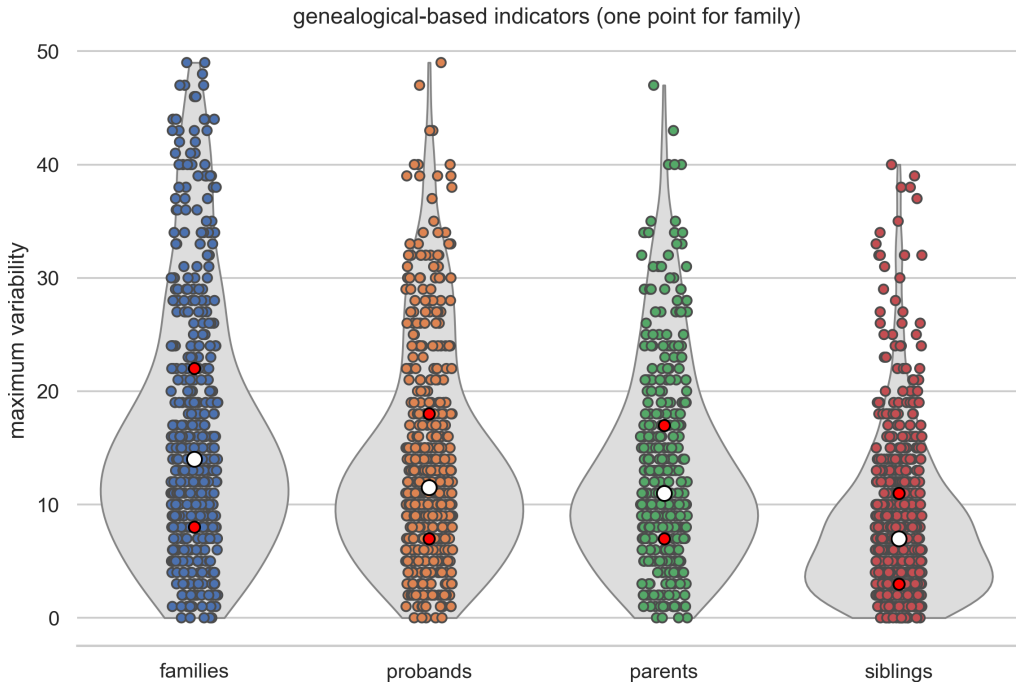


Figure 4.7: Violin plots depict the distribution of maximum age of onset variability across familial, proband, siblings, and parents-based relations (one record per family). The central white dot represents the 50% quantile, while the upper and lower red dots signify the 75% and 25% quantiles, respectively. These plots, along with the quantile positions, highlight that siblings exhibit the least overall variability within the studied ATTRVal30M families.

4.5 Discussion

Comparing the groups (Q1, Q2, Q3, and Q4) reveals several differences.

From the bagging procedure utilizing the diagnosis years of symptomatic patients, the most significant years were identified as 1985, 1998, 2008, and 2023. As the first patient was diagnosed in 1939, the initial three groups were diagnosed within a span of 46, 13, and 10 years, respectively. The final patient group, corresponding to the most recent 25% of cases, received their diagnoses over the past 15 years.

Overall, in clinical terms, Q4 represents the most pertinent group as it encompasses the most current and active patients. The varying time spans leading to the establishment of each group (46, 13, 10, and 15 years) indicate that the peak of patient diagnoses was in the 1990s, and significant changes transpired over recent decades. One such change is the reduction in newly diagnosed patients, with the final 25% of patients requiring an additional 33% of time, when compared to the previous set. Furthermore, a decrease in diagnosis time was observed in the last group compared to the previous ones. In clinical terms, it's important to reference those significant changes that occurred in the diagnosis procedure over time. Across the long period of patient observations, diagnostic methods were refined, making it easier to confirm the diagnosis. Currently, the criteria consider a combination of key symptoms, neurophysiological and cardiac exams, alongside a positive demonstration of amyloid deposition. This, combined with a significant number of cases

of a specific variant (ATTRVal30Met), contributes to a smaller bias in confirmed diagnosis results.

Interestingly, the years of 1985 and 1998 hold significance in the disease's history. In 1986, the introduction of the genetic test marked a pivotal moment, enabling the diagnosis of carriers even before symptoms emerged. Additionally, in 1992—six years prior to the start of the third group—a significant milestone was reached in Portugal, with the first liver transplant performed in symptomatic patients. This marked the inception of available treatments for symptomatic patients.

It is noteworthy that when analyzing each individual group (refer to figure 4.3), one reaches a series of additional conclusions.

When considering patient variability and focusing in Q4, results suggest a balance between men and women, both being diagnosed and having symptoms onset in later stages in life. Also, the profile of proband patients has been evolving, due partly to interaction of neurology with other clinical specialties such as nephrology ([66], [73]) and cardiology [115].

Considering familial variability across the four patient groups, several changes stand out. In Q1, the emergence of early onset families is evident. In Q2, a balance is observed between newly diagnosed families and older families, the latter experiencing an increase in diagnoses, following the introduction of the genetic test. Q3 witnesses a surge in diagnoses among old family cases, possibly attributed to the introduction of the initial treatment. In Q4, a rise is noted in diagnoses of mixed and late-onset families within active families, potentially linked to the increased average life expectancy in recent decades. Lastly, in Q4, a noticeable growth in the diagnosis of new families is seen, spurred by a rise in late-onset male diagnoses.

Furthermore, it's worth highlighting that concerning older families, the reduction in patients within Q4 could be attributed to multiple factors. These include a decline in the birth rate in Portugal, and a broader distribution of clinical follow-up resulting partly from a larger geographical dispersion of families due to mobility-related immigration and emigration patterns [69].

Collectively, these recent findings demonstrate the diversity in age-of-onset differences among symptomatic patients, encompassing hereditary considerations (parents and siblings' relations) as well as familial aggregation (relations involving all patients and the proband). Late and mixed-onset families exhibit a heightened average maximum variability, driven by the broader range of values observed.

4.5.1 Conclusions

This consists in the first study on ATTRv Amyloidosis families that analyses familial data with this degree of depth. Overall, this study reveals a consistent increase in elderly cases over the last few decades, which suggests the ever-growing need of further studies in new approaches to diagnose and treat the ever-elderly symptomatic patients. The analysis of familial data shows that symptomatic patients exhibit less variability when compared to siblings than other family members, particularly in terms of the transmitting ancestors' age of onset.

The genealogy-based indicators studied generally have high standard deviation (SD) values, highlighting the heterogeneity in the distribution of clinical indicators. It's also noteworthy that the results show that when feasible, the presence of symptomatic siblings can serve as a valuable

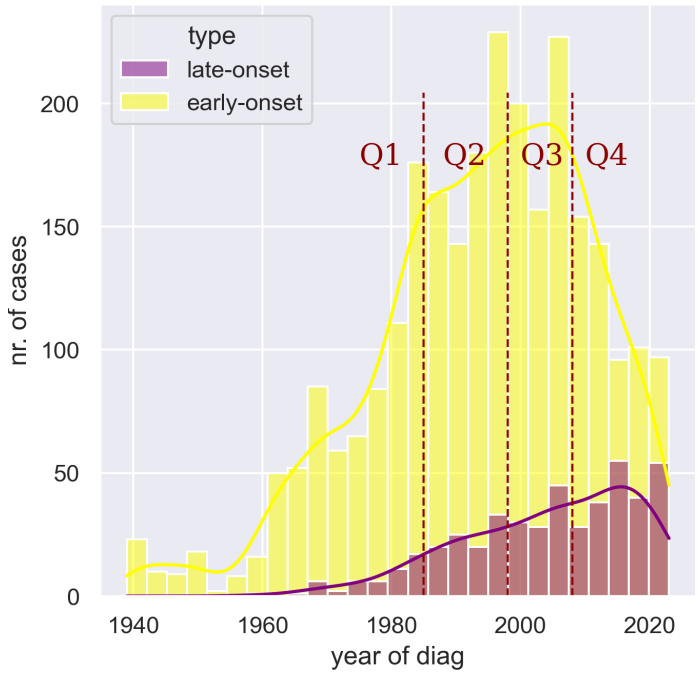


Figure 4.8: Distribution of early and late onset diagnosis over time. Yellow shows early-onset diagnosis while pink shows late-onset.

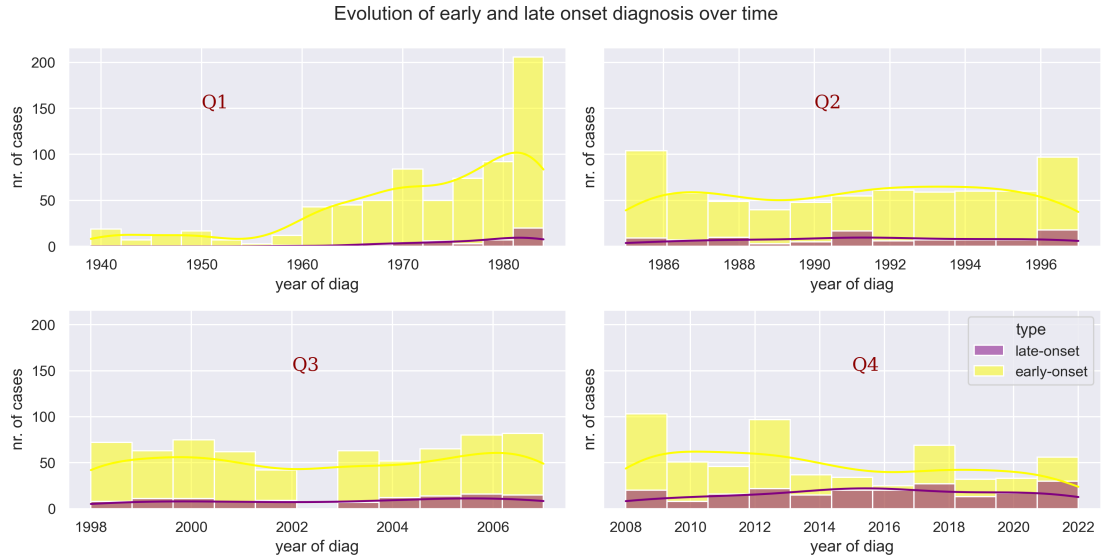


Figure 4.9: Distribution of early and late onset diagnosis over the different Q landmarks. Yellow shows early-onset while pink shows late-onset diagnosis.

indicator to evaluate the likelihood of a patient developing symptoms in the near future. This is shown by the distribution of maximum age of onset variability across familial, proband, siblings, and parents-based relations, and the smaller value of the 50% quantile of the siblings relation (refer to figure 4.7).

Family-wise, the number of new, late or mixed-onset families has shown a tendency to grow. In these cases, due to the variable nature of symptoms, trends show that the profile of future symptomatic patients will most likely continue to change.

Regarding the patient diagnosis, figures 4.8 and 4.9 assist in transmitting the distribution of the early and late onset diagnosis over time. In this case it shows that the diagnosis of late onset patients suffered from a delay in Q1, during Q2-Q3 it had a less steep growth when compared to early-onset (yellow bars) and that in the last landmark it has been decreasing. Also, and as stated previously, the four groups Q1, Q2, Q3, Q4 are the result of sorting on year of diagnosis. Within them, there was a distinction between early-onset and late-onset. However, the likelihood of going to the doctor increased over the years, first because of improvements in the medical system, and then also because people are more likely to go to the doctor when they get older. This suggests a variation in the likelihood of observing early-onset increases with the calendar year and, that the age at which a person was diagnosed as positive is likely to decrease with the calendar year.

Lastly, it is important to note that the current database comprises 934 families, totaling 33608 individuals with diverse familial relationships. These relationships range from close ones, such as parents, grandparents, siblings, and children, as well as to more extended relations like uncles, cousins, as well as 2nd, 3rd or 4th-degree relatives. From a clinical perspective, even if an individual within a family is not a carrier, they may still be affected by the disease due to their role of caretakers. This is a common situation, particularly among children or spouses of individuals with highly debilitating diseases, and underscores the importance of involving psychiatrists and psychologists in family support and care.

4.5.2 Study limitations

The present study has some limitations. Firstly, it centers on families monitored at a single facility, although this institution constitutes the largest hub for the disease globally. Secondly, the results exhibit a notable standard deviation, underscoring the significant variability that patients diagnosed with the condition tend to inherent. Lastly, the results can be affected by various factors, including the dimensions of the subsets and the number of patients selected by each indicator. Whenever it was deemed possible, the decisions aimed to minimize these effects.

Table 4.4: Genealogical age-of-onset Generalized Linear Models (GLM) for family, proband, parent and sibling relationships. We considered male as the reference class.

model	**variables**	**avg coef**	**stdev.coef**	**R^2**	**stdev.R^2**	**n. models**
family relation	Intercept	25.26	3.39	0.18	0.09	14
	sex[T.women]	6.09	0.84			
	sex_other[T.women]	-1.75	3.18			
	ageonset_other	0.22	0.11			
proband relation	Intercept	15.17	-	0.31	-	1
	sex[T.women]	4.84				
	sex_proband[T.women]	-1.5				
	ageonset_proband	0.49				
parent relation	Intercept	17.15	-	0.51	-	1
	sex[T.women]	4.45				
	sex_parent[T.women]	-2.72				
	ageonset_parent	0.39				
sibling relation	Intercept	9.13	5.09	0.66	0.12	3
	sex[T.women]	3.48	0.95			
	sex_sib[T.women]	-2.26	0.55			
	ageonset_sib	0.75	0.21			

Chapter 5

Preliminary attempt to answer the medical problem

Complex medical issues often call for a dynamic research strategy, where multidisciplinary teams continuously reassess and re-adapt. This leads to the need to, from a computational side, entail an iterative and multifaceted development approach. This was the case for this work, which was marked by a series of pre processing data sets assessments, each one with its own data quality evaluations, cleansing procedure phases, relatively complex data extraction procedures to entail entity deduplication methods, among others. This process was delineated to allow for the continual refinement, optimization, and tailoring of responses to meet specific ever changing needs.

5.1 Introduction

This chapter describes the first attempt to answer the medical problem of predicting the age-of-onset of a subject diagnosed as a carrier of ATTRv Amyloidosis and is based on the publication presented in [84]. It focuses on data cleaning, presents a few of the most relevant data preparation issues, highlights key considerations regarding the baseline model based on clinical feedback, and continues to identify a set of key considerations for the modeling and validation methodology. Then it presents the experiments conducted and the results achieved, before concluding with main discussion considerations.

5.2 Data cleaning

In this work, from early data processing, it was important to carefully undertake a set of measures to evaluate and correct for data quality problems.

A few of the most relevant data quality issues found were:

- changes in the naming between clinical data sets and genealogical information, where the incorporation or not of husbands' surname changed over time;
- changes in the designation of a patient's process number, which made it relevant to create deduplication procedures capable to insure a correct identification of each patient in the tree based on the clinical data;
- incorrect identification of a patient's status, which made it important to correct the information in the genealogical trees with the information found in the clinical processes;
- existence of information in the genealogical trees with possible double meaning, which made it important to enhance the deduplication procedures so as to validate the clinical processes. In this case the problem was related with 4 digit numbers which could identify important years or process numbers;
- patients referenced with the wrong process numbers, which lead to the development of specific validation procedures based on name, gender and close-relatives data.

To deal with previously set of problems it was important to define a set of rules to identify and correct the errors. Over time, the need to process new as well as older genealogical trees lead to the need to change this process, and to continuously update it. In the end and to ensure quality results, it was relevant to manually correct near 10% of the patient records.

5.3 Key data preparation concerns

To answer the medical problem for predicting the age-of-onset of ATTRv patients, and after carefully cleaning a set of data sets released by medical professionals, assembling entity deduplication procedures, and ensuring that the data extraction procedures focused on the most relevant (heterogeneous) data records, it was possible to create an entirely new, cleaned, predictive data set, populated by clinical and genealogical-based features.

The initial approach to generate the feature set focused on three sets of features namely patient, close relatives and extended family data. The primary process only entailed patients with a positive diagnosis, thus discarding currently asymptomatic patients.

Using the newly created features, the endeavors focused on creating an initial basic predictive pipeline based on a group of state of the art machine learning regression methods. The initially chosen ones were Decision Tree (DT), Elastic Net (EN), Lasso (LA), Linear Regression (LR), Random Forest Regressor (RF), Ridge Regression (RI) and Support Vector Machine Regressor (SV).

Later, and to compare the predictive results with current medical practice, the focus was on defining a rule-based baseline model capable to express the clinical practice known at the time.

After assembling the data sets, delineating the evaluation schema, and running the prediction models, an initial set of results was obtained. These results indicated a clear improvement when

using a machine learning-based approached, at least when comparing the initial set of results with the clinical baseline.

A general overview of the main steps regarding data preparation is presented next.

5.3.1 Data set

The work uses data collected from patients hierarchical family structures, also known as *genealogical trees* or *pedigree charts*, and patients clinical information available in *electronic health records* (EHR).

At the clinical facility, patients are stratified into one of three categories, depending on their diagnosis status. These correspond to:

- symptomatic patients, which defines patients that already received a positive diagnosis of the disease and which had a medical professional assure that the symptoms onset had already commenced;
- pre symptomatic patients, which consist on the cohort of patients that although showing a few irregular symptoms have the diagnosis not determined with 100%;
- finally, asymptomatic patients, that define the group of patients that have not expressed signs of the disease yet.

For simplification purposes the modeling will rely on patients that already expressed signs of the disease and that are, presently classified as symptomatic, despite the data will reference moments in time when they were classified as asymptomatic. This way the results are of a **supervised** nature.

It is noteworthy to refer that if resorting to survival analysis one could rely both on patients with and without diagnosis, but since the lost for follow-up set of patients correspond to a very small amount of subjects, the regression-based setting is considered as the most suitable for this clinical problem.

Next, key concepts associated with the feature modeling approach are presented.

5.3.2 Feature modeling

Overall, the extraction process that was outlined generated three distinct sets of features:

- patient;
- first level;
- expanded family.

For the set of **patient features**, the pipelines processed information regarding a patient's birth, death, and onset events. For the **first level** family structure, the focus was on information regarding a patient's relatives, namely siblings, children, parents, and uncles.

The genealogical features, classified as **expanded family** variables, were assembled to account for potential dimensional variations between families. For example, some genealogical trees, have a higher number of generations, patients and individuals, than others.

The extraction process also created features to distinguish between patients that were clinically classified as early or late onset. By early onset, one is interested in family relatives that started symptoms before 50 years old, while late onset classifies patients that started symptoms onset at or after 50 years old. These correspond to actively relevant clinical classes. The gender of patients was used as a further filter rule.

The full set of 58 features are listed in table 5.1. They are all numeric, with the exception of *patients' gender* and *born before mother and father had symptoms*, which correspond to **dichotomous** variables.

Table 5.1: Manually extracted features

Type	**Description**	**N(Features)**
Patient	Patients' Gender	1
	Father and Mother Age of Onset	1
	Born before Father or Mother had Symptoms	1
	Age of Onset (Target Variable)	1
First Level	Number of Sons (by gender)	2
	Number of Siblings (by gender)	2
	Number of Uncles (by gender)	2
	Number of Sons with the disease (by gender)	2
	Number of Siblings with the disease (by gender)	2
	Number of Uncles with the disease (by gender)	2
	Avg, Max and Min Age of Onset of Sons (by gender)	6
	Avg, Max and Min Age of Onset of Siblings (by gender)	6
	Avg, Max and Min Age of Onset of Uncles (by gender)	6
Extended	Avg, Max and Min Age of Onset of Patients followed at the Clinical Unit (by gender)	6
	Avg, Max and Min Age of Onset of Patients in the Family Tree (by gender)	6
	Avg, Max and Min Age of Onset of Early Onset Patients (Patients with Age of Onset < 50 years)	6
	Avg, Max and Min Age of Onset Late Onset Patients (Patients with Age of Onset ≥ 50 years)	6

Once the feature data set was fully compiled, cleaned and separated, the next step went to leverage the model construction phase. The approach involved defining a set of virtual scenarios that correspond to various patient ages. These virtual scenarios allow for the combination of patients from different decades, by aligning them according to their age. This method enables the simulation of age-of-onset prediction for patients throughout their lifespan. It also allows for a focus on the most relevant predictive characteristics at each individual age. Prediction at each age is seen as a specific and individual regression prediction task. This means that for each patient, a set of individual prediction instances are created, one for each of the considered ages. Then,

depending on the age, a different model is created.

5.3.2.1 How to deal with time-related data, data quality issues and main considerations for data leakage prevention

This preliminary study lead to the creation of individual groups of data, one for each age value. In this case, an important modeling premise, is that, any patient's data set only compiles information known at that time. For instance, if the process is gathering information about a patient in 1950, it cannot incorporate information related to the patient's uncles if their symptoms only manifested in 1960. If not properly managed, this could lead to a data leakage issue.

To achieve this, it was necessary to manage the information considered at each point in time, as if various snapshots of the genealogical trees were accessible over the years. This allowed for the precise combination of one patient's data with newly available information from other patients or individuals. This process involved a data engineering task that needed to be refactored over time, to accommodate the integration of various types of clinical information.

Another important step was to evaluate the data sets in terms of missing data, so it was possible to compare the machine learning models with the baseline results. To this end it was important to take into account the status and availability of each column of data, both in the clinical and the family data records.

All the different conditions related with data quality and completeness made it pertinent to represent the data into two disjoint sets, namely **data set A**, constituted by patients having a known ancestor with the disease, and both were or are being followed at the clinical unit; and **data set B**, for patients that don't have a known ancestor with the disease, or one of them hasn't been followed at the clinical unit.

The pre-processing stage assisted in having data sets with information regarding each patient which evolved over time. In this case, in this early attempt, data set A had 1103 patients while data set B had 1709.

In the final experiments both data sets were also combined thus giving birth to a third set data set entitled **full data set**.

5.4 Baseline, methodology, modeling and validation

This work resorts to a supervised learning strategy for which models use information from previously positive, diagnosed, and currently symptomatic patients. A few clarifications regarding the modeling strategy are provided next, as well as a discussion about the full prediction results. For its importance, this section starts with an explanation over the baseline model.

5.4.1 Baseline model

A general clinical rule-based model was extrapolated from the work of [64], so as to compare the machine learning approach with current medical practice.

In [64], authors considered valuable to analyze the distribution of age-of-onset of Val30Met patients, while sampling the patients regarding the gender of both the transmitting parent and affected patient children. From a statistical point of view this is considered interesting if one admits there is a linear trend between a patient and his ancestors' age of onset, as well as an anticipation phenomenon that distinguishes between the gender of a patient and that of his ancestor.

In the case of the baseline model it is relevant to consider four different groups, namely:

- father and son;
- father and daughter;
- mother and son;
- and, mother and daughter.

With these four groups in mind and by measuring the average age-of-onset in each of them, a predicted value for a new asymptomatic patient can be seen as $target = parent(ageonset) - \overline{X}$, with $\overline{X}$ representing the average age of onset of the patients within that subgroup.

Thus, the overall baseline model is instantiated as follows:

- if (father and son): target = parent(ageonset) - 6.06 ;
- if (father and daughter) : target = parent(ageonset) - 1.23;
- if (mother and son): target = parent(ageonset) - 10.43;
- if (mother and daughter): target = parent(ageonset) - 7.43.

Since the standard deviation (SD), for each subgroups is above 9 years [64], the final error amplitude within each group can already be considered high. Even with this handicap, and as this model represents the medical practice as of 2018, this is considered as a viable reference.

