Clinical Decision Support Systems for Palliative Care Referral: Design and Evaluation of Frailty and Mortality Predictive Models

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Universitat Politècnica de València
Programa de Doctorado en
Tecnologías para la Salud y el Bienestar

DOCTORAL THESIS

Presented by
Vicent Blanes Selva

Directed by
Prof. Juan Miguel García Gómez
Dr. Ascensión Doñate Martínez

Valencia, Spain
September 2022

Agradecimientos
Acknowledgements

> You have to die a few times before
> you can really live
>
> Charles Bukowski

Es posible que si tuviera que enumerar todo lo que me ha ocurrido durante el transcurso de esta tesis la mayoría de las frases empezarían por *"he aprendido"*. Sería de una tremenda falsedad dar a entender que este proceso de evolución, académico y personal, ha dependido únicamente de mí y de la *meritocracia*. Nada más lejos de la realidad, el poder hoy presentar una tesis se ha parecido más a un trabajo en equipo que a otra cosa.

En primer lugar, expresar mi mas sincero agradecimiento a mis dos directores. Primero al Dr. Juan Miguel García Gómez, al que pude convencer en su momento para que me dejara entrar en su grupo de investigación, el Biomedical Data Science Lab (BDSLab), y desde entonces no he hecho más que aprender de una persona con una visión científica increíble. Y en segundo lugar, a la Dr. Ascensión Doñate Martínez, con la que tuve el placer de trabajar durante el proyecto europeo que ha supuesto el núcleo de mi tesis y que ha tenido la paciencia suficiente para, entre otras cosas, enseñarme a escribir mejor y a desenvolverme con más soltura en proyectos europeos. Sois el mejor ejemplo que podía haber tenido, en lo académico y en lo personal, y espero algún día poder llegar a vuestra altura.

Quiero dar las gracias también a mis compañeros en el BDSLab, siento que os debo mucho del crecimiento personal que he experimentado en estos cinco años. Muchas gracias al Dr Carlos Sáez, el Dr Elies Fuster, la Dr. Sabina Asensio, Pablo Ferri, Marta Durá, Ángel Sánchez, María del Mar Álvarez, Javier Salvador y Paula Balaguer, que además me ha asesorado con la *bonicor* de las figuras. Mi agradecimiento también para el Dr Jose Vicente Manjón, el Dr Salvador Tortajada y *mis dos hermanos mayores* el Dr Javier Juan y el Dr Jose Enrique Romero.

Una investigación como la que se ha desarrollado en esta tesis es imposible sin instituciones y estructuras que la sustenten. Es por ello que quiero agradecer su contribución a esta tesis a los miembros del consorcio del proyecto InAdvance, especialmente al Instituto de Investigación La Fe y al instituto Polibienestar de la Universidad de Valencia. Quiero agradecer especialmente sus aportaciones al Dr Vicente Ruíz y al Dr Gordon Linklater. Agradecer al instituto ITACA y a la Universitat Politècnica de València (UPV), la estructura proporcionada, que ha sido necesaria para el desarrollo

de una tesis.

Por último, pero no menos importante, quiero agradecer a mi familia toda su ayuda y apoyo. Sin ellos jamás hubiera tenido la posibilidad de llegar hasta aquí. Gracias a mi padre Vicent, a mi madre Francisca, a mis abuelos, especialmente a mi abuelo Vicent que siempre creyó en mí y a mi tio Xavier, sin el que seguramente no hubiera empezado mis andaduras universitarias. Gracias a Carmen por *parchearme* y arreglarme, a Juan Joya y a todos los amigos que me habéis acompañado este tiempo, sois muchos, pero a cada uno os corresponde un pedacito de mí y de esta tesis.

Abstract

Palliative Care (PC) is specialized medical care that aims to improve patients' quality of life with serious illnesses. Historically, it has been applied to terminally ill patients, especially those with oncologic diagnoses. However, current research results suggest that PC positively affects the quality of life of patients with different conditions. The current trend on PC is to include non-oncological patients with conditions such as Chronic Obstructive Pulmonary Disease (COPD), organ function failure or dementia. However, the identification of patients with those needs is complex, and therefore alternative tools based on clinical data are required.

The growing demand for PC may benefit from a screening tool to identify patients with PC needs during hospital admission. Several tools, such as the Surprise Question (SQ) or the creation of different indexes and scores, have been proposed with varying degrees of success. Recently, the use of artificial intelligence algorithms, specifically Machine Learning (ML), has arisen as a potential solution given their capacity to learn from the Electronic Health Records (EHRs) and with the expectation to provide accurate predictions for admission to PC programs.

This thesis focuses on the creation of digital tools based on ML for the identification of patients with palliative care needs at hospital admission. We have used mortality and frailty as the two clinical criteria for decision-making, being short survival and increased frailty our targets to predict. Next, we have focused on their implementation in clinical settings and studied their usability and acceptance in clinical workflows.

Then, first, we studied and compared ML algorithms for one-year survival in adult patients during hospital admission. To do so, we defined a binary variable to predict, equivalent to the SQ and defined the set of predictive variables based on literature. We compared models based on Support Vector Machine (SVM), k-Nearest Neighbours (kNN), Random Forest (RF), Gradient Boosting Machine (GBM) and Multilayer Perceptron (MLP), attending to their performance, especially to the Area under the ROC curve (AUC ROC). Additionally, we obtained information on the importance of variables for tree-based models using the GINI criterion.

Second, we studied frailty measurement of Quality of Life (QoL) in candidates for PC intervention. For this second study, we narrowed the age of the population to elderly patients ($\geq$ 65 years) as the target group. Then we created three different models: 1) for the adaptation of the one-year mortality model for elderly patients, 2) a regression model to estimate the number of days from admission to death to complement the results of the first model, and finally, 3) a predictive model for frailty status at one year. These models were shared with the academic community through

a web application [a] that allows data input and shows the prediction from the three models and some graphs with the importance of the variables.

Third, we proposed a version of the 1-year mortality model in the form of an online calculator. This version was designed to maximize access from professionals by minimizing data requirements and making the software responsive to the current technological platforms. So we eliminated the administrative variables specific to the dataset source and worked on a process to minimize the required input variables while maintaining high the model's AUC ROC. As a result, this model retained most of the predictive power and required only seven bed-side inputs.

Finally, we evaluated the Clinical Decision Support System (CDSS) web tool on PC with an actual set of users. This evaluation comprised three domains: evaluation of participant's predictions against the ML baseline, the usability of the graphical interface, and user experience measurement. A first evaluation was performed, followed by a period of implementation of improvements and corrections to the platform detected during the interview. Then, the second round of evaluation was executed with another set of participants. The platform passed the usability test and excelled in user experience in both rounds. During the validation sessions, the participants indicated that they felt important to have a predictive tool for palliative care referral in their daily workflow.

The results of this thesis comprehend part of the technological results of the European project InAdvance and have been published in five scientific contributions, including journals and conferences in the field of medical informatics, information systems and integrated care.

[a] https://demoiapc.upv.es/

Resumen

Los Cuidados Paliativos (PC) son cuidados médicos especializados cuyo objetivo es mejorar la calidad de vida de los pacientes con enfermedades graves. Históricamente, se han aplicado a los pacientes en fase terminal, especialmente a los que tienen un diagnóstico oncológico. Sin embargo, los resultados de las investigaciones actuales sugieren que la PC afecta positivamente a la calidad de vida de los pacientes con diferentes enfermedades. La tendencia actual sobre la PC es incluir a pacientes no oncológicos con afecciones como la EPOC, la insuficiencia de funciones orgánicas o la demencia. Sin embargo, la identificación de los pacientes con esas necesidades es compleja, por lo que se requieren herramientas alternativas basadas en datos clínicos.

La creciente demanda de PC puede beneficiarse de una herramienta de cribado para identificar a los pacientes con necesidades de PC durante el ingreso hospitalario. Se han propuesto varias herramientas, como la Pregunta Sorpresa (SQ) o la creación de diferentes índices y puntuaciones, con distintos grados de éxito. Recientemente, el uso de algoritmos de inteligencia artificial, en concreto de Machine Learning (ML), ha surgido como una solución potencial dada su capacidad de aprendizaje a partir de las Historias Clínicas Electrónicas (EHR) y con la expectativa de proporcionar predicciones precisas para el ingreso en programas de PC. Esta tesis se centra en la creación de herramientas digitales basadas en ML para la identificación de pacientes con necesidades de cuidados paliativos durante el ingreso hospitalario. Hemos utilizado mortalidad y fragilidad como los dos criterios clínicos para la toma de decisiones, siendo la corta supervivencia y la mayor fragilidad nuestros objetivos a predecir. Después, nos hemos centrado en su implementación en entornos clínicos y hemos estudiado su usabilidad y aceptación en los flujos de trabajo clínicos.

En primer lugar, hemos estudiado y comparado los algoritmos de ML para la supervivencia a un año en pacientes adultos durante el ingreso hospitalario. Para ello, definimos una variable binaria a predecir, equivalente a la SQ y definimos el conjunto de variables predictivas basándonos en la literatura. Comparamos modelos basados en Support Vector Machine (SVM), k-Nearest Neighbours (kNN), Random Forest (RF), Gradient Boosting Machine (GBM) y Multilayer Perceptron (MLP), atendiendo a su rendimiento, especialmente al Área bajo la curva ROC (AUC ROC). Además, obtuvimos información sobre la importancia de las variables para los modelos basados en árboles utilizando el criterio GINI.

En segundo lugar, estudiamos la medición de la fragilidad de la calidad de vida (QoL) en los candidatos a la intervención en PC. Para este segundo estudio, redujimos la franja de edad de la población a pacientes ancianos ($\geq$ 65 años) como grupo objeti-

vo. A continuación, creamos tres modelos diferentes: 1) la adaptación del modelo de mortalidad a un año para pacientes ancianos, 2) un modelo de regresión para estimar el número de días desde el ingreso hasta la muerte para complementar los resultados del primer modelo, y finalmente, 3) un modelo predictivo del estado de fragilidad a un año. Estos modelos se compartieron con la comunidad académica a través de una aplicación web [b] que permite la entrada de datos y muestra la predicción de los tres modelos y unos gráficos con la importancia de las variables.

En tercer lugar, propusimos una versión del modelo de mortalidad a un año en forma de calculadora online. Esta versión se diseñó para maximizar el acceso de los profesionales minimizando los requisitos de datos y haciendo que el software respondiera a las plataformas tecnológicas actuales. Así pues, se eliminaron las variables administrativas específicas de la fuente de datos y se trabajó en un proceso para minimizar las variables de entrada requeridas, manteniendo al mismo tiempo un ROC AUC elevado del modelo. Como resultado, este modelo conservó la mayor parte del poder predictivo y sólo requirió siete variables de entrada obtenibles durante visitas a pie de cama.

Por último, evaluamos la herramienta web del sistema de apoyo a las decisiones clínicas (CDSS) en el PC con un conjunto real de usuarios. Esta evaluación comprendía tres ámbitos: la evaluación de las predicciones de los participantes frente a la línea de base del ML, la usabilidad de la interfaz gráfica y la medición de la experiencia del usuario. Se realizó una primera evaluación, seguida de un periodo de implementación de mejoras y correcciones en la plataforma detectadas durante la entrevista. A continuación, se ejecutó la segunda ronda de evaluación con otro conjunto de participantes. La plataforma superó la prueba de usabilidad y destacó en experiencia de usuario en ambas rondas. Durante las sesiones de validación, los participantes indicaron que consideraban importante contar con una herramienta predictiva para la derivación de cuidados paliativos en su flujo de trabajo diario.

Los resultados de esta tesis forman parte de los resultados tecnológicos del proyecto europeo InAdvance y han sido publicados en cinco contribuciones científicas, incluyendo revistas y conferencias en el campo de la informática médica, los sistemas de información y la atención integrada.

[b] https://demoiapc.upv.es/

Resum

Les Cures Pal · liatives (PC) són cures mèdiques especialitzades l'objectiu de les quals és millorar la qualitat de vida dels pacients amb malalties greus. Històricament, s'han aplicat als pacients en fase terminal, especialment als quals tenen un diagnòstic oncològic. No obstant això, els resultats de les investigacions actuals suggereixen que les PC afecten positivament a la qualitat de vida dels pacients amb diferents malalties. La tendència actual sobre les PC és incloure a pacients no oncològics amb afeccions com la malaltia pulmonar obstructiva crònica, la insuficiència de funcions orgàniques o la demència. No obstant això, la identificació dels pacients amb aqueixes necessitats és complexa, per la qual cosa es requereixen eines alternatives basades en dades clíniques.

La creixent demanda de PC pot beneficiar-se d'una eina de garbellat per a identificar als pacients amb necessitats de PC durant l'ingrés hospitalari. S'han proposat diverses eines, com la Pregunta Sorpresa (SQ) o la creació de diferents índexs i puntuacions, amb diferents graus d'èxit. Recentment, l'ús d'algorismes d'intel · ligència artificial, en concret de Machine Learning (ML), ha sorgit com una potencial solució donada la seua capacitat d'aprenentatge a partir de les Històries Clíniques Electròniques (EHR) i amb l'expectativa de proporcionar prediccions precises per a l'ingrés en programes de PC. Aquesta tesi se centra en la creació d'eines digitals basades en ML per a la identificació de pacients amb necessitats de cures pal · liatives durant l'ingrés hospitalari. Hem utilitzat mortalitat i fragilitat com els dos criteris clínics per a la presa de decisions, sent la curta supervivència i la major fragilitat els nostres objectius a predir. Després, ens hem centrat en la seua implementació en entorns clínics i hem estudiat la seua usabilitat i acceptació en els fluxos de treball clínics.

En primer lloc, hem estudiat i comparat els algorismes de ML per a la supervivència a un any en pacients adults durant l'ingrés hospitalari. Per a això, definim una variable binària a predir, equivalent a la SQ i definim el conjunt de variables predictives basantnos en la literatura. Comparem models basats en Support Vector Machine (SVM), k-Nearest Neighbours (kNN), Random Forest (RF), Gradient Boosting Machine (GBM) i Multilayer Perceptron (MLP), atenent el seu rendiment, especialment a l'Àrea sota la corba ROC (AUC ROC). A més, vam obtindre informació sobre la importància de les variables per als models basats en arbres utilitzant el criteri GINI.

En segon lloc, estudiem el mesurament de la fragilitat de la qualitat de vida (QoL) en els candidats a la intervenció en PC. Per a aquest segon estudi, vam reduir la franja d'edat de la població a pacients ancians ($\geq$ 65 anys) com a grup objectiu. A continuació, creem tres models diferents: 1) l'adaptació del model de mortalitat a un any per a pacients ancians, 2) un model de regressió per a estimar el nombre de dies des

de l'ingrés fins a la mort per a complementar els resultats del primer model, i finalment, 3) un model predictiu de l'estat de fragilitat a un any. Aquests models es van compartir amb la comunitat acadèmica a través d'una aplicació web [c] que permet l'entrada de dades i mostra la predicció dels tres models i uns gràfics amb la importància de les variables.

En tercer lloc, vam proposar una versió del model de mortalitat a un any en forma de calculadora en línia. Aquesta versió es va dissenyar per a maximitzar l'accés dels professionals minimitzant els requisits de dades i fent que el programari responguera a les plataformes tecnològiques actuals. Així doncs, es van eliminar les variables administratives específiques de la font de dades i es va treballar en un procés per a minimitzar les variables d'entrada requerides, mantenint al mateix temps un AUC ROC elevat. Com a resultat, aquest model va conservar la major part del poder predictiu i només va requerir set variables d'entrada obtenibles durant visites a peu de llit.

Finalment, avaluem l'eina web del sistema de suport a les decisions clíniques (CDSS) en el PC amb un conjunt real d'usuaris. Aquesta avaluació comprenia tres àmbits: l'avaluació de les prediccions dels participants enfront de la línia de base del ML, la usabilitat de la interfície gràfica i el mesurament de l'experiència de l'usuari. Es va realitzar una primera avaluació, seguida d'un període d'implementació de millores i correccions en la plataforma detectades durant l'entrevista. A continuació, es va executar la segona ronda d'avaluació amb un altre conjunt de participants. La plataforma va superar la prova d'usabilitat i va destacar en experiència d'usuari en totes dues rondes. Durant les sessions de validació, els participants van indicar que consideraven important comptar amb una eina predictiva per a la derivació de cures pal · liatives en el seu flux de treball diari.

Els resultats d'aquesta tesi formen part dels resultats tecnològics del projecte europeu InAdvance i han sigut publicats en cinc contribucions científiques, incloent revistes i conferències en el camp de la informàtica mèdica, els sistemes d'informació i l'atenció integrada.