It's important to note that the results from the baseline model are only comparable for the experiments conducted with data set A. This is because the model requires information about the age-of-onset for a patient's ancestor.

5.4.2 Modeling schema

After the full set of features was processed and consolidated, the next step was to create a method that, for a given numerical age, generates an instance of an algorithm after going through training and tuning phases.

For selection purposes, and due to the low number of available cases, a k-fold cross validation procedure was transformed to a nested cross validation (see sub-section 3.4.2.3). The technical issues were reviewed previously and an algorithmic clarification between k-fold cross validation and nested cross validation can be found in [88].

In this work, the experimental setup, for both of the two k-fold cross validation procedures (single + nested) takes into account a 10 fold partitioning of the available data.

5.4.3 Validation schema

To validate the work it was considered relevant to compare, for each set of algorithms, for each of the experiments (e.g., data set A, data set B, full data set), the MAE score obtained in the outer folds. The comparison was performed using a **Friedman rank test**. This corresponds to a non parametric rank hypothesis test that ranks the algorithms, for each individual data set experiment, separately. The usage of Friedman rank test has been extensively explored and recommended in [31].

5.4.3.1 Hypothesis

The null hypothesis (H_0) was considered as stating that the average results for each algorithm can be considered equivalent, while in the alternative hypothesis (H_1) it states that they are not. If only two algorithms are taken into account, the hypothesis can be represented as:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

As the Friedman rank test compares the average series of results of the different algorithms, and the evaluation resorted to the outer cross validation, the comparison is defined as having 10 different data sets. In this case H_0 states that the algorithms are equivalent, meaning that their ranking positions are equivalent. As such, the statistic that needs to be measured consists of:

$$F_F = \frac{(N-1)\chi_F^2}{N(A-1)-\chi_F^2}$$

$$\chi_F^2 = \frac{12N}{A(A+1)}\left[\sum_j R_j^2 - \frac{A(A+1)^2}{4}\right]$$

In the Friedman rank test, when the H_0 hypothesis is rejected, it indicates that there are differences in the results. To get the indication of which algorithms are different one can resort to a nemenyi post-hoc test [31].

5.5 Experiments and results

This section report the main results for the preliminary set of experiments.

The comparison has a three layer approach.

The first layer discusses the MAE results for all of the three sets of experiments, that is, the results for data set A, data set B and full data set are explored (see table 5.2).

The second layer consists in plotting the distribution of the scores of the outer k-fold results, for each of the models tested, in each of the ages considered.

Finally, in the last layer, the results of a Friedman rank test are presented and, when there is a difference, the clusters of results are further analyzed with a nemenyi post-hoc critical difference diagram.

Of note that as referenced before, the comparison of the results with the baseline model is only viable for the data set A. Also, to deal with missing data and possible differences among the amplitude of the features, during the pre processing step a missing-indicator was included in the feature set. This method indicates that, for each column with missing values, a binary auxiliary column with 1 when there is information and 0 when there is no available information, should be added. In the case there is a missing data this should be encoded with a -100 value [49]. Also it is note worthy that for the SV it is important to standardize its input features, and that the regularization algorithms need the input features normalized. All the experiments were performed using *scikit-learn* framework [1].

Globally, from table 5.2 and in the case of **data set A**, the results expressed by Elastic Net (EN), Lasso (LA), Ridge (RI), Linear Regression (LR), Decision Tree (DT), Random Forest (RF) and Support Vector Regressor (SV) are superior to the results of the Baseline (BL) approach. In this case it is observable that the results indicate that MAE and SD results are lower, for each of the outer folds results. If one takes into account the 20 years as opposed to the 30 years it is visible that there is a tendency for the mean absolute error to decrease. Also, the global standard deviation results show that the MAE results don't vary significantly, what suggests that the models tend to generalize accordingly.

For **data set B**, experiments indicate that the MAE results are almost four times as large than those achieved by data set A, and that these values tend to get smaller as the ages grow. This does not happen in the case of data set A, even though from a conceptual perspective it is an expected result.

Regarding the **full data set** experiments, the results have improved with respect to data set B even though they continue to be worse then the values achieved by data set A. In this case, the results have a wider dispersion, suggesting greater fluctuations in the prediction error.

The poor results of the data set B and full data set experiments, when compared to data set A, are related with the lack of available information. Actually this is a common issue in most models since if in a set of experiments the pre processing changes slightly even thought the model construction phase does not change, it can appear as an improvement. This is not true since even if the overall set of results appear to improve they aren't related with an actual factual improvement in the machine learning model construction phase.

The distributions of the MAE for each of the outer folds (see figures 5.1, 5.2 and 5.3) indicate that:

- in case of data set A, the amplitude of the errors is greater for the baseline than for the machine-learning based approach;
- and the previous remarks are corroborated.

Table 5.2: Mean absolute error (MAE) results for each algorithm, regarding ages = [20 and 30 years]

Model	Data set	20 Years	30 Years
Baseline (BL)	A	5.19 ± 0.40	5.79 ± 0.81
Elastic Net (EN)		2.32 ± 0.32	2.05 ± 0.31
Lasso (LA)		2.28 ± 0.31	2.00 ± 0.26
Ridge (RI)		2.19 ± 0.26	1.90 ± 0.27
Linear Regression (LR)		2.24 ± 0.25	1.98 ± 0.25
Decision Tree (DT)		2.17 ± 0.40	1.94 ± 0.37
Random Forest (RF)		2.62 ± 0.32	2.57 ± 0.37
Support Vector Machine (SV)		3.55 ± 0.18	3.38 ± 0.40
Elastic Net (EN)	B	8.83 ± 0.40	8.11 ± 0.61
Lasso (LA)		8.88 ± 0.40	8.27 ± 0.59
Ridge (RI)		8.84 ± 0.39	8.15 ± 0.59
Linear Regression (LR)		8.90 ± 0.41	8.22 ± 0.51
Decision Tree (DT)		8.78 ± 0.47	7.97 ± 0.63
Random Forest (RF)		8.56 ± 0.41	7.78 ± 0.68
Support Vector Machine (SV)		8.71 ± 0.51	7.94 ± 0.83
Elastic Net (EN)	Full	6.87 ± 0.28	6.75 ± 0.62
Lasso (LA)		6.90 ± 0.27	6.86 ± 0.60
Ridge (RI)		6.79 ± 0.25	6.71 ± 0.59
Linear Regression (LR)		6.82 ± 0.27	6.71 ± 0.59
Decision Tree (DT)		6.70 ± 0.32	6.67 ± 0.60
Random Forest (RF)		6.66 ± 0.30	6.43 ± 0.55
Support Vector Machine (SV)		7.01 ± 0.37	6.70 ± 0.48

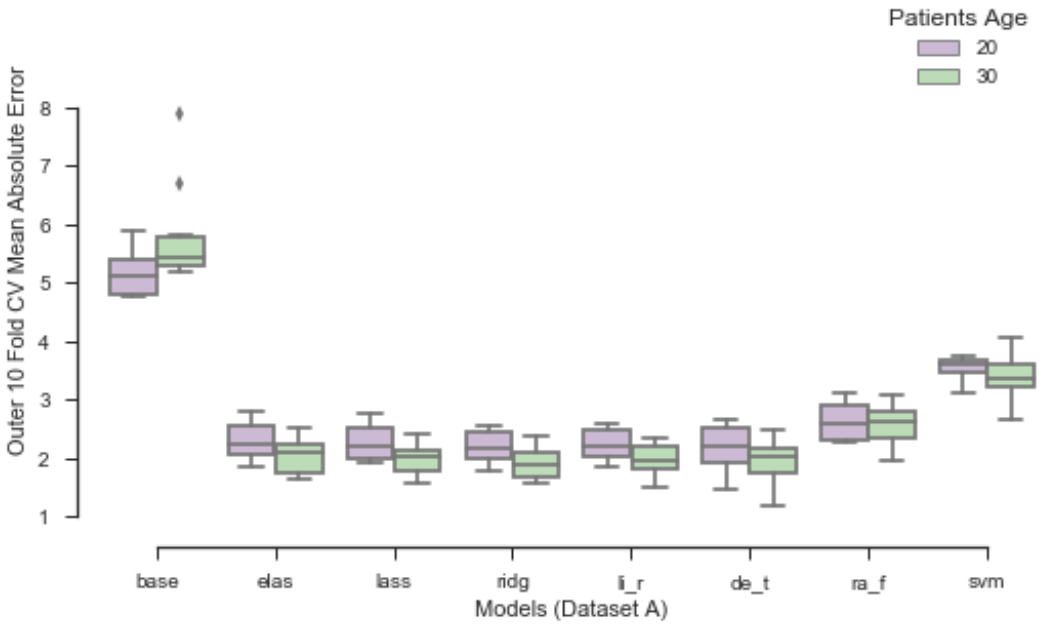


Figure 5.1: Box-plot for all of the experiments performed for data set A

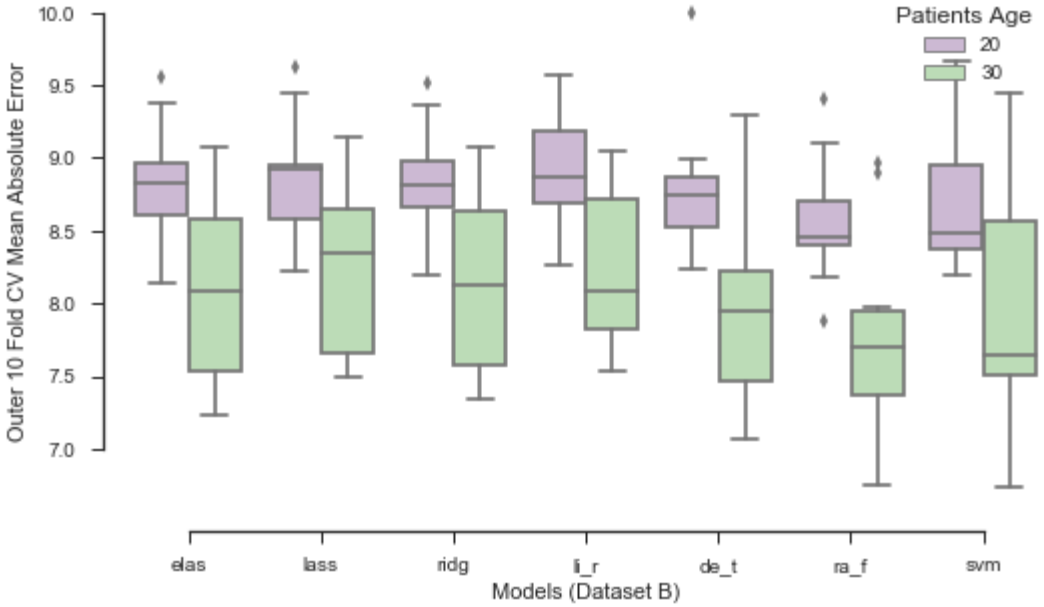


Figure 5.2: Box-plot for all of the experiments performed for data set B

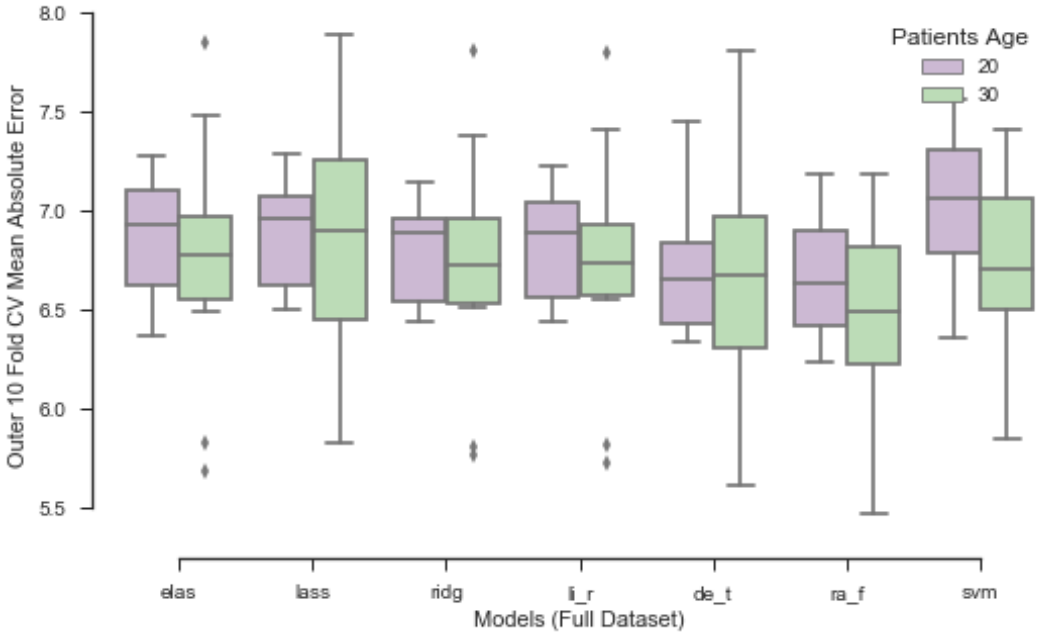


Figure 5.3: Box-plot for all of the experiments performed for full data set

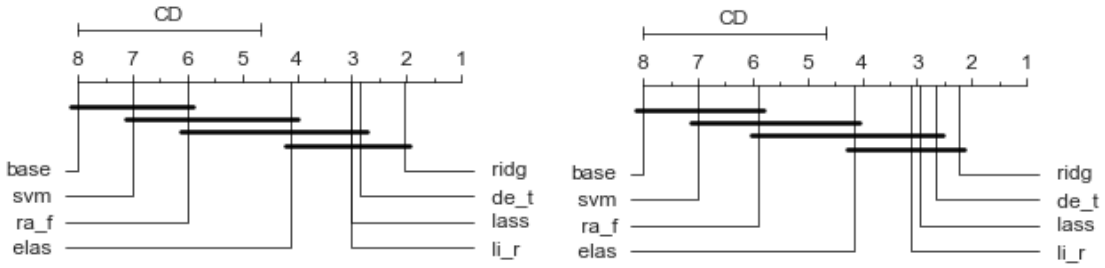


Figure 5.4: Nemenyi critical diagrams for 20 and 30 (data set A)

5.5.1 Results of the hypothesis testing procedures

All of the statistical tests performed considered as significance $\alpha = 0.05$.

Overall, all the hypothesis testing results leads to the rejection of all of the H_0 hypothesis as there are significant differences, namely with the p-value results being below α. The critical difference (CD) diagrams for the nemenyi post-hoc tests are depicted in figures 5.4, 5.5 and 5.6. These diagrams show which clusters of algorithms exist in the results.

In **data set A** results (see figure 5.4) it is evident that even though the baseline model has the worst results, it does not present any statistical difference when compared with the SV model. The best results are achieved by RI and DT algorithms.

In the case of **data set B** (see figure 5.5) there are two individual clusters, for both the 20 as well as 30 years experiments. In particular, the 20 year results show that LA and LR show the worst rank results, while RF and DT achieved the best rank results. For the 30 year results the worst results are obtained by LA and RI, while RF and DT have the best ranks.

Finally, in the case of the **full data set**, (see figure 5.6), the depicted set of results show two clusters of algorithms, with the best set of results, in both cases, being achieved by RF, DT and RI.

In all of the aforementioned results, when experiments show a higher amount of missing data (data set B and full data set), RF achieved the top ranked position. When there aren't any missing data concerns (data set A), DT and RI regression models show the lowest overall error estimation.

A discussion of the preliminary set of results on the prediction approach for the predition of the age-of-onset of ATTRVal30M carriers is presented next.

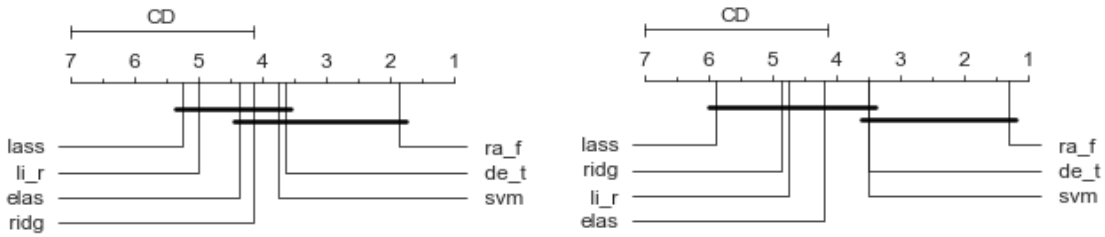


Figure 5.5: Nemenyi critical diagrams for 20 and 30 years (data set B)

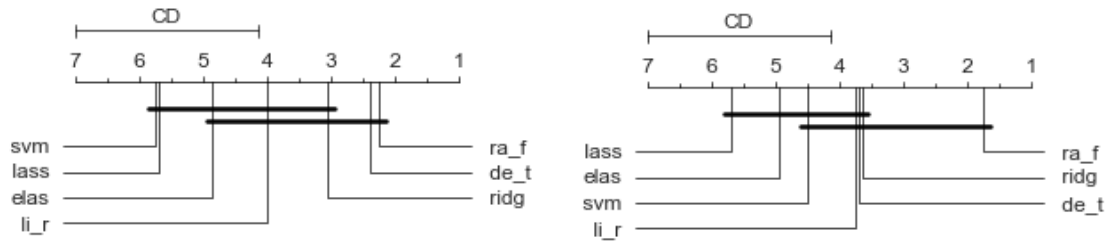


Figure 5.6: Nemenyi critical diagrams for 20 and 30 (full data set)

5.6 Discussion

This work exploits data already being collected in a research center that works under the umbrella of Centro Hospitalar Universitário de Santo António (CHUdSA). This machine-learning based approach resorts to the distribution of the age-of-onset of diagnosed patients and their individual familial genealogical trees to model changes over time. The final purpose is to use this information to predict their risk of developing symptoms onset, and compare, when applicable and possible, the results to a clinical baseline that represents the medical practice by 2018.

With this approach it is possible to make predictions for patients to whom the baseline model does not apply. This can represents a valuable contribution for the medical facility.

From the preliminary results it is possible to conclude that when there is a high degree of missing values (data set B), the best algorithm is RF. This follows previously reviewed background which states RF works well with missing data. In the case of data set A it is clear that RI, DT, LA, and LR improve over the baseline results - in other words we enable the improvement of the current medical practice.

Regarding the research questions explored in chapter 1 the results presented here directly impacts (RQ1) and (RQ2), since it presents results of a summarizing approach focused on extracting meaningful features, by transforming the genealogical trees into networks of data records. These results are further explored next.

Chapter 6

What is the importance of feature selection?

When addressing prediction problems it is crucial to have an integrative and iterative process that assists in a continuous refinement and adjustment of strategies, based on newly found information or ideas. This way it is possible to ensure personalized and optimized healthcare outcomes.

In this chapter the focus changes to evaluate the importance of genealogical data, when clinical records have no information related with the ancestor of the patient, namely its gender or age of disease onset. This is evaluated with a comparison of the estimated predictive error results after combining first and extended level features, with the initial set of patient features.

For a clear assessment of each type of features please refer to table 5.1, depicted in the previous chapter. For reference purposes, this chapter involves the results presented and published both in [84] and [85].

6.1 Introduction

Feature selection [101] plays a crucial role in machine learning models. It involves indicating both the importance of sets of features to specific business problems as well as, from a more academical point of view, what are the types of features that work best with specific sets of algorithms. As such, feature selection represents an important step in machine learning model development since having irrelevant features, e.g., noise, can decrease general accuracy. Also, allowing a model to learn based on irrelevant information can lead to irregular performance values. This suggests that a careful evaluation of the results presented previously needs to be addressed.

Prediction algorithms can approach feature selection from a different perspective. In this case it is noteworthy that Ridge (RI), Elastic Net (EN) and Lasso (LA) can contribute in cases were the results are irregular. RI [55] is a technique used when the data suffers from multicollinearity

(independent variables are highly correlated). It works by adding a degree of bias to the coefficient estimates, so as to reduce the standard errors.

LA (Least Absolute Shrinkage and Selection Operator) adds "absolute value of magnitude" of coefficient as penalty term to the loss function. Its main advantage is to reduce over fitting, not only help during feature selection phases.

EN can be seen as working as a equilibrium between RI and LA, since it combines the penalties of RI and LA to get the best of both worlds. It aims at minimizing the complexity of the regression model, improving overall models' explainability, which is why it is advantageous when dealing with a data set with many features with possible correlation issues.

6.2 Chapter research hypothesis

In this study, the research hypothesis are:

- (H1) whether extending the basic set of patient features with first level genealogical features improves the results of the best performing regression algorithms;

- (H2) whether the further extension of the set of features with extended genealogical features improves the results of the regression algorithms.

6.3 Approach

This chapter answers which group (or groups) of genealogical-based features best serves to predict the AOO of ATTRv Amyloidosis patients (please refer to table 5.1).

Since the data set was compiled over several decades, it has information regarding disease carriers, that were diagnosed as well as observed by different professionals. So, it can be greatly impacted by possible data problems (missing or erroneous data).

Actually, it is noteworthy that in medical data sets, missing information constitutes one of the most common issues [49]. Another one, maybe with greater impact, deals with the risk of errors, due to tedious manual data compilation techniques. This fact underscores the importance of assessing possible variations in prediction results, specially when incorporating specific feature sets.

As such, this chapter represents a small evolution or follow-up over the previous chapter. It is important to note that, as previously, there are three distinct data sets. Firstly, **data set A** includes patients with a previously diagnosed ancestor. Secondly, **data set B** comprises records of patients for whom we do not have a measurable age of onset for the transmitting ancestor. Lastly, the **full data set** encompasses records from both data set A and data set B. To deal with missing data the approach used previously was maintained.

6.3.1 Approach and regression algorithms used

As previously stated, this prediction approach builds a set of virtual scenarios, each one focused on a specific patient's age. This allows for the simulation of the task of predicting the age-of-onset of each patient along with their lifespan. This means that, for each of the patients, and considering their asymptomatic phase (since birth until symptoms onset) each patient was unfolded into a virtual representation of one of its valid ages. This way, prediction at each age can be seen as a specific and individual regression predictive task.

The resulting virtual scenarios aligned each patient, independently of their actual year of birth or year of onset.

In the final evaluation procedures, the patients where aligned accordingly to their year of onset. This way it is possible to prevent data leakage in the full prediction process.

6.3.2 Modeling schema and experiments setup

To estimate the predictive value of each group of features (please refer to table 5.1), it was important to create a predictive pipeline with model tuning, selection and evaluation phases. After the experiments, the results were compared while resorting exclusively to the mean absolute error (MAE) metric.

The **model selection** schema was defined in a top-level procedure that, for a given numerical age, generates an instance of a model. The chosen evaluation schema consisted on a nested cross validation, where for each training phase, there is another cross validation, to find the set of best tuning parameters. This way it is possible to have a model selection between different algorithms, and tune their parameters towards a specific domain. The full algorithm is presented in [88].

After both the tuning of the parameters and the model selection are completed, the predictive accuracy for each model is estimated by focusing on the MAE present in the outer folds.

In this chapter the experimental setup states that for each individual age group (20 and 30 years-old data sets), and for each individual data set (data set A, data set B, and full data set), a total of 4 experiments need to be performed.

The final experiments consisted on:

- patient features, hereafter designated as experiment 1;
- patient + first level features, hereafter designated as experiment 2;
- patient + extended features, hereafter designated as experiment 3;
- patient + first level + extended features, hereafter designated as experiment 4.

6.4 Results

The results for both the 20 and 30 age experiments are expressed in figure 6.1.

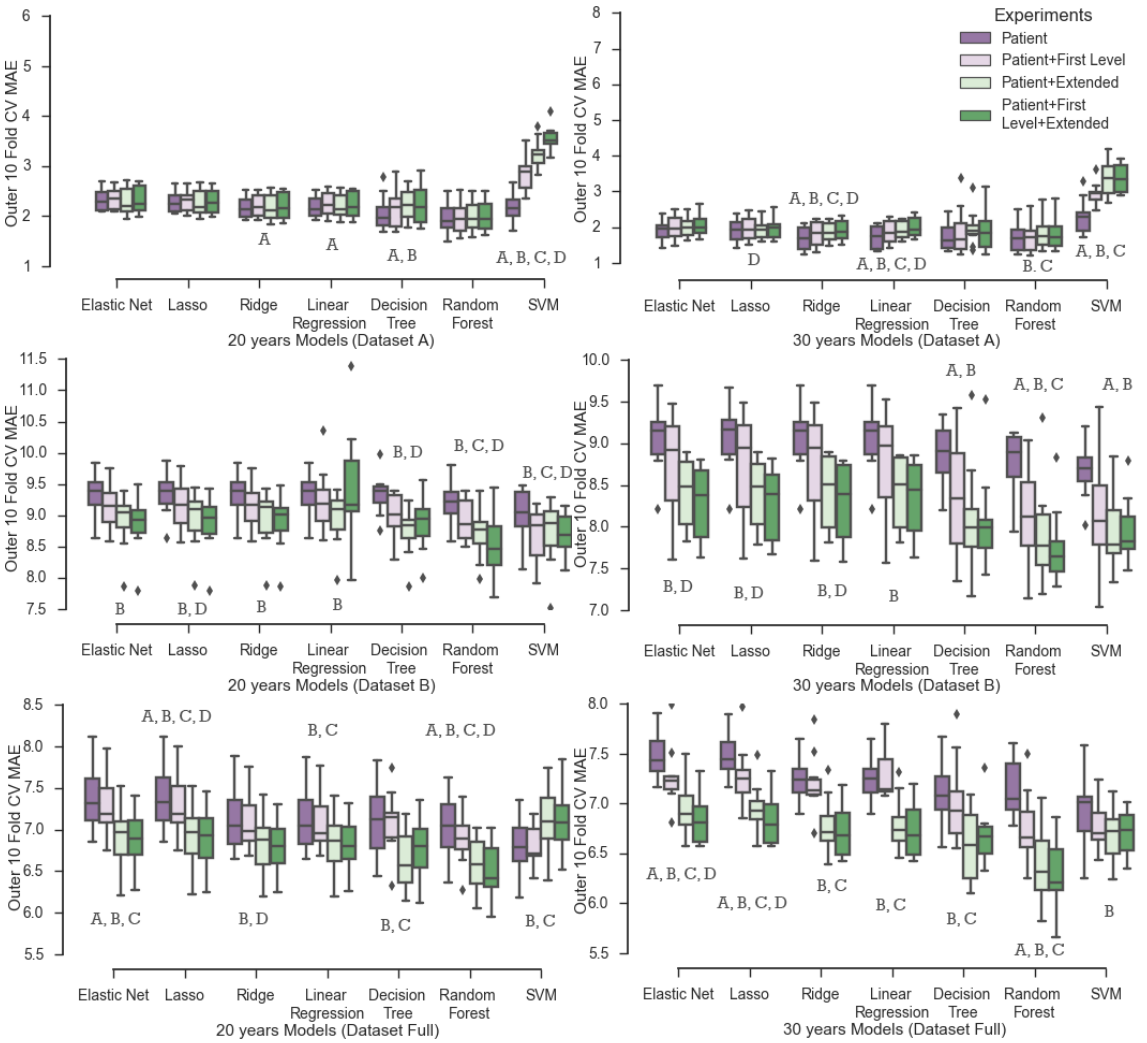


Figure 6.1: 20 years (left) and 30 years (right) experiments, for data set A, data set B and full data set, from top to bottom. Each algorithm has a set of 4 experiments. When there's a letter (A, B, C, D) near to the box-plot this means that there is a significant difference in the set of results between: (A): experiment 1 versus experiment 2; (B): experiment 1 versus experiment 3; (C): experiment 2 versus experiment 4; and (D) experiment 3 versus experiment 4.

An analysis of the individual box-plots shows there is a clear improvement of the predictive accuracy of **data set B** and **full data set** when there is an increase in the number of features.

To check for statistical significance, the chosen procedure was a wilcoxon hypothesis rank test. This lead to the need to perform 4 individual statistical comparisons, namely:

- (A): Experiment 1 versus Experiment 2;
- (B): Experiment 1 versus Experiment 3;
- (C): Experiment 2 versus Experiment 4;
- (D) Experiment 3 versus Experiment 4.

Wilcoxon hypothesis rank test consists on a non-parametric statistical hypothesis test used to compare if two related samples, or repeated measurements on a single sample mean rank values differ. When there are significant statistical differences, these are represented in the graphs next to each individual algorithm as (A), (B), (C), and (D). Each comparison (i.e., A, B, C and D) is considered individually.

The results depicted previously show that when there isn't a letter in the graph it is not possible to reject the null hypothesis. On the other hand if there is a letter, then, based on the statistical test, one can be confident, with a 95% confidence interval, of the existence of significant differences in the ranks. In this case the main observations consist on:

- (data set A): when extra information is integrated to the patients features there is a general loss of accuracy for the SV results;
- (data set A): the error amplitude is smaller for shrinkage methods, which leads to the conclusion that these methods are important for cases with high number of features and that, have possible correlations between them;
- (data set B and full data set): there is an increase in the amplitude of the errors for most algorithms when results are compared with data set A;
- (data set B and full data set): and, finally, there is a general increase in the number of outliers present, which leads to the conclusion that there is a significant difference in the distribution of the training and testing sets for each fold.

With these sets of results and regarding the previously defined research hypothesis (H1) and (H2) it is possible to conclude that:

- in the data set A, results tend to get worse when first level and extended level features are added, which is understandable as this data set is expected to have the more relevant characteristics of the patient feature set;
- in the case of data set B and full data set, it is evident that there is an improvement in the predictive accuracy, by adding first level features;
- and, finally, there are high differences in the reaction of the algorithms to each set of features.

6.5 Discussion

This chapter compares the predictive accuracy of three sets of features (please refer to table 5.1) used in the preliminary results for the prediction of the age of onset of ATTRv Amyloidosis patients. A set of experiments were defined to determine if, on average, the rank results of a set of nested cross validation experiments changed.

When considering each individual algorithm, results show that there are large differences in the results under each of the set of experiments. This signals it may be important to perform a

thorough evaluation to determine the behavior of the algorithms to each feature. This may be even more relevant for the RI, LA and EN results.

Regarding the research hypothesis, experiments verified that the addition of first level genealogical features does indeed improve the predictive accuracy of the tested regression algorithms for the data set B and for the full data set experiments (H1). It was also verified that the further extension of features with extended genealogical characteristics also lead to an enhancement of the predictive results, in both data set B and full data set experiments (H2).

As for the overall research questions in this thesis, this chapter expresses an approach that takes advantage of the methodological approach for time-to-event prediction. In this case it shows how it is possible to change tightly coupled data assembled for time-to-event episodes prediction in order to measure the value of different sets of features (RQ1). It also reinforces that the sumarization approach signals it can be an effective approach to extract effective time based feature sets. Notwithstanding it is worth mentioning that since the feature spaces is large, there may be curse of dimensionality issue, especially in data set B experiments. This should have been assessed with an analysis regarding the confidence interval of the regressor besides a test against the NULL hypothesis that says that the outcome is independent of the features considered.

The next chapter further enhances (RQ1) and (RQ2) results by presenting a new feature construction methodology based on embedding feature construction.

Chapter 7

Is important information being lost?

The two previous chapters presented the preliminary results of a predictive approach that takes into account genealogical, as well as clinical data, from patients being followed at Unidade Corino de Andrade (UCA) from Centro Hospitalar Universitário de Santo António, in order to predict their age-of-onset. In general terms, the previously presented approach summarizes the information collected in the genealogical trees. Then, by taking into account the amount of available ancestor data, it resorts to one of three models to indicate the risk of a patient experiencing symptoms in the future.

Since the feature construction methods studied so far have provided summaries of the overall information, a question emerged: ***if when resorting to genealogical tree summaries, is important information being lost?***

This chapter presents a new approach, based on feature embeddings, and evaluates whether this new, less data loss prone, genealogical feature construction approach can improve the previous results. For comparison purposes, it is deemed important to continue to compare these results with the medical baseline.

7.1 Introduction

This chapter presents a new feature construction approach focused on transforming genealogical family trees into numerically dense vectors, called embeddings in machine learning. The resulting set of features are then handled as input features into the previously assembled prediction pipeline.

In simpler terms, the purpose of this chapter is to evaluate if the summarizing approach handled so far is the best (or even the only) possible path for transforming the information present in genealogical trees for prediction based inputs. In this case, it is also deemed important to check if there is another approach that can later be used to improve the regression-based setting assembled so far, for TTR-FAP age-of-onset prediction.

7.2 Chapter research questions

This chapter has a set of focused research questions related to further exploring possible answers for RQ1 and RQ2. These are:

- (RQ7.1) can we improve predictive results for TTR-FAP age of onset, using a representation learning approach?
- (RQ7.2) does the combination of learned and pre-defined features obtains useful predictive results?
- and (RQ7.3) in what conditions is it useful to employ a representation-learning approach to the prediction of age of onset of TTR-FAP patients?

The experimental methodology corroborates these hypotheses in what was, at that moment, as far as it is possible to check, the first work exploiting embeddings as a viable feature construction methodology to represent heterogeneous sets of family trees.

7.3 Problem overview

This chapter's objective remains in predicting the age of onset of asymptomatic TTR-FAP patients, with a dataset based on over 900 individual genealogical trees, as well as a set of clinical information of patients. To this purpose the genealogical trees and the patient data records, namely gender, age-of-onset, birth and death dates were integrated over a set of different age instances. In this case each patient is unfolded into a set of age centered replicas, with each feature corresponding to the known value at the date the patient was at a specific age.

With a more computational view in mind, in conceptual terms, to model this approach, it is important to define a function f that takes as input a set of n manual predictors p, and a feature vector e, of d dimensions, and outputs an embedding of the genealogical tree for a specific i year. These representations are fed to a prediction pipeline that conceptualizes a virtual scenario accordingly to the patient's age (see Equation 7.1 and figure 7.1) to predict his future age of onset.

It is expected that, even thought each patient has a distinct disease progression from other patients, they should have a comparable evolution. So, depending on the characteristics, the prediction approach defines that the prediction task, at each age, corresponds to a specific and individual regression prediction task.

$$y = f(p_1, p_2, ..., p_{n-1}, p_n, e_1, e_2, ..., e_{d-1}, e_d) \tag{7.1}$$

7.3.1 Defining and using embeddings

To obtain an optimized vector representation from genealogical trees and relations between individuals, it is important start from the individual's life events, namely birth and death, so as to represent the **genealogical trees** over their lifetime. Disease diagnosis and symptom onset are also

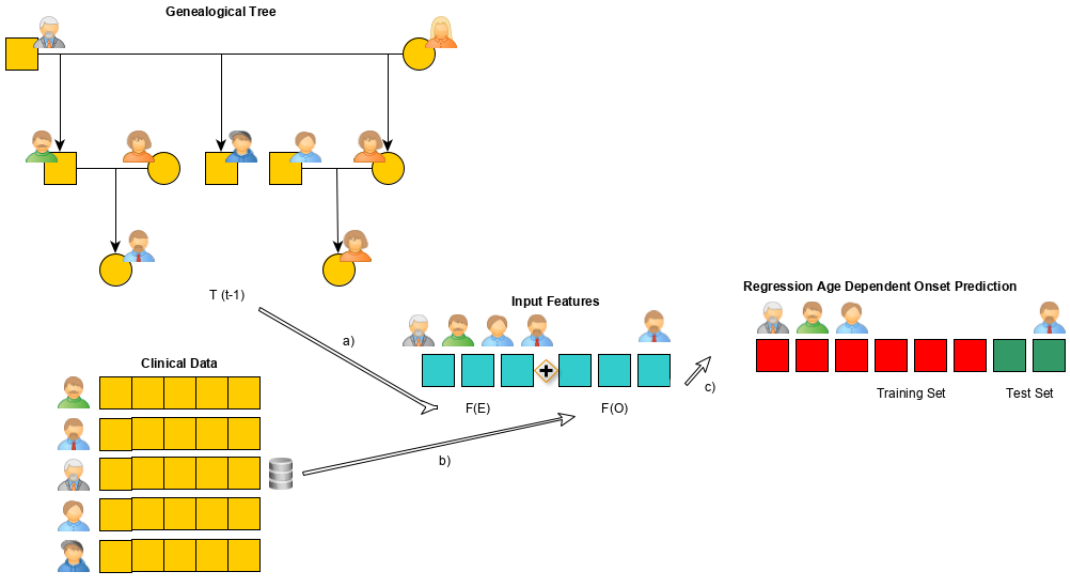


Figure 7.1: Conceptual Graph-based Embedding Prediction Pipeline

taken into account, namely when the *node* of a tree consists of a clinical identifiable patient. Each patient that is matched to a genealogical tree node v has a set of input and output familial relations r, where relevant relations are depicted in equation 7.2. Then, it is up to the embedding algorithm to select and maintain the most representative structural data.

It is important to note that, in this chapter, the interest relies on the relations between biological parents and their children, and doesn't explore second degree or higher familial relations, such as uncles or cousins. To combine this goal with the work presented in the two previous chapters, and related with the two previous publications [84, 85], the approach continued by, for each genealogical tree, generating the embedding representation, on a yearly basis, from 1952 [8] until the current year. In this way, it is possible to account for when the first patient that belongs to the tree (proband) is positively diagnosed, kicking-off familial follow-up of relatives. Then, and for each tuple (patient, prediction scenario), where scenario pertains to data set a, data set b, or data set f (previously referenced as full data set), the method considered the genealogical embedding that corresponded to the year prior to the patient's age instance.

As an example, if the prediction task at hand is to predict the age of onset of asymptomatic patients at 20 years of age, and if the current patient was 20 years old in 1982, then the familiar embedding representation of the patient corresponds to the year 1981. The embedding generation mechanism was implemented with the help of application programming interface (API) presented in [95].

With this data alignment, the embedding representation of the tree serves as a meta-model of prevailing familial relations.

$$father_{of}(A,C) \tag{7.2a}$$
$$mother_{of}(B,C) \tag{7.2b}$$
$$child_{of}(C,A) \tag{7.2c}$$
$$child_{of}(C,B) \tag{7.2d}$$

7.3.2 Experimental setup

In this chapter, the experimental setup was outlined so as to measure the predictive gain of resorting to graph-embeddings over manually engineered features, while also evaluating possible bias associated with the missing data imputation approach, and the chosen graph-embeddings dimension.

To achieve that aim, three important sets of features were delineated, namely:

- a set of original features, henceforth designated as **Original**, combining the features presented in previous works in ([84], [85]);
- a set of embedding based features to be used as predictors, henceforth designated as **E**;
- and, a set of original and embedding features combined, from now on designated as **O.E**.

To accomplish the task of evaluating possible bias in the missing imputation approach (k), as well as the dimensional length (d) of embeddings, the framework developed resorted to varying the two parameters between:

- 5, 10, 15, 20 (parameter ***k***);
- 64, 128 (parameter ***d***).

As to the individual regression tasks themselves, the chosen problem can be defined as *to predict the age of onset of asymptomatic patients on ages ranging from 22 to 46 with a step of 3 years (22, 25, 28, ..., 46)* (parameter ***a***), while testing on linear regression (LR), XGBoost (XG), elastic net (EN), random forest (RF), lasso (LA), support vector regression machine (SV), ridge (RI), and decision tree (DT).

The experiments were separated into different models, according to whether the patients have information on their ancestors' age of onset or not, thus creating a trio of experimental data sets henceforth referenced as:

- data set a, for patients with at least one direct ancestor who was previously diagnosed and referenced as symptomatic;
- data set b, for patients without at least one direct ancestor previously diagnosed as symptomatic carrier with a known age of onset;

- data set f, that combines all patients, independently of the information of their ancestor;

This separation means that most patients will disappear from *data set(b)* and appear in *data set(a)* as the modeled age increases.

With regards to the estimation and evaluation of the error for the different experiments, the focus was on the distribution of the mean absolute error (MAE) calculated with a simple holdout strategy, where the training set holds 80% of the records and the test set considers the most recent 20% of diagnosed patients. The choice of using holdout strategy was mainly for pragmatic reasons, since a train-test-validation strategy using cross-validation, or even nested cross validation, lead to unpractical execution times.

7.4 Results

This section outlines and presents the results.

7.4.1 Evaluating embedding variants

If the best variant in each individual regression task is considered - i.e. at each age of prediction -, it is noticeable that these vary according to the data sets that served as inputs (see table 7.1 and figure 7.2).

Table 7.1: Rank of the top(3) best variants. E.O.64 represents the variant that combines embedding and original features with 64 dimensions, while E.O.128 represents the embedding and original features for 128 dimensions. Original represents the variant that only considers original features.

r_i	age		variant	k	alg	mae ± sd		variant	k	alg	mae ± sd		variant	k	alg	mae ± sd
	22	dataset(a)	Original	5	ri	2.7 ± 3.2	dataset(b)	E.O.128	20	xg	4.2 ± 5.3	dataset(f)	Original	20	la	2.8 ± 3.6
	25		E.O.128	10	en	2.4 ± 2.7		E.O.128	5	xg	3.9 ± 4.9		E.O.64	20	la	2.7 ± 3.4
	28		E.O.128	10	en	1.9 ± 2.3		E.O.128	5	xg	3.7 ± 4.8		Original	5	la	2.7 ± 3.1
	31		E.O.128	5	la	2.1 ± 2.5		E.O.64	20	xg	3.7 ± 4.8		Original	10	la	2.6 ± 3.3
1	34		Original	5	la	1.8 ± 2.4		E.O.128	15	xg	4.2 ± 5.4		E.O.128	20	en	3.1 ± 3.9
	37		E.O.128	5	en	2.3 ± 2.4		E.O.128	10	xg	4.5 ± 5.4		Original	5	ri	4.1 ± 4.9
	40		E.O.64	10	la	2.3 ± 2.5		Original	5	xg	4.6 ± 5.1		E.O.128	20	ri	4.6 ± 6.0
	43		E.O.128	10	en	1.7 ± 2.3		E.O.64	15	ri	5.2 ± 5.6		Original	15	xg	3.9 ± 4.4
	46		Original	20	en	2.1 ± 2.6		Original	15	ri	5.2 ± 6.0		Original	15	la	3.5 ± 3.9
	22		Original	20	ri	2.8 ± 3.1		E.O.128	10	xg	4.2 ± 5.3		Original	10	la	2.9 ± 3.6
	25		E.O.64	10	en	2.4 ± 2.7		E.O.128	20	xg	4.0 ± 5.0		E.O.64	15	la	2.7 ± 3.4
	28		E.O.64	5	en	1.9 ± 2.3		E.O.128	10	xg	3.7 ± 4.8		Original	10	la	2.7 ± 3.1
	31		E.O.64	5	la	2.1 ± 2.5		E.O.64	10	xg	3.8 ± 4.6		Original	15	la	2.6 ± 3.3
2	34		Original	5	en	1.8 ± 2.4		E.O.64	15	xg	4.4 ± 5.6		E.O.64	20	en	3.1 ± 3.9
	37		E.O.64	5	en	2.3 ± 2.4		E.O.128	15	xg	4.5 ± 5.5		Original	20	ri	4.1 ± 5.0
	40		E.O.128	10	la	2.3 ± 2.5		Original	15	xg	4.9 ± 5.1		E.O.128	15	ri	4.7 ± 6.2
	43		E.O.64	10	en	1.7 ± 2.3		E.O.64	20	ri	5.3 ± 5.8		E.O.64	20	xg	4.1 ± 4.5
	46		E.O.128	10	en	2.1 ± 2.6		Original	20	ri	5.3 ± 6.1		Original	20	la	3.5 ± 3.9
	22		Original	15	ri	2.8 ± 3.2		E.O.128	15	xg	4.3 ± 5.2		Original	15	la	2.9 ± 3.6
	25		E.O.128	5	en	2.4 ± 2.7		E.O.128	10	xg	4.1 ± 5.1		E.O.64	20	en	2.7 ± 3.3
	28		E.O.128	15	en	1.9 ± 2.3		E.O.128	15	xg	3.8 ± 4.9		Original	15	la	2.7 ± 3.1
	31		E.O.128	5	en	2.1 ± 2.5		E.O.64	20	en	3.8 ± 4.8		Original	20	la	2.6 ± 3.3
3	34		Original	10	la	1.8 ± 2.4		E.O.64	5	ri	4.4 ± 6.1		E.O.64	15	en	3.2 ± 3.9
	37		Original	5	en	2.3 ± 2.4		E.O.128	5	xg	4.5 ± 5.2		Original	15	ri	4.2 ± 5.0
	40		E.O.64	5	la	2.3 ± 2.5		E.O.128	5	xg	5.2 ± 5.5		E.O.64	20	xg	4.7 ± 4.9
	43		Original	10	en	1.7 ± 2.3		E.O.128	20	xg	5.5 ± 5.0		Original	20	xg	4.2 ± 4.7
	46		E.O.64	10	en	2.1 ± 2.6		Original	5	ri	5.3 ± 6.2		Original	10	en	3.5 ± 3.9

The top(3) ranking results suggest that:

- in all the top performers, the best result is achieved using only the embedding-based features, and these do not surpass the baselines of human engineered features;
- that there is an improvement, especially in data set(b), when the focus is on the E.O.128 or E.O.64 alternative,

This leads to the conclusion that, *when a patient does not have enough information regarding its ancestors, there is predictive gain in generating and resorting to family tree embeddings as part of the predictors*. This is most clear to observe in the fluctuation of the original best results in data set(b), and the similarity between the E.O.64 variant and the original results in data set(f), when these are taken into consideration (see figure 7.2).

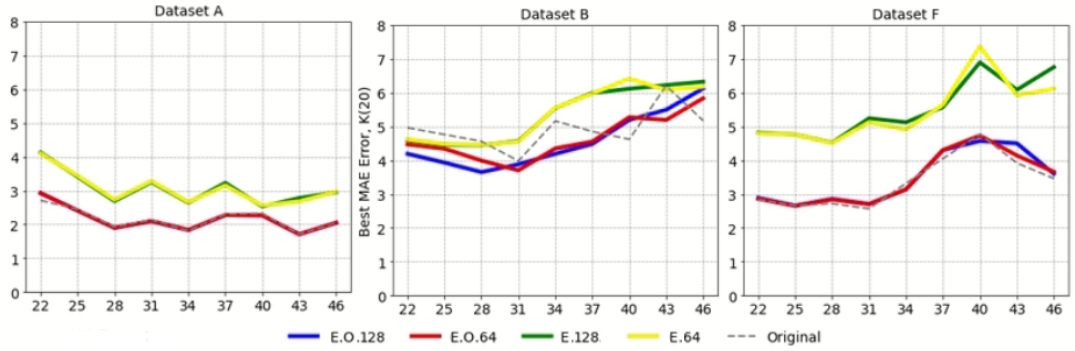


Figure 7.2: Best results, for each prediction task, for K(5) experiments. E.O.128 represents the variant that combines embedding and original features for a dimension of 128, while E.O.64 has a dimension of 64. Note that, in each age and according to each variant, it is possible to have different *best* algorithms. For the case of predictions at 22 years, the *best* algorithms are *EN* for E.O variants, *SV* for E variants and *RI* for Original. In data set(a) all the records have information regarding the transmitting parent age of onset, while in data set(b) this is unknown. Data set(f) combines the patients of data set(a) and data set (b). X axis represents the age models while y axis represents the average mean absolute error (MAE) values.