[c] https://demoiapc.upv.es/

Glossary

Acronyms

ACP Advanced Care Planning

AI Artificial Intelligence

ANN Artificial Neural Network

API Application Programming Interface

AUC ROC Area under the ROC curve

AUTH Aristotle University of Thessaloniki

BDSLab Biomedical Data Science Lab

BER Balanced Error Rate

BUN Blood Urea Nitrogen

CDSS Clinical Decision Support System

CER Comparative Effectiveness Research

CI Confidence Interval

COPD Chronic Obstructive Pulmonary Disease

CPOE Computerized Provider Order Entry

CRP C-Reactive Protein

DL Deep Learning

DNN Deep Neural Network

DRG Diagnosis Related Group

EHR Electronic Health Record

FI Frailty Index

GBM Gradient Boosting Machine

GUI Graphical User Interface

HP Healthcare Provider

HULAFE Hospital Universitario La Fe

ICIC International Conference on Integrated Care

ICU Intensive Care Unit

kNN k-Nearest Neighbours

MAR Missing at Random

MCAR Missing Completely at Random

MLP Multilayer Perceptron

ML Machine Learning

MNAR Missing Not At Random

OYM One-Year Mortality

PC Palliative Care

PU Positive-Unlabeled

QA/QI Quality Assurance and Quality Improvement

QoL Quality of Life

RCT Randomized Control Trial

RFE Recursive Feature Elimination

RF Random Forest

RL Reinforcement Learning

RMSE Root Mean Squared Error

ReLU Rectified Linear Unit

SQ Surprise Question

SUS System Usability Scale

SVM Support Vector Machine

SaaS Software as a Service

SoA State-of-the-Art

UEQ-S User Experience Questionnaire (short version)

UPV Universitat Politècnica de València

UX User eXperience

WHO World Health Organization

Contents

Abstract **iii**

Resumen **v**

Resum **vii**

Glossary **ix**

1 Introduction **1**
1.1 Motivation 1
1.2 Research questions and objectives 2
1.3 Thesis contributions 3
1.3.1 Main contributions 3
1.3.2 Scientific publications 4
1.3.3 Software 5
1.3.4 Other contributions 6
1.4 Projects and partners 6
1.4.1 Other projects 7
1.5 Thesis outline 8

2 Rationale **11**
2.1 Palliative Care and Advance Care Planning 11
2.1.1 Definitions 11
2.1.2 Brief history: oncological and non-oncological care 12
2.1.3 Clinical Routine and evidence 12
2.2 Machine Learning 14
2.2.1 Concepts and definitions 14
2.2.2 Overview of Machine Learning (ML) pipeline 14
2.2.3 Preprocessing for EHR tabular data 14
2.2.4 ML taxonomy 17
2.2.5 Evaluation 27
2.3 Clinical Decision Support Systems 31
2.3.1 Definition 31
2.3.2 Taxonomy: attributes and examples 32

3 Comparative study of ML methods to predict one-year mortality 35
3.1 Background and significance . . . 35
3.2 Materials . . . 37
3.3 Methods . . . 37
3.3.1 Development of the models . . . 37
3.3.2 Feature importance . . . 40
3.3.3 Validation of the state-of-the-art models . . . 40
3.3.4 Evaluation of the models . . . 40
3.4 Results . . . 40
3.5 Discussion . . . 41
3.6 Conclusion . . . 45

4 Frailty and mortality predictive models for older patients 47
4.1 Introduction . . . 47
4.2 Materials . . . 49
4.2.1 Basic description . . . 49
4.2.2 Mortality target variables . . . 49
4.2.3 Frailty target variable . . . 50
4.2.4 Data censoring and distributions . . . 51
4.3 Methods . . . 51
4.3.1 Predictive models . . . 51
4.3.2 Hyperparameters and variable selection . . . 53
4.3.3 Evaluation . . . 53
4.3.4 Comparison with baseline models . . . 53
4.3.5 Software . . . 54
4.4 Results . . . 54
4.4.1 Associations between distributions . . . 54
4.4.2 One-year mortality classifier . . . 54
4.4.3 Survival regression . . . 54
4.4.4 One-year frailty . . . 54
4.4.5 GINI Importances . . . 55
4.5 Discussion . . . 55
4.6 Conclusion . . . 59

5 Responsive and minimalist bedside mortality calculator 61
5.1 Introduction . . . 61
5.2 Materials and Methods . . . 63
5.2.1 Data . . . 63
5.2.2 Feature Selection and Modelling . . . 64
5.2.3 Explainability Layer . . . 65
5.2.4 APP Implementation and Software . . . 65
5.3 Results . . . 66
5.4 Discussion . . . 68
5.5 Conclusions . . . 71

6 Validation of a clinical decision support platform for palliative care: The Aleph **73**
6.1 Introduction . . . 73
6.2 Materials and methods . . . 75
6.2.1 The Aleph CDSS Platform . . . 75
6.2.2 Recruitment process . . . 75
6.2.3 Study structure . . . 76
6.3 Results . . . 78
6.3.1 Model evaluation . . . 78
6.3.2 Qualitative results . . . 78
6.3.3 Performance of tasks . . . 79
6.3.4 Usability . . . 79
6.3.5 User experience . . . 80
6.4 Discussion . . . 81
6.5 Conclusions . . . 86

7 Concluding remarks and recommendations **87**
7.1 Concluding remarks . . . 87
7.2 Recommendations . . . 88

Bibliography **91**

A Appendix: variables and hyperparameters **107**

Chapter 1

Introduction

1.1 Motivation

Clinical data availability has been improving since the start of the century due to the increase in the implementation and adoption of the EHR technology across the globe [Liang et al., 2021]. EHRs' the primary purpose is to support individual patient care by facilitating access to clinical evidence. However, there are secondary uses to this information, such as the conduction of projects that use this information to explore alternatives to standard care using retrospective information.

Among these projects and studies using retrospective data stand out, the ones using Artificial Intelligence (AI). The increased use of computerised systems in the clinical practice since the 1990s [Hughes, 2003] and the steady increment of computational power [Mack, 2011] have allowed researchers to handle increasing amounts of data and therefore create more complex and accurate predictive models.

The first application using clinical data and AI were the ones known as expert systems; this application implemented the knowledge available as a set of rules to be applied to the system's input [Compton and Jansen, 1988]. Probably the best example of this kind of system, due to its impact, successful implementation and validity, is the Computerized Provider Order Entry (CPOE) [Khajouei and Jaspers, 2010]. CPOE systems allowed the physicians to introduce the different medications prescribed to patients. The system checked their correctness and adherence in different dimensions: allergies, incompatibility and dosage, among others.

However, expert systems are based on a compilation of the available knowledge and clinical guidelines, and despite their contribution to improving the healthcare pathway, they did not provide new knowledge. With the development of Machine Learning (ML) algorithms and their ability to process several inputs simultaneously, the possibility to discover new patterns in the data, until now limited to the human cognitive capacity, exponentially increased. This meant that the availability of the EHR not only provided the required amount of data for the system to learn from but also the possibility to explore the effect of the different clinical features on a given problem. ML algorithms historically had been divided into two broad categories[d], depending on the output of the problem. Those problems where the output is known are called supervised problems,

e.g., predicting 30 days survival status after ICU admission. Problems where the output is unknown and the subjects or cases are clustered using their clinical features are known as unsupervised problems.

Despite the significant amount of research produced by the intersection of AI and medicine, there is a lack of implementation of these results into clinical practice. Implementing machine learning models as part of the Healthcare Provider (HP) software stack, primarily as CDSS, present a series of challenges. Usability, ease of use and minimal disruption of the actual clinical workflow are common desirable properties to maximise the probability of adoption. Some socio-cultural factors also intervene in the implementation, such as the perception and confidence of the HP and their organisation on the software. Therefore it is necessary to study these difficulties to design a CDSS with chances to be used in the clinical setting.

With the estimated increasing demand for PC in the near future [Etkind et al., 2017] it is necessary to use tools that help HP to identify patients with these needs. Admission to PC is currently based on the prognosis provided by the medical team and techniques such as the Surprise Question (SQ) that rely on the subjective opinion of the medical team. Introducing data-driven approaches to the detection and referral of patients with PC could help make decisions more objective and, sometimes, more precise. In addition, using the ML models within the information systems could contribute to better resource management.

Given this context, we felt the need to research ML applications in the realm of PC. We focused the content of this thesis on helping HPs decide which patients admitted to the hospital could benefit from entering the PC programs. Our effort did not stop with constructing predictive models; we also researched designing and implementing properly usable software integrated within the clinical workflow.

1.2 Research questions and objectives

The use of patient data from the EHR to evaluate the appropriateness of the admission to PC programs poses a number of challenges that need to be addressed. First of all it is needed to determine which of the variables available in the EHR systems are relevant to the problem, if any. This task includes the selection of proxy problems, i.e., the secondary and more objective prediction targets that could work as an equivalent for PC needs. We have expressed these knowledge gaps as research questions, which will be answered during this thesis.

RQ1 Is it possible to predict One-Year Mortality (OYM) on hospital admitted patients using ML algorithms by means of their admission profile?

RQ2 Is it possible to estimate the survival time of an admitted patient?

RQ3 Can we contribute to a better understanding of the factors that help predicting mortality within the year?

[d] Currently, there are some intermediate categories such as semi-supervised learning or PU learning, but their distinction is not as relevant as supervised/non-supervised.

RQ4 Is it possible to predict a patient frailty status one year after admission?

RQ5 What value adds the the mortality and frailty prediction models to a CDSS?

RQ6 Will HPs validate and accept a CDSS based on mortality and frailty models?

The research work conducted in this thesis aims to provide solutions to these questions by empirically validated scientific methods applied to the study of all-cause mortality and frailty using data from their admission profiles. To this end, the following objectives were defined:

O1 Review the state-of-the-art to analyse current works in mortality and frailty and which variables are relevant to the problem.

O2 Develop and compare ML models to predict OYM on all-cause admitted adult patients.

O3 Adapt the OYM models to older patients (age ≥ 65).

O4 Develop regression models to estimate the survival time from admission on older all-cause admitted patients.

O5 Research and select a QoL indicator to work alongside the mortality criteria to help select which older patients could benefit from PC programs.

O6 Develop a predictive model for QoL useful to decide referral to PC interventions.

O7 Implement a CDSS to help HP decide about PC referral.

O8 Evaluate the usability, user-experience and perception that HPs have of the CDSS for PC decision-making.

1.3 Thesis contributions

This section presents the main contributions of this thesis. First, a summary of the most relevant aspects of each contribution is presented. Next, the scientific publications in high impact journals and conferences are listed. Finally, the technological and software results are compiled.

1.3.1 Main contributions

C1 - Comparative study of machine learning methods to predict OYM
In this study, we compared different supervised ML algorithms for the task of predicting all-cause mortality after hospital admission for adult patients. This study started from the clinical status of the patient at admission that clinical experts considered relevant and compared the performance of the different ML algorithms using evaluation metrics. The importance of the different variables was also assessed to extract information data-driven information from the problem. This work was published as a journal article **P1** [Blanes-Selva et al., 2021a]

C2 - Frailty and Mortality predictive models for older patients

In this other study, we focused our efforts on older patients (age ≥ 65); to explore the criteria used as a proxy for all-cause hospital admitted older patients. In addition to the one-year mortality we estimated the need of a mortality regression, to approximate the length of the survival, specially for those with a life expectancy under a year due to its effects on the palliative pathway and resources management. Besides the mortality criteria, we determined an age-specific metric by proposing frailty as a third PC criteria. We used Deep Neural Networks (DNNs) and Gradient Boosting Machines (GBMs) to create the predictive models and extracted knowledge from them in the form of variable importance. This work was published as a journal article **P2** [Blanes-Selva et al., 2022a]. The models were incorporated into a predictive service that was registered to the UPV software catalogue **S1**.

C3 - Responsive and minimalist bedside mortality calculator
When working with predictive models for difficult tasks such as mortality, researchers tend to include as much information as possible to maximise the model's predictive power. Usually, these required variables include administrative information such as the department where the patient was admitted, if the patient was admitted from the emergency room or other codes that may not be adaptable to other EHRs. We addressed these problems by designing a minimalist OYM model, removing EHR specific variables and including the minimum number of variables to minimise the required input by the HP while maximising its performance. The model was also implemented in a responsive website as a calculator, with the main aim to help HP during bedside consultations. The study was published as a journal article **P3** [Blanes-Selva et al., 2021b]

C4 - PC CDSS user-centred validation
As the culmination of the previous studies, our team developed a web CDSS platform including the predictive models of **C2** and **C3**. This demonstrator was evaluated using a user-centred validation. Validation sessions with 21 HPs from 6 different countries[e] were performed, collecting thoughts, questions and opinions about the platform as well as usability and user experience results. The results of these evaluations were published as a journal article **P4** [Blanes-Selva et al., 2022b]. Afterwards, we implemented the final version of the platform and presented the results in the International Conference on Integrated Care (ICIC) conference **P5**. The platform was registered in the UPV software catalogue **S2**.

1.3.2 Scientific publications

The main scientific contributions of this thesis have been published of three different peer-review journals and one conference of medical informatics, user interaction and integrated care specialized disciplines. The publications are listed as follows:

[e]Italy, Brazil, Spain, Greece, Scotland and Portugal

P1 - **Vicent Blanes-Selva**, Vicente Ruíz-García, Salvador Tortajada, José-Miguel Benedí, Bernardo Valdivieso, Juan M. García-Gómez. *'Design of 1-year mortality forecast at hospital admission: A machine learning approach'*. Health Informatics Journal, 27(1), 1460458220987580. January 2021. [Blanes-Selva et al., 2021a].

IF: 2.932 (JCR 2019): 12/27 SCIE Medical Informatics (Q2)

P2 - **Vicent Blanes-Selva**, Ascensión Doñate-Martínez, Gordon Linklater, and Juan M. García-Gómez. *'Complementary frailty and mortality prediction models on older patients as a tool for assessing palliative care needs'*. Health Informatics Journal 28(2), 14604582221092592. June 2022. [Blanes-Selva et al., 2022a].

IF: 2.681 (JCR 2020): 19/30 SCIE Medical Informatics (Q3)

P3 - **Vicent Blanes-Selva**, Ascensión Doñate-Martínez, Gordon Linklater, Jorge Garcés-Ferrer and Juan M. García-Gómez. *'Responsive and Minimalist App Based on Explainable AI to Assess Palliative Care Needs during Bedside Consultations on Older Patients'*. Sustainability, 13(17), 9844. September 2021. [Blanes-Selva et al., 2021b].

IF: 3.251 (JCR 2020): 124/274 SCIE Environmental Sciences (Q2)

P4 - **Vicent Blanes-Selva**, Sabina Asensio-Cuesta, Felipe Pereira Mesquita, Ascensión Doñate-Martínez and Juan M. García-Gómez. *'User-centred Design of a Clinical Decision Support System for Palliative Care: Insights from Healthcare Professionals'*. [Blanes-Selva et al., 2022b]

Published as a preprint on medRxiv. Under consideration for SAGE Digital Health.

P5 - **Vicent Blanes-Selva**, Ascensión Doñate-Martínez, Gordon Linklater, Sabina Asensio-Cuesta, Felipe Pereira Mesquita, Jorge Garcés-Ferrer, Ángel Sánchez-García and Juan M. García-Gómez. *'The Aleph: A Multi-Purpose Clinical Decision Support Platform for Palliative Care Screening'*

Presented in the 22nd International Conference on Integrated Care (ICIC 2022) as abstract with presentation and future publication at the International Foundation for Integrated Care. IF: 5.120.

1.3.3 Software

The research line followed during this thesis has produced the Aleph Platform (thealeph.upv.es). Aleph is a multi-purpose CDSS platform that includes different predictive services and allows predictive services to *'plug and play'* to the platform benefiting from a consistent and validated interface. At this moment, a dedicated service of Aleph for PC includes the compact mortality model published in **P3** and a service containing the three predictive models published in **P2**. This last service and the Aleph platform were registered as software in the UPV technological catalogue. The following list summarizes the software contributions produced during this thesis.

S1 - **Vicent Blanes-Selva** and Juan M. García-Gómez. *'S-106-2022 - Módulo de predicción de mortalidad y fragilidad para cuidados paliativos'*. Explora $I+D+i$ Registry of the Universitat Politècnica de València. In process.

S2 - **Vicent Blanes-Selva**, Sabina Asensio-Cuesta, Ángel Sánchez-García and Juan M. García-Gómez. *'S-105-2022 - The Aleph'*. Explora *I+D+i* Registry of the Universitat Politècnica de València. In process.

1.3.4 Other contributions

During the development of this doctoral thesis, the author participated in other projects that, despite not having a direct relationship with the PC topic, contributed to the knowledge on other fields related to this thesis, such as ML predictive models trained with data from EHR and development of systems focused on the final user and their acceptance.

The first group was related to using technology to help in the public policy decision-making for the obesity problem. The author participated in the European project **CrowdHEALTH**, where the main goal was to implement a predictive service-connected to CrowdHEALTH's platform that helped to screen for unidentified obese patients. As a direct result of this work, two conference papers were published: the first described the models used in the obesity predictive service [Blanes-Selva et al., 2020] and the second one described the whole architecture of the CrowdHEALTH platform [Mavrogiorgou et al., 2020].

In parallel to that work, the author implemented a Telegram chatbot to gather health habits, mostly eating and exercise, from the general population. This project started as a CrowdHEALTH deliverable but soon evolved to its own project due to the chatbot's complexity and possibilities. Several designs were proposed, including colour scheme, images and application name. After pooling more than 400 members of the UPV community, the final design was selected, including the name of the chatbot: **Wakamola**. Wakamola is registered in the UPV technical catalogue and produced three research articles: the first one presented the chatbot and the data collected [Asensio-Cuesta et al., 2021a], the second analysed the lifestyle changes registered in the chatbot due to the COVID-19 confinement [Asensio-Cuesta et al., 2021b]. Finally the third article explored the differences between the three pilots performed using Wakamola [Asensio-Cuesta et al., 2021c]. The software was released as free software under the GNU Public License v3. https://github.com/bdslab-upv/Wakamola.

The author contributed to the development of the predictive models to aid dispatching emergency medical calls by classifying them in terms of life-threatening level, admissible response delay and emergency system jurisdiction [Ferri et al., 2021]. In addition, the author is participating in the CANCERLESS project, which is expected to develop micro-simulation models to evaluate the best policies to help the homeless population on cancer prevention and treatment.

1.4 Projects and partners

The development of this thesis has occurred in the context of the InAdvance Project, funded by the European Commission.

InAdvance *Patient-centred pathways of early palliative care, supportive ecosystems and appraisal standard* Funded by the European Commission (Grant agreement number 825750).

Objectives: The overall aim of the InAdvance project is to improve the benefit of the palliative care interventions for patients, families and caregivers, and professionals through the design of effective, replicable and cost-effective early palliative care interventions focused on and oriented by the patients.

Partners: Polibienestar, Universitat de València (Valencia, Spain), Hospital Universitario La Fe (HULAFE) (Valencia, Spain), NHS Highland (Inverness, Scotland), Erasmus MC (Netherlands), Aristotle University of Thessaloniki (AUTH) (Greece), University of Leeds (United Kingdom), Salumedia Labs (Seville, Spain), Sabien, UPV (Valencia, Spain), Santa Casa de Misericórdia Amadora (Portugal), AGE Platform Europe (Brussels, Belgium), Wita Care (Italy) and Biomedical Data Science Lab (BDSLab), UPV (Valencia, Spain).

1.4.1 Other projects

During this period other projects funded by the European Commission and the Agencia Valenciana de Seguridad y Respuesta a las Emergencias have been carried out with the participation of partners from several countries:

CrowdHEALTH *Collective wisdom driving public health policies* Funded by the European Commission (Grant agreement number 727560).

Objectives: CrowdHEALTH proposal is to create Holistic Health Records. These are structured health records that may include several types of information that are relevant to a patient's health status, such as laboratory medical data, clinical data, lifestyle data collected by the patient or related people, social care data, or physiological and environmental data collected by medical devices and sensors.