Next, an evaluation of the prediction of the age of onset of asymptomatic patients is conducted. It focuses on the 22 years old results for 5 nearest neighbors. This hyper-parameter value for the missing imputation approach was chosen by analytically observing the evolution of the prediction results; in this case, by focusing on the best variant, as opposed to the original set of experiments. From the results section (see figure 7.3), it is clear that k(5) experiments are stable in all experimented ages.

7.4.2 Studying the effect of predicting AOO for 22 years old asymptomatic carriers

For a clear picture of the variance difference between all the models, for all the individual experiments, a comparison between the error results of all the comparable test sets was conducted. This was done by resorting to a Friedman rank test, with a level of significance α of 0.05.

The hypothesis results leads to reject all of the H_0 hypothesis, as there are significant differences (p-value below α and near 0). To have a clear view of the magnitude of the difference

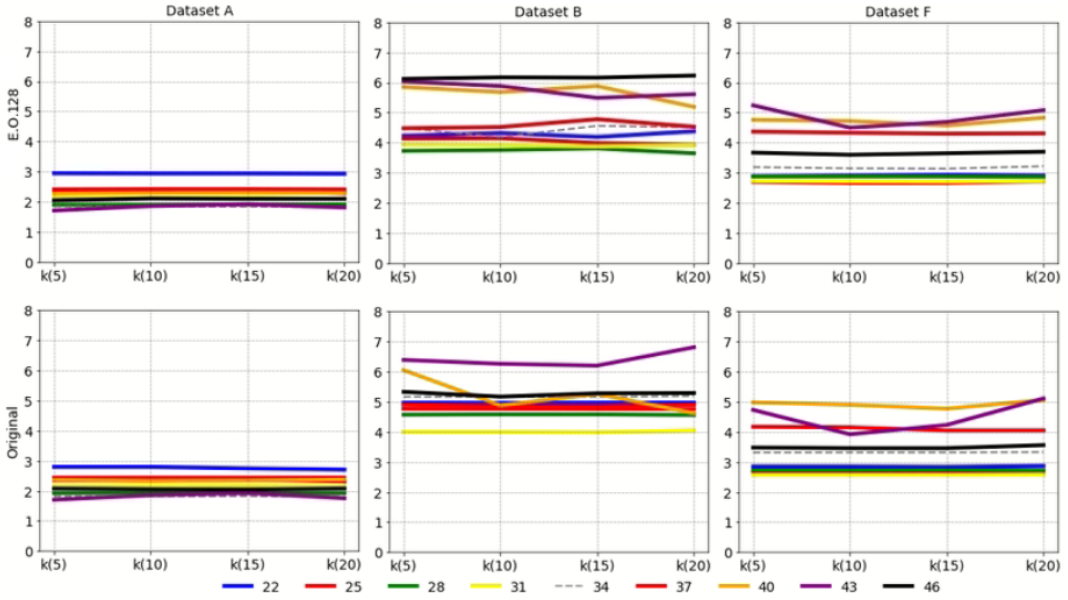


Figure 7.3: Influence of *k* parameter on the best variants: E.O.128 and Original. Each line represents the best values for specific patient age prediction task and the different missing data imputation results are represented by k(n), where n is the number of closest neighbors averaged for imputation. In data set(a), these values are constant, while in data set(b) and data set(f) there is some fluctuation for models starting from the age of 43 years, with values decreasing between k(5) and k(10), and increasing between k(15) and k(20).

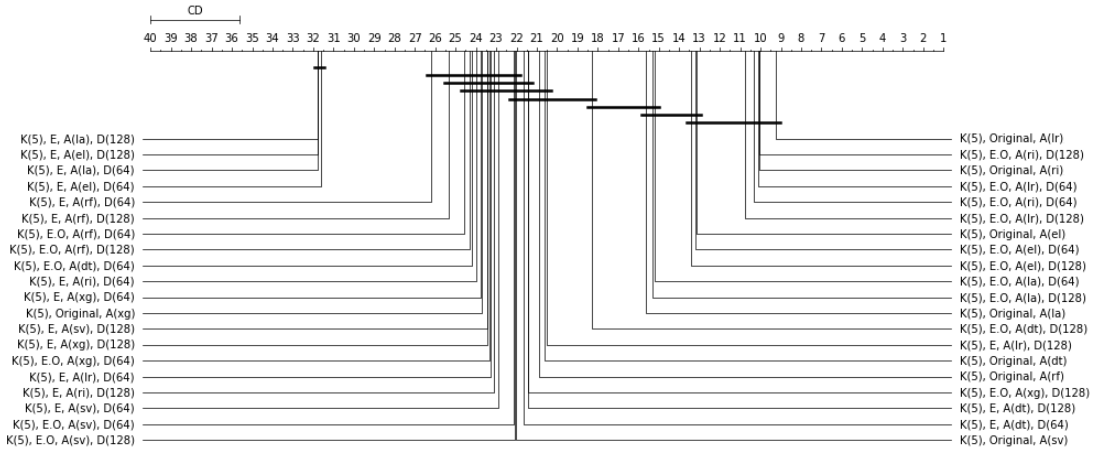


Figure 7.4: Nemenyie results for each prediction model for age = 22, k(5) and data set(a) parameters. D(128) represents results for a dimension of 128, while D(64) represents a dimension of 64. E.O represents the experiments for combining embedding and original features, while E represents experiments only with embeddings features and original represents experiments with only manual features. Each individual prediction algorithm is represented in the A(x) parameter with: LR → linear regression, EL → elastic net, LA → lasso, RI → ridge, RF → random forest, DT → decision tree, SV → support vector machine regressor, and XG → XGBoost. Results show that the combination of original and embedding features statistically performs as well as original experiments.

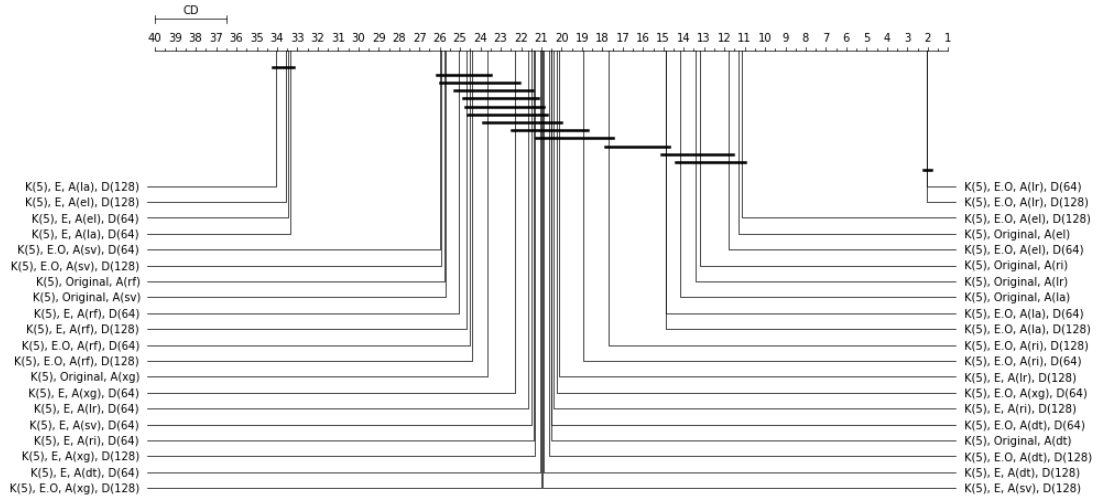


Figure 7.5: Nemenyie results for each prediction model for age = 22, k(5) and data set(b) parameters. Results show that combining original and embedding features statistically performs better than original experiments.

between all the experiments, it is relevant to follow-up with a nemenyi post-hoc test and evaluate the critical difference diagrams (see figures 7.4, 7.5 and 7.6). Results show that, for all the experiments, there are different levels of similar variance results, observable on the improvement of O.E experiments for data set(b) and data set(f).

7.5 Discussion

This chapter presents a new methodology for feature construction, focused on using embeddings to transform a set of clinical indicators present in genealogical hereditary trees. As such, it presents an approach that takes advantage of applying network embeddings as a feature construction method, thus transforming genealogical trees into numerical feature vectors.

The prediction results were compared with manually engineered features, (**Original**), with two new set of results, namely individual embedding feature vectors, (**E**), and combined human engineered and embedding feature vectors, (**O.E**).

Results show that there is evidence that the application of embedding feature construction methods improves current state of the art prediction approaches for the age of onset of TTR-FAP patients (RQ7.1), and that by combining manually engineered features with graph-embeddings feature vectors it is possible to have predictive improvement (RQ7.2).

It is also possible to conclude that the O.E approach has greater impact when there is no complementary information regarding the immediate familial history regarding the disease (RQ7.3). When this is not the case, manually engineered features show better results.

As for the main research questions presented in chapter 1, the work presented here seeks to answer (RQ1), (RQ2) and (RQ3) by presenting a new methodological approach for feature construction of tightly coupled data records (RQ1), which is then compared with the manual based

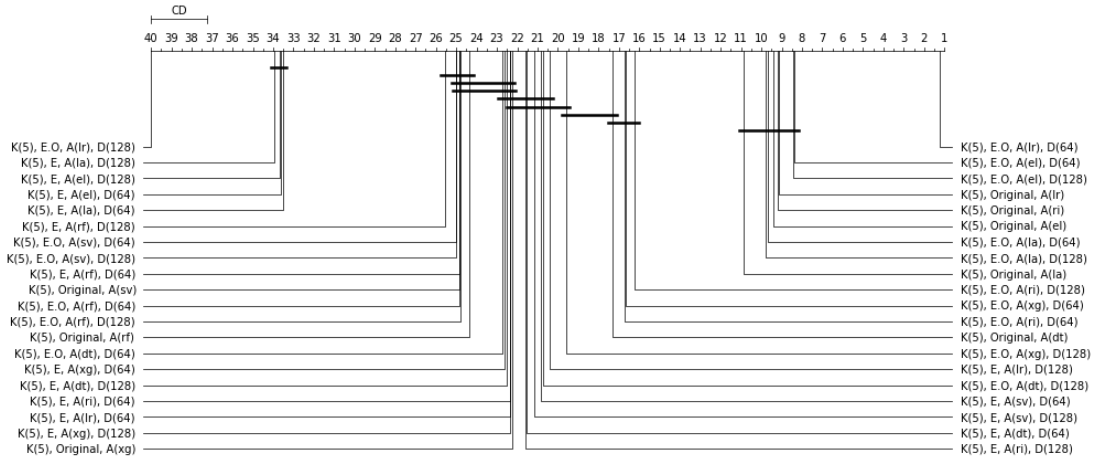


Figure 7.6: Nemenyie results for each prediction model for age = 22, k(5) and data set(f) parameters. Results show that original and embedding features statistically performs better than original experiments.

feature construction approach presented in the two previous chapters (RQ2). The comparison methodology, which uses the Friedman rank test to compare results, allows to robustly evaluate the impact of using genealogical data for time-to-event prediction (RQ3). Results confirm that it is indeed possible to resort to such data to robustly evaluate time-to-event prediction.

Notwithstanding since the feature spaces are large and there is a risk to enter into the curse of dimensionality issue, especially in dataset B, it would be relevant to assess the confidence interval of the regressors as well as if there are any patients for which the imputation mechanism is biased.

Chapter 8

Is it possible to blend heterogeneous models?

Predicting the age-of-onset for a given patient is relevant for management and clinical issues. Independently of the methodologies used, the results can be biased towards algorithms characteristics. So, it can be interesting to assert as to possible bias of prediction results.

This chapter represents an evolution of the previous ones, as it takes advantage of an established practice for predicting age of onset of ATTRv Amyloidosis patients which resorts to genealogical and clinical data and evaluates if it is possible to take individual predictions and blend them, while triggering an improvement in prediction accuracy, as well as its acceptance by clinical professionals. Of note that the intended improvement may not rely solemnly on accuracy but on other qualitative metrics, such as a decrease in results variability. As such, this chapter intends to study what approaches are easily integrated with standard regression or classification based data-oriented prediction methodologies so as to improve state of the art results.

Of note that the approach here presented works while blending the information that exists in neighboring time models. Its' purpose consists in combining the results based on each other capabilities.

The blending mechanism employed corresponds to different methodological strategies, namely **Symmetric** (Sym) and **Asymmetric** (Asym) models. These consist on two different averaging approaches, which were further extended with a simple weighting mechanism. In this case the models created were: **Symmetric** (Sym), **Symmetric-weighted** (Sym-w), **Asymmetric** (Asym), and **Asymmetric-weighted** (Asym-w). These ensemble models are compared to the original approach which is focused on individual regression base learners responses. The algorithms consist of: Baseline (BL), Decision Tree (DT), Elastic Net (EN), Lasso (LA), Linear Regression (LR), Random Forest (RF), Ridge (RI), Support Vector Regressor (SV) and XGBoost (XG).

Of note that the final results show that by aggregating predictions from neighbor models, the average mean absolute error obtained by each base learner decreases. Thus, the best results are

achieved by regression-based ensemble tree models as base learners.

8.1 Introduction

Blending regression results with ensemble approaches is a powerful technique in machine learning. It focuses on developing and evaluating combination strategies that together create more accurate and robust predictions. This is a useful approach when dealing with complex data sets, where a single model may not capture all the underlying patterns and relationships.

In the context of regression, blending involves training multiple regression models on the same data set and then combining their predictions. Each model in the ensemble bag contributes to the final prediction. This way, the way their outputs are combined can vary. Some common methods include simple averaging, weighted averaging, or more complex meta-learning approaches, where the outputs of the individual models are used as inputs to a higher-level model.

While ensemble learning is considered an important approach, it needs careful tuning and validation strategies. They also need to be carefully implemented since such strategies can lead to over complex implementations that can ultimately, produce results that suffer from data leakage.

8.2 Chapter research questions

In this chapter, the research questions consist of:

- (RQ8.1) can we improve the predictive results by combining different age-neighbor models?
- (RQ8.2) what is the impact of incorporating a weighting strategy to these bags of predictions?
- (RQ8.3) and, how can we combine our sets of base learner prediction results?

Regarding the last research question, it is important to study combination strategies focused on the prediction phase, so as to work towards combining base learners developed in geographically distinct clinical reference centers. This means it is important to develop ways to combine multiple models, created in different locations, so as to improve the general accuracy of predictions, without compromising the privacy and security of patients' personal information. This way this work ensures that patient data remains confidential, thus agreeing with guidelines for data protection.

8.3 Single learning approach and combination strategies

This section overviews the single learning approach and its' transformation to output combined results. This way it is possible to easily compare both.

8.3.1 Prediction problem and single learning approach

As stated previously, this work focuses on predicting the age-of-onset of ATTRv patients. In a more conceptual view, one starts with a given patient X known to be a carrier of the ATTRv mutation, at a given moment in time T. The medical purpose is to estimate the predicted value of the variable Y that represents the age when symptoms will manifest by the patient (onset).

The base predictive models are obtained from one baseline (BL) and several state-of-the-art regression methods, namely: a Decision Tree (DT) model chosen by its interpretability; an Elastic Net (EN), Lasso (LA), and Ridge (RI) approach due to temporal relations present in the features; a Linear Regression (LR) for its simplicity, eventhough this is a multivariate methodology (resorts to several features); a Random Forest (RF) for its results when working with problematic data sets, i.e. noisy, irregular and outlier prone data; Support Vector Regression Machine (SV) for assisting in the course of generating models of non-linear relationships between variables; and finally XGBoost (XG) for its efficient boosting approach.

The BL was implemented following state of the art current medical practice [64] regarding the predicted age-of-onset of symptomatic patients. In this case, and while assuming a linear trend between the age of onset of a patient and that of his/her transmitting ancestor the rules state that:

- if (father and son) $\rightarrow$ target = parent(ageonset) - 6.06 ;
- if (father and daughter) $\rightarrow$ target = parent(ageonset) - 1.23;
- if (mother and son) $\rightarrow$ target = parent(ageonset) - 10.43;
- if (mother and daughter) $\rightarrow$ target = parent(ageonset) - 7.43.

8.3.2 Data and evaluation strategy

The data was supplied by Centro Hospitalar Universitário de Santo António (ChuSA) at two different time moments. These were:

- before 2018 with this subset of data being used for training and validating individual algorithms;
- after 2018 with this subset being used for subsequent testing regarding the original and ensemble scenarios. For the set of features please check [84] and [85] as well chapter 5.

8.3.2.1 Training phase

The training phase takes into account information from all the patients diagnosed previously to 2018 and unfolds each of them into a set of age-specific projections. For each age and patient, the data is gathered up to the year when the patient reached that age and until the patient becomes symptomatic. Each of these patient-age projections become one training example described by clinical as well as genealogical features. The value of the respective age of onset is the target value. These sets of results consider ages from 22 to 64 with a step of 3 years (22, 25, 28, . . . , 64).

After feature construction, the training pipeline runned, for each training example, an imputing missing data mechanism focused on resorting to a 5-nearest neighbor mechanism. It also resorts to hyper-parameter tuning using 5-fold-cross-validation. Then, for each age and learning algorithm, a prediction model is finally obtained from the age-specific patient projections.

8.3.2.2 Testing phase

In the testing phase, and to evaluate the prediction pipeline, patients diagnosed after 2018 are selected, and, after performing the same pre processing activities as in the training data, the models predict their age of onset for each model individually.

So, by taking the data obtained previously to 2018 and for each current symptomatic patient the process considered the information known at the time at which they were 22, 25, 28, 31, 34, 37, 40, 43, 46, 49, 52, 55, 58, 61, and 64 years-old (as long as they were still asymptomatic), and instantiated an individual prediction model. This means that for each age group there is a single experiment with each of the 9 different algorithms (i.e. BL, DT, EN, LA, LR, RF, RI, SV and XG). Later, in the validation phase, patients diagnosed after 2018 were selected and they had their age of onset predicted, individually, for each algorithm.

This phase includes the ensemble combination strategies, defined next.

8.3.3 Combination strategies

As stated, this chapter studies the effect of combining different models to obtain a new, stable, less variance-prone, generic model. In this case, the selected approach concatenates models that share the same algorithm, while focusing on neighboring age instances.

So, considering a patient x with an age a (Equation 8.1), for whom the system has the purpose of delivering an age of onset prediction, there are two aggregation scenarios. In the first scenario, there is a combination of the results of an *age*-based model with the models from the two previous time frames, namely $age-3$ and $age-6$ ages. This corresponds to the **asymmetric scenario** with a set of predictions (Equation 8.4) that averaged (Equation 8.6) generate an ensemble result. When combined with a γ weight vector (Equation 8.2) they generate a weighted set of results (Equation 8.8).

Alternatively, there is a combination of the current *age* model's prediction with the predictions for age $age-3$ and the results of age $age+3$. In this case, the set of results (Equation 8.5) are later combined with a weighting feature vector (Equation 8.3) to generate a weighted (Equation 8.9) or simple ensemble (Equation 8.7) set of results. This corresponds to the **symmetric scenario**.

$$y_b = model(x,a) \tag{8.1}$$

$$\gamma_{asym} = (0.14, 0.29, 0.57) \tag{8.2}$$

$$\gamma_{sym} = (0.25, 0.5, 0.25) \tag{8.3}$$

$$\vec{y}_{asym}(x,a) = (y_b(x,a-6), y_b(x,a-3), y_b(x,a)) \tag{8.4}$$

$$\vec{y}_{sym}(x,a) = (y_b(x,a-3), y_b(x,a), y_b(x,a+3)) \tag{8.5}$$

$$asym(x,a) = mean(\vec{y}_{asym}(x,a)) \tag{8.6}$$

$$sym(x,a) = mean(\vec{y}_{sym}(x,a)) \tag{8.7}$$

$$asym_w(x,a) = \frac{1}{3}\vec{y}_{asym}(x,a) \cdot \gamma_{asym} \tag{8.8}$$

$$sym_w(x,a) = \frac{1}{3}\vec{y}_{sym}(x,a) \cdot \gamma_{sym} \tag{8.9}$$

8.3.3.1 Weighting mechanism

To obtain the weighting vectors referenced in equations 8.2 and 8.3 a generic weight formula presented in algorithm *WeightingPredictions* was taken into consideration. This method takes as input a set of time points.

Time points refer, for example, to the sym_w by a vector with references t-3, t, t+3. In the algorithm presented, Lines 1 to 3 generate the vector of weights, while lines 5 to 7 scale this vector in order for its *sum* to be equal to one.

As an example, to generate the weights for sym_w as presented in equation 8.3 the input must be [-1, 0, +1]. For the $asym_w$ the same function generates the results in equation 8.2 with the input [-2,-1,0]. These experiments did not have into account the age gap difference.

Algorithm 1 Weighting Predictions

Input *agevec*
Output *weightvec*

1: **for** i in agevec **do**
2: $w[i] = 0.5^{abs(i)}$
3: **end for**
4: $total = sum(weightvec)$
5: **for** i in range(len(agevec)) **do**
6: $w[i] = w[i]/total$
7: **end for**=0

8.3.4 Evaluation

Both with traditional and ensemble modeling strategies the performance was measured by averaging a set of prediction results regarding a specific error metric. In this chapter the chosen metric corresponds to the mean absolute error (MAE) (Equation 8.10). Of note that MAE is expressed in the same unit as the target variable and for which larger errors have the same impact as smaller ones. Other metrics, such as mean squared error (MSE) (Equation 8.11) and root mean squared error (RMSE) (Equation 8.12) differentiate between larger and smaller errors. In this case, since the use case corresponds to a medical problem, this is not seen as a requirement. Thus, the interest relies on MAE for the greater interpretability of absolute errors in special for non-technical users.

$$MAE = \frac{1}{V}\sum_{i=1}^{V}|x_i - y_i| \tag{8.10}$$

$$MSE = \frac{1}{V}\sum_{i=1}^{V}(x_i - y_i)^2 \tag{8.11}$$

$$RMSE = \sqrt{\frac{1}{V}\sum_{i=1}^{V}(x_i - y_i)^2} \tag{8.12}$$

8.4 Results

As is, for each tuple (age, algorithm) this chapter presents a set of five macro-level prediction results:

- original prediction results where these are given by BL, DT, EN, LA, LR, RF, RI, SV, XG models (Equation 8.1);
- asymmetric prediction results;
- symmetric prediction results;
- asymmetric-w prediction results;
- and symmetric-w prediction results.

To clarify, the original results correspond to the single learning approach results explained in section 8.3.1 while the symmetric, asymmetric-w and symmetric-w correspond to the combination strategies referenced in section 8.3.3. The overall experimental setup and the different testing combination strategies are schematically referenced in figure 8.1.

After each aggregation, a clipping operation was implemented so as to transforms invalid results, namely those where the prediction result is smaller than the corresponding age. The correction post-processing operator changes the predicted result to two years from the current age a.