Partners: ATOS (Spain), Engineering - Ingegneria Informatica Spa (Italy), Siemens (Romania), Singular Logic Cyprus Ltd (Cyprus), EOPYY (Greece), The National Institute of Public Health (Slovenia), Karolinska Institutet (Sweden), German Research Center for Artificial Intelligence (Germany), BioAssist (Greece), Information Catalyst (UK), Care Across (UK), LeanXcale (Spain), University of Piraeus Research Center (Greece), Universitat Politécnica de Madrid (Madrid, Spain), IT Innovation Centre (UK), Jožef Stefan Institute (Slovenia), University of Ljubljana (Slovenia), European Federation for Medical Informatics (Switzerland), Taiwan Medical University (Taipei, Taiwan) and HULAFE (Valencia, Spain) including BDSLab as a third party.

CANCERLESS *Cancer prevention and early detection among the homeless population in Europe: Co-adapting and implementing the Health Navigator Model.* Funded by the European Commission (Grant agreement number 965351).

Objective: The aim of the CANCERLESS project is to promote timely access to primary and secondary cancer prevention services among people experiencing homelessness though the adaptation and implementation of the patient navigator model.

Partners: Medical University of Vienna (Austria), Polibienestar, Universitat de València (Valencia, Spain), Kveloce (Spain), Consejeria de Familia, Juventud y Política Social (Madrid, Spain), Servicio Madrileño de Salud (Madrid, Spain), International Foundation for Integrated Care (Netherlands), Praksis (Greece), FEANTSA (Belgium), Prolepsis (Greece), Anglia Ruskin University (UK) and Biomedical Data Science Lab (BDSLab), UPV (Valencia, Spain).

112 Emergency Service *Servicio de desarrollo de un método experto de ayuda a la clasificación de la demanda sanitaria de urgencias, emergencias extrahospitalarias y llamada sanitaria 112.* Funded by Agencia Valenciana de Seguridad y Respuesta a las Emergencias.

Objective: The objective of the project is to assess the possibility of improving the classification of out-of-hospital emergency classification out-of-hospital emergency incidents using techniques based on machine learning, applied to both structured clinical data and unstructured free text fields.

Partners: Conselleria de Sanitat Universal i Salut Pública (Valencia, Spain), Intelligent Data Analysis Laboratory, Universitat de València (Valencia, Spain) and BDSLab, UPV (Valencia, Spain).

1.5 Thesis outline

This thesis is divided into seven chapters that describe the work carried on during the different studies. Chapter 1 has introduced the motivations, research objectives and main contributions. Chapter 2 describes the thesis rationale, introducing the clinical background and the methodology used during the thesis. Chapter 3 presents the first study composing this thesis where we compared different ML algorithms to predict OYM on adults. Chapter 4 describes our work with older patients, including two new PC inclusion criteria with the mortality regression and the use of frailty status. Chapter 5 describes the creation of a minimalist and portable model implemented in a responsive web application and designed to be used during bedside consultation. In Chapter 6 we described Aleph in its demonstrator version and the user-centred validation in terms of usability and user experience that led to its first release version. Finally, Chapter 7 ends this dissertation with the concluding remarks and recommendations to continue with the research developed in this thesis.

For a graphical overview of the thesis, and its associate contributions, Figure 1.1 outlines the thesis contributions structured among the thesis chapters, along with the publications, research projects, transfer actions, patents and the software developed during this study.

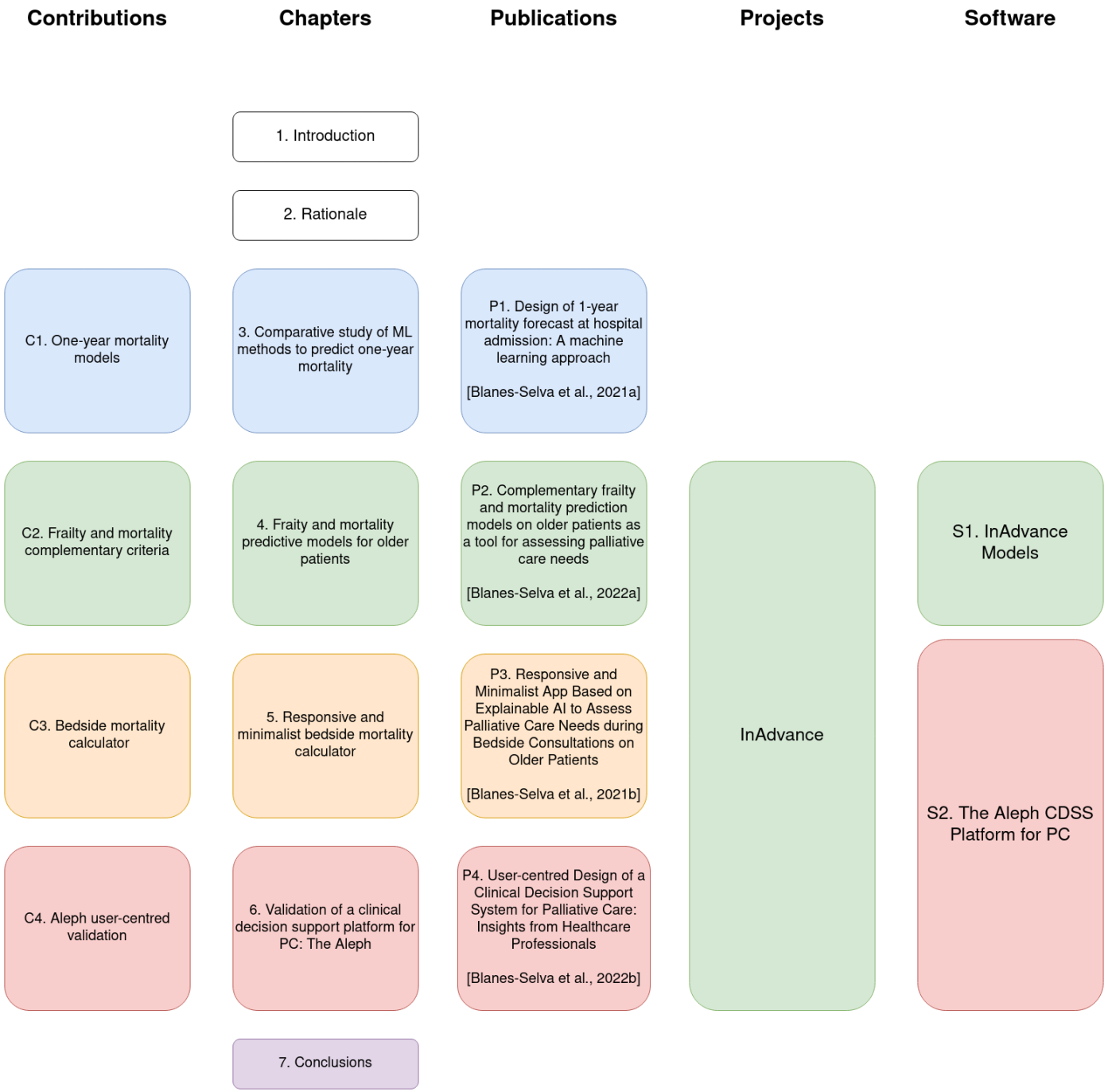


Figure 1.1: Outline of the thesis contributions, chapters, publications, projects, transference actions, patents and software

Chapter 2

Rationale

This chapter describes the thesis rationale divided into three sections. First, the concepts of Palliative Care (PC) and Advanced Care Planning (ACP) are introduced including their definitions, brief history and its current status and evidences supporting their implementation. Secondly, we will introduce the ML framework used during the thesis, the conceptual ideas between the models as well as their mathematical interpretation. Finally, we will introduce the concept of Clinical Decision Support System (CDSS) and review their taxonomy.

2.1 Palliative Care and Advance Care Planning

2.1.1 Definitions

According to the World Health Organization (WHO), *PC is an approach that improves the QoL of patients (adults and children) and their families who are facing problems associated with life-threatening illness. It prevents and relieves suffering through the early identification, correct assessment and treatment of pain and other problems, whether physical, psychosocial or spiritual* [WHO, 2020].

While PC programmes aim to address the suffering of these patients, their scope is not limited only to physical symptoms. PC aims to support patients and their caregivers, and focus on help addressing practical needs, bereavement counselling and offering support systems to help patients live as actively as possible until death [WHO, 2020]. Therefore, PC should be provided through person-centred and integrated health services by multidisciplinary teams including physicians, nurses, psychologists and other professionals able to help with spiritual distress.

On the other hand, Advanced Care Planning (ACP) is defined as *conversations that cover an individual's specific health conditions, their options for care, and what care best fits their personal wishes, including at the end of life* [Fulmer et al., 2018]. Despite their conceptual difference, both concepts, PC and ACP are related in practice. When the possibility of near-death is presented to the patients, PC appears as one option to be presented to the patient. Therefore, ACP should always exist alongside PC.

2.1.2 Brief history: oncological and non-oncological care

The first global definition of PC was issued by the WHO in 1990. However, prior to then, there are documented historical events that contributed to the appearance of the field, such as the creation of the L'Association des Dames du Calvaire in 1984, the foundation of the Society for the Prevention and relief of cancer in 1911 and the acknowledgement of the abandonment of dying people from medical professionals in the 1950s [Milligan and Potts, 2009]. In 1967 Cicely Saunders, one of the pioneers of the PC culture, founded St. Christopher Hospice the world's first modern hospice [Richmond, 2005].

In 2002 the WHO updated their definition of PC, acknowledging that PC should be available when required to everyone with a life-limiting illness, regardless of their diagnosis. This definition removed the notion that PC is exclusive for those diagnosed with cancer. In 2004, The National Institute of Clinical Excellence[f] suggested that this general level of PC should be an integral clinical skill for all caring professionals, as most PC will be provided by the dying person's day-to-day professional carers. In 2006, the Scottish Partnership for Palliative Care[g] included in their definition of PC the involvement of the family and caregivers. Despite this scope expanding definitions, professional careers treated PC as equivalent for terminal or end-of-life care [Milligan and Potts, 2009].

For many years, clinical procedures and research around PC have been focused on oncological patients. However, since the WHO's 2002 definition of PC, research has also focused on PC programmes for non-oncological patients. PC research has been developed significantly for diseases such as heart failure, renal failure, chronic obstructive pulmonary disease (COPD) or Neurodegenerative diseases such as dementia [Mahtani-Chugani et al., 2010].

2.1.3 Clinical Routine and evidence

In practice, PC is usually provided upon exhaustion of disease-modifying treatments; this means that patients in need are admitted late if at all to these programmes [Bakitas et al., 2015]. However, several studies have found benefits of performing a timely referral to PC. Bakitas *et al.*. [Bakitas et al., 2009] performed a Randomized Control Trial (RCT) comparing PC against usual care that reported a statistically significant improvement on QoL and mood in patients that underwent early PC. In 2015, the same author [Bakitas et al., 2015] reported improved 1-year survival in patient enrolled in PC. Temel *et al.*in 2010 [Temel et al., 2010] also found the improvement on QoL and mood in their RCT for patients included in PC.

There is no clear consensus on the concrete timing for PC to be considered early PC. This is due to the different trajectories of the many diseases to which PC has been applied. In their article, Temel *et al.* [Temel et al., 2010] recruited patients with non-small-cell lung cancer within eight weeks after diagnosis; however, using the diagnosis as the sole sentinel point may not be appropriate for other types of cancer or non-cancer diseases.

[f] https://www.nice.org.uk/

[g] https://www.palliativecarescotland.org.uk/

Hospital admission is a great checkpoint to consider and start addressing PC needs [Fischer et al., 2006]. During this time, the patient is a captive audience and exacerbation of an illness may promote self-reflection, which could be a great trigger to initiate ACP conversations. According to the same authors, the most important barrier to start PC earlier in the disease's progression is the failure to identify limited life expectancy. Therefore, medical criteria to identify patients in need of PC are required. This criteria are often referred in the scientific literature as *triggers*, *screening tools* or *referral tools*.

The Surprise Question (SQ) (*"Would you be surprised if this patient died within the next x months?"*) has been a used as a referral tool for PC programs. The SQ is meant to be asked to themselves by the physicians in charge of the patient. If the answer is negative, then a short survival is expected, and therefore ACP conversations may be triggered. This method has been included as part of the Gold Standard Framework (GSF) proactive identification guidance tool in the UK [Thomas et al., 2017]. As main benefits, the SQ does not require clinicians to collect clinical data, use scoring algorithms or estimate life expectancy [White et al., 2017]. However, the performance of this method has been found to be *poorly to modest* in the systematic review and meta-analysis by Downar *et al.* [Downar et al., 2017]. Therefore, there is a need to find more accurate alternatives to this tool.

The search for a PC timely-referral screening tool has led to several research studies trying to evaluate the current criteria and propose new alternatives. Nelson *et al.* [Nelson et al., 2013] reviewed the PC inclusion criteria for patients admitted to Intensive Care Unit (ICU), were the most predominant criteria were: diagnosis for terminal illness, a short expected survival and concerns/petitions by the patient or their family or caregivers during the ACP. This article concluded that the use of specific screening tools for PC could help reduce the ICU resource utilization. Fischer *et al.* [Fischer et al., 2006] proposed the CARING criteria (Primary diagnosis of cancer, Admissions ≥ 2, Resident in nursing home, ICU admission with multi-organ failure, ≥ 2 Non-cancer hospice guidelines [Organization, 1996]). Clinical scores have also been proposed as predictive tools, some examples are: the Palliative Prognostic Score [Pirovano et al., 1999], the Palliative Prognostic Index [Morita et al., 1999], Palliative Performance Scale [Lau et al., 2009], the Supportive and Palliative Care Indicators Tool [Highet et al., 2014] or the PROFUND index [Bernabeu-Wittel et al., 2011]. These scores are a set of rules that add points if certain conditions are present in the patient and then assess the situation depending on the number of points obtained.

More recent scientific works have focused on constructing mortality predictive models to be used as a referral tool for PC programs. The most prominent examples are the different versions of the HOMR model [van Walraven et al., 2015]: HOMR-Now! [van Walraven and Forster, 2017], m-HOMR [Wegier et al., 2021] and the Deep Learning (DL) model proposed by Avati *et al.* [Avati et al., 2018]. However, these models target adult population (≥ 18 years old) and some of them incorporate variables that are hard to obtain during the first admission hours.

Despite the different scientific efforts present in the literature, there is not a decisive solution in the search for the *definitive* PC referral tool yet. Indexes are easier to validate externally since they habitually require less data but at the expense of being

less powerful than predictive models. Also, the lack of adoption of these technologies in clinical practice presents a problem that should be carefully studied. Besides, other criteria than *bad outcome* or *poor survival expectation* should be explored in order to find a better approximation to the PC criteria.

2.2 Machine Learning

2.2.1 Concepts and definitions

As seen on the previous section, the field of PC could benefit from the use of data-driven techniques to improve the detection of patients in need of these programs. Retrospective data availability alongside AI techniques offer the potential to create predictive tools that could help HPs in this task. Among the different options, ML algorithms are the candidates to create the more accurate referral tools.

The first ML definition was proposed by Arthur Samuel in 1959: *"Field of study that gives computers the ability to learn without being explicitly programmed"*. In other words, ML is a discipline that comprises algorithms that use large sets of data inputs and outputs to recognise patterns and effectively learn to solve problems based on data examples [Helm et al., 2020]. A subtype of these algorithms generates and adjusts mathematical predictive models that retain relevant information about the dataset distribution to make predictions when new data is presented. Other subtypes just run on the input data and find patterns and connections among them.

2.2.2 Overview of ML pipeline

Data scientists and ML practitioners follow a linear process. After the data is obtained, usually from EHRs or downloaded as public datasets, it has to suffer a series of transformations, which usually receive the name of preprocess (subsection 2.2.3) and includes dimensionality reduction, treatment of categorical variables and strategies to deal with missing. After that, a ML algorithm should be selected and trained on the data (subsection 2.2.4). Finally, a evaluation process is needed to estimate the performance of the model (subsection 2.2.5). Usually this pipeline of process is iterative and explore different strategies regarding algorithms and hyper-parameters in order to maximize performance metric of the final model. Figure 2.1 shows visually the overview of a common ML pipeline.

2.2.3 Preprocessing for EHR tabular data

Categorical variables

ML algorithm implementations use only numeric values as inputs. This means that a determined kind of information from the EHR should be mapped into a numerical representation to be used on the models. The most common case is the categorical variables. Categorical variables can take a set of defined and limited values. For example, a diagnosis admission expressed on their ICD-10 code, in this scenario, the variable *admission_diagnosis* could have any of the ICD-10 available codes such as:

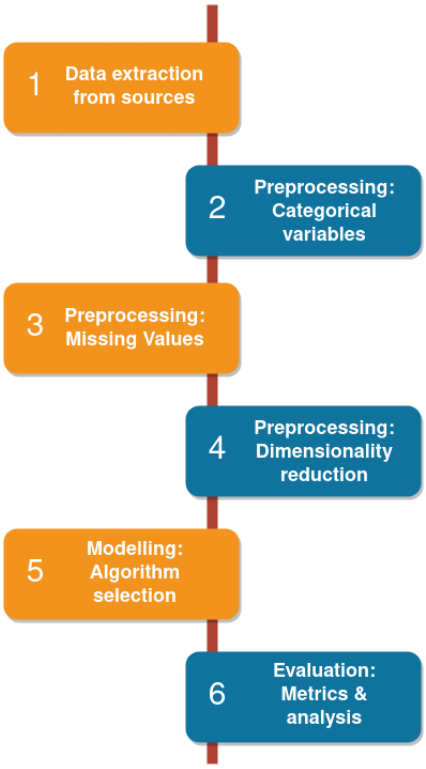


Figure 2.1: Overview of the classical ML pipeline. Usually the pipeline is executed several time with alternative algorithms and hyper-parameters on each step to find the best performing strategy.

K22.1 (Ulcer of esophagus with bleeding), N17.9 (Acute kidney failure) or C70.9 (Malignant neoplasm of meninges). Another typical example could be the department where the patient is admitted, e.g., Cardiology, Gynaecology or Neurosurgery. A naive and useful approximation to solve the problem is to assign a positive integer to each possible category. However, there are some pitfalls to this strategy. First, if the total number of possible values is not defined and after shipping the predictive model to *production*[h] environment, a variable presents a new value. A solution to this problem is implementing the "unknown" category or use the One Hot Encoder, where the unknown value is implemented as zeros in all the categories. Each strategy has its perks, depending on the context and the predictive model used, and should be chosen carefully.