To analyze each scenario the average MAE of the original approach was compared with the asymmetric and symmetric scenarios (table 8.1).

An analysis of these results shows that:

- there is a clear improvement of the predictive results by combining different prediction moments, namely by considering the symmetric scenario, both with and without the weighting mechanism (RQ8.1);
- in the asymmetric scenario there is a predictive improvement when adding the weighting mechanism, i.e. when adding a vector of weights [0.14, 0.29, 0.57], for the bag of current and neighbor predictions (RQ8.1);
- and, in the symmetric scenario there isn't a predictive improvement when adding the weighting mechanism, i.e. the results of $(Sym >> Sym-w)$ when adding a vector of weights [0.25, 0.5, 0.25] (RQ8.2).

Figure 8.2 shows the average error of the BL and the top MAE performers, which correspond to DT, SV and XG. Here it is noticeable that:

- the machine learning approach improves prediction results when compared with the medical baseline;
- there is some interpolation between the original and the symmetric results with the original results being quite irregular over very close ages;
- combining the predictions leads to a smoothing effect over the different age models;
- late onset age models (models for ages greater than 50 years) there the smoothing effect is accompanied with an overall improvement in the results for DT and XG, while in the SV it is clear that this does not happen due to the very large peaks in ages around 55 and 58 years.

In is noticeable that these results are impacted by the **original** individual results of each model. This fact leads to the question *what could the effect be to combine models not only over the ages but also over the algorithms since some models work better with heterogeneous data sets.*

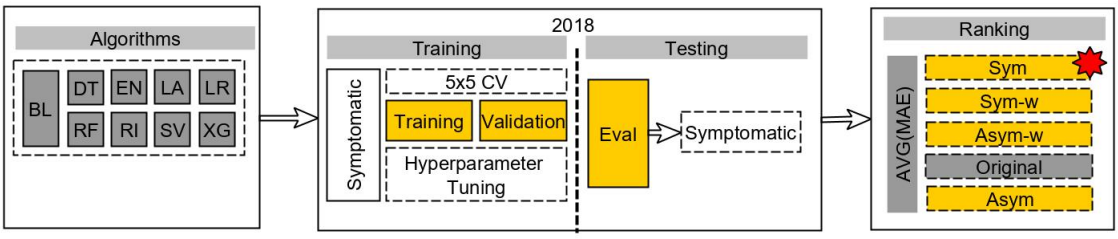


Figure 8.1: Conceptual Models for the Aggregation Scenarios

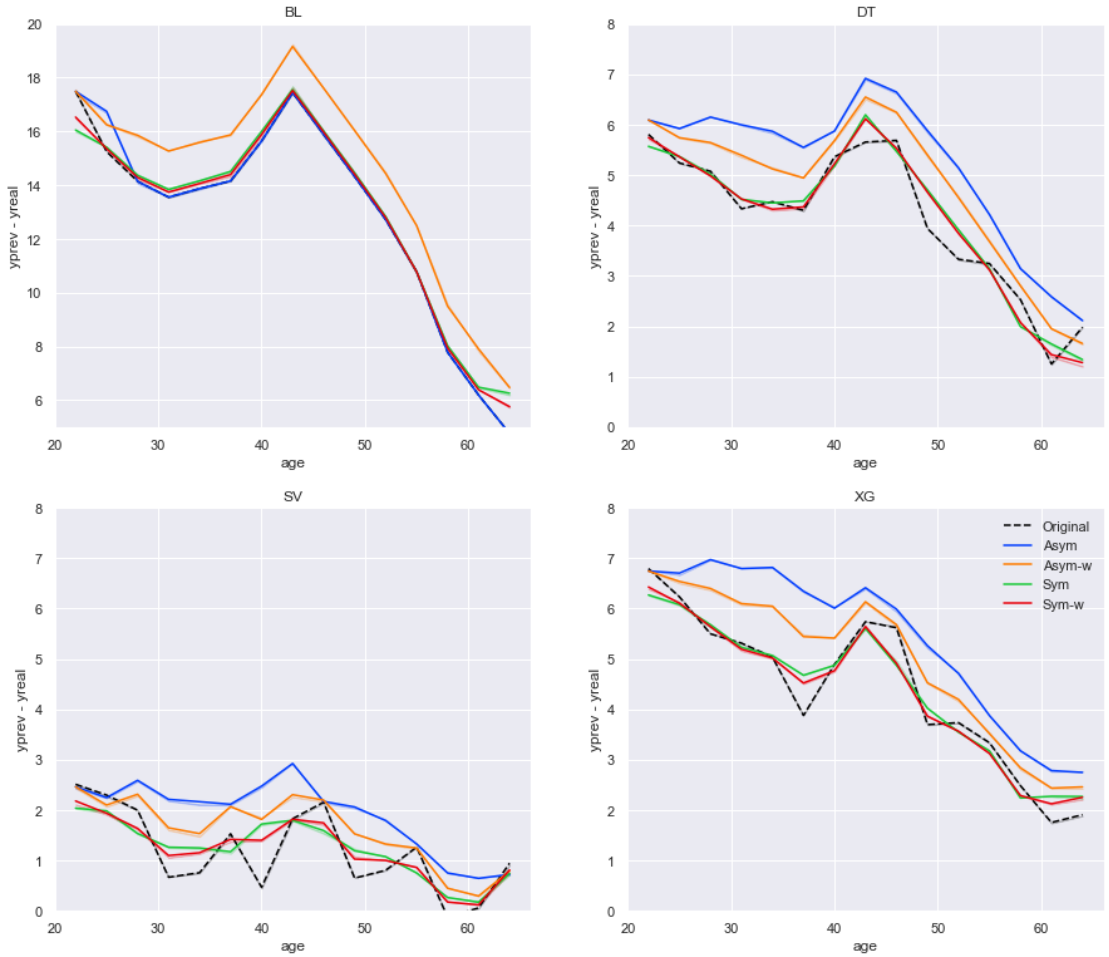


Figure 8.2: Results of the prediction error over the different ages for BL and the three best ML methods - DT, SV and XG - for the test data. Each graph contains the results for the original and each of the different combination strategies (Asym, Asym-w, Sym, Sym-w)

Table 8.1: Average MAE for each pair (algorithm, scenario). With a grey background, we have the approaches that beat the original or base learners.

	original	**asym**	**asym-w**	**sym**	**sym-w**
BL	14.39	14.60	15.65	14.30	14.30
DT	5.64	6.13	5.85	5.34	5.38
EN	7.43	8.15	7.87	7.26	7.30
LA	8.68	9.23	8.96	8.54	8.58
LR	7.30	7.93	7.69	7.07	7.10
RF	6.01	6.62	6.29	5.77	5.81
RI	7.33	8.02	7.76	7.15	7.17
SV	3.29	3.31	3.25	3.14	3.16
XG	5.91	6.56	6.22	5.68	5.72

8.5 Discussion

For this chapter the predictive approach first introduced in [84] and [85] was transformed, so as to determine if single algorithm results would outperform different combination scenarios (RQ8.3). In this case, two different aggregation scenarios were experimented. One is symmetric, centered on the age of the patient. The other is asymmetric with the age of the patient at the right. These scenarios were further enhanced by resorting to a weighting mechanism.

Regarding the research questions it is clear that (RQ8.3) it is possible to take different aggregation approaches and achieve different prediction results. Here, if on the one hand, it was possible to improve the predictive accuracy of an aggregation scenario when compared to the single stand-alone algorithms (RQ8.1), it was also possible to improve upon these results by adding a weighting factor in the asymmetric case.

For reference purposes, in the asymmetric case, it was not possible to improve the original single-algorithm runs (RQ8.2). The most noticeable fact of these sets of results comes from the smoothing effect that the different combination strategies have when compared with the SV results. This fact can help non-data experts grasp the overall results of the machine learning prediction based approach when compared with the BL results.

When account for the overall thesis research questions this work focuses on (RQ3) since it enhances the base learner methodology for predicting time-to-event for ATTRv patients, and (RQ4) since it compares the base learner approach and the blending approach and reaches to a set of quantitative and qualitative improvements. The latter, by decreasing the variability between sequential age models.

Chapter 9

Clinical model for time-to-event prediction

This chapter represents the merging of the works presented previously at conferences and constitutes a significant milestone in disseminating the findings of this thesis work. It is noteworthy that the data gathered at the medical facility was continuously gathered over the different publications, and was target of distinct data pre processing and cleaning procedures, that had to evolve as new people were assigned to the process. As such, this chapter resorts to information referenced in [83], and uses patient data not previously used in [84, 85]. Also it consists of data records that were target of several data correction procedures.

9.1 Introduction

As stated, this work follows upon the work referenced previously which correspond to early attempts at answering an intricate medical problem. It acts as an enhancement, since the volume of experiments as well as of manual constructed features increased from the previously 58 to a total of 77 features. It also performs a more in-depth explanation of the full methodology assembled to predict the age-of-onset of ATTRv Amyloidosis patients.

9.2 Materials and methods

This section formulates the problem, defines the current medical practice by presenting a baseline model, and finishes with a short overview of the cost modeling approach assembled.

9.2.1 Background

For a given patient X known to be a carrier of TTR-FAP mutation, at a given moment in time T, the interest is on estimating the variable Y that represents the age when symptoms will manifest (onset).

The prediction task is modeled both in a classification and a regression setting. The resulting models can be used by medical practitioners during a medical appointment to better understand the clinical case of each patient.

9.2.2 Baseline model

For TTR-FAP, the incorporation of a predicted age of onset to insure a more accurate asymptomatic patient follow-up of is not an established practice. Relatively recent works that study this need and acknowledge its importance, in the Portuguese population, can be found in [64] and [27].

In the first work, authors studied the evolution of the age of onset in a group of TTR-FAP families and found the existence of an anticipation mechanism [41]. This fact has been studied for TTR-FAP in other geographical areas with the works of [34] and [106]. This work, which is focused, sound and resorts to a large enough pool set of patient records, can foster the development of a wide range of analytical inferences focused in the distribution of the age-of-onset of TTR-FAP patients and its genealogical influence.

On a later work [27], a team of researchers with a wide knowledge in treating and following patients diagnosed with TTR-FAP debated the importance of establishing the predicted age of onset of asymptomatic disease (PADO). As authors acknowledge, the definition of a prediction value can insure the correct monitoring of known TTR-FAP asymptomatic patients *until such time that onset is medically observed* with less costs for the public health system. In their work authors mention that PADO depends on patient mutation, typical age of onset for that mutation, and age of onset in family members with ATTR amyloidosis, specifically its propositus (e.g. family first diagnosed case). This work lacks the study and experimentation of a data-driven approach, or at the very least a basic computational basic transmission of the rules one should follow to established their methods predicted age-of-onset. With this in mind, in this work, and observing that there isn't all of the aforementioned information from the second study, the definition of a medical baseline will rely on the work of [64].

So, the baseline rule model assumes a linear trend between the age of onset of a patient and that of the transmitting ancestor, and states that:

if (father and son) **then**

$$\hat{y} = parent(ageonset) - 6.06;$$

if (father and daughter) **then**

$$\hat{y} = parent(ageonset) - 1.23;$$

if (mother and son) **then**

$$\hat{y} = parent(ageonset) - 10.43;$$

if (mother and daughter) **then**

$$\hat{y} = parent(ageonset) - 7.43;$$

These rules, which have different constant values that represent the average difference in the age of onset of all previously diagnosed known pairs of (parent, child), are based on the existence of an anticipation inheritance mechanism. This genetic anticipation, a common factor in most hereditary diseases, acknowledges that a patient has a higher risk to show symptoms of the disease prior to that of the parent(s). In this case one takes into account this anticipation mechanism and for each known different pair of (patient, ancestor) gender the overall differences of each known element are averaged.

To produce comparable results in all experiments a special rule was added. Its' intended purpose was to consider the cases where there isn't enough information regarding the gender or the age of onset of the transmitting ancestor. In this case, for simplicity purposes, one states that *the patient will most likely feel symptoms of the disease in the two years after the prediction time point*. Thus the final rule states that:

if (data unknown) **then**

$$\hat{y} = patient(current\_age) + 2$$

As is, this baseline suggests that there is a need to continuously assess and study of familiar inheritance ([84], [85]), what represents a common practice in hereditary diseases studies.

9.2.3 Data

As previously stated the data comes from a medical facility in Portugal. From their records one started by selecting information from 2259 patients, diagnosed and followed before 2018. Then, it was considered relevant to add a validation data set for the overall prediction approach and include data from 213 patients, diagnosed between 2018 and the end of 2021 (see table 9.1). By then, there were 511 patients classified as TTR-FAP asymptomatic carriers that have since then experienced symptoms of the disease, or are expected to, in the next few years. These numbers have a tendency to increase over time, as new families are diagnosed.

9.2.4 Cost model

To allow for an earliest as possible treatment time, asymptomatic patients known to be carriers of a disease should be monitored for symptoms [103]. For TTR-FAP, these may include a wide range of highly heterogeneous symptoms (progressive symmetric sensory-motor neuropathy, autonomic

Table 9.1: 2472 Patients diagnosed with TTR-FAP in Unidade Corino de Andrade as of December 2021

Symptoms	**Gender**	
Symptomatic (before 2018)	Male	1258
	Female	1001
Symptomatic (after 2018)	Male	110
	Female	103
Asymptomatic or Pre-Symptomatic	Male	178
	Female	333

dysfunction, GI complaints, weight loss, among others) [51]. Current patient cost evaluations focus on the estimation of annual healthcare costs associated with management of patients according to disease stage ([56], [57]). These stages occur after the symptoms onset, which begins with patients walking unaided and ends when they are fully bedridden [3]. If left untreated, patients have a decreased mean life-expectancy of 12 years with mean healthcare costs of 125.645€ per patient ([56], [57]).

To estimate the potential gain of the application of a prediction based approach to coordinate the triage of asymptomatic patients, this chapter presents an **annual cost model estimate** and compares the prediction error estimate that takes into account both the predicted ($\hat{y}$) and true values (y) against the current medical follow-up of TTR-FAP asymptomatic patients ([78], [51]).

To model the cost error appropriately, it is important to separate between anticipation and delay errors. In either case, one starts the follow up approach 5 years prior to the prediction of the onset of symptoms. Of note, in [27] authors debate the importance of predicting the age of onset of TTR-FAP patients in order to improve the asymptomatic patient follow-up without defining a prediction-based approach and studying its error distribution. Nonetheless, they define the pre-prediction follow-up period as 10 years.

9.2.4.1 Delay

The delay scenario occurs when the patient presents symptoms before the predicted time event. This originates *late diagnosis costs*, which includes loss of medical treatments. Although it presents a decrease in asymptomatic patients follow-up costs, the loss of treatment options by itself represents a non-negligible set of costs that have direct impact on a patient's life options. To account for that, the model treats this error type I case as:

$$\hat{y} > y, Cost = CDPT$$

where CDPT stands for **Cost Delay Prediction Treatment** and represents an impact factor for final cost values.

9.2.4.2 Anticipation

If one encounters an anticipation error, then the patient presents symptoms after the predicted time event. Here one is faced with a set of costs related with the annual revision of the patient status. These may include exams and medical appointments, both of which are currently included in the asymptomatic patient follow-up procedure. For this, the error model cost (type II error) is defined as:

$$\hat{y} \leq y, e = y - \hat{y}$$

$$Cost = e * CAPA$$

where CAPA stands for **Cost Asymptomatic Prediction Approach**.

It is important to note that as the medical costs of reviewing a patient process are already included in the current patient follow-up procedure, one expects a decrease of the overall costs by supporting a predictive-based asymptomatic patients follow-up policy. This decrease depends on the type of prediction error the chosen model prediction generates (if type I or type II).

9.3 Approach and technical issues

This section describes in depth all the key issues regarding the methodology assembled to predict the age-of-onset of ATTRv Amyloidosis patients, while resorting to genealogical data for feature construction. The chosen feature construction methodology is exclusive to the manual synopsis methodology.

9.3.1 Instance feature construction

To construct the patient feature set, the focus was on information regarding important life events, namely the time of *birth*, *death*, and *onset*. From those, a set of clinical age related variables were extrapolated, namely age of onset, current age, and age of death. As these variables delimit the time horizons where each patient is at, and represent quite different life stages, it was deemed interesting to represent each patient for a set of specific ages, that will henceforth be referenced as time marks. These *age* time marks correspond to: 22, 25, 28, 31, 34, 37, 40, 43, 46, 49, 52, 55, 58, 61, and 64.

Of note that the representation of a patient at a specific time mark is only viable for the prediction pipeline if the patient was still alive, both for the classification and regression case, or if the patient was still considered asymptomatic by the medical team, in the regression case.

As an example, figure 9.1 shows, for three different patients, their time marks, and their valid prediction time horizons.

The **patient A**, consists on a patient with age of onset (30) and age of death (39). To create all valid *patient A* instances it is important to focus on: (i) the valid time horizons, which go from 22

to 30 years; (ii) that currently this is a symptomatic patient, with an age of onset at 30 years; (iii) that the asymptomatic ages range from 22 to 28 years old; and that this patient died when he was 39 years old. As is, this patient can be incorporated in the prediction pipeline, and can be used in the model construction phase. Of note that he can't be used in models for ages greater then 39, since he died when he was 39 years old.

As for **patient B**, he can be present in models until he reaches his age of onset (39). In order to create each *patient B* instance one has that: (i) the valid time horizons range from 22 to 39; (ii) this is currently a symptomatic patient, with an age of onset of 39; (iii) the asymptomatic ages have an interval from 22 to 37 years old. Currently this patient is still alive but is already showing symptoms of the disease. As is, this patient may be incorporated in the model construction phase.

Finally, **patient C**, has as life event his current age of 39. His relevant data is that: (i) the valid time horizon range from 22 to 39 ages; (ii) this is an asymptomatic patient; (iii) the asymptomatic ages range from 22 to 39 years; (iv) this patient can only be used in the model validation phase, where this chapter performs asymptomatic patient evaluation as one can not assess his prediction error.

In conclusion, for a specific patient, independently of his current state, and for each time horizon, this chapter calculates three sets of features (see table 9.2): Patient Feature Set, First Level Feature Set and Extended Feature Set. When creating each feature one resorts only to the information known at the time of the event, thus needing to enforce a set of validation constraints.

Most of the first and extended level feature set features are defined by sets of different mathematical operators, and depend on familiar connections between patients and their family members. In this case it was important to define and create the lowest common ancestor algorithm, so as to calculate for each age model, the valid relations for each patient [12].

After constructing the set of valid instances for each patient, both for the training and validation sets, a set of blocks were conceptualized and implemented to assist in the definition of the model construction and evaluation phases (see Figure 9.1). But before discussing it, it is important to discuss data separation issues.

9.3.2 Data separation

To validate the approach and compare it with the medical baseline, the methodology resorts to a temporal validation block [11, 104]. So, the prediction pipeline starts by splitting the data set into two parts: one part containing early treated patients that can be used to develop the model, and another part containing the most recently treated patients, so as to assess the overall validity of the process. The algorithm started with a non-random split [104] in which patients that showed symptoms and were diagnosed prior to 2018 were used in the training set and to check the robustness of the model; while the patients diagnosed after 2018 and the patients that currently haven't shown symptoms, but are expected to, were bagged together and used in the validation set. This corresponds to the horizontal data split.

To have comparable results between the approach and the current medical baseline, the cohort suffered a further vertical break, and separated between the patients with known age of onset of

Table 9.2: List of 77 original and generated features classified between Patient, First Level, and Extended Level

Type	Description	N.Feat.
Patient	Patient's Gender	1
	Father and Mother Age Onset	1
	Born before Father or Mother had Symptoms	1
	Year of Birth, Death and Onset	3
	Age Current and Age of Death	2
	Survival Current and Survival Length	2
	Death and Onset Status	2
	Age of Onset (Target Variable)	1
First Level	Number of Children (by gender)	2
	Number of Siblings (by gender)	2
	Number of Uncles and Aunts	2
	Number of Nephews and Nieces	2
	Number of Children with the disease (by gender)	2
	Number of Siblings with the disease (by gender)	2
	Number of Uncles with the disease (by gender)	2
	Number of Nephews with the disease (by gender)	2
	Avg, Max and Min Age Onset of Children (by gender)	6
	Avg, Max and Min Age Onset of Siblings (by gender)	6
	Avg, Max and Min Age Onset of Uncles and Aunts	6
	Avg, Max and Min Age Onset of Nephews and Nieces	6
Extended	Avg, Max and Min Age Onset of Patients followed at the Clinical Unit (by gender)	6
	Avg, Max and Min Age of Onset of Patients in the Family Tree (by gender)	6
	Avg, Max and Min Age of Onset of Early Onset Patients (Patients with Age of Onset < 50)	6
	Avg, Max and Min Age of Onset Late Onset Patients (Patients with Age of Onset ≥ 50)	6

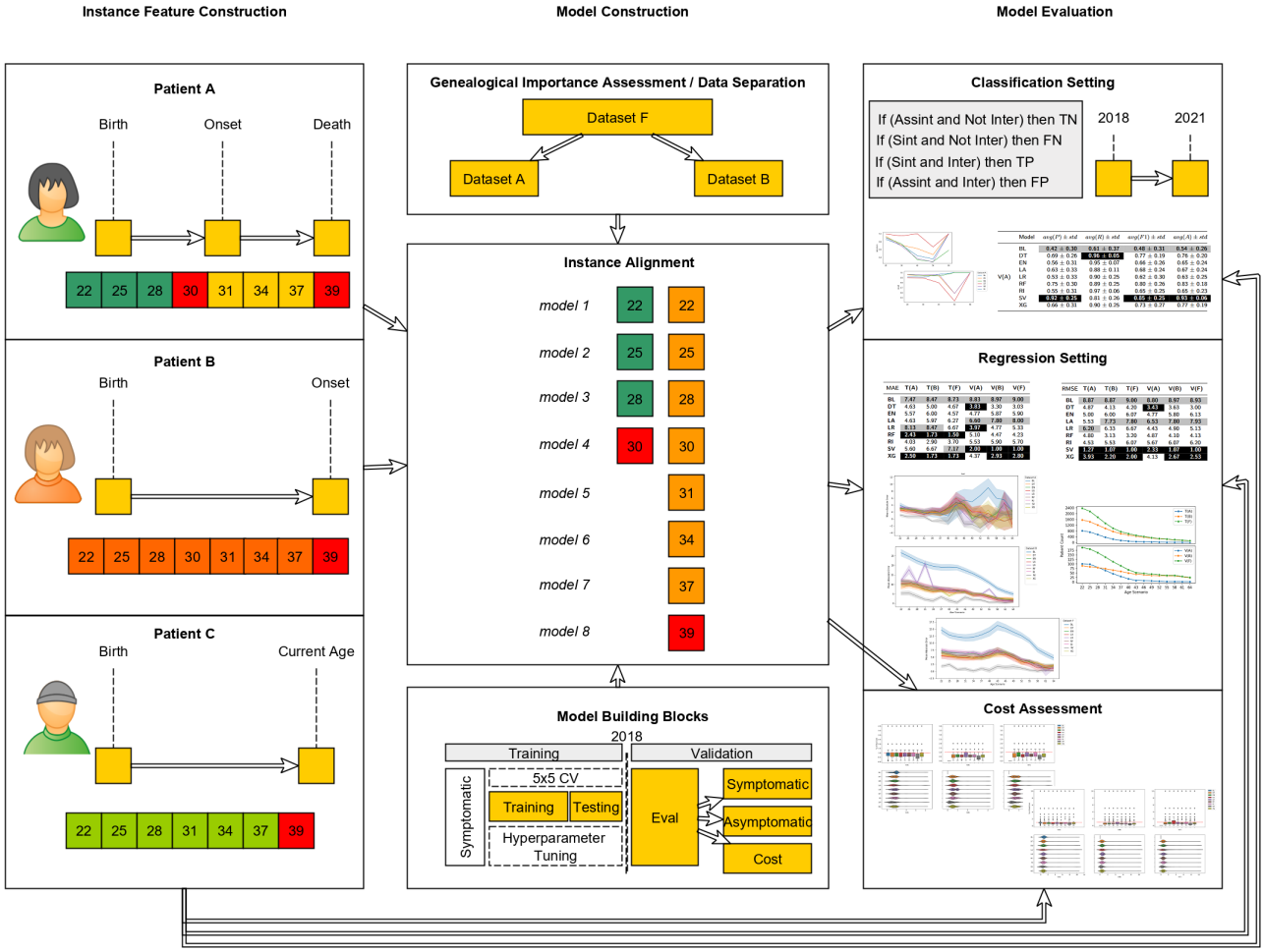


Figure 9.1: Conceptual model of Virtual Age Scenarios

parent(s), and those without. To that end, three types of models were developed: **data set A**, containing patients with previously diagnosed ancestor(s); **data set B**, which contains those for whom we do not have a measurable age of onset for the ancestor; and **data set F**, composed of all available records.