Another possible downfall is to ignore possible relations between the values. For example, if the health status after the examination is coded from excellent" to "very frail" in five different categories, assigning successive values can help to simplify discriminant frontiers. These strategy choices are vital to the correct learning of the model since a predictive model can only be as good as the data available to train it

[h]A term from software development that means that the product has passed the appropriate tests and is being exploited

and belong to the Feature Engineering field.

Missing values

Missing values are a common problem when data from EHR should be used to train predictive models. The term missing values refer to variables whose value is not present in the data. Three different missing patterns have been identified: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not At Random (MNAR). MCAR correspond to completely aleatory errors, such as information system errors that wipe random fields and therefore are very unlikely to happen. MAR occurs when the missing pattern is not random but can be fully accounted for by variables where there is complete information. MNAR is the most predominant in the EHR. For example, laboratory results are missing for certain patients because the physician in charge did not consider its condition required lab tests.

As with the previous situation, missing values are also represented by non-numerical values, and therefore it is necessary a preprocessing strategy on the dataset previous to training the model. The naïve strategy in this case is to remove features or cases containing missing values use a specific value to indicate that the value is missing or imputation. This first strategy has substantial drawbacks, especially when dealing with MNAR. First, bias could be introduced, e.g., healthier patients did not get blood analysis and therefore, the model is only observing sicker patients. Furthermore, because removing the amount of available data is going to affect the performance of the model, as more data is removed, more information is lost, and therefore the function learnt by the ML usually have a more significant difference from the reality. Marking the missing values with a specific numerical value could work if the variable is categorical, e.g., the admission department or if the variable has impossible values, e.g., using -1 if the age value is missing. However, this strategy is not consistent with the different data types and need particular caution to be performed.

The final option is the **imputation**, in which the main drawback is that the data fed to the model is not real and could introduce bias. However, it is usually the best option to avoid losing too much information and predictive power. There are multiple imputation algorithms that we could classify on two different groups: imputation by statistics of the variable and imputation by likelihood of the samples.

Imputation by statistics consists in taking all the available values for a given variable and calculating the mean, the median, the most frequent value or any other simple statistic and then filling the missing values in that variable with the result. Using statistics is not very accurate and can introduce noise to the models. However, it is computationally cheap and works fairly well in practice.

The imputation by likelihood comprehends complex algorithm that try to fill the missing values using other available input samples. For that, ML algorithms are often used, for example, using regression or classification models to predict the missing feature or use models based on similarity such as kNN (subsection 2.2.4).

Dimensionality reduction

In general, ML practitioners try to keep the maximum amount of information. However, there are situations where we will want to reduce the dimensionality of our data, i.e., reduce the number of input variables. For example, when the dimensionality is very high, and the number of available samples is relatively low, we can fall into the *curse of dimensionality* or the extraordinarily rapid growth in the difficulty of problems as the number of dimensions increases [Kuo and Sloan, 2005]. Another situation could be due to data being introduced manually by the users when it is desirable to keep only the most relevant or informative variables and find an equilibrium between the number of dimensions and the model's performance.

Within this field, there are two prominent families of algorithms. On the one hand, the feature projection algorithms, which the main objective is to move the data points to another representation space with fewer dimensions, some of the foremost exponents of this field are: Principal Component Analysis [Szlam et al., 2014], Linear Discriminant Analysis [Balakrishnama and Ganapathiraju, 1998], Autoencoders [Wang et al., 2016] or Uniform manifold approximation and projection [McInnes et al., 2018].

On the other hand, we find the feature selection group of algorithms, which keep the data on the original representation space but try to eliminate the less relevant dimensions. The selection can be performed based on multiple criteria, for example, removing the variables with low variance. Variable selection could also be based on some measure of variable importance, e.g., GINI importance [Nembrini et al., 2018] on decision tree models or the weights on a regression. The selection could be univariate consider interactions between variables. A common multivariate algorithm is the Recursive Feature Elimination (RFE). In short, RFE follows these steps: 1) train a model with a precise importance measure (decision trees, regressions, Artificial Neural Networks...), 2) Obtain the importance measure for every variable 3) Remove the k less important variable and 4) return to the first step until the desired number of variable is reached.

2.2.4 ML taxonomy

As staged by the definitions, the ML field contains different algorithms, which may be divided into different categories and subcategories. This taxonomy is changing as the field develops and new studies appear. Also, different authors present different taxonomies, depending on their vision of the field. In this section, we will focus on a broad, classical and widely accepted set of categories.

The first category split between the fields is how these models learn. In classic taxonomies, there were two main categories: *supervised* and *unsupervised learning*. The difference between them is the presence (or absence) of the target variables or *ground truth*. Supervised learning focuses on finding a function to map the data input (X) to their output/target (Y). Meanwhile, unsupervised learning focuses on finding patterns in the data and grouping the different data points ($x \in X$) by similarity. Some categories represent a mix between supervised and unsupervised learning, such as semi-supervised learning or its subcategory Positive-Unlabeled (PU) Learning [Fusilier et al., 2015]. However, most of the algorithms in this category are adaptations from the main

categories tweaked to deal with missing targets on parts of the data. In recent years, a new type of learning has appeared and gained popularity. In this case, the model acts as an *agent*, and its actions into the *environment* and other agents produce a reaction that is used a feedback to be learnt from. This kind of learning has received the name of *Reinforcement Learning (RL)*.

Within the supervised learning category we could find two main subdivisions: **classification** and **regression**. Classification algorithms map the input data into a category or *classes*, e.g., given the picture of a skin lesion, determine the diagnosis. Whereas regression models map their inputs into a continuous value, e.g., given clinical information about a patient, determine their tumour burden in the sentinel node. As well as other taxonomy levels, there are other ML applications falling on distinct and minority categories. For example, there are supervised applications which output is structured instead of a category or a continuous value, such as Part of Speech Tagging [Marquez et al., 2000] used to analyse clinical notes on the EHR.

Unsupervised learning aggregates those data-driven methods that do not use a target variable. Among their subcategories, we could find clustering, which consists on aggregate the different data points into categories, based on a measure of similarity measure, e.g., the euclidean distance, calculated as the difference between the dimensional values[i] of two data point (x and d) (see Equation 2.1). Another important subfield of unsupervised learning is dimensionality reduction. These techniques aim to move the data from a high dimensional space to reduced dimensionality keeping as much information as possible. Sometimes these methods are used as part of the data preprocessing steps before applying another ML algorithm.

$$d(x, d) = ||x - d|| = \sqrt{\sum_{i=1}^{D}(x_i - q_i)^2} \tag{2.1}$$

However, and despite the multiple learning algorithms present in the ML taxonomy, this thesis is going to focus on supervised learning, concretely on classification and regression algorithms. We found this kind of learning is the best fit to answer our research questions due to the availability of retrospective cases containing labels. To be more specific, we will make use of the following methods.

Support Vector Machines

The Support Vector Machine (SVM) was proposed by Vladimir Vapnik and their team [Guyon et al., 1993]. In its origins, the SVM was designed as a non-probabilistic binary linear classifier. Intuitively, the SVM algorithm works by placing the data points in the representation space and then finding a linear frontier that separates the data into two classes. In addition to this, the SVM tries to maximize the width of the gap between the two classes. Figure 2.2 show a graphical representation of the algorithm.

The mathematical formulation of the binary SVM is the following: given the training vectors $X_i \in \mathbb{R}^p, i = 1...n$ and the label vector $y \in \{1, -1\}^n$, the method objective

[i]Each dimension is a variable. The set of dimensions is also known in the ML field as features

is to find a direction vector $w \in \mathbb{R}^p$ and the bias $b \in \mathbb{R}$ such that the prediction given by $\text{sign}(w^t\phi(x) + b)$ is correct for most of the classes.

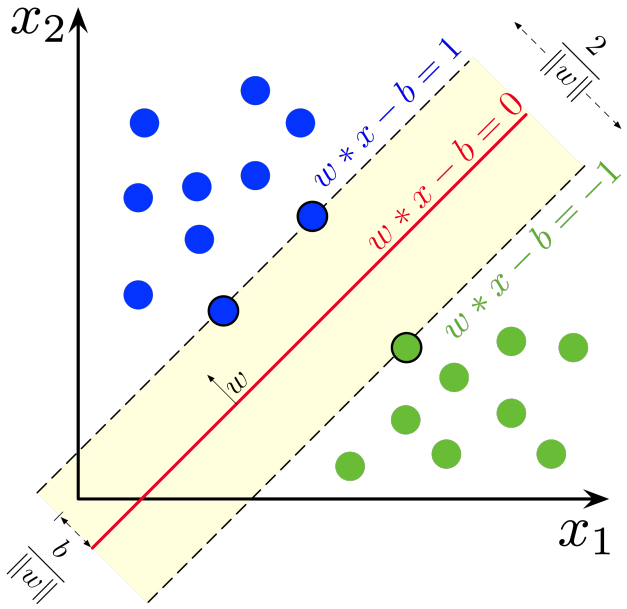


Figure 2.2: Graphical representation of the frontier and margin obtained through the SVM algorithm on separable data.

$$\begin{aligned} \min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i \\ \text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i, \\ \zeta_i \geq 0, i = 1..n \end{aligned} \tag{2.2}$$

Equation 2.2 is the formulation for the SVM soft margin problem. This means that it accepts miss-classified samples, making the algorithm feasible for non linearly separable data. ζ is the slack variable introduced to obtain the soft margins, and C is a regularization parameter for ζ. The optimization problem can be transformed into a dual problem:

$$\begin{aligned} \min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\ \text{subject to } y^T \alpha = 0 \\ 0 \leq \alpha_i \leq C, i = 1...n \end{aligned} \tag{2.3}$$

Where e is the vector of all ones, and Q is an n by n positive semidefinite matrix $Q_{ij} \equiv y_i y_j K(x_i, x_j)$ where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is a kernel, designed for non linearity. The α_i terms are called the dual coefficients they are upper-bounded by C. Once the optimization problem is solved, the decision function for a given sample is:

$$\sum_{i \in SV} y_i \alpha_i K(x_i, x) + b \tag{2.4}$$

K-Nearest Neighbours

The main idea behind the kNN is to place the training samples in a representation space, and then, when new samples are required to be classified, the algorithm searches its closest neighbour and assign the same label to the new sample. Euclidean distance (equation 2.1) is often used to determine the proximity between the data points. However, any distance measure between two points ($x, d \in \mathbb{R}^p$) can be implemented, and popular implementations of the algorithm usually include several options by default, such as Manhattan or Mahalanobis (equations 2.5 and 2.6). Also, the algorithm allows using multiple (K) neighbours to assign their value to the new sample. The most common methods to use are simple voting or weighted voting using the inverse of their distance to the new sample.

$$d_{\text{manhattan}}(x, d) = ||x - d||_1 = \sum_{i=1}^{n} |x_i - d_i| \tag{2.5}$$

$$d_{mahalanobis}(x, d) = \sqrt{(x - d)^T S^{-1} (x - d)} \tag{2.6}$$

where S is the nonsingular covariance matrix of the data

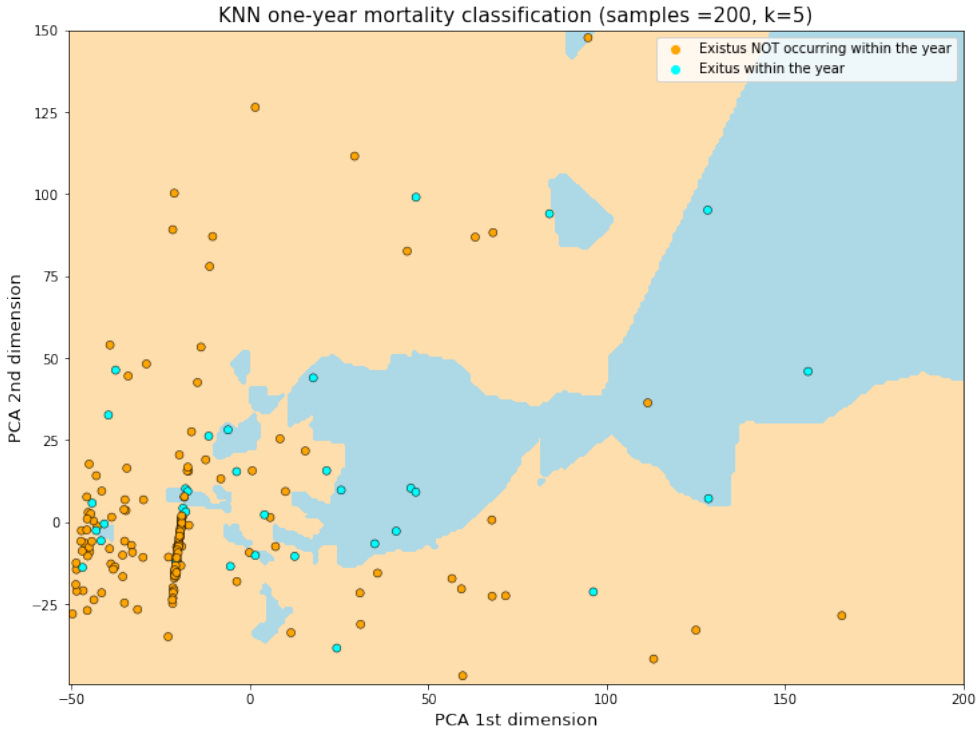


Figure 2.3: kNN method for a binary classification problem: one-year mortality using hospital admission data. Data has been projected to two dimensions before the classification. The representation space is divided into two regions based on the distances to the training samples.

Decision Trees

Decision Trees are ordered lists of conditions following the structure *if X then Y else Z*. This kind of abstraction has been present in collective thought for generations and therefore, the invention of the concept can't be traced to an specific time and location. Some scholars have suggested that the first reference of a decision tree as we know it came in 1959 [Belson, 1959]. However, the reference publication regarding these models is Breiman's book [Breiman et al., 2017], which first version appeared in 1984.

The goal of Decision Trees is to predict a target variable value based on a set of rules inferred from the training data features. This set of rules are represented in the tree nodes (see Figure 2.4) when the optimization algorithm chooses a variable and a threshold value to divide the space into two groups. Once the tree is trained, predicting a new sample consists of following the tree from the root (upper node) to one of the leaves (terminal nodes), using the variables of that sample to decide the path on the splits. The class assigned to the new data point is the majority class, if not unanimous, of the training set. Decision Trees can also be used for regression problems. In this case, the value predicted by the tree is an statistic of the training samples on that leaf such as the median or the mean.

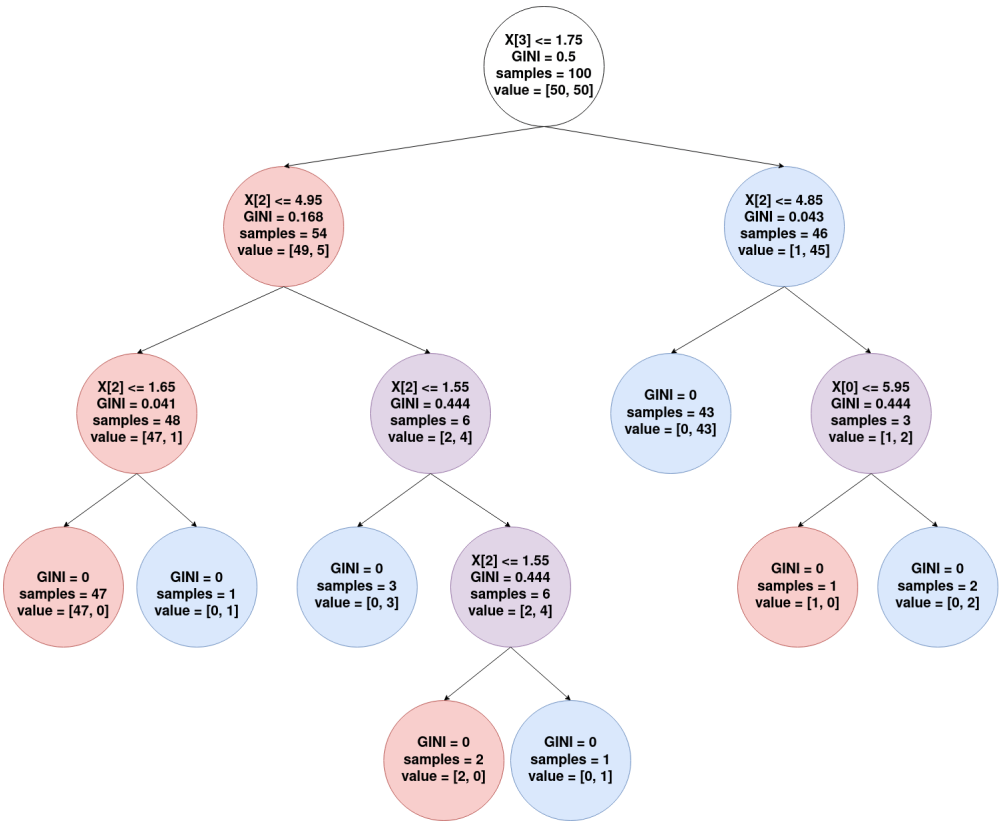


Figure 2.4: Example of Decision Tree trained with 100 samples from two different classes. Nodes contain the variable and the feature used to make the split, the GINI impurity values and how many samples of each class are left on the node.

The mathematical process to create the tree splits is as follows: Q_m represents the training data at node m, For each candidate split $\theta = (j, t_m)$ where j is one of

the features and t_m is a threshold value, partition the data into $Q_m^{\text{left}}(\theta)$ and $Q_m^{\text{right}}(\theta)$ subsets.

$$\begin{aligned} Q_m^{\text{left}}(\theta) &= \{(x, y)|x_j \leq t_m\} \\ Q_m^{\text{right}}(\theta) &= Q_m \setminus Q_m^{\text{left}}(\theta) \end{aligned} \tag{2.7}$$

The quality of the split candidate is then evaluated using a loss function $H()$, which offers different alternatives for classification and regression trees.

$$\begin{aligned} \theta^* &= \arg\min_{\theta} G(Q_m, \theta) \\ G(Q_m, \theta) &= \frac{N_m^{\text{left}}}{N_m} H(Q_m^{\text{left}}(\theta)) + \frac{N_m^{\text{right}}}{N_m} H(Q_m^{\text{right}}(\theta)) \end{aligned} \tag{2.8}$$

The algorithm is recursive for subsets $Q_m^{\text{left}}(\theta^*)$ and $Q_m^{\text{right}}(\theta^*)$, having different stop criteria: reaching the maximum defined depth, N_m being inferior to the minimum number of samples defined to split a node or reaching the value of 1.