9.3.3 Model construction

After dealing with the horizontal and vertical break of the data set, i.e. the separation between different sets of features and instances to train and validate the work, each patient-instance was aligned (see Figure 9.1) according to the age it represented. This alignment allowed for the construction of different models that could be trained according to the age instance of each patient. This is independent of their actual decade of existence.

With the age alignment process, each patient has a set of prediction results, which correspond to a risk factor. Also, of note, that since the prediction of the age of onset of a patient at each age corresponds to an independent task, the bags of patients can be bagged together, even though they correspond to different temporal decades.

To evaluate a large set of approaches, the pipeline model incorporated a model selection phase. This consists of, for each numerical age, generating an instance of a model with the help of a **nested cross-validation** operator. In this specific case, and for each training phase, one has an outer cross-validation method that optimizes the tuning meta-parameters that each algorithm has.

With this model selection approach, one ends up with an independently tuned model for each time horizon, and can select the best set of hyper-parameters and algorithms for relatively small data sets [88]. In addition, this step allows for a first evaluation regarding the robustness of the model when comparing the results among the different cross validation folds. Of note is that in this case, the focus was on using a random split, since the overall evaluation is only performed with respect to the validation set.

9.3.4 Model evaluation

For the model evaluation assessment, the focus was on developing different diagnostic procedures according to the current state of the patient, i.e. symptomatic or asymptomatic, as well as conceptualizing, implementing, and presenting an independent cost assessment procedure that follows the current asymptomatic patients guidelines presented in [78] and [51].

In one case the choice was to measure the predictive ability in a *classification-based setting*, while in the other the focus on the *regression-based setting*.

In the classification-based setting the purpose is to infer the capability of the approach to answer the question of: "if a patient previously diagnosed with the genetic error will show symptoms of the disease in the next 4 years". In this case, the interest is not in the patient's age of onset, but rather on whether he will be accurately diagnosed as showing symptoms of the disease in the period between 2018 and 2021. The metrics that will be considered as most important are Precision (P), Recall (R), F1 and Accuracy (A).

In terms of the regression-based setting, the purpose is to infer the capability of the approach in answering the question of: "when will a patient show symptoms of the disease". Here the interest is in the difference between the $\hat{y}$ and the y values. The mean absolute error (MAE) and the root mean squared error (RMSE) correspond to the two chosen metrics to evaluate the rank of each algorithm. These were chosen based on the fact that one is well-known for its interpretability easiness (MAE), while the other is better suited for robustness issues (RMSE).

Regarding the measure of the cost gain achieved by changing the current asymptomatic patients follow-up procedure, this value was estimated as the difference between the current and cost of the prediction-based approach, for each patient.

All these evaluations are subject of further explanation next.

9.4 Results

This section presents the experimental setup and the results.

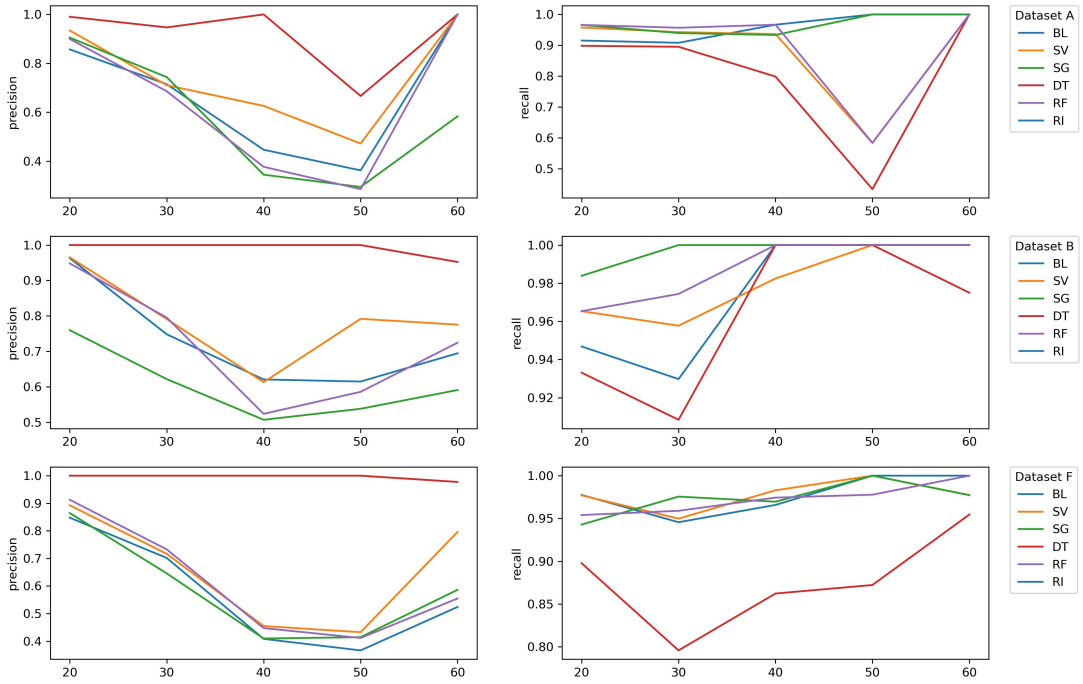


Figure 9.2: Evolution of Precision and Recall for V(A), V(B) and V(F) data set

9.4.1 Experimental setup

As discussed previously, each asymptomatic patient had his information compiled at different ages, namely 22, 25, 28, 31, 34, 37, 40, 43, 46, 49, 52, 55, 58, 61, and 64 years-old (as long as they were still asymptomatic). Then the pipeline instantiates an individual model for each type of data set. This means that for each age group, and each data set, a single experiment is performed with each of the 8 different algorithms (i.e. LR, EN, LA, RI, SV, DT, RF, XG). In total, this pipeline ran 405 experiments to compare the results for different patients' ages.

The results are presented in three layers.

For the classification-setting, the validation results are analyzed, for both the asymptomatic and symptomatic patients. This consists in analyzing, the last prediction age result for each patient and verify if the result fell in the period set between 2018 and 2021 or not. This results in each patient having a single prediction value, which allows for them to be bagged together, independently of the disease stage (see table 9.3 and Figure 9.2).

As for the regression-setting, the average rank of each tuple *(algorithm, data set)* is analyzed, so as to have a clear and simple comparison of how each model performs (see table 9.4 and Figure 9.3). In this case, the average rank in the training set as well as the validation set are also target of analysis. This assists in diagnosing over fitting issues, as well as intuiting the robustness of the approach. Finally, it assists in labeling each model into **top** and **worst** performers.

Regarding the cost assessment, it is important to mention that the cost modeling approach leads to two main types of cost results. These consist of: (i) real cost (CR) that is calculated

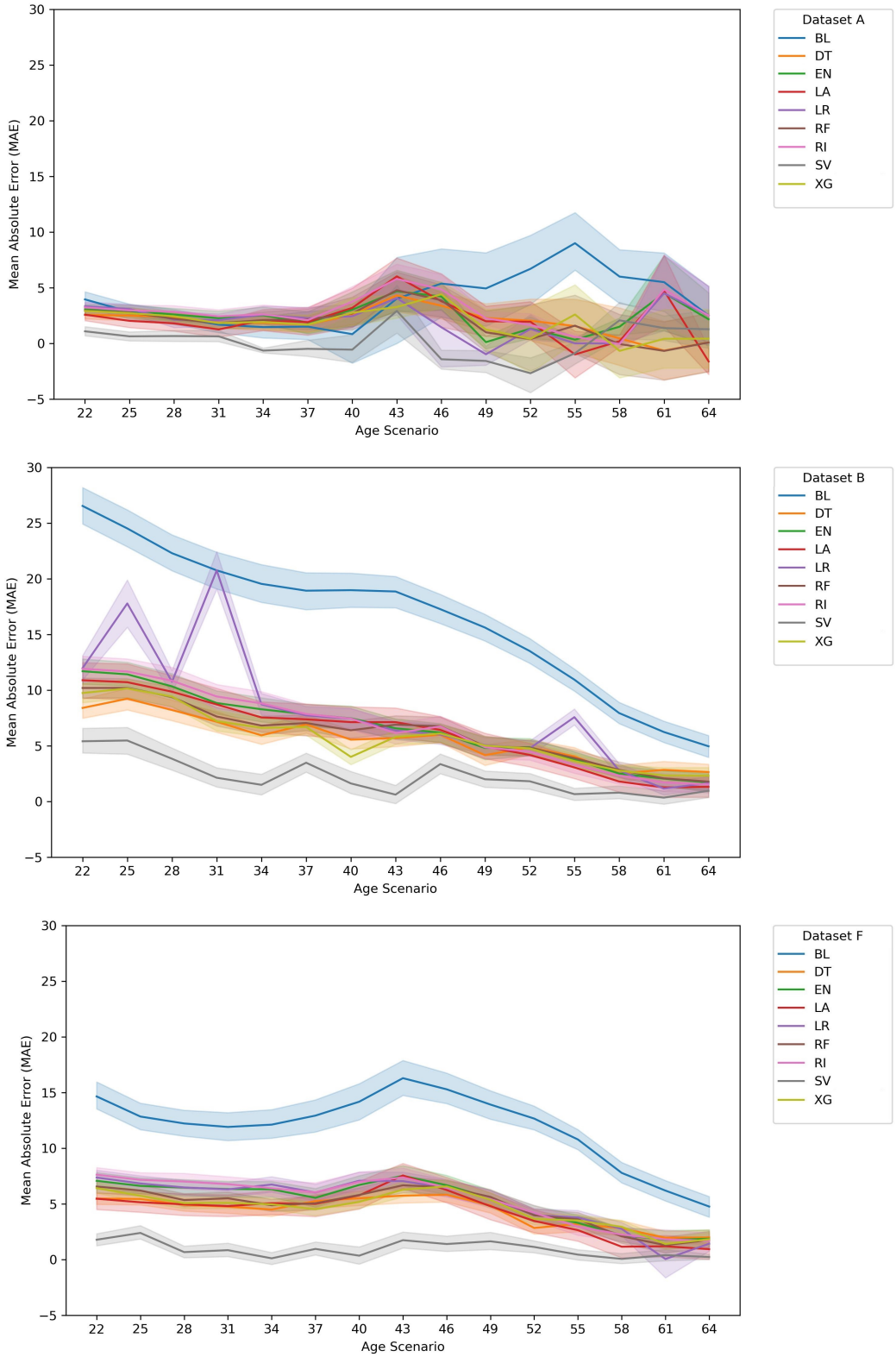


Figure 9.3: Evolution of MAE for Val(A), Val(B), and Val(F) data sets

by taking into account an asymptomatic follow-up approach of all patients from when they are 20 years old; (ii) predicted real cost (CPR) that corresponds to the cost by taking into account the negative impact of cost of delay prediction treatment (CDPT), i.e. the impact of having a prediction that results in an unused follow-up result; and (iii) the cost of asymptomatic prediction follow-up approach (CAPA).

9.4.2 Results

9.4.2.1 Classification setting

For the classification-based evaluation, and by considering the interval between 2018 and 2021, each patient was aligned at the beginning of 2018. In this case the prediction age was transformed to a year instance. Since the prediction results only consider patients older than 21, 9 patients were excluded at this point, since they have less than 22 years and are currently asymptomatic.

Overall it is clear that these results are impacted by the rather small prediction period that ranges between 2018 and 2021. One factor relates with the early age results identifying mostly negative samples. This effect tends to decrease for the late onset cases (e.g., before 50 years old as opposed to after 50 years old).

The ranks of the top and worst performers show that, although there is some fluctuation, the overall top performer is the SV algorithm for most of the chosen metrics. Although it may not be the case for the recall (R) results, this was deemed less important due to the overall high scores of most of the algorithms. In fact, when having a closer look thought the different ages, it is clear the consistency of the SV algorithm.

9.4.2.2 Regression setting

The validation results of the MAE error distribution (see Figure 9.3) shows that, for the data set A case, and at the early ages, the approach ranks top when compared with the BL, although in later ages this isn't as clear. This fact is likely affected by the decrease of the validation set size, over the different age groups. This can be due to the lack of new asymptomatic patients as the ages increase, and in the change between the training and validation sets (see table 9.1).

As for data set B and data set F, it is clear that the machine learning based approach shows the best results, with SV being the best chosen algorithm, and XG the second best.

Regarding the **average ranks** (see tables 9.4 9.5), and when comparing the dispersion and difference between the top-2 performers, two main observations standout. First, that on the one hand, the different MAE ranks aren't stable between the training and validation set. In this case and when looking at the RMSE average ranks it is visible that this does not happen. These ranks difference between T and V sets in MAE results suggests that there is a difference in large errors distribution in SV results. This can indicate some future problems, since clinical models need assurances as to the future performance of the **top** selected approaches. As is, this point will be the focus of future work developments.

Table 9.3: Classification based metrics evaluation. With a black background we show the top performers, and in light gray we show the worst performers. (P - Precision, R - Recall, A - Accuracy, F1- harmonic mean of the precision and recall)

	Model	$avg(P) \pm std$	$avg(R) \pm std$	$avg(F1) \pm std$	$avg(A) \pm std$
	BL	**0.42 ± 0.30**	**0.61 ± 0.37**	**0.48 ± 0.31**	**0.54 ± 0.26**
	DT	0.69 ± 0.26	**0.96 ± 0.05**	0.77 ± 0.19	0.76 ± 0.20
	EN	0.56 ± 0.31	0.95 ± 0.07	0.66 ± 0.26	0.65 ± 0.24
	LA	0.63 ± 0.33	0.88 ± 0.11	0.68 ± 0.24	0.67 ± 0.24
V(A)	LR	0.53 ± 0.33	0.90 ± 0.25	0.62 ± 0.30	0.63 ± 0.25
	RF	0.75 ± 0.30	0.89 ± 0.25	0.80 ± 0.26	0.83 ± 0.18
	RI	0.55 ± 0.31	0.97 ± 0.06	0.65 ± 0.25	0.65 ± 0.23
	SV	**0.92 ± 0.25**	0.81 ± 0.26	**0.85 ± 0.25**	**0.93 ± 0.06**
	XG	0.66 ± 0.31	0.90 ± 0.25	0.73 ± 0.27	0.77 ± 0.19
	BL	**0.55 ± 0.08**	**1.00 ± 0.00**	**0.71 ± 0.07**	**0.55 ± 0.08**
	DT	0.71 ± 0.15	0.98 ± 0.03	0.81 ± 0.10	0.74 ± 0.16
	EN	0.62 ± 0.12	0.99 ± 0.02	0.76 ± 0.09	0.65 ± 0.12
	LA	0.99 ± 0.06	**0.88 ± 0.11**	0.92 ± 0.08	0.92 ± 0.08
V(B)	LR	0.57 ± 0.10	0.98 ± 0.06	**0.71 ± 0.07**	0.57 ± 0.10
	RF	0.77 ± 0.17	0.98 ± 0.03	0.86 ± 0.11	0.80 ± 0.16
	RI	0.59 ± 0.10	1.00 ± 0.01	0.74 ± 0.07	0.62 ± 0.10
	SV	**0.99 ± 0.03**	0.97 ± 0.05	**0.98 ± 0.03**	**0.97 ± 0.03**
	XG	0.70 ± 0.16	0.99 ± 0.02	0.81 ± 0.11	0.73 ± 0.16
	BL	0.55 ± 0.12	0.91 ± 0.04	**0.67 ± 0.09**	0.58 ± 0.08
	DT	**0.55 ± 0.18**	**0.98 ± 0.03**	0.69 ± 0.14	0.57 ± 0.17
	EN	0.59 ± 0.19	0.95 ± 0.06	0.71 ± 0.12	0.62 ± 0.15
	LA	0.99 ± 0.03	**0.77 ± 0.13**	0.86 ± 0.08	0.88 ± 0.08
V(F)	LR	0.56 ± 0.21	0.96 ± 0.05	0.68 ± 0.15	**0.56 ± 0.19**
	RF	0.65 ± 0.20	**0.98 ± 0.03**	0.76 ± 0.14	0.69 ± 0.19
	RI	0.57 ± 0.17	**0.98 ± 0.03**	0.70 ± 0.13	0.60 ± 0.16
	SV	**0.99 ± 0.02**	0.88 ± 0.08	**0.93 ± 0.04**	0.94 ± 0.04
	XG	0.59 ± 0.18	**0.98 ± 0.03**	0.72 ± 0.13	0.63 ± 0.16

Table 9.4: Average MAE ranks of each tuple (algorithm , dataset) for model-based imputation with parameter k=10. T(dataset) corresponds to training results. V(dataset) corresponds to validation results. With a black background we have top(2) **Top** performers. With grey coloring we have top(2) **Worst** performers.

	T(A)	T(B)	T(F)	V(A)	V(B)	V(F)
BL	**7.47**	**8.47**	**8.73**	**8.83**	**8.97**	**9.00**
DT	4.63	5.00	4.67	**3.83**	3.30	3.03
EN	5.57	6.00	4.57	4.77	5.87	5.90
LA	4.63	5.97	6.27	**6.60**	**7.80**	**8.00**
LR	**8.13**	**8.47**	6.67	**3.97**	4.77	5.33
RF	**2.43**	**1.73**	**1.50**	5.10	4.47	4.23
RI	4.03	2.90	3.70	5.53	5.90	5.70
SV	5.60	6.67	**7.17**	**2.00**	**1.00**	**1.00**
XG	**2.50**	**1.73**	**1.73**	4.37	**2.93**	**2.80**

Second, that, there is a stability of the ranks between each horizontal data set separation (e.g. between data set A, data set B and data set F). As is, in this case, and seeing that the MAE ranking is most stable for the XG algorithm, and that this is the top(2) performer of both training and validation data sets, the approach produces good results for the XG algorithm.

The results of the XG algorithm might be influenced by its built-in capabilities to prevent over fitting, while working well with small data, as well as with data with subgroups. Of note that the differences in size between training and validation set can impact the generalization of each algorithm, and can lead to differences between training and validation results (see table 9.1).

9.4.2.3 Cost assessment

The cost model results (see Figures 9.4 and 9.5) show a possible decrease in cost. This can be relatively high for the cases where the patient follow-up is sustained by the prediction results and the patients has late onset symptoms. This can be observed by: (i) the rank of SV, independently of the type of data set; and (ii) the larger dispersion of BL in data set A results. It is important to note that there are several variables that impact these results which might alter the effectiveness on similar problems, namely: the CDPT value, and the capability of a specific algorithm to have higher type I or type II errors (i.e. errors by delay as opposed to errors by anticipation, the cost of the later also being biased by the CDPT value). In these results, the costs for the BL in the Val(B) and Val(F) experiments were removed, seen these are biased towards type II errors.

9.4.3 Discussion

This work explores genealogical features in modeling and predicting the Age of Onset of TTR-FAP patients. The definition of different sets of data explores pre-existing information regarding the patients' ancestry.

In this classification setting results show the supremacy of the SV algorithm, even thought they are impacted by the rather small prediction period. In the regression problem, these show

Table 9.5: Average RMSE ranks of each tuple (algorithm , dataset) for model-based imputation with parameter k=10. T(dataset) corresponds to training results. V(dataset) corresponds to validation results. With a black background we have top(2) **Top** performers. With grey coloring we have top(2) **Worst** performers.

	T(A)	T(B)	T(F)	V(A)	V(B)	V(F)
BL	**8.87**	**8.87**	**9.00**	**8.80**	**8.97**	**8.93**
DT	4.87	4.13	4.20	**3.43**	3.63	3.00
EN	5.00	6.00	6.07	4.77	5.80	6.13
LA	5.53	**7.73**	**7.80**	**6.53**	**7.80**	**7.93**
LR	**6.20**	6.33	6.67	4.43	4.90	5.13
RF	4.80	3.13	3.20	4.87	4.10	4.13
RI	4.53	5.53	6.07	5.67	6.07	6.20
SV	**1.27**	**1.07**	**1.00**	**2.33**	**1.07**	**1.00**
XG	**3.93**	**2.20**	**2.00**	4.13	**2.67**	**2.53**

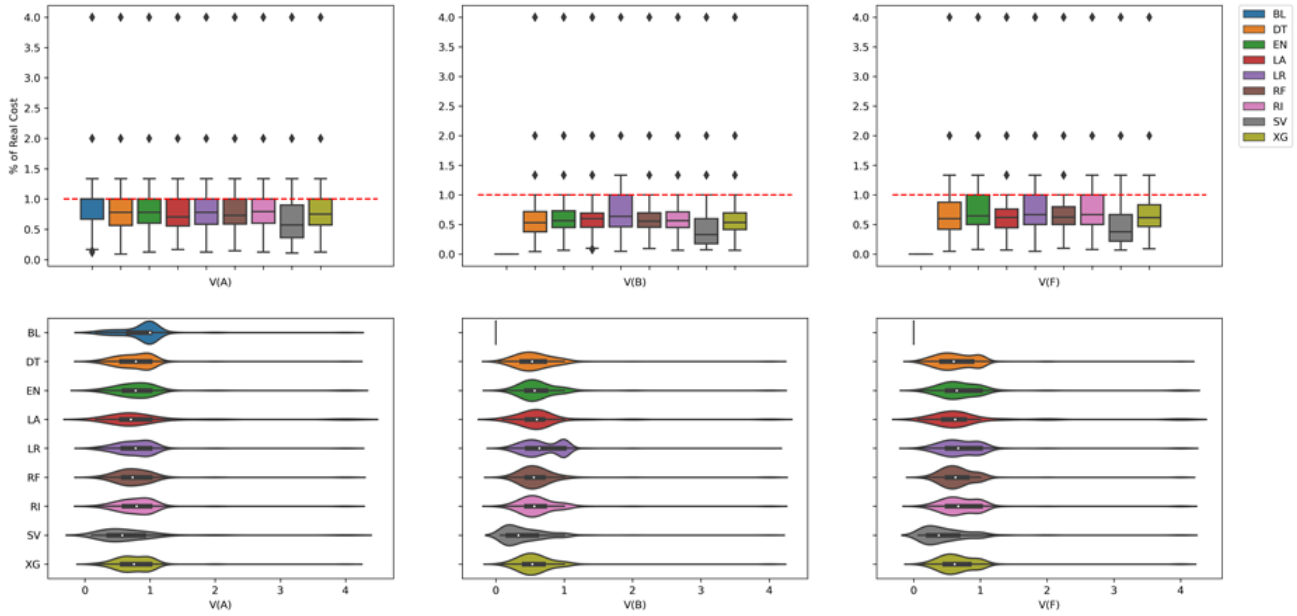


Figure 9.4: Cost Estimation for the difference between the prediction-based approach and current annual asymptomatic patients follow-up, for Val(A), Val(B) and Val(F) if we consider a pre follow-up period of 3 and a CDPT of 4

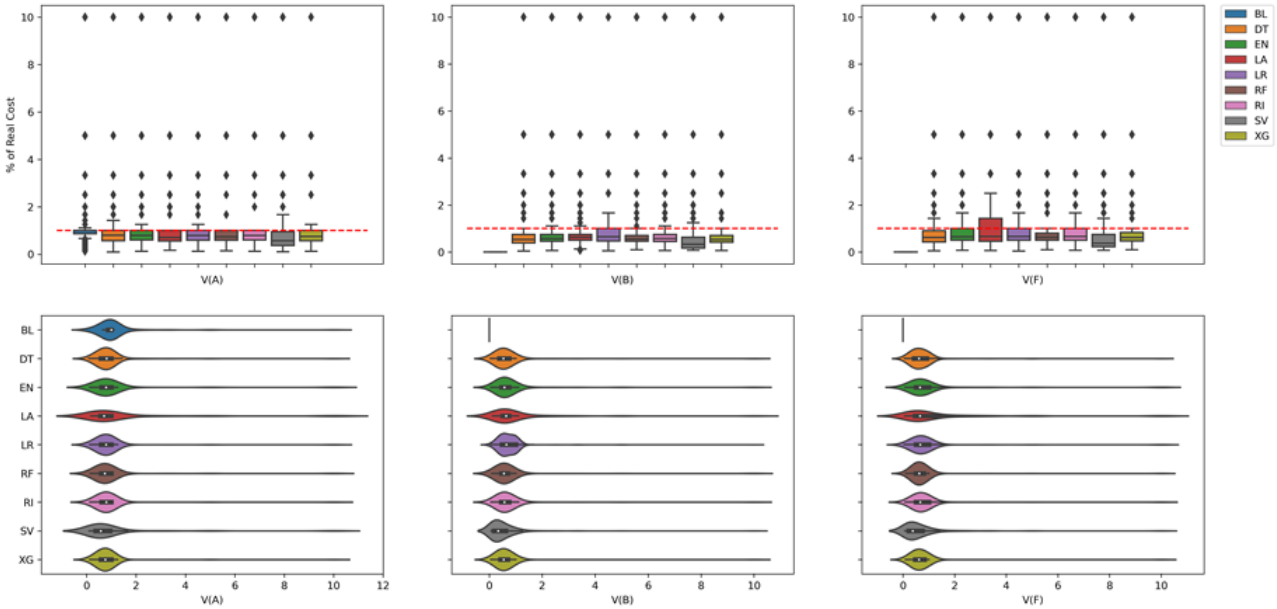


Figure 9.5: Cost Estimation for the difference between the prediction-based approach and current annual asymptomatic patients follow-up, for Val(A), Val(B) and Val(F) if we consider a pre follow-up period of 3 and a CDPT of 10

the supremacy of tree based methods: in particular, XG is ranked as top performer in most of the regression-based experiments, showing stable results when comparing training and validation results. This was expected due the match between the algorithms strengths and the intrinsic characteristics of the problem namely: the existence of missing values, complex data, and small sized available inputs. Worth noting that since some of the tried methods show some instability, from a cautionary perspective, the usage of DT is also suggested, since it was consistently classified as **worst** performer, and it is considered as a more intelligent, but still human perceptible, baseline. In practice, this method represents a simpler approach that, regardless of age group and data set, can function as a computational baseline.

As for the cost modeling results, these show the gain that can be achieved by the medical professionals in establishing a prediction based approach to follow patients diagnosed with TTR-FAP.

On a final note, this work also defined an adequate experimental approach which enables the exploration of different sets of genealogical features and their comparison, in a time-to-event prediction setting.

Chapter 10

Conclusions

This section summarizes the key findings and insights drawn from this thesis. It contributes with a reflection on the results.

10.1 Synthesis and Main Contributions

This work explores the topic of Time-to-Event prediction applied on a real medical problem. Specifically, it presents actionable results capable of impacting the job of professionals working with patients afflicted by a genetic hereditary disease known as ATTRv Amyloidosis [15]. From a clinical perspective, these results can impact the care received by patients, since ranking them according to the risk of showing symptoms can ultimately improve their quality of life by, for example, optimizing treatments afforded by health service budgets.

All the different approaches that are summarized next were substantiated on an age oriented data-driven methodology (see chapters 5 to 8), assembled to assist medical professionals tailor clinical treatments towards patient needs. From a technical perspective, this work focused mostly on the assembly of approaches based on feature construction, model construction and model evaluation (see chapter 3 for background). From a clinical perspective, it focused on extracting knowledge from clinical and genealogical data sets by generating a set of outputs capable of enhancing current medical knowledge.

10.1.1 Feature construction

With regards to feature construction, this work focused on solving a set of implementation challenges, namely on creating and evaluating a synopsis approach focused on extracting meaningful sets of features from heterogeneous data sets (genealogical trees), and comparing it with a more automated approach that takes advantage of transforming the genealogical trees using a graph embedding-based approach (see chapters 5, 6 and 7).

From a technical side, results showed it is possible to methodologically define a set of steps capable of ensuring, with a high degree of reliability, that it is feasible to extract statistically meaningful features from genealogical trees. Of note is that these characteristics, by having clear concepts behind their definition (average age of onset, maximum age of onset), can be easily grasped and interpreted by medical professionals. The same does not apply to graph or node embedding features, since these have a more purely, low level, abstract mathematical conception whose results aren't as easily relatable to observable characteristics.

Also, from a technical perspective, feature construction for genealogical data sets can look like an interesting path, but it demands taking into consideration a few key aspects. Firstly, genealogical data sets are hard to find, specially in association with the medical field. Secondly, since doctors spend a lot of time collecting and manually introducing data in heterogeneous data silos, they have a high degree of ownership over the data records. Anecdotally, it is possible to have a strong feeling that the amount of time spent collecting the data records is highly proportional to the feeling of ownership they have. Thirdly, these data sets, when being collected while resorting to undefined or unknown scenarios, may suffer from highly variable degrees of data completeness. All of this leads to a high number of problems which need assistance from business owners (for this work, medical professionals) so that solving the problems doesn't get lost in unmanageable paths.

10.1.2 Model construction

As for model construction, this work presents a multivariate data-oriented regression methodology assembled with the purpose of performing time-to-event prediction, specially focused on resorting to genealogical data. In this case, the approach tracks patients' risk of developing disease onset over different ages. Also on the topic of model construction, it explores the impact of combining different age models of near time windows (see chapters 8 and 9).

In the end, results show both a clear improvement over baseline results, and an increase on their interpretability (how results are accepted by non-technical users) - since the blending mechanism is able to reduce the effect of the heterogeneity of patient clinical data, which in turn assists in making the results more acceptable to non-technical users.

Notwithstanding the above, there are some caveats when evaluating results. In the first place, the fact that the data sets used are focused on a single medical facility can lead to biased results. Secondly, since the assembled approach focused on combining the models over time windows, and not on blending different models (algorithms), it wasn't (yet) possible to take advantage of the special characteristics of each individual base learner.

10.1.3 Model evaluation

With regards to model evaluation, the main contribution deals with the development of an approach to measure the return of investment (ROI) of triggering patient clinical follow-up based on the output of a prediction model (see chapter 9). The results show it is possible to reduce the costs if

the patients are followed on a regular basis while in the asymptomatic stage. If this is not the case, possibly due to the need to reduce overall costs, or due to a shift of concept for the distribution of age of onset over the years, then the improvement is reduced.

There is also a methodological contribution given by the different robust approaches that measure the implications of the prediction methodology, specially when compared with the clinical baseline. Here results tend to assure the models have a high degree of predictive accuracy on unseen data, and they also provide statistical assistance in model selection.

In the end, this work also addressed the lack of knowledge that clinical teams face while working with non-integrated data sets by presenting the results of a descriptive and analytical longitudinal study capable of showing the evolution of genealogical based indicators over time. This study is presented in chapter 4 and annex A, and it is published in a more current version in [82].

10.2 Discussion

Time-to-Event prediction faces significant challenges, specially when applied to medical data. A focus on genealogical data, even though can be seen as interesting from an academic perspective, issues several hardships that can be difficult to overcome.

A few of the most important issues that were dealt with here concerned on how to obtain computationally usable information from raw genealogical data: from data validation, through data transformation, to result interpretation.

It is also important to mention that a good prediction approach depends on, and thus must first evaluate, the type of data, its quality characteristics, and accurate clinical adoption. These can not simply rest upon computational or AI based improvements.

As for future work directions, one should start by noting that this work has contributions with different degrees of complexity. As such, it can have continuation in several directions, with a few of the most relevant ones being discussed next.

10.2.1 Recommendations for future work directions

This study concentrated on predictions at specific time points. Given the heterogeneity inherent in ATTRv Amyloidosis, it could be beneficial to develop models that focus on predicting in time intervals, specifically using conformal prediction approaches [99]. Such an endeavor would ideally begin with the design of a clinical trial involving a diverse group of patients from various geographical locations. This would enable the study to evaluate the impact of different variants, which could potentially lead to heightened clinical interest on this path.

Another significant aspect to consider is that the current contributions primarily focused on traditional regression methods. While this is a research path with consistent contributions, a potential extension of this work could involve comparing the methodologies developed in this thesis, specifically the multivariate regression-base prediction approach, with traditional survival analysis

approaches. This would necessitate the introduction of new genealogical-based medical datasets, which has proven to be a challenging endeavor.

On a final note, a consistent time-to-event prediction, which could be considered as an essential tool in the course of clinical trials or when dealing with the healthcare ecosystem, can assist in, for example, stratifying patients according to their risk of developing certain conditions. It can also, for example, help clinicians comprehend the effectiveness of a treatment or intervention, and lead to the adoption of new therapies and other improvements in patient care. These approaches help to show ways for this work to lead to different follow-up work directions independent from prediction results, which can assist in gathering interest from clinical professionals.

Appendix A

Chapter 4: Study on the heterogeneity and variability of Portuguese families diagnosed with ATTRv Amyloidosis

Hereditary transthyretin amyloidosis, clinical named ATTRv amyloidosis, is an inherited disease with significant variability, where the study of family history holds importance in baseline evaluations and clinical follow-up.

This annex supplies a set of tables that present genealogical-based clinical indicators that demonstrate profile changes that occurred in the families that have been target of clinical follow-up at Unidade Corino de Andrade, Centro Hospitalar Universitário de Santo António, since the first patient was diagnosed in the "outpatients facility of Hospital Geral de Santo António" [8].

The initial format of this study is incorporated in this thesis in chapter 4, but since the first submission it was accepted and published in Amyloid, The Journal of Protein Folding Disorders. This is a medical journal which consistently publishes articles on amyloid based diseases and the full work can be seen in [82].

Table A.1: Extended familiar evolution for newly diagnosed families. Bold results show the variables which show, on average, a smaller genealogical variability. Asterisk (*) indicates a significant difference with Q4 (due to Mann-Whitney having a p-value result below 0.05 which shows there is a significant difference in age of onset). Only significant differences are shown.

new families		**Q1**	**Q2**	**Q3**	**Q4**
total families with active patients		366	213	97	129
early onset families		316	142	46	44
male / female patients		417 / 285	118 / 103	31 / 40	27 / 26
largest AOO difference	average	8.7*	7.5	8.1	3.7
	stdev	6.3	6	6.6	1.7
	coef.var	0.7	0.8	0.8	0.5
	median	8	6	6	4
	range	[0,27]	[0,24]	[2,22]	[1,6]
	statistic	814			
	p-value	0.038			
largest AOO difference proband-patient	average	7.6	6.1	5.3	**3.4**
	stdev	5.5	4.7	2.4	1.7
	coef.var	0.7	0.8	0.5	0.5
	median	7	5	5.5	4
	range	[0,27]	[0,24]	[2,9]	[1,6]
largest AOO difference parent-child	average	9.1	18.3	22.5	
	stdev	6.2	6	7.5	
	coef.var	0.7	0.3	0.3	-
	median	8	21	22.5	
	range	[0,30]	[10,24]	[15,30]	
largest AOO difference siblings	average	**5.9**	**5.4**	**5**	3.6
	stdev	4.4	3.6	2.5	1.9
	coef.var	0.7	0.7	0.5	0.5
	median	5	4.5	4.5	4
	range	[0,22]	[1,13]	[2,9]	[1,6]
late onset families		21	44	35	64
male / female patients		16 / 6	26 / 21	21 / 15	57 / 22
largest AOO difference	average	**6**	**9**	**6.5**	11.4
	stdev	0	0	5.5	4.7
	coef.var	0	0	0.8	0.4
	median	6	9	6.5	10.5
	range	-	-	[1,12]	[3,18]
largest AOO difference proband-patient	average	**6**	**9**	**6.5**	9.6
	stdev	0	0	5.5	3.9
	coef.var	0	0	0.8	0.4
	median	6	9	6.5	9
	range	-	-	[1,12]	[3,18]
largest AOO difference parent-child	average				
	stdev				
	coef.var	-	-	-	-
	median				
	range				
largest AOO difference siblings	average	10	**9**		**8.6**
	stdev	4	0		3.2
	coef.var	0.4	0	-	0.4
	median	10	9		9
	range	[6,14]	-		[3,13]

Table A.2: Extended familiar evolution over time for newly diagnosed mixed onset families. Bold results show the variables which show, on average, a smaller genealogical variability.

new families		**Q1**	**Q2**	**Q3**	**Q4**
total families with active patients		366	213	97	129
mix onset families		17	21	14	12
male / female patients		10 / 10	11 / 16	4 / 12	8 / 8
largest AOO difference	average	22.6	26.4	25.9	19.1
	stdev	11	8.5	9.9	10.4
	coef.var	0.5	0.3	0.4	0.5
	median	21	26	28.5	17
	range	[2,44]	[9,43]	[11.40]	[5,34]
largest AOO difference proband-patient	average	18.6	26.4	25.3	20.2
	stdev	8	9.5	8.9	7.2
	coef.var	0.4	0.4	0.4	0.4
	median	19	26	25	19
	range	[2,35]	[9,43]	[11,40]	[13,32]
largest AOO difference parent-child	average	22.1	24.9	30.4	27.8
	stdev	6.4	7	5.7	6.4
	coef.var	0.3	0.3	0.2	0.2
	median	21	24	30	32
	range	[11,35]	[15,43]	[22,40]	[17,34]
largest AOO difference siblings	average	**10.4**	**11.9**	**9.7**	**13.5**
	stdev	7.2	10.7	6.6	4.4
	coef.var	0.7	0.9	0.7	0.3
	median	8	8.5	11	15.5
	range	[1,21]	[2,38]	[1,17]	[4,17]

Table A.3: Extended familiar evolution over time for old diagnosed families. Bold results show the variables which show, on average, a smaller genealogical variability. Asterisk (*) indicates a significant difference with Q4 (due to Mann-Whitney having a p-value result below 0.05 which shows there is a significant difference in age of onset). Only significant differences are shown.

old families (OMR / NMR)		**Q2**	**Q3**	**Q4**
total families with active patients		205	266	291
early onset families		188	223	221
male / female patients		225 / 197	255 / 228	183 / 195
largest AOO difference	average	7.9 / 10.9	7.2 / 12.6	8.1 / 13.4
	stdev	4.9 / 5.9	4.9 / 5.7	5.5 / 5.8
	coef.var	0.6 / 0.5	0.7 / 0.5	0.7 / 0.4
	median	7 / 10	6 / 12	7 / 13
	range	[0,24] / [0,27]	[0,23] / [2,27]	[0,22] / [0,27]
largest AOO difference proband-patient	average	- / 10.5*	- / 12.4	1 / 13.6
	stdev	- / 7	- / 7.5	0 / 7.8
	coef.var	- / 0.7	- / 0.6	0 / 0.6
	median	- / 9	- / 11	1 / 12
	range	- / [0,40]	- / [2,39]	- / [0,40]
	statistic	/ 14168		
	p-value	/ < 0.001		
largest AOO difference parent-child	average	15.7 / 9.9	18.8 / 10.9	- / 12
	stdev	8.2 / 7.5	7 / 7.2	- / 7.4
	coef.var	0.5 / 0.8	0.4 / 0.7	- / 0.6
	median	14 / 8	16.5 / 10	- / 10
	range	[5,33] / [0,35]	[10,31] / [0,40]	- / [0,35]
largest AOO difference siblings	average	**5.4/ 7.6***	**4.9 / 4.9**	**5.2 / 5.2**
	stdev	3.7 / 4.9	3.5 / 3.5	3.7 / 3.7
	coef.var	0.7 / 0.6	0.7 / 0.7	0.7 / 0.7
	median	5 / 7	4 / 4	5 / 5
	range	[0,16] / [0,31]	[0,16] / [0,16]	[0,15] / [0,15]
	statistic	/ 3112		
	p-value	/ 0.008		
late onset families		9	18	24
male / female patients		6 / 6	9 / 12	11 / 18
largest AOO difference	average	**13 / 11.5**	**6.5 /7.9**	8.3 / **10.4**
	stdev	0 / 2.6	1.5 / 5.2	5.8 / 6.2
	coef.var	0 / 0.2	0.2 / 0.7	0.7 / 0.6
	median	13 / 13	6 / 7	6 / 9
	range	- / [7,13]	[5,9] / [1,15]	[3,18] / [1,22]
largest AOO difference proband-patient	average	- / 20	- / 18.6	- / 20.3
	stdev	- / 7.4	- / 14.4	- / 11.3
	coef.var	- / 0.4	- / 0.8	- / 0.6
	median	- / 24	- / 15	- / 16
	range	- / [7,27]	- / [1,49]	- / [5, 47]
largest AOO difference parent-child	average	- / 12	- / 15.3	- / 15.9
	stdev	- / 6.7	- / 8.8	- / 11.7
	coef.var	- / 0.6	- / 0.6	- / 0.7
	median	- / 10	- / 15	- / 14
	range	- / [5,21]	- / [2,28]	- / [3,47]
largest AOO difference siblings	average	- / 16.2	- / 14.1	**4** / 18.4
	stdev	- / 8.2	- / 9.2	1 / 11.7
	coef.var	- / 0.5	- / 0.7	0.3 / 0.6
	median	- / 16.5	- / 13	4 / 16
	range	- / [7,25]	- / [1,32]	[3,5] / [1,40]

Table A.4: Extended familiar evolution over time for old diagnosed families. Bold results show the variables which show, on average, a smaller genealogical variability.

old families (OMR / NMR)		**Q2**	**Q3**	**Q4**
total families with active patients		205	266	291
mix onset families		8	25	43
male / female patients		4 / 9	13 / 25	25 / 30
largest AOO difference	average	34 / 35.1	24.2 / 33.8	24.6 / 33.1
	stdev	6.6 / 5.9	12.3 / 8.4	10.2 / 8.7
	coef.var	0.2 / 0.2	0.5 / 0.2	0.4 / 0.3
	median	33.5 / 34.5	25 / 33.5	24 / 34.5
	range	[26,47] / [26,47]	[4,43] / [10,46]	[3,47] / [17,49]
largest AOO difference proband-patient	average	- / 27	- / 26.6	- / 23.8
	stdev	- / 4.5	- / 8.9	- / 7.7
	coef.var	- / 0.2	- / 0.3	- / 0.3
	median	- / 28	- / 26	- / 25
	range	- / [21,32]	- / [5,43]	- / [10,43]
largest AOO difference parent-child	average	24.7 / 22.9	29 / 24.3	21.4 / 20.9
	stdev	3.4 / 8.4	2.8 / 9.8	4.4 / 10.3
	coef.var	0.1 / 0.4	0.1 / 0.4	0.2 / 0.5
	median	25.5 / 25.5	29 / 25	20 / 20
	range	[19,29] / [3,32]	[25,33] / [6,43]	[16,28] / [0,43]
largest AOO difference siblings	average	**8.8 / 15.1**	**8.9 / 20.2**	**6.8 / 17.5**
	stdev	6 / 8.7	4.7 / 8.4	5.7 / 9.1
	coef.var	0.7 / 0.6	0.5 / 0.4	0.8 / 0.5
	median	5 / 20	9 / 19	3.5 / 16
	range	[2,17] / [3,26]	[2,19] / [2,38]	[1,20] / [3,38]

References

[1] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[2] David Adams, Haruki Koike, Michel Slama, and Teresa Coelho. Hereditary transthyretin amyloidosis: a model of medical progress for a fatal disease. *Nature Reviews Neurology*, 15(7):387–404, 7 2019.

[3] David Adams, Marie Théaudin, Pierre Lozeron, Clovis Adam, Guillemette Beaudonnet, Céline Labeyrie, Catherine Lacroix, and Cécile Cauquil. Management of stage 1 TTR FAP: French experience. *Orphanet Journal of Rare Diseases*, 10(S1), 12 2015.

[4] Flora Alarcon, Violaine Planté-Bordeneuve, and Grégory Nuel. Study of the parent-of-origin effect in monogenic diseases with variable age of onset. Application on ATTRv. *PLoS ONE*, 18(8 August), 8 2023.

[5] Douglas G Altman and J Martin Bland. Time to event (survival) data. *British Medical Journal (BMJ)*, 317, 1998.

[6] Per Kraghand Andersen, Ørnulfand Borgan, Richard D. Gill, and Niels Keiding. Introduction. In *Statistical Models Based on Counting Processes*, pages 1–44. Springer US, New York, NY, 1993.

[7] Yukio Ando, Teresa Coelho, John L Berk, Márcia Waddington Cruz, Bo-Göran Ericzon, Shu-ichi Ikeda, W David Lewis, Laura Obici, Violaine Planté-Bordeneuve, Claudio Rapezzi, Gerard Said, and Fabrizio Salvi. Guideline of transthyretin-related hereditary amyloidosis for clinicians. *Orphanet Journal of Rare Diseases*, 2013.

[8] Corino Andrade. A peculiar form of peripheral neuropathy; familiar atypical generalized amyloidosis with special involvement of the peripheral nerves. *Brain: a journal of neurology*, 75(3):408–27, 1952.

[9] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[10] Vladimir Batagelj and Andrej Mrvar. Analysis of kinship relations with Pajek. *Social Science Computer Review*, 26:224–246, 2008.

[11] Danielle Beaulieu, James D. Berry, Sabrina Paganoni, Jonathan D. Glass, Christina Fournier, Jonavelle Cuerdo, Mark Schactman, and David L. Ennist. Development and validation of a machine-learning ALS survival model lacking vital capacity (VC-Free) for use

in clinical trials during the COVID-19 pandemic. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 22(S1):22–32, 2021.

[12] Michael A Bender, Martín Farach-Colton, Giridhar Pemmasani, Steven Skiena, and Pavel Sumazin. Lowest Common Ancestors in Trees and Directed Acyclic Graphs 1. *Journal of Algorithms*, 57(2):75–94, 2005.