Once the Decision Tree is fit to the data, we could determine the importance of the different features using the GINI importance criterion [Nembrini et al., 2018, Menze et al., 2009], which indicates how often a feature j was selected and their discriminative power. To obtain this criterion first we need to calculate the impurity (i) for each of the tree nodes (τ). On the binary classification scenario (class $\in 0, 1$), the node impurity is defined as

$$i(\tau) = 1 - p_0^2 - p_1^2 \tag{2.9}$$

Splitting the node into two different sub-nodes (τ_l and τ_r) with its different number of samples $p_l = \frac{n_l}{n}$ and $p_r = \frac{n_r}{n}$ causes impurity decrease

$$\Delta i(\tau) = i(\tau) - p_t i(\tau_l) - p_r i(\tau_r) \tag{2.10}$$

The impurity decrease created by splitting a node by the j feature is added together for all the tree nodes, obtaining the GINI importance (I_G) of that feature

$$I_G(j) = \sum_{\tau} \Delta i_j(\tau) \tag{2.11}$$

The GINI importance can also be extended to models composed from different Decision Trees (T)

$$I_G(j) = \sum_{T} \sum_{\tau} \Delta i_j(\tau, T) \tag{2.12}$$

Random Forests

Random Forest (RF) model appeared the first time in the publication by Ho *et al.*in 1995 [Ho, 1995]. In 2001 Breiman *et al.*extended the algorithm, combining the idea of a random selection of features introduced by Ho and the bagging procedure.

Intuitively, a RF is an **ensembles** of multiple Decision Trees that learn from the same training set and produce a prediction on new data by voting with their predictions. Each tree is trained using a random subset of the training data points and a subset of the features. The goal for this is to reduce the variance and, therefore, the over-fitting on the training set due to the nature of the single decision trees. A perk of these design is the possibility to train this meta-model in parallel since the Decision Trees has no dependence on each other nor the ensemble model. Figure 2.5 shows a simple representation of the method.

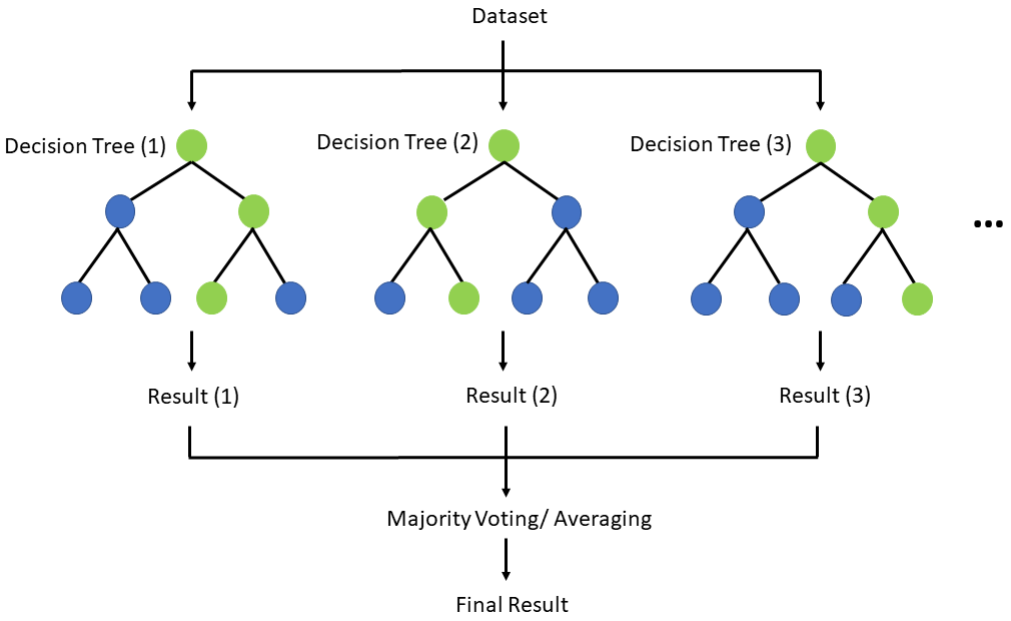


Figure 2.5: Random Forest basic functioning schema for classification.

Gradient Boosting Machines

The gradient boosting concept is derived from the observations that boosting can be interpreted as an optimization algorithm by Leo Breiman. The first gradient boosting algorithms were developed by Friedman [Friedman, 2001, 2002]. Despite sharing the same basic intuition as the RF, a model composed by an ensemble of decision trees, the mathematical formulation is very different due to the boosting greedy algorithm that incorporates each tree into the Gradient Boosting Machine (GBM). The output of the GBM for regression tasks is computed as follows:

$$\hat{y} = F_M(x) = \sum_{m=1}^{M} h_m(x) \tag{2.13}$$

Each of the M decision trees h_m provides an output for each data point (x). Similarly to other boosting algorithms, the GBM (F_M) is generated by greedily adding Decision Trees:

$$F_m(x) = F_{m-1}(x) + h_m(x) \tag{2.14}$$

The newly added tree (h_m) is fitted to minimize the sum of the losses (L_m), given a previous set of trees (F_{m-1})

$$h_m = \arg\min_h L_m = \arg\min_h \sum_{x \in X} l(y_i, F_{m-1}(x) + h(x)) \tag{2.15}$$

The sum of trees $F_M(x_i) = \sum_m h_m(x_i)$ produce a continuous value and therefore can not be used to predict classes. To perform the mapping to a classification version, the probability that x_i belongs to the positive class is modeled as $p(y_i = 1|x_i) = \sigma(F_M(x_1))$ where σ is the sigmoid function.

Artificial Neural Network

Artificial Neural Networks (ANNs) are connectionist computational models inspired by the biological model of the brain. ANNs use computational units (neurons) and connections between them to process the data input. The first concept of the ANN appeared in 1943 with the computational model of a biological neuron [McCulloch and Pitts, 1943]. Later, in 1958 and 1967 appeared the perceptron, which is the basis for the ANN and the first multilayer ANN [Rosenblatt, 1958, Ivakhnenko et al., 1967].

MLPs are universal function appropriators that try to learn functions defined as $f(\cdot) : R^m \rightarrow R^o$, where m is the number of input features and o is the number of dimensions on the output. Given a set of features and a target, the MLP algorithm can learn a non-linear function that approximates to the real function of the dataset in either classification and regression problems.

Figure 2.6 present an example of a simple MLP, where the left part corresponds to the input layer, which works as an entry point of the data to the model. The middle layer at the figure is the hidden layer composed of three different neurons (s_1^1, s_2^1, s_3^1), each of the neurons is performing a weighted linear summation, where θ_{ij}^m correspond to the weight of the m layer, j refers to the input neuron id and i correspond to the end-side neuron. X_m defines the values entering the model by the input layer.

Usually, a non-linear function is applied on the value of each neuron after the weighted linear summation is performed: $g(\cdot) : R \rightarrow R$. Some common functions are the hyperbolic tangent or the sigmoid function (equations 2.16 and 2.17). Finally, the output layer receives the values from the last hidden layer, performs the weighted linear summation and outputs the result. In some cases, a final activation function is applied to the output, for example, the softmax function (equation 2.18), which normalize the output values so each of the single output could be interpreted as a class probability.

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{2.16}$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.17}$$

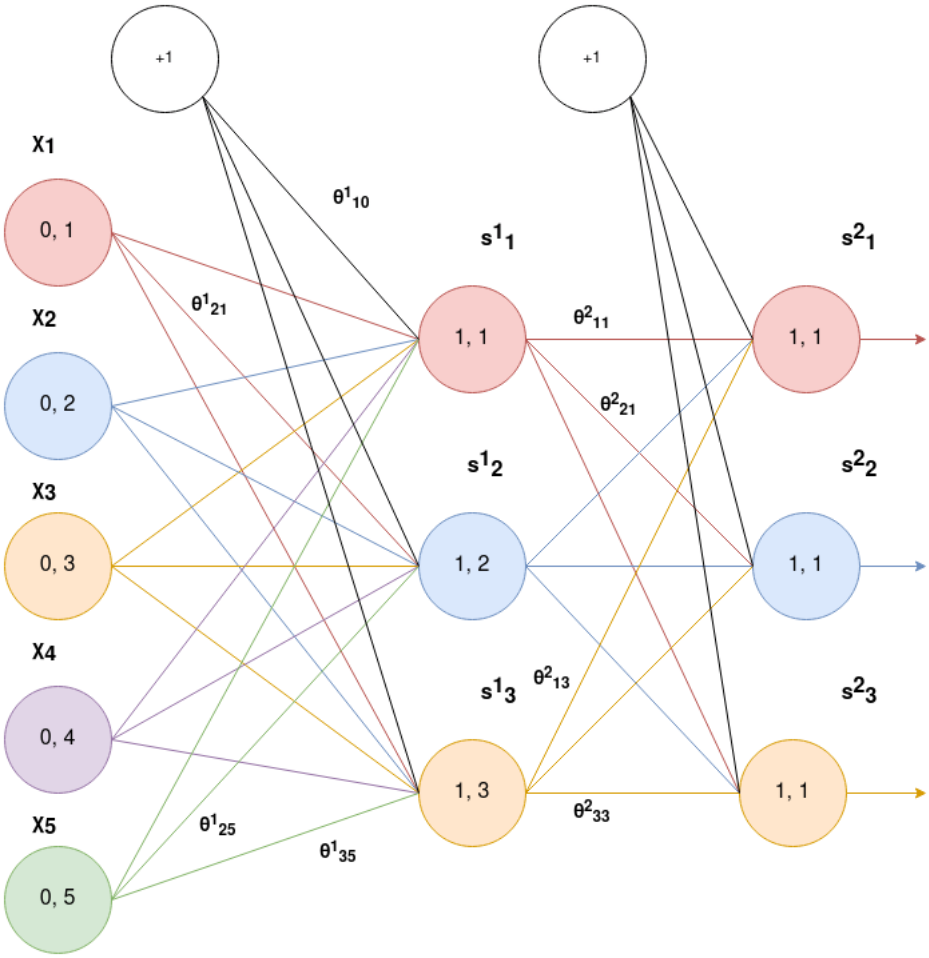


Figure 2.6: Visual representation of a MLP with a single hidden layer and three output units. Color on each connection matches the source neuron

$$\text{softmax}(z)_i = \frac{\exp(z_i)}{\sum_{l=1}^{k} \exp(z_l)} \tag{2.18}$$

The mathematical formulation for a binary classification MLP is defined as follows: given a training dataset $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ where $x_i \in \mathbb{R}^n$ and $y_i \in \{0, 1\}$, a one hidden layer perceptron computes the following function on its hidden layer's ith neuron:

$$s_i^1(x; \Theta) = g(\sum_{j=0}^{M_0} \theta_{ij}^1 x_j) \tag{2.19}$$

While the function computed on the ith neuron from the output layer is defined as:

$$s_i^2(x; \Theta) = g(\sum_{j=0}^{M_1} \theta_{ij}^2 s_j^1(x)) \tag{2.20}$$

For a neural network to adjust its parameters (θ_{ij}^m) and learn the target function, the most commonly used algorithm is the **Backpropagation**. This algorithm computes the gradient of the loss function regarding the network's weights. The weight actualization on the output layer ($1 \leq i \leq M_2$, $0 \leq j \leq M_1$) is defined as:

$$\begin{aligned} \Delta\theta_{ij}^2 &= -\rho \frac{\partial q_S(\Theta)}{\partial \theta_{ij}^2} = \frac{\rho}{N} \sum_{n=1}^{N} \sigma_i^2(x_n) s_j^1(x_n) \\ \sigma_i^2(x_n) &= (t_{ni} - s_i^2(x_n)) g'(\phi_i^1(x_n)) \end{aligned} \tag{2.21}$$

Where $q_S(\Theta)$ represents the loss function used, ρ is a hyper-parameter called learning rate, that is used to ponder the amount of change introduced in the weights at each backpropagation iteration, and σ_i^3 refers to the error, or difference, between the actual target t_{ni} and the network output value s_i^3.

The weight actualization on the hidden layers ($1 \leq i \leq M_1$, $0 \leq j \leq M_0$) follows the next formulation:

$$\begin{aligned} \Delta\theta_{ij}^1 &= -\rho \frac{\partial q_S(\Theta)}{\partial \theta_{ij}^1} = \frac{\rho}{N} \sum_{n=1}^{N} \sigma_i^1(x_n) s_j^1(x_n) \\ \sigma_i^1(x_n) &= (\sum_{r=1}^{M_2} \sigma_r^2(x_n) \theta_{ri}^2) g'(\phi_i^1(x_n)) \end{aligned} \tag{2.22}$$

Deep Learning

Deep Learning (DL) is the family of algorithms and applications that make use of DNNs. The most basic definition of a DNN is a ANN with multiple hidden layers. However, this definition is blurry, and there is no hard threshold differencing between a *deep* and a *shallow* ANN such as the MLP. Some studies such as [Winkler and Le, 2017] used one hidden layer as threshold, labeling as "deep" networks with more than one hidden layer. While in practice, the two-hidden layer perceptron was a common application. Due to the increase in computing power through the years [Mack, 2011] and therefore the capacity to increase the free parameters on the models, some practitioners use higher thresholds such as six hidden layers.

DL is the most recent chapter of the ANN's history, and the apparition of the field can be traced to some milestones. The first of them was the discovery of the gradient vanishing problem and how to solve it [Hochreiter, 1998]. When a neural network had more than two layers, the gradients, and the weight variations, calculated by the backpropagation algorithm decreased significantly. This meant that the first layers of the network could not learn. Another important landmark was the apparition of the first convolutional neural network [Fukushima and Miyake, 1982] and the publication of the Long short-term memory (LSTM) networks [Hochreiter and Schmidhuber, 1997]. Finally, due to the development of the Graphical Processing Units (GPUs) and the algorithms to take advantage of their high parallelism in 2012 appeared AlexNet [Krizhevsky et al., 2012], a convolutional DNN that outperformed the other competitors during the ImageNet Large Scale Visual Recognition Challenge.

Several mechanisms have been incorporated to DNNs to make them feasible to train and overcome the gradient vanishing problem. The most notorious is the use of the Rectified Linear Unit (ReLU) activation functions [Maas et al., 2013] (equation 2.23). ReLU is a non-linear function, so it allows the network to model complex functions by stacking layers. Moreover, the function is ranged between $[0, +\infty)$, hence allowing the gradient not to vanish as there is no saturation in any range of the function ·

$$\text{ReLU}\,(x) = \max\,(0, x) \tag{2.23}$$

Another important incorporation to solve the gradient vanishing problem is the Batch Normalization (equation 2.24). Batch Normalization is a technique employed to normalize the inputs of a layer which provides a couple of benefits to the training process. First, since the inputs are always normalized in the same numerical range, the layer can focus on learning the relations between the inputs rather than the adjacent layers. Second, batch normalization also significantly accelerates the training process, ensuring higher learning rates do not produce outlier activations.

$$\begin{aligned} \mu_{\mathcal{B}} &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^i \\ \sigma^2_{\mathcal{B}} &= \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}^i - \mu_{\mathcal{B}}\right)^2 \\ \hat{\mathbf{x}}^i &= \gamma\frac{\mathbf{x}^i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}}}} + \beta \end{aligned} \tag{2.24}$$

The last two deep-learning used during our research were: the leaky ReLU activation function and the Dropout method. Leaky ReLU (equation 2.25) presents a slightly slope on the negative side (ϵ) to address the "dying ReLU", when a negative activation become 0 and its unlikely to recover a non-zero value.

$$\text{LReLU}(x) = \max(0, x) + \epsilon * \min(0, x) \tag{2.25}$$

The Dropout method is applied to individual layers to deactivate them during the training (Figure 2.7). This makes the process noisier and forces the nodes to take on more responsibility for the inputs. Also, Dropout simulates a sparse activation, which encourages the network to learn a sparse representation as a side-effect. Therefore this mechanism can be interpreted as a regularization method.

2.2.5 Evaluation

Evaluating predictive models is a crucial part of the ML workflow and allows the researchers to estimate how good the models are fitted to predict the actual data. There are two basic types of validation, depending on the origin of the data, external and internal validations. External validation is referred to the use of new participant-level data, external to those used for model development [Riley et al., 2016]. This external data could be classified as *retrospective*, if the data already has been recorded

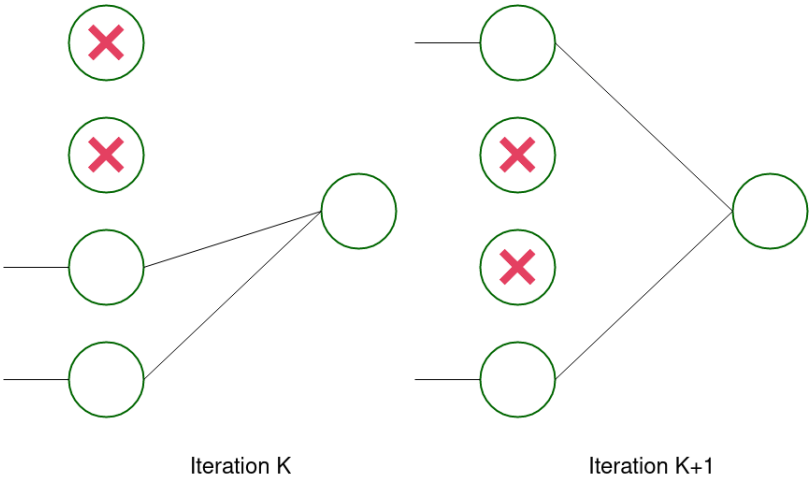


Figure 2.7: Dropout example over the last two layers of a ANN in two different iteration. Dropout rate is set to 0.5 so half of the neurons are deactivated in each iteration

or *prospective* if a pilot or a clinical protocol is in place to gather the data required for the validation.

In contrast, internal validation refers to evaluating the models using the development's dataset. Internal validation is always the first step towards model validation since it can estimate their predictive power. However, it cannot ensure the model's applicability to other data sources due to the difference in populations and data distributions. It is widely known that models should not be evaluated with the same data points used to train them, due to the *overfitting*[j], which produce over-optimist results. Several methods have been developed to split the dataset into train and test sets in order to perform more accurate results, some of them are: 1) **Holdout**, a simple method to split the data into two sets, usually a proportion of 80%/20%; depending on the number of data availability, using the bigger split to train the model and the other to evaluate it. 2) **K-Folding**, the data is split in k sets, usually between 4 and 10, and over k iterations, one of the sets is chosen to be validation and the others are used to train the models. This ensures that the model is evaluated with every data point (see Figure 2.8). 3) **Resampling** methods, such as bootstrap [Efron and Tibshirani, 1994], that performs a data split similar to holdout but uses the test split to sample different new sets of data and evaluate the model with them.