[13] Abdulbari Bener, Frss Ffph, Elnour E Dafeeah, and Nancy Samson. Does consanguinity increase the risk of schizophrenia? Study based on primary health care centre visits. *Mental Health in Family Medicine*, 2012.

[14] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 8 2013.

[15] Merrill D. Benson, Joel N. Buxbaum, David S. Eisenberg, Giampaolo Merlini, Maria J.M. Saraiva, Yoshiki Sekijima, Jean D. Sipe, and Per Westermark. Amyloid nomenclature 2020: update and recommendations by the International Society of Amyloidosis (ISA) nomenclature committee. *Amyloid*, 27(4):217–222, 10 2020.

[16] Elfriede Bollschweiler. Benefits and limitations of Kaplan-Meier calculations of survival chance in cancer surgery. In *Langenbeck's Archives of Surgery*, volume 388, pages 239–244, 9 2003.

[17] C J Brackenridge and Betty Teltscher. Estimation of the age at onset of Huntington's disease from factors associated with the affected parent. *Journal of Medical Genetics*, 12(64), 1975.

[18] L Breiman, Jerome H Friedman, Richard A Olshen, and C J Stone. *Classification and Regression Trees*. Taylor & Francis, 1984.

[19] Leo Breiman. Random Forests. *Machine Learning*, 45:5–32, 10 2001.

[20] Maria E. Briseño-Godínez, Karla Cárdenas-Soto, Rosa X. Domínguez-Vega, and Alejandra González-Duarte. What a neurologist must know of hereditary ATTRv amyloidosis. *Revista Mexicana de Neurociencia*, 24(2), 2 2023.

[21] Gavin C Cawley and Nicola L C Talbot. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

[22] Haochen Chen, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. A Tutorial on Network Embeddings. *arXiv preprint arXiv:1808.02590*, 8 2018.

[23] Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 785–794, New York, New York, USA, 2016. ACM Press.

[24] Ching-Fan Chung, Peter Schmidt, and Ana D. Witte. Survival analysis: A survey. *Journal of Quantitative Criminology*, 7(1):pp 59–98, 1991.

[25] Eugenia Cisneros-Barroso, Juan González-Moreno, Adrian Rodríguez, Tomas Ripoll-Vera, Jorge Álvarez, Mercedes Usón, Antonio Figuerola, Cristina Descals, Carles Montalá, Maria Asunción Ferrer-Nadal, and Ines Losada. Anticipation on age at onset in kindreds with hereditary ATTRV30M amyloidosis from the Majorcan cluster. *Amyloid*, pages 254–258, 2020.

[26] Maria Teresa Pardal Monteiro Coelho. *Disease Modifying Therapies for ATTR Amyloidosis: Clinical Development of New Drugs and Impact on the Natural History of the Disease*. PhD thesis, University of Porto, Porto, 2019.

[27] Isabel Conceição, Thibaud Damy, Manuel Romero, Lucía Galán, and et al. Early diagnosis of ATTR amyloidosis through targeted follow-up of identified carriers of TTR gene mutations*. *Amyloid*, 26(1):3–9, 1 2019.

[28] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 9 1995.

[29] Paula Coutinho. *Doença de Machado-Joseph : Tentativa de definição*. PhD thesis, Universidade do Porto, Porto, 1992.

[30] Abhijit Dasgupta, Yan V. Sun, Inke R. König, Joan E. Bailey-Wilson, and James D. Malley. Brief review of regression-based and machine learning methods in genetic epidemiology: The Genetic Analysis Workshop 17 experience. *Genetic Epidemiology*, 2011.

[31] Janez Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[32] Angela Dispenzieri, Teresa Coelho, Isabel Conceição, Márcia Waddington-Cruz, Jonas Wixner, Arnt V. Kristen, Claudio Rapezzi, Violaine Planté-Bordeneuve, Juan Gonzalez-Moreno, Mathew S. Maurer, Martha Grogan, Doug Chapman, Leslie Amass, Pablo Garcia Pavia, Ivaylo Tarnev, Jose Gonzalez Costello, Maria Alejandra Gonzalez Duarte Briseno, Hartmut Schmidt, Brian Drachman, Fabio Adrian Barroso, Taro Yamashita, Olivier Lairez, Yoshiki Sekijima, Giuseppe Vita, Eun Seok Jeon, Mazen Hanna, David Slosky, Marco Luigetti, Samantha LoRusso, Francisco Munoz Beamud, David Adams, and et al. Clinical and genetic profile of patients enrolled in the Transthyretin Amyloidosis Outcomes Survey (THAOS): 14-year update. *Orphanet Journal of Rare Diseases*, 17(1), 12 2022.

[33] Maike F. Dohrn, Michaela Auer-Grumbach, Ralf Baron, Frank Birklein, Fabiola Escolano-Lozano, Christian Geber, Nicolai Grether, Tim Hagenacker, Ernst Hund, Juliane Sachau, Matthias Schilling, Jens Schmidt, Wilhelm Schulte-Mattler, Claudia Sommer, Markus Weiler, Gilbert Wunderlich, and Katrin Hahn. Chance or challenge, spoilt for choice? New recommendations on diagnostic and therapeutic considerations in hereditary transthyretin amyloidosis with polyneuropathy: the German/Austrian position and review of the literature. *Journal of Neurology*, 268(10):3610–3625, 10 2021.

[34] U. Drugge, R. Andersson, F. Chizari, M. Danielsson, G. Holmgren, O. Sandgren, and A. Sousa. Familial amyloidotic polyneuropathy in Sweden: A pedigree analysis. *Journal of Medical Genetics*, 30(5):388–392, 1993.

[35] Duc Duy Nguyen, Christoph Lohrmann, and Pasi Luukka. A Comparison of Feature Construction Methods in the Context of Supervised Feature Selection for Classification. 1969.

[36] Julia Efremova. *Mining social structures from genealogical data*. PhD thesis, 2016.

[37] Flora Alarcon Gregory Nuel Intissar Anan Farida Gorram Malin Olsson and Violaine Planté-Bordeneuve. New data on the genetic profile and penetrance of hereditary Val30Met transthyretin amyloidosis in Sweden. *Amyloid*, 28(2):84–90, 2021.

[38] Michael Fire and Yuval Elovici. Data Mining of Online Genealogy Datasets for Revealing Lifespan Patterns in Human Population. *ACM Transactions on Intelligent Systems and Technology*, 2015.

[39] Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001.

[40] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[41] Judith E Friedman. Anticipation in hereditary disease: the history of a biomedical concept. *Human Genetics*, 130(6):705–714, 2011.

[42] João Gama, André Ponce de Leon Carvalho, Katti Faceli, Ana Carolina Lorena, and Márcia Oliveira. *Extração de Conhecimento de Dados*. 3 edition, 2017.

[43] Yan Gao and Yan Cui. Clinical time-to-event prediction enhanced by incorporating compatible related outcomes. *PLOS Digital Health*, 1(5):1–10, 5 2022.

[44] Zheng Gao, Gang Fu, Chunping Ouyang, Satoshi Tsutsui, Xiaozhong Liu, Jeremy Yang, Christopher Gessner, Brian Foote, David Wild, Ying Ding, and Qi Yu. Edge2vec: Representation learning using edge semantics for biomedical knowledge discovery. *BMC Bioinformatics*, 20(1), 6 2019.

[45] Héloïse Gauvin, Jean-François Lefebvre, Claudia Moreau, Eve-Marie Lavoie, Damian Labuda, Hélène Vézina, and Marie-Hélène Roy-Gagnon. GENLIB: an R package for the analysis of genealogical data. *BMC Bioinformatics*, 16(1):160, 5 2015.

[46] Pablo Gay, Beatriz López, Albert Plà, Jordi Saperas, and Carles Pous. Enabling the use of hereditary information from pedigree tools in medical knowledge-based systems. *Journal of Biomedical Informatics*, 2013.

[47] Alexander E. Gorbalenya, Susan C. Baker, Ralph S. Baric, Raoul J. de Groot, Christian Drosten, Anastasia A. Gulyaeva, Bart L. Haagmans, Chris Lauber, Andrey M. Leontovich, Benjamin W. Neuman, Dmitry Penzar, Stanley Perlman, Leo L.M. Poon, Dmitry V. Samborskiy, Igor A. Sidorov, Isabel Sola, and John Ziebuhr. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2, 4 2020.

[48] Palash Goyal and Emilio Ferrara. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78 – 94, 2018.

[49] Rolf H.H. Groenwold, Ian R. White, A. Rogier T. Donders, James R. Carpenter, Douglas G. Altman, and Karel G.M. Moons. Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis. *CMAJ*, 2012.

[50] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-August-2016, pages 855–864. Association for Computing Machinery, 8 2016.

[51] Hartmut H-j Schmidt, Fabio Barroso, Alejandra Gonz Alez-duarte, Isabel Conceiç, Laura Obici, Denis Keohane, and Leslie Amass. Invited review management of Asymptomatic Gene Carriers of Transthyretin Familial Amyloid Polyneuropathy. Technical report, 2016.

[52] Frank E Harrell. Cox Proportional Hazards Regression Model. In *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, pages 465–507. Springer New York, New York, NY, 2001.

[53] Frank E. Harrell. *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer; Corrected edition (January 10, 2001), 2001.

[54] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Springer New York, 2009.

[55] Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 1970.

[56] M Inês, T Coelho, I Conceicao, M Borges, M Carvalho, and J Costa. Costs Associated With Transthyretin Familial Amyloid Polyneuropathy Progression. *Value in Health*, 20(9):A553–A554, 10 2017.

[57] M Inês, T Coelho, I Conceicao, P Saramago, M Carvalho, and J Costa. Life Expectancy And Costs of Transthyretin Familial Amyloid Polyneuropathy. *Value in Health*, 20(9):A553, 10 2017.

[58] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *Annals of Applied Statistics*, 2(3):841–860, 9 2008.

[59] I.T. Jolliffe. *Principal Component Analysis*. Springer New York, NY, 2002.

[60] David L. Katz, Rachna Govani, Kieran Anderson, Lauren Q. Rhee, and Dina L. Aronson. The Financial Case for Food as Medicine: Introduction of a ROI Calculator. *American Journal of Health Promotion*, 36(5), 2022.

[61] David G. Kleinbaum and Mitchel Klein. Survival Analysis, A Self-Learning Text, Third Edition. In *Survival Analysis: A Self-Learning Text*, pages 289–361. Springer New York, New York, NY, 2012.

[62] Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 1995.

[63] Haruki Koike, Ken-ichiro Misu, Shu-ichi Ikeda, Yukio Ando, Masamitsu Nakazato, Eiko Ando, Masahiko Yamamoto, Naoki Hattori, and Gen Sobue. Type I (transthyretin Met30) familial amyloid polyneuropathy in Japan: early- vs late-onset form. *Archives of neurology*, 59 11:1771–6, 2002.

[64] Carolina Lemos, Teresa Coelho, Miguel Alves-Ferreira, Ana Martins-Da-Silva, Jorge Sequeiros, Denisa Mendonça, and Alda Sousa. Overcoming artefact: anticipation in 284 Portuguese kindreds with familial amyloid polyneuropathy (FAP) ATTRV30M. *J Neurol Neurosurg Psychiatry*, 2014.

[65] Sébastien Lerique, Jacob Levy Abitbol, and Márton Karsai. Joint embedding of structure and features via graph convolutional networks. *Applied Network Science*, 5(1), 12 2020.

[66] Luísa Lobato and Ana Rocha. Transthyretin amyloidosis and the kidney, 8 2012.

[67] Paulo José Lorenzoni, Vinicius Riegel Giugno, Renata Dal Prá Ducci, Lineu Cesar Werneck, Claudia Suemi Kamoi Kay, and Rosana Herminia Scola. Genetic screening for transthyretin familial amyloid polyneuropathy to avoid misdiagnosis in patients with polyneuropathy associated with high protein in the cerebrospinal fluid, 8 2023.

[68] Marzvanyan A and Alhawaj AF. Physiology, Sensory Receptors, 8 2023.

[69] Maria Filomena Mendes. *Como Nascem e Morrem os Portugueses — nascimentos, natalidade, fecundidade, óbitos, mortalidade, causas de morte*. Fundação Francisco Manuel dos Santos, 2 2020.

[70] João Mendes-Moreira, Carlos Soares, Alípio Mário Jorge, and Jorge Freire De Sousa. Ensemble approaches for regression. *ACM Computing Surveys*, 2012.

[71] Melinda Mills. *Introducing Survival and Event History Analysis*. SAGE Publications Ltd, december 21, 2010 edition, 2014.

[72] António Rodrigues Morais. *Perseguindo um gene... Que caminhos?* PhD thesis, Minho, 2004.

[73] Carla Leal Moreira, Ana Rocha, Josefina Santos, Marta Santos, Idalina Beirão, Teresa Coelho, and Luísa Lobato. The ever-growing understanding of transthyretin amyloidosis nephropathy. *Amyloid*, 24(sup1):117–118, 2017.

[74] Hiroshi Motoda and Huan Liu. Feature Selection, Extraction and Construction. *Communication of IICM (Institute of Information and Computing Machinery, Taiwan)*, pages 67–72, 2002.

[75] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswa. graph2vec: Learning Distributed Representations of Graphs. 2017.

[76] Alice Nevone, Giampaolo Merlini, and Mario Nuvolone. Treating Protein Misfolding Diseases: Therapeutic Successes Against Systemic Amyloidoses, 7 2020.

[77] Todd G Nick and J Michael Hardin. Regression Modeling Strategies: An Illustrative Case Study From Medical Rehabilitation Outcomes Research. *The American Journal of Occupational Therapy*, 1998.

[78] Laura Obici, Jan B. Kuks, Juan Buades, David Adams, Ole B. Suhr, Teresa Coelho, and Theodore Kyriakides. Recommendations for presymptomatic genetic testing and management of individuals at risk for hereditary transthyretin amyloidosis. *Current Opinion in Neurology*, 29:S27–S35, 2016.

[79] Alexandra Alves Oliveira, Rita Ana Gaio, and Luís Paulo Reis. *Modelling Reporting Delays for Infectious Diseases: An Application to the Portuguese HIV-AIDS Surveillance Data*. PhD thesis, University of Porto, 2018.

[80] Oxford. The Oxford Dictionary of Phrase and Fable.

[81] Yesim Parman, David Adams, Laura Obici, Lucía Galán, Velina Guergueltcheva, Ole B. Suhr, and Teresa Coelho. Sixty years of transthyretin familial amyloid polyneuropathy (TTR-FAP) in Europe. *Current Opinion in Neurology*, 2016.

[82] Maria Pedroto, Teresa Coelho, Joana Fernandes, Alexandra Oliveira, Alípio Jorge, and João Mendes-Moreira. Heterogeneity in families with ATTRV30M amyloidosis: a historical and longitudinal Portuguese case study impact for genetic counselling. *Amyloid*, 0(0):1–11, 2024.

[83] Maria Pedroto, Teresa Coelho, Alípio Jorge, and João Mendes-Moreira. Clinical model for Hereditary Transthyretin Amyloidosis age of onset prediction. *Frontiers in Neurology*, 14, 2023.

[84] Maria Pedroto, Alípio Jorge, João Mendes-Moreira, and Teresa Coelho. Predicting Age of Onset in TTR-FAP Patients with Genealogical Features. In Jaakko Hollmén, Carolyn McGregor, Paolo Soda, and Bridget Kane, editors, *31st IEEE International Symposium on Computer-Based Medical Systems,CBMS 2018, Karlstad, Sweden, June 18-21, 2018*, pages 199–204. IEEE Computer Society, 2018.

[85] Maria Pedroto, Alípio Jorge, João Mendes-Moreira, and Teresa Coelho. Impact of Genealogical Features in Transthyretin Familial Amyloid Polyneuropathy Age of Onset Prediction. In Fdez-Riverola Florentino, Mohd Saberi, Mohamad, and Rocha Miguel, and De Paz Juan F., and and González Pascual, editors, *Practical Applications of Computational Biology and Bioinformatics, 12th International Conference*, pages 35–42. Springer International Publishing, 2019.

[86] Maria Pedroto, Alípio Jorge, João Mendes-Moreira, and Teresa Coelho. Improving the prediction of Age of Onset of TTR-FAP patients using graph-embeddings. In *Progress in Artificial Intelligence: 21st EPIA Conference on Artificial Intelligence, EPIA 2022*, pages 183–194, 2022.

[87] Maria Pedroto, Alípio Jorge, João Mendes-Moreira, and Teresa Coelho. Combining neighbor models to improve predictions for ATTRv carriers. 2023.

[88] Christian. Petersohn. *Temporal video segmentation*. Jörg Vogt Verlag, 2010.

[89] Violaine Planté-Bordeneuve, Farida Gorram, Malin Olsson, Intissar Anan, Anna Mazzeo, Luca Gentile, Eugenia Cisneros-Barroso, Juan Gonzalez-Moreno, Ines Losada, Marcia Waddington-Cruz, Luiz Felipe Pinto, Yeşim Parman, Pascale Fanen, Flora Alarcon, and Gregory Nuel. A multicentric study of the disease risks and first manifestations in hereditary transthyretin amyloidosis (ATTRv): insights for an earlier diagnosis. *Amyloid*, 30(3):313–320, 7 2023.

[90] Sarah Pungitore and Vignesh Subbian. Assessment of Prediction Tasks and Time Window Selection in Temporal Modeling of Electronic Health Record Data: a Systematic Review. *Journal of Healthcare Informatics Research*, 7(3):313–331, 9 2023.

[91] Enislay Ramentol, Yailé Caballero, Rafael Bello, and Francisco Herrera. SMOTE-RSB *: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. *Knowledge and Information Systems*, 33(2):245–265, 11 2012.

[92] J. Ranstam and J. A. Cook. Kaplan–Meier curve, 3 2017.

[93] Debashish Roy, Rajeev Srivastava, Mansi Jat, and Mustafa Said Karaca. A Complete Overview of Analytics Techniques: Descriptive, Predictive, and Prescriptive. In

EAI/Springer Innovations in Communication and Computing, pages 15–30. Springer Science and Business Media Deutschland GmbH, 2022.

[94] Amelie Royer and Christoph H Lampert. Classifier Adaptation at Prediction Time. 2015.

[95] Benedek Rozemberczki, Oliver Kiss, and Rik Sarkar. Karate Club: An API Oriented Open-Source Python Framework for Unsupervised Learning on Graphs. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 3125–3132, New York, NY, USA, 2020. Association for Computing Machinery.

[96] Davis C. Ryman, Natalia Acosta-Baena, Paul S. Aisen, Thomas Bird, Adrian Danek, Nick C. Fox, Alison Goate, Peter Frommelt, Bernardino Ghetti, Jessica B.S. Langbaum, Francisco Lopera, Ralph Martins, Colin L. Masters, Richard P. Mayeux, Eric McDade, Sonia Moreno, Eric M. Reiman, John M. Ringman, Steve Salloway, Peter R. Schofield, Reisa Sperling, Pierre N. Tariot, Chengjie Xiong, John C. Morris, and Randall J. Bateman. Symptom onset in autosomal dominant Alzheimer disease: A systematic review and meta-analysis. *Neurology*, 2014.

[97] Yvan Saeys, I. Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. In *Bioinformatics*, 2007.

[98] Gerard Said and Violaine Planté-Bordeneuve. Familial amyloid polyneuropathy. Technical report, 2011.

[99] Glenn Shafer and Vladimir Vovk. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

[100] Alex J Smola and Bernhard Sc. A Tutorial on Support Vector Regression *. *Statistics and Computing*, 14(3):199–222, 2004.

[101] Saúl Solorio-Fernández, J Ariel Carrasco-Ochoa, and José Francisco Martínez-Trinidad. A survey on feature selection methods for mixed data. *Artificial Intelligence Review*, 55(4):2821–2846, 2022.

[102] Alda Maria Botelho Correia Sousa. *A variabilidade fenotípica da Polineuropatia Amiloidótica Familiar: um estudo de genética quantitativa em Portugal e na Suécia (Phenotypic variability of Familial Amyloid Polyneuropathy: a quantitative genetics study in Portugal and in Sweden)*. PhD thesis, University of Porto, 1995.

[103] E. W. Steyerberg. Frank E. Harrell, Jr., Regression Modeling Strategies: With Applications, to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd ed. Heidelberg: Springer., 2016.

[104] Ewout W Steyerberg. *Statistics for Biology and Health Clinical Prediction Models A Practical Approach to Development, Validation, and Updating Second Edition*. Second edition edition, 2019.

[105] Y. Sun, L. Garcia-Pueyo, J. B. Wendt, M. Najork, and A. Broder. Learning Effective Embeddings for Machine Generated Emails with Applications to Email Category Prediction. *IEEE International Conference on Big Data (Big Data)*, pages 1846–1855, 2018.

[106] Kazuhiro Tashima, Yukio Ando, Yoshiya Tanaka, Makoto Uchino, and Masayuki Ando. Change in the Age of Onset in Patients with Familial Amyloidotic Polyneuropathy Type I. *The Japanese Society of Internal Medicine - Internal Medicine*, 34(8):748–750, 1995.

[107] E. Ya Tetushkin. Genetic aspects of genealogy. *Russian Journal of Genetics*, 47(11):1288–1306, 11 2011.

[108] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 1996.

[109] Robert Tibshirani. What is Cox's proportional hazards model? *Significance*, 19(2):38–39, 2022.

[110] VCV000013417.64. National Center for Biotechnology Information. ClinVar, 2023.

[111] Ping Wang, Yan Li, and Chandan K. Reddy. Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys*, 1(1), 2018.

[112] Ping Wang, Virginia Tech, Yan Li, and Chandan K Reddy. Machine Learning for Survival Analysis: A Survey. Technical report.

[113] Michaël Charles Waumans and Hugues Bersini. Genealogical trees of scientific papers. *PLoS ONE*, 2016.

[114] Waxenbaum JA, Reddy V, and Varacallo M. Anatomy, Autonomic Nervous System. *StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing*, 7 2023.

[115] Hiroyuki Yamamoto and Tomoki Yokochi. Transthyretin cardiac amyloidosis: an update on diagnosis and treatment, 12 2019.

[116] Guolei Yang, Ying Cai, and Chandan K Reddy. Spatio-Temporal Check-in Time Prediction with Recurrent Neural Network based Survival Analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 2976–298. International Joint Conferences on Artificial Intelligence Organization, 2018.

[117] Shelemyahu Zacks. System Effectiveness. In *Introduction to Reliability Analysis: Probability Models and Statistical Methods*, pages 1–12. Springer New York, New York, NY, 1992.

[118] Chongsheng Zhang, Changchang Liu, Xiangliang Zhang, and George Almpanidis. An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications*, 82:128–150, 10 2017.

[119] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. Network Representation Learning: A Survey. *IEEE Transactions on Big Data*, 6(1):3–28, 12 2020.

[120] Jianfei Zhang, Lifei Chen, Yanfang Ye, Gongde Guo, Rongbo Chen, Alain Vanasse, and Shengrui Wang. Survival neural networks for time-to-event prediction in longitudinal study. *Knowledge and Information Systems*, 62(9):3727–3751, 9 2020.

[121] Liang Zhao. Event Prediction in the Big Data Era: A Systematic Survey. *ACM Computing Surveys*, 54(5), 6 2021.

[122] Sida Zhou. Empirical effect of graph embeddings on fraud detection/ risk mitigation. Technical report.

[123] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67(2):301–320, 2005.