Classification metrics

Alongside these evaluation techniques, several metrics are needed to estimate the predictive power of the models in different dimensions. During this thesis we have focused on binary classification methods, as opposed to multi-class classification. Some of the most relevant metrics for binary classification are 1) **Sensitivity** (equation 2.26), also called true positive ratio, and measures the percentage of correctly classified samples among the ones predicted by the model to be positive. 2) **Specificity** (equation 2.27),

[j]The production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably. - Oxford dictionary

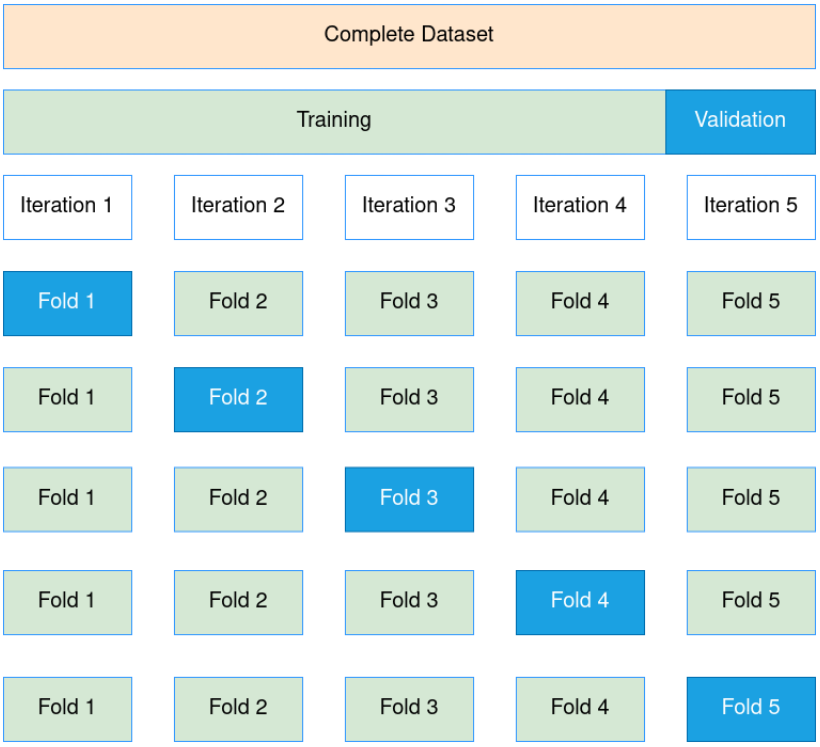


Figure 2.8: Visual representation of K-folding on a dataset using the parameter $K = 5$

or true negative ratio, which is the sensitivity counterpart for the samples predicted as negative. 3) **Balanced accuracy**, or Balanced Error Rate (BER) in its negative form, is the arithmetic mean of the sensitivity and the specificity (equation 2.28). All these metrics are based on the Confusion Matrix obtained when comparing the output of the classifier with the real class (see Figure 2.9).

		Predicted condition	
	Total Population (P + N)	Predicted Positive (PP)	Predicted Negative (PN)
Actual condition	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

Figure 2.9: Confusion matrix for a binary classification problem

$$\text{SEN} = \frac{TP}{TP + FN} \tag{2.26}$$

$$\text{SPE} = \frac{TN}{TN + FP} \tag{2.27}$$

$$\begin{aligned} \text{BAR} &= \frac{1}{2}\Big(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\Big) \\ \text{BER} &= 1 - \text{BAR} \\ &= \frac{1}{2}(\frac{FN}{TP + FN} + \frac{FP}{TN + FP}) \end{aligned} \tag{2.28}$$

The ROC curve is also a very relevant performance measurement in binary classification. Usually, the curve is plotted using the Sensitivity in the y-axes and $1 -$ specificity on the x-axes. From the curve is possible to obtain the AUC ROC statistic, also called C-statistic which is the area delimited by the curve in a 1x1 plot (Figure 2.10). The main property of this metric is that it doesn't depend on the classification threshold and represents how much the model can distinguish between two classes.

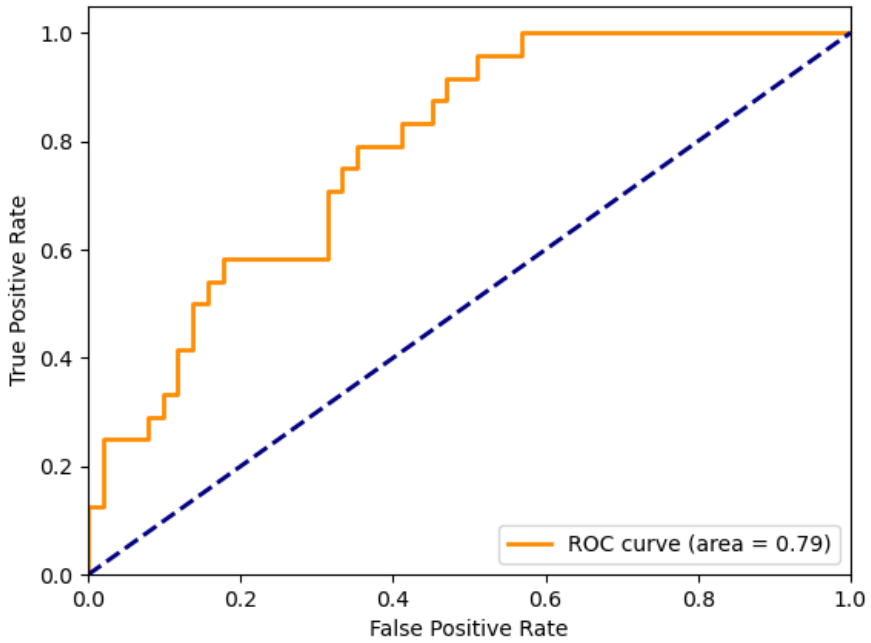


Figure 2.10: Example of ROC curve and AUC ROC statistic

The AUC ROC is closely related to the Mann-Whitney U statistic, and can be approximated using the formula in Equation 2.30

$$\text{AUC ROC} = \frac{1}{|P| * |N|} * \sum_{p \in P} \sum_{n \in N} S(p, n) \tag{2.29}$$

$$\tag{2.30}$$

Where P and N are the sets containing the probability assigned to the model to the true positive and true negative samples and the function S is defined as

$$S(p,n) = \begin{cases} 0 & p < n \\ \frac{1}{2} & p == n \\ 0 & p > n \end{cases} \tag{2.31}$$

Another classification metric used during this thesis is the **accuracy**, which measures the percentage of data points correctly classified (equation 2.32). This metric is also compatible with multi-class classification without complementary strategies such as pairwise analysis [Landgrebe and Duin, 2007].

$$\text{accuracy}(y,\hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}[y = \hat{y}] \tag{2.32}$$

Regression metrics

Regression problems have their own metrics, during this thesis we have focused on two of the most classical: **mean squared error** and **mean absolute error** (equation 2.33). Both metrics measure the average error between the real values and model prediction among all test data points.

$$\begin{aligned} \text{MSE} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ \text{MAE} &= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \end{aligned} \tag{2.33}$$

2.3 Clinical Decision Support Systems

2.3.1 Definition

Clinical Decision Support System (CDSS), also known as Computerized Decision Support System, is a widely used term within the medical informatics field. CDSS can be traced to the 1970s [Shortliffe and Buchanan, 1975]. Like many other concepts, there are different relevant definitions from various authors. Kawamoto *et al.*in 2005 defined a CDSS as *any electronic system designed to aid directly in clinical decision making, in which characteristics of individual patients are used to generate patient-specific assessments or recommendations that are then presented to clinicians for consideration* [Kawamoto et al., 2005]. Another definition states, *A Clinical Decision Support System (CDSS) is intended to improve healthcare delivery by enhancing medical decisions with targeted clinical knowledge, patient information, and other health information* [Sutton et al., 2020]. Therefore, we can determine that CDSSs are a vast category of information systems; however, they share common traits: a CDSS is a software, or

a software module, that using available information is meant to help HPs with the decision-making process.

Classic CDSS include features such as alerts, reminders, automatic calculations, or care summary dashboards that provide performance feedback on quality indicators and information retrieval tools for context-specific knowledge [Bright et al., 2012]. Their positive effect on the adherence to clinical guidelines increased patient safety, improved service quality, increased service time dedicated directly to the patients and decreased costs, among others [Sutton et al., 2020, Tundjungsari et al., 2017].

Regarding the human-system interaction, there is no single interaction process with a CDSS. Their usage is very context-and implementation-dependent. However, there are general theoretical approaches to the information flow between all the parts involved in the process [Yu et al., 2018]. In Figure 2.11 we can observe the three most common scenarios: a conventional clinical practice, a clinical practice where clinicians make use of EHR data and CDSS to get feedback which they can incorporate into their reasoning process and a final scenario where the CDSS is integrated within the EHRs.

2.3.2 Taxonomy: attributes and examples

CDSS could be classified in different axes, depending on the analysed dimension [Sim and Berlin, 2003]. However, the most common classification refers to how the information is represented internally. Sim and Berlin name this axis as *Reasoning Method*. The methods present in this axis can be split into two different categories: knowledge-based and non-knowledge based [Sutton et al., 2020]. Knowledge-based CDSS have the clinical information used to make decisions codified as "if-else" conditions. These systems represent the real information as states or *facts*, then an inference engine tries to match the available facts to the different rules composing the ruleset. The inference engine stops when no more rules can be applied or the trigger of a specific rule halts the process explicitly. The rule set incorporated into the system is usually created to match expert consensus from clinical guidelines and attention protocols. Due to its nature, this kind of system helps HPs to have better adhesion to standard practices. However, the approach is not free from pitfalls. The two most relevant to implementing and using the systems are the complexity derived from mapping clinical guidelines into a coherent and concise ruleset. Even when fuzzy logic inference engine [Anooj, 2012] is used on the system, a significant amount of effort designing and testing the system is required. The other main drawback of this approach is that no new information is extracted from the use of the system.

Probably the most successful knowledge-based CDSS are the subcategory known as Computerized Provider Order Entry (CPOE) [Khajouei and Jaspers, 2010]. CPOE systems allow physicians to introduce the prescribed medication to patients, and the system then performs a list of checks to ensure the safety and adhesion to the clinical protocol. These checks usually include but are not limited to dosage control, patients' allergies to prescribed medications and incompatibility or interactions between drugs. According to the different studies reviewed by [Khajouei and Jaspers, 2010] CPOEs have a significant impact on quality of drug management, reduced serious medical

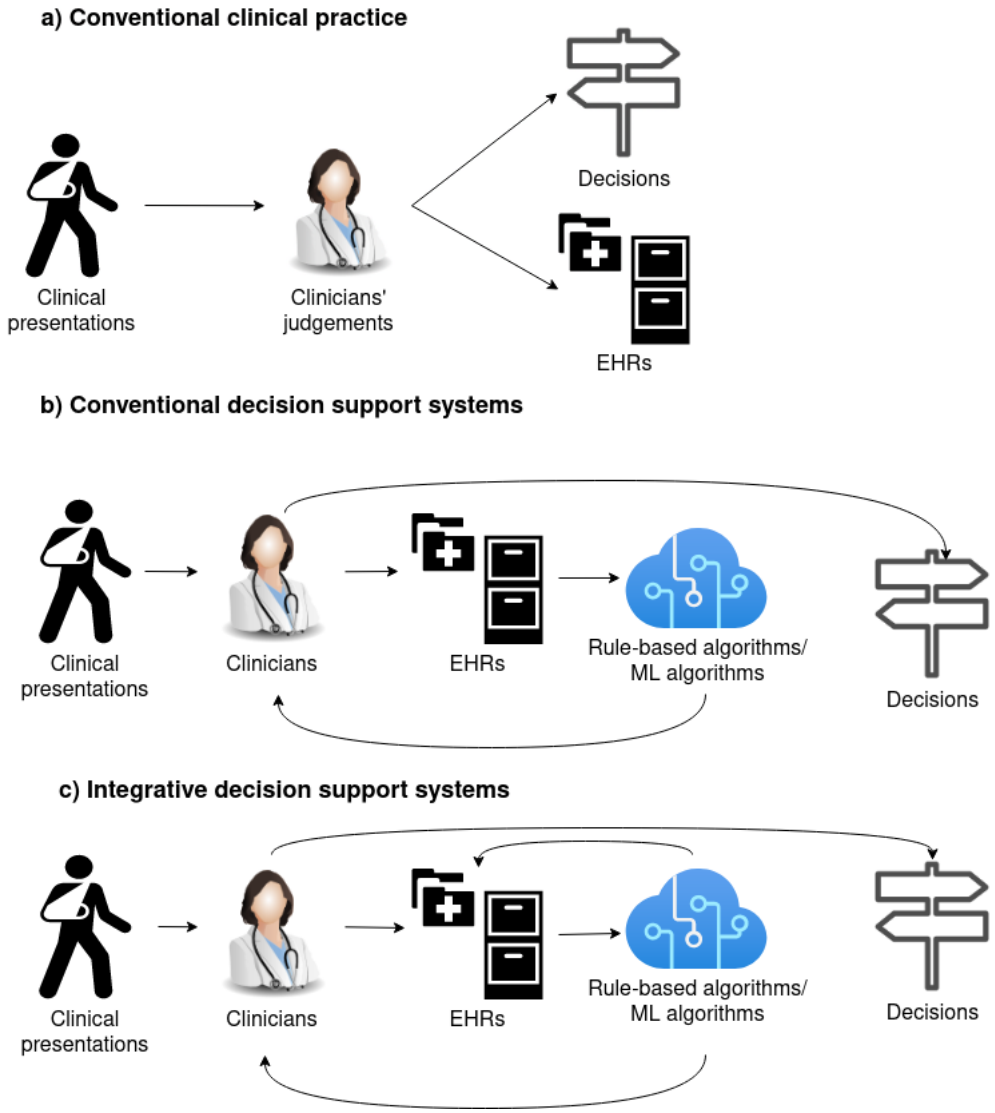


Figure 2.11: Model of information flow in clinical practice, as described by [Yu et al., 2018]. a) In a convention setting, the clinical case is presented to the physicians, which make their decision based on their reasoning and expertise. The information derived from this process is also registered on the EHR. b) In the conventional scenario using CDSSs, the information from the EHR is introduced to an algorithm, and their output is provided to the physicians, which use that output/prediction as part of their reasoning in order to make a decision. c) In this scenario, the CDSS is completely integrated, and it registers their outcomes on the EHR in addition to providing them to the physicians

.

errors and incomplete or inappropriate prescription and even decreased length of stay and costs.

On the other hand, ML-based CDSS are not dependent on hand-crafted rules adapted from the current medical knowledge, but they depend on rules extracted from clinical data. These models are trained using clinical data on specific problems and allow the different models to infer the relationship within the datasets, as seen in the previous section. However, these models are not free from pitfalls in their design and implementation. The system designers should ensure that the model has indeed

generalized appropriately from the data instead of using patterns derived from noise and bias collecting the data. Related to this problem, a very recommended practice when developing the models is to perform an external validation using data from other sources. The main objective of this testing is to ensure that the model has learned relevant medical data instead of patterns introduced by the protocols or idiosyncrasy from the original centres. Therefore, the process required to implement and ensure the correction of the models also requires a significant amount of resources.

There have been some successful and notorious examples of non-knowledge based CDSSs. One of the most successful fields for ML in clinical practice is the medical image. Samala *et al.*in 2016 developed a method to detect masses in breast tomosynthesis using convolutional neural networks and transfer learning [Samala et al., 2016]. In 2018 Poplin *et al.*created a system to predict cardiovascular risk factors using images from the retinal fundus and DL [Poplin et al., 2018]. In 2019 Khumancha *et al.*developed an application to detect lung cancer from computerized tomography scans also using convolutional neural networks [Khumancha et al., 2019]. Other applications based on data from the EHR have also been successful such as the short term prediction of mortality by Makar and colleagues [Makar et al., 2015] or the development of a score to predict adverse events such ICU transfer or death [Churpek et al., 2014].

Chapter 3

Comparative study of ML methods to predict one-year mortality

Palliative Care (PC) programs aim to improve the Quality of Life in patients with dire prognoses. However, the decision to start the palliative approach, alongside or detaching curative treatment, is not clear and therefore needs well-defined criteria. PC programs usually have limited resources; this fact makes the decision even more difficult. Currently, one of the most used criteria to detect poor outcomes and PC needs is the Surprise Question (SQ). Nevertheless, this method is based on the subjective opinion of the healthcare professional evaluating the case at a given moment. In this chapter, we propose the use of a data-driven approach, in the form of Machine Learning (ML) model constructed using data from the Electronic Health Record (EHR) from hospital admission, to provide a more objective alternative to the SQ. In addition, we compare the performance of different ML algorithms to the mortality prediction task.

The contents of this chapter were published in a scientific journal article by Blanes-Selva et al, (2021b) - thesis contribution ***P1***

3.1 Background and significance

An increasing number of people have multiple morbidities and conditions in the final moments of their lives, current medicine tries to maintain a quality of life of these people, including their needs in the final moments. In this situation, PC tries to facilitate the life of people in these conditions from a patient perspective.

PC is a multidisciplinary care that aims to grant comfort to the patient, avoid painful and/or aggressive treatments, alleviate pain, other symptoms, psychological and spiritual distress [Kelley and Morrison, 2015]. In addition, there are some studies which prove that patients receiving early PC present a better quality of life, mood, satisfaction with the treatment [Bakitas et al., 2009, Zimmermann et al., 2014, Temel et al., 2010] and even a longer survival when compared to patients whose PC was delayed [Bakitas et al., 2015].

A criterion for the PC inclusion is desirable as early as possible. An adverse event such as a hospital admission could be considered a convenient episode to check this

Approach	No. of features	AUC ROC	Reference
Buurman	4	0.77 (CI 95%, 0.72 - 0.82)	[Buurman et al., 2008]
PROFUND	9	0.70 (CI 95%, 0.67 - 0.74	[Bernabeu-Wittel et al., 2011]
HOMR	11	0.89 (CI 95%, 0.87 - 0.91) to 0.92 (CI 95%, 0.91–0.92)	[van Walraven et al., 2015]
Deep learning	13,654	0.93 (all); 0.87 (admitted patients)	[Avati et al., 2018]

Table 3.1: Summary of previous mortality studies

criterion. Nowadays, the main indicator to include a patient in PC is the clinical criterion of a potential *exitus* within the next 12 months. An example of that is the surprise question described in Moss et al [Moss et al., 2008].

Mortality forecast has been previously studied by other groups. Buurman *et al.* [Buurman et al., 2008] proposed a method for predicting 90-day mortality risk using few clinical features: Barthel index [Mahoney et al., 1965], Charlson score [Charlson et al., 1994], and Malignancy and Blood Urea Nitrogen (BUN) (mmol/L). The authors of this study calculated how modifications on the features affect the outcome. The study reported AUC ROC = 0.77 (CI 95%, 0.72 - 0.82). Bernabeu-Wittel *et al.* [Bernabeu-Wittel et al., 2011] proposed a method for detecting 1-year mortality for polypathological patients. That model computed the PROFUND score, based on some features to assign a mortality risk to the patient: Age, Hemoglobin, and Barthel index, No caregiver or caregiver other than the spouse, hospital admissions ≥ 4 in last 12 months and positive for few diseases. The PROFUND score is mapped into mortality (in less than a year) probability. The reported validation result was AUC ROC = 0.7 (CI 95%, 0.67 - 0.74). Van Walraven *et al.* [van Walraven et al., 2015] reported a 1-year mortality forecast model based on patient demographics, health burden, and severity of acute illness. The model uses a binomial logistic regression. The AUC ROC ranged from 0.89 (CI 95%, 0.87 - 0.91) to 0.92 (CI 95%, 0.91 - 0.92). Recently, Avati *et al.* [Avati et al., 2018] presented a deep neural network for 1-year mortality prediction by using 13,654 features, corresponding to the different ICD9 codes in different time windows through the year. They reported and 0.93 of AUC ROC for all validation patients but only 0.87 for admitted patients. A summary of previous studies is provided in Table 3.1.

Based on the promising results in the literature we have addressed the design of a high-performance predictive model of 1-year mortality exclusively based on observations at hospital admission. The overall aim of our study was to provide quantitative methods to healthcare caregivers to decide the inclusion of patients in the PC program during the hospital admission. To this aim, we have designed and evaluated five predictive models from the state-of-the-art machine learning discipline.

These models are meant to be more complex, in terms of algorithm, parameters and

amount of data needed than the first studies we presented but with less requirements that Avati's deep learning approximation, being the most adequate option for our dataset size.

The models presented in this work are continued in the InAdvance project (http://www.inadvanceproject.eu/) along with other kinds of models such as frailty and resource consumption models to create a complete CDSS [Mazzaglia et al., 2016] for the PC inclusion decision. This CDSS is meant to join other information systems created to improve the PC process, such as O'Connor *et al.* [O'Connor et al., 2009] and Dy *et al.* [Dy et al., 2011]

3.2 Materials

The data of the study was extracted from the Electronic Health Records from Hospital La Fe. We gathered all the hospitalization episodes of adult patients ($\geq$ 18 years old), excluding those related to mental health, gynecology and obstetrics, from January 2014 to December 2017 (a total number of 114,393 cases) that have been discharged from the hospital. All the patients received standard care, so no effects like a prolonged survival from PC [Bakitas et al., 2015] affect the data. To guarantee independent observations, we selected a random single episode for each patient, reaching a total of 65,279 episodes.

The dataset contains information about the previous and current admission (seven features), laboratory test results (seven features) and a list of 28 selected diseases for which the patient is positive or negative. Sex, age, Charlson index, and Barthel tests result are also available. This adds up a total of 36 features which can be obtained straightforwardly in the first hours of admission. Some of these features were used, with positive results, in previous studies.

Target variable was exitus after 1-year from the admission date. The number of patients that have died in less than a year (positive cases) was 8113 (~12.43%), the number of negative cases is 57,166 (~87.57%). The whole variable description can be seen in Table 3.2. The distributions for Admission Destination and Service are represented in Figures 3.1 and 3.2 respectively.

3.3 Methods

3.3.1 Development of the models

Five machine learning techniques were employed for developing our predictive models: Gradient Boosting Classifier [Friedman, 2001], Random Forest [Breiman, 2001], K-Nearest Neighbors [Cover and Hart, 1967], Multilayer Perceptron (MLP) [Hinton and Carbonell, 1990], and Support Vector Machine [Guyon et al., 1993]. The implementation of the scikit-learn toolkit [Pedregosa et al., 2011] was employed in all of them except in the MLP which uses Keras and TensorFlow [Abadi et al., 2016]. Moreover, the optimization tool TPOT [Olson et al., 2016] was used in order to find a good model to fit the data.

Variable	Types	Missing	Distribution
Sex	CAT	0	Males: 51.86%
Age	INT	0	61.33 ± 18.38
Urgent admission	BOOL	0	Yes: 56.97%
Admission destination	CAT	0	-
Service	CAT	0	-
Admission cause	CAT	0	-
Prev. stays	INT	0	6.119 ± 9.502
Barthel test	INT	56,214	67.268 ± 37.919
Prev. admissions	INT	0	0.300 ± 0.789
Prev. emergency room	INT	0	0.935 ± 1.691
Charlson score	INT	0	4.233 ± 3.238
Albumin (g/dL)	REAL	46,857	2.955 ± 0.677
Creatinine (mg/dL)	REAL	16,920	0.505 ± 1.063
Hemoglobin (g/dL)	REAL	14,434	11.703 ± 2.228
Leucocytes (Cel/mL)	REAL	14,434	9.457 ± 7.389
C-Reactive Protein (CRP) (mg/L)	REAL	30,285	63.083 ± 84.48
Sodium (mEq/L)	REAL	17,183	139.672 ± 4.35
Urea (mg/dL)	REAL	18,459	46.255 ± 34.63
Acute myocardial infarction	BOOL	0	Yes: 3.09%
Congestive heart failure	BOOL	0	Yes: 6.14%
Peripheral vascular disease	BOOL	0	Yes: 4.88%
Cerebrovascular disease	BOOL	0	Yes: 6.76%
Dementia	BOOL	0	Yes: 1.5%
Chronic pulmonary disease	BOOL	0	Yes: 10.03%
Rheumatic disease	BOOL	0	Yes: 1.6%
Peptic ulcer disease	BOOL	0	Yes: 1.57%
Mild liver disease	BOOL	0	Yes: 5.77%
Diabetes without complications	BOOL	0	Yes: 13.19%
Diabetes with complications	BOOL	0	Yes: 1.27%
Hemiplegia paraplegia	BOOL	0	Yes: 1.28%
Renal disease	BOOL	0	Yes: 7.46%
Malignancy	BOOL	0	Yes: 18.2%
Moderate severe liver disease	BOOL	0	Yes: 1.49%
Metastasis	BOOL	0	Yes: 3.27%
AIDS	BOOL	0	Yes: 0.57%
Delirium	BOOL	0	Yes: 0.12%

Table 3.2: Features information.

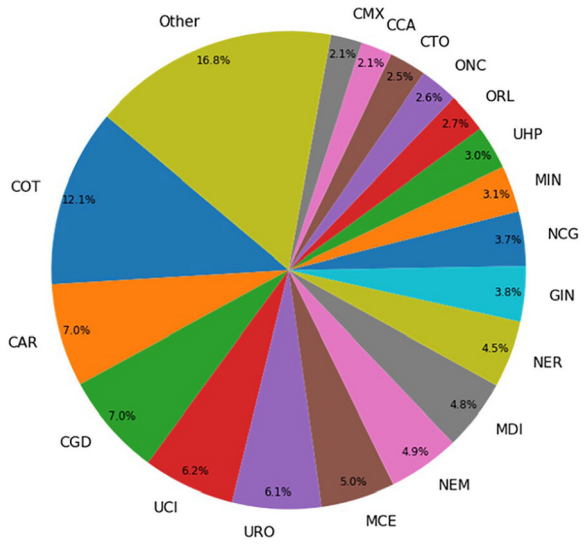


Figure 3.1: Distribution for admission destination. Legend: CMX = Maxilofacial Surgery, CCA = Cardiovascular Surgery , CTO = Thoracic Surgery, ONC = Oncology, ORL = Otorhinolaryngology, UHP = Hepato-pancreato-biliary Surgery, MIN = Internal Medicine, GIN = Gynaecology, NER = Neurology, MDI = Autoinmune Diseases, NEM = Pneumology, MCE = Short Stay Unit, URO = Urology, UCI = Intensive Care Unit, CGD = General Surgery, CAR = Cardiology and COT = Orthopedic Surgery and Traumatology.

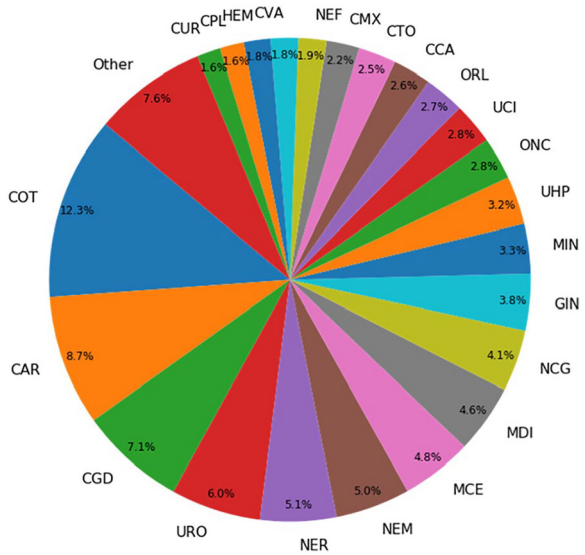


Figure 3.2: Distribution for service. Share the same codes as Figure 3.1, including CUR = Brachytherapy, CPL = Plastic Surgery, HEM = Hematology and Hemotherapy, CVA = Angiology and Vascular Surgery and NCG = Neurosurgery.

3.3.2 Feature importance

We studied the relevance of each feature for the final prediction by calculating the GINI importance provided by the Gradient Boosting Classifier. The GINI importance measures the average gain of purity by splits of a given variable. If the variable is discriminant for the problem, it tends to split mixed labeled nodes into pure single class nodes [Breiman et al., 2017].

3.3.3 Validation of the state-of-the-art models

As a first step, we have compared our model with the PROFUND and Buurman's model using the same evaluation method. For the Buurman's model, a clinical committee led by Vicente Ruiz-García at Hospital La Fe adapted the Buurman's proposal as 1-year mortality index, using a linear regression with the 1-year mortality target variable. Besides, we evaluated the original PROFUND model proposed in Bernabeu-Wittel *et al.* [Bernabeu-Wittel et al., 2011] The validation of the other models in state of the art was not possible due to the lack of part of their features in our data system.

3.3.4 Evaluation of the models

First, we have computed the ROC Curve [Guyon et al., 2008] for each model and calculated the optimum probability threshold (minimum probability to assign the positive class to a sample) running our models using a random split of the data from separating train and test. We iterated over all the different values that could change the specificity and the sensitivity of the model and kept the threshold that minimizes the balanced error rate (BER) [Guyon et al., 2008].

Once the threshold is established for each model, we internally validated them using a 100-repetition stratified hold-out (80% of the data in order to train the model and 20% for test it). The missing values have been imputed using an iterative approach that models each feature with missing values as a function of other features, available in scikit-learn as IterativeImputer [Van Buuren and Groothuis-Oudshoorn, 2011]. Five metrics have been stored for each experiment (accuracy, AUC ROC, specificity, sensitivity, and balanced error rate) [Breiman et al., 2017, Guyon et al., 2008]. For each metric, the mean and the 95% confidence interval have been computed.

In addition to this process, we performed a second round of experimentations, where we trained the models using a balanced set of data by oversampling the positive class using the SMOTE [Chawla et al., 2002] technique. This will help us to determine which are the effect of class imbalance in the ML algorithms for this problem.

3.4 Results

Following the above methodology, this section analyses the results of the proposed machine learning models, including a validation from Buurman's modified model and PROFUND from the literature. In terms of AUC ROC, the model based on GBM achieves the higher score with 0.91, being followed by the Random Forest and the

Multi-Layer Perceptron. Table 3.3 contains means and confidence interval at 95% for all proposed models and selected metrics. Table 3.4 provides the same experiments as staged before but using SMOTE as an oversampling technique for the positive class. Once again, the Gradient Boosting Classifier is the model with higher AUC ROC (0.908), followed by the Random Forest. Results using SMOTE are slightly worse on the selected metrics. As for the variable importance, Service, Urea and Leucocytes obtained the higher scores (10.6%, 10.32%, 8.65%), followed by CRP (7.65%) and Age (6.99%). Table 3.5 lists every variable and its Gini Importance.

3.5 Discussion

The prediction of death before 1 year could be a relevant criterion to admit the patients into palliative care programs [Moss et al., 2008]. Also, the prediction of the death at admission of the patient would help the hospital management to better manage its resources in a more accurate way.

We used the area under the ROC curve as the comparison metric because is the common metric to all other works in the State-of-the-Art (SoA). We also chose the threshold for considering a sample into the positive class taking the value that minimizes the balanced error rate.

The Buurman's modified model and the PROFUND index have been validated, the models described in our work outperform them in terms of AUC ROC, sensitivity and specificity. Whereas our models presented a bigger number of features (36) than the mentioned articles (four for Buurman's modified model and nine for PROFUND index).

Comparing with the most recent work, Avati *et al.* [Avati et al., 2018] that presented a neural network with 18 hidden layers of 512 neurons each was trained with 177,011 patients. The models in our approach are trained with 52,223 patients. The network used 13,654 features as input, our model uses only 36. Finally, they achieved an AUC ROC of 0.93 for all their patients but it only achieved 0.87 when only admitted patients are considered. We achieved better results using a significative smaller amount features, this led to a more compact model that also is more interpretable since the best performing model is based on decision trees.

The results obtained using the oversampling technique in the training set are very similar to the ones obtained in the original experiment. The main differences are in the optimal thresholds, which increased in all cases due to the change in proportion in the classes.

The best results in our models achieved the interval reported by van Walraven *et al.* [van Walraven et al., 2015]: 0.89–0.92 AUC ROC. Despite the number of final features is smaller in HOMR (10) two of their features are composed: "charlson comorbidity index score" (15 items) and "diagnostic risk score" (70 items), so at the end, HOMR requires more information about the patients than our models. The performance comparison with Avati *et al.* [Avati et al., 2018] and HOMR have been made using their reported results which implies the use of different evaluations and datasets.

We obtained consistent results compared to other studies. In HOMR the features that are capable to add more points to the index are the admitting service (up to

Model	Thres.	Accuracy [CI 95%]	AUC ROC [CI 95%]	Specificity [CI 95%]	Sensitivity [CI 95%]	BER [CI 95%]
Baseline: Buurman's mod.	0.101	0.794 [0.766, 0.821]	0.681 [0.680, 0.683]	0.838 [0.804, 0.873]	0.483 [0.460, 0.505]	0.340 [0.333, 0.346]
Baseline: PROFUND	0.058	0.569 [0.534, 0.603]	0.739 [0.738, 0.740]	0.534 [0.493, 0.575]	0.814 [0.800, 0.828]	0.326 [0.313, 0.339]
Gradient Boosting Machine	0.1	0.814 [0.812, 0.816]	0.91 [0.909, 0.912]	0.808 [0.806, 0.81]	0.858 [0.855, 0.861]	0.167[0.165, 0.169]
Random Forest	0.145	0.802 [0.801, 0.804]	0.9 [0.899, 0.902]	0.795 [0.793, 0.797]	0.851 [0.848, 0.855]	0.177 [0.175, 0.178]
k-Nearest Neighbours	0.08	0.755 [0.752, 0.757]	0.87 [0.868, 0.872]	0.742 [0.739, 0.745]	0.844 [0.839, 0.849]	0.207 [0.205, 0.209]
Support Vector Machine	0.105	0.686 [0.613, 0.76]	0.857 [0.855, 0.859]	0.664 [0.576, 0.753]	0.842 [0.811, 0.872]	0.247 [0.217, 0.277]
Multilayer Perceptron	0.15	0.797 [0.788, 0.806]	0.9 [0.897, 0.903]	0.789 [0.778, 0.8]	0.854 [0.845, 0.862]	0.179 [0.175, 0.182]

Table 3.3: Results for the machine learning proposed models.

Model	Thres.	Accuracy [CI 95%]	AUC ROC [CI 95%]	Specificity [CI 95%]	Sensitivity [CI 95%]	BER [CI 95%]
Baseline: Buurman's mod.	0.429	0.486 [0.351, 0.621]	0.688 [0.686, 0.691]	0.45 [0.28, 0.62]	0.737 [0.626, 0.847]	0.407 [0.377, 0.436]
Baseline: PROFUND	0.337	0.556 [0.5, 0.611]	0.765 [0.762, 0.769]	0.516 [0.449, 0.583]	0.835 [0.81, 0.86]	0.324 [0.303, 0.346]
Gradient Boosting Machine	0.142	0.813 [0.811, 0.815]	0.908 [0.906, 0.909]	0.808 [0.806, 0.81]	0.849 [0.845, 0.853]	0.171 [0.169, 0.174]
Random Forest	0.225	0.812 [0.811, 0.814]	0.895 [0.894, 0.896]	0.81 [0.808, 0.812]	0.832 [0.828, 0.836]	0.179 [0.177, 0.181]
k-Nearest Neighbours	0.39	0.765 [0.763, 0.767]	0.87 [0.869, 0.872]	0.756 [0.753, 0.759]	0.828 [0.824, 0.833]	0.208 [0.206, 0.21]
Support Vector Machine	0.462	0.772 [0.77, 0.774]	0.869 [0.867, 0.871]	0.765 [0.763, 0.767]	0.819 [0.815, 0.824]	0.208 [0.205, 0.21]
Multilayer Perceptron	0.42	0.776 [0.747, 0.805]	0.867 [0.855, 0.879]	0.772 [0.737, 0.807]	0.805 [0.785, 0.825]	0.212 [0.199, 0.224]

Table 3.4: Results using oversampling strategy for the positive class

Variable	Importance (%)	Variable	Importance (%)
Service	10.60	Malignancy	0.90
Urea	10.32	Sex	0.68
Leucocytes	8.65	Congestive heart failure	0.51
CRP	7.88	Renal disease	0.42
Age	7.65	Dementia	0.40
Creatinine	6.99	Chronic pulmonary disease	0.35
Albumin	6.66	Diabetes without complications	0.34
Prev. Stays	5.83	Acute myocardial infarction	0.30
Hemoglobin	5.67	Moderate severe liver disease	0.26
Sodium	5.07	Cerebrovascular disease	0.26
Charlson score	4.55	Mild liver disease	0.24
Admission destination	3.65	Peripheral Vascular Disease	0.22
Barthel test	3.20	Rheumatic Disease	0.21
Prev. emergency room	1.98	Hemiplegia Paraplegia	0.20
Cause of admission	1.71	Peptic Ulcer Disease	0.20
Prev. admissions	1.70	Diabetes with complications	0.17
Urgent admission	1.10	Delirium	0.16
Metastasis	0.95	AIDS	0.05

Table 3.5: Variable importance provided by GBC sorted by decreasing importance.

28) and the age × comorbidity (other 28 points). We agree with the most important variable (real service code) and the fifth one in importance order (age). Our second most important variable, nitrogen in urea, is included among the Buurmans model. Moreover, creatinine in blood is related to the BUN and is a variable also associated with mortality is our results.

It is known from the scientific literature cited above that morbidities, together with age and functionality (in our case measured by the Charlsons index) are the main predictors of mortality. However, our study on the importance of the variables shows that morbidities such as COPD score low, which is consistent with the van Walraven *et al.* [van Walraven et al., 2015] rule, which gives it only 2 points out of 35. Something similar happens with dementia, which scores only 3points.

The value of Urea and Creatinine indicate renal failure measured in two different ways, and we also know that these are predictors in other mortality prediction rules such as Buurman's *et al.* [Buurman et al., 2008] rule, where they appear as predictive with once again the functional situation.

Our models confirm that the disease variables weight the decision process, but they are less important when no other disease is present. The models also guide us to consider that administrative variables such as previous urgent admissions, previous stays and destination of admission may have an additional weight not considered until now that contribute more to mortality than the fact of suffering a specific disease. Having this information updated permanently will allow for the adjustment of additional risks for patients who are acutely admitted to hospital through the Emergency Services.

We compared the method we have used to assigning importance to the different variables in the dataset, the GINI importance provided by the Gradient Boosting Classifier, to other methods based on features ablation. We used a technique called Feature ranking with RFE from scikit-learn [Pedregosa et al., 2011]. The RFE starts

training a model and obtains the variable importance. The least important variables are removed from the dataset, and the process continues recursively until a minimal set of variables is obtained. The results obtained were almost identical to the ones presented in this work, so we conclude that, in addition to its reputation in other areas such as machine learning interpretability, GINI importance is a robust indicator to feature's significance in the model.

The clinical features included in our work have clinical relevance and appear in other clinical prediction rules. They appears in the records of our hospital databases in Spain and allow the creation of alerts for the clinicians to address patients, to palliative care programs not only for advanced oncology patients but for other chronic pathologies as dementia (a critical literature review exploring the challenges of delivering effective palliative care to older people with dementia, cardiac failure, or COPD or end-stage renal disease)[Klinedinst et al., 2019, Bostwick et al., 2017, Birch and Draper, 2008].

This study has caused a direct impact on Hospital La Fe since the model based on the Gradient Boosting Classifier has been implemented in the pre-production information systems and it is on a test stage. Once in the day, a program gathers all the admitted patients' data and extracts the features, this information is passed to the model who gives a posteriori probability and a label prediction, this information is stored on a separated table of the same database including the timestamp.

The main limitation of the study was the use of data from only one hospital, we can't ensure that the models learned with the study population are effective with patients of another country/region, or another type of hospital, Hospital La Fe is a tertiary Hospital a referral in the Valencia region, with different patients and severity.

In addition, the models had only an internal derivation, so we need to refine and validate this model to reproduce the findings with different settings (smaller hospital and with less severity illness) perhaps outside the same city or Valencian community, where we can have a population with different habits such diet or lifestyle. It is necessary to work on additional criteria for palliative care admission besides mortality, for example, introducing the available resources in the decision-making process. Also, an inclusion criterion for chronic patients is needed since their illness trajectories are different from other patients [Murray et al., 2005].

The work presented in this article is being continued in the InAdvance project, a European project about palliative care which aims to research and standardize some palliative care procedures as well as other aspects such as admission criteria. Our work in this project is to develop predictive models for mortality, frailty and resource consumption which integrated into CDSS could be proven as reliable and objective inclusion criteria for palliative care. This PC criterion is intended to be unbiased and independent from the place the care takes place, so other institutions such as hospices or nursing homes can take advantage of them with the only restriction of having access to the variables the model needs to run.

3.6 Conclusion

This work proposes machine-learning forecast of 1-year exitus using data from hospital admission. Our forecast achieved an area under ROC curve of 0.9 and a BER of

0.17, being the Gradient Boosting Classifier the best model. The features used in the models correspond to basic demographic and administrative information, some laboratory results and a list of positives or negatives for certain diseases. The presented models could have an instant impact on every hospital, only the feature extraction module and the table for results need to be adapted to the particular information system of every hospital, the rest of the components are ready to set in production. Our results have reached the best results in the state-of-the-art, corresponding to the HOMR index which validation in few Canadian hospitals produces AUC ROC from 0.89 to 0.92.

Chapter 4

Frailty and mortality predictive models for older patients

With the current trend toward global ageing, it is not surprising that the need for PC has also increased. Since most people in need of PC are over 65 years old, it makes sense to design a series of tools that focus on that group of patients, trying to improve the precision among that segment. Moreover, in this chapter, we continued our previous work but focused solely on older patients. In addition to developing 1-year mortality models, we tried to estimate better the survival and create a regression model to complement the prediction. However, our main contribution is the creation of a frailty index to develop ML predictive models to assess frailty status and use it alongside mortality to help on the PC referral.

The contents of this chapter were published in a scientific journal article by Blanes-Selva et al, (2022a) - thesis contribution ***P2***

4.1 Introduction

Palliative Care (PC) is a holistic approach that improves patients' quality of life with life-limiting diseases. It is recommended to incorporate early in the disease trajectory, even in conjunction with potentially curative treatments [Callaway et al., 2018]. PC can improve quality of life [Temel et al., 2010], mood [Bakitas et al., 2009], symptom control [Yennurajalingam et al., 2011], reduce emergency department visits and hospitalisation [Quinn et al., 2020], and even increase 1-year survival [Bakitas et al., 2015].

PC services have traditionally been mainly accessed by cancer patients, but there is growing consensus about the importance of promoting access for patients with non-malignant disease at earlier stages [McIlfatrick, 2007, Higginson and Addington-Hall, 2001, Kingston et al., 2020]. Patients' prognoses and functional decline are two crucial elements in decision-making to be considered by healthcare professionals in the introduction of PC need assessment and PC conversations with older people.

On the one hand, it is estimated that at least 75% of patients would benefit from access to PC during their terminal illness [Etkind et al., 2017]. Nevertheless, uncertainty about prognostication is cited as a common barrier to PC referral, particularly for

patients with non-malignant diseases [Murray et al., 2015]. On the other hand, frailty in older patients is defined as a state characterised by reduced physiological reserve and loss of resistance to stressors caused by accumulated age-related deficits [Clegg et al., 2013]. Two of the most popular frailty dimensions are the frail phenotype by Fried *et al.* [Fried et al., 2001], which describes frailty as a biological syndrome; and the Frailty Index (FI) by Mitnitski *et al.* [Mitnitski et al., 2001], which is based on health deficits accumulations, also, frailty has been defined since a more comprehensive approach taking into consideration a holistic understating of the person. In this sense, frailty can be experienced by a decrease in human functioning at the physical level and psychological and social domains [Gobbens et al., 2010]. Raudonis *et al.* [Raudonis and Daniel, 2010] suggest in their study that frail older adults could benefit from involvement in PC programs as frailty is associated with poor health outcomes and death [Koller and Rockwood, 2013].

Different strategies have been used to try to aid prognostication. Clinical intuition was harnessed with the Surprise Question ("Would I be surprised if this patient died in the next year?") which, has been promoted as a tool to prompt clinicians to recognise patients with a limited prognosis [Moss et al., 2008]. However, in 2017 Downar *et al.* [Downar et al., 2017] published a systematic review of the surprise question, concluding that more accurate tools are required given its poor to modest performance as a mortality predictor. Also, it has been demonstrated that the risk of death increases with lower performance levels and with falling performance levels, but survival data varied across different healthcare systems [Linklater et al., 2012]. In this line, the Supportive and Palliative Care Indicators Tool (SPICT) proposes a set of clinical indicators of poor prognosis developed through a consensus of expert opinion [Highet et al., 2014], which has shown to have a predictive accuracy of up to 78% [Woolfield et al., 2019]. Other studies have used data analysis to propose alternative tools to predict short-term mortality. Bernabeu-Wittel in 2010 developed the PROFUND index [Bernabeu-Wittel et al., 2011], a predictive model for patients with multimorbidity. Van Walraven *et al.*in 2015 proposed HOMR [van Walraven et al., 2015], a tool for predicting one-year mortality in adults ($\geq$18 years and $\geq$ 20 years for the different cohorts). In 2018 Avati *et al.* [Avati et al., 2018] proposed a deep learning approach to identify patients with a survival between 3 and 12 months a, in 2019 Wegier *et al.* [Wegier et al., 2019] proposed a version of HOMR but using only variables available at the admission. In 2021, our team also presented a one-year mortality model for adults [Blanes-Selva et al., 2021a].

Additionally, and as stated before, quantifying frailty is important since as patients become frail, advance care planning conversations should be prioritised to establish patient goals and wishes in advancing serious illness [Porter et al., 2020], which may include the involvement in PC programmes. A wide array of frailty indexes has been proposed to assess the health status of older adults. The frailty index has been used to predict mortality and poor health outcomes [Shamliyan et al., 2013]. Some studies have tried to predict frailty status: Babič *et al.*in 2019 [Babič et al., 2019] use a clustering approach to identify clusters considering the prefrail, non-frail and frail status using ten numerical variables for adults over 60 years old. Sternberg *et al.* [Sternberg et al., 2012] in 2012 tried to identify frail patients with their methods against the VES frailty

score [Saliba et al., 2001] for patients over 65 years old. Bertini *et al.* [Bertini et al., 2018] in 2018 created two predictive models for patients over 65 years old: one to assess frailty risk using the probability of hospitalisation or death within the year and a second one to assess worsening risk to each subject in the lower risk class.

Based on these previous results, our aim in this work is to propose a set of machine learning tools capable of making predictions about mortality and frailty for older patients, oncological and non-oncological, so healthcare professionals can benefit from quantitative approaches on data-driven evidence when deciding advance care planning. In this sense, we propose the creation of three different but complementary models: a) a one-year mortality classifier that will work as a binary predictor; b) a survival regression model aimed to obtain a prediction in days from admission to death; and c) a one-year frailty classifier to predict the health status, assessed by the Frailty Index, of a patient one year after admission. The authors consider that the combination of both mortality and frailty criteria, working as complementary information sources, can positively impact detecting needs to start PC conversations.

4.2 Materials

4.2.1 Basic description

Data was extracted from the system on Nov 1 2019. The dataset contained hospital admissions records for older patients (age $\geq$ 65) from Jan 1 2011, to Dec 31 2018. Patients admitted to psychiatry and obstetrics services were excluded from the study.

Data contains a total of 39,310 hospitalisation episodes corresponding to 19,753 unique patients. The cohort was composed of 9780 males and 9973 females with a mean age of 80.75 years (see Table 4.1).

Sex	N	Mean Age (years)	STD Age (years)
Female	9973	80.75	8.67
Male	9780	77.44	8.24
All	19753	79.11	8.62

Table 4.1: Patient demographic information.

4.2.2 Mortality target variables

Mortality target variables were extracted from administrative admission data and the recorded death date of regional civil registration. Patients alive during data extraction were censored for the regression problem due to our inability to know their survival time from admission. However, patients alive with an admission date prior to Nov 1, 2018 (one year prior to the extraction) could be included since we could determine their mortality status within the year.

4.2.3 Frailty target variable

As for the frailty target, following the work of Searle *et al.* [Searle et al., 2008], we calculated the FI of every episode (admission frailty) and sorted them chronologically. The target FI of a given episode was the admission frailty of the following episode if this next episode happened within the year. We used the most recent episode as the target if a patient had multiple admissions during the following year. Otherwise, target frailty was set to the same value as the current admission frailty. Most recent episodes and patients with only one episode were removed because no posterior data was available, so we considered them as censored data. Figure 4.1 presents an example of target FI calculation for each possible situation.

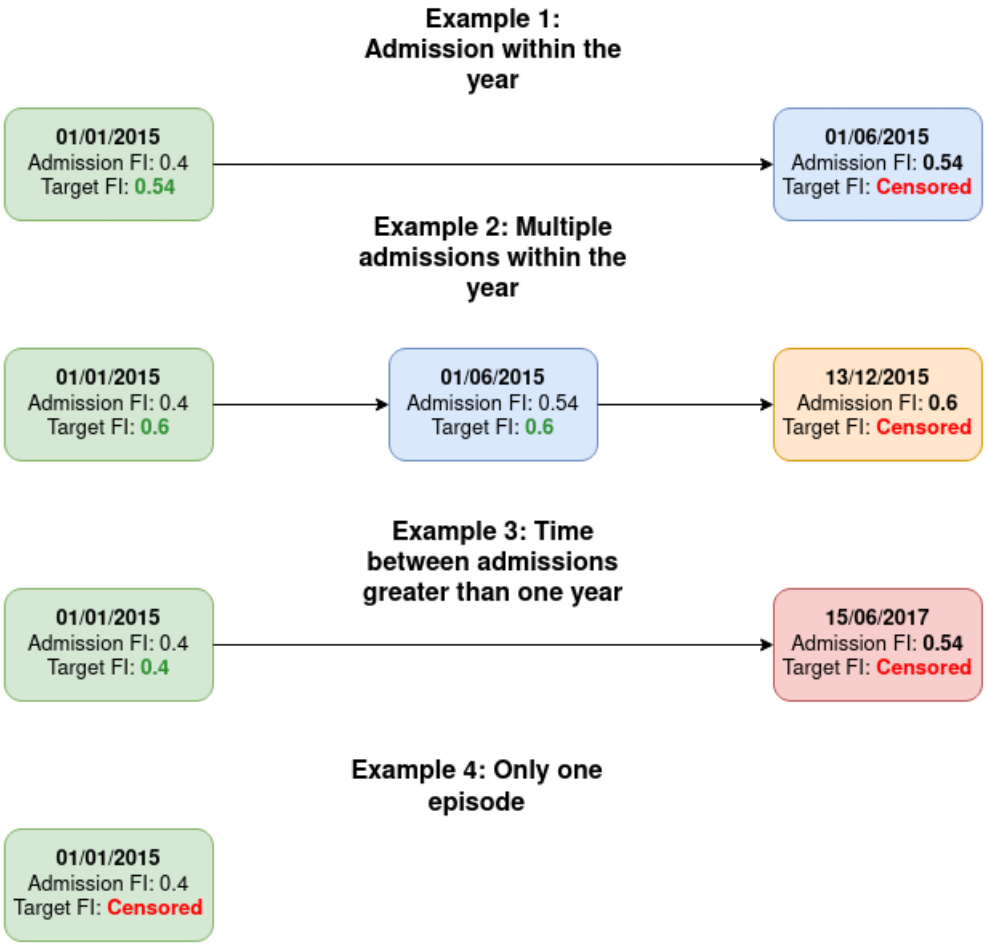


Figure 4.1: Visual representation of the algorithm to calculate the target FI in all four possible situations.

Finally, we stratified the FI into four categories according to the work of Hoover *et al.* [Searle et al., 2008] and aggregated the two less severe frailty conditions (Non-Frail + Vulnerable) and the two more frail statuses (Frail + Most Frail). Variables used in the frailty index are listed in Table 4.2 and were extracted as part of the original 147 variables.

Variables	N		Variables	N	
	Yes	No		Yes	No
Difficulties in dressing	3829	35481	Difficulties in urinating	3683	35627
Difficulties in bathing	5999	33311	Difficulties in stooling	3121	36189
Difficulties in grooming	5242	34068	Difficulties in eating	2965	36345
Difficulties in moving	3398	35912	Hypertension	30975	8335
COPD	9724	29586	Heart Failure	13228	26082
Stroke	9828	29482	Parkinson	1655	37655
Thyroid Disorders	4538	34772	Diabetes Mellitus	15910	23400
Gastro. or liver disease	27401	11909	Musculoskeletal Diseases	24330	14980
Dementia	4479	34831	Malnutrition	2718	36592
Pressure Ulcers	1886	37424	Anaemia	12546	26764
Hear impairment	6777	32533	Gastrointestinal problems	12567	26743
Chronic renal failure	8679	30631	Depression	587	38723
Cancer	16536	22774	Constipation	5088	34222
Atrial fibrillation	12434	26876	Visual impairment	20100	19210
Psychiatric disease	19436	19874			

Table 4.2: List of variables included in the frailty index and their distribution. All variables are binary, and their distribution represents the condition's presence (Yes) or absence (No).

4.2.4 Data censoring and distributions

After data censoring, the one-year mortality target variable distribution was: 24985 (65.83%) episodes were negative cases (time to exitus > 365 days), and 13431 (34.17%) episodes were positive (time to exitus <= 365 days) as shown in Figure 4.2A. The survival regression target variable (20959 episodes; mean 368.59; range [0 to 3033]) presents a right-skewed shape, as can be observed in its density plot in Figure 4.2B.

The admission FI (mean 0.27; std 0.12) and the FI target variable (22859 episodes; mean 0.32; std 0.14), resembled a slightly skewed normal distribution (plot in Figure 4.2C and Figure 4.2D), while the distributions of the different categories are: Non-Frail 986 (2.2%), Vulnerable 10911 (24.34%), Frail 25638 (57.19%), Most Frail 7294 (16.27%). As aggrupation of two categories: Non-Frail + Vulnerable 11897 (26.54%), Frail + Most Frail 32932 (73.46%), data represented in Figure 4.2E.

4.3 Methods

4.3.1 Predictive models

As the first approach for predictive models, we have selected the Gradient Boosting Machines (GBM) [Friedman, 2001], which can be used for classification and regression. GBMs are ensemble models composed of decision trees. This model follows an iterative training algorithm. In each step, the tree that minimises the selected loss function is added to the ensemble until the hyperparameter setting the number of trees is reached. The GBM models are known for their notable performance on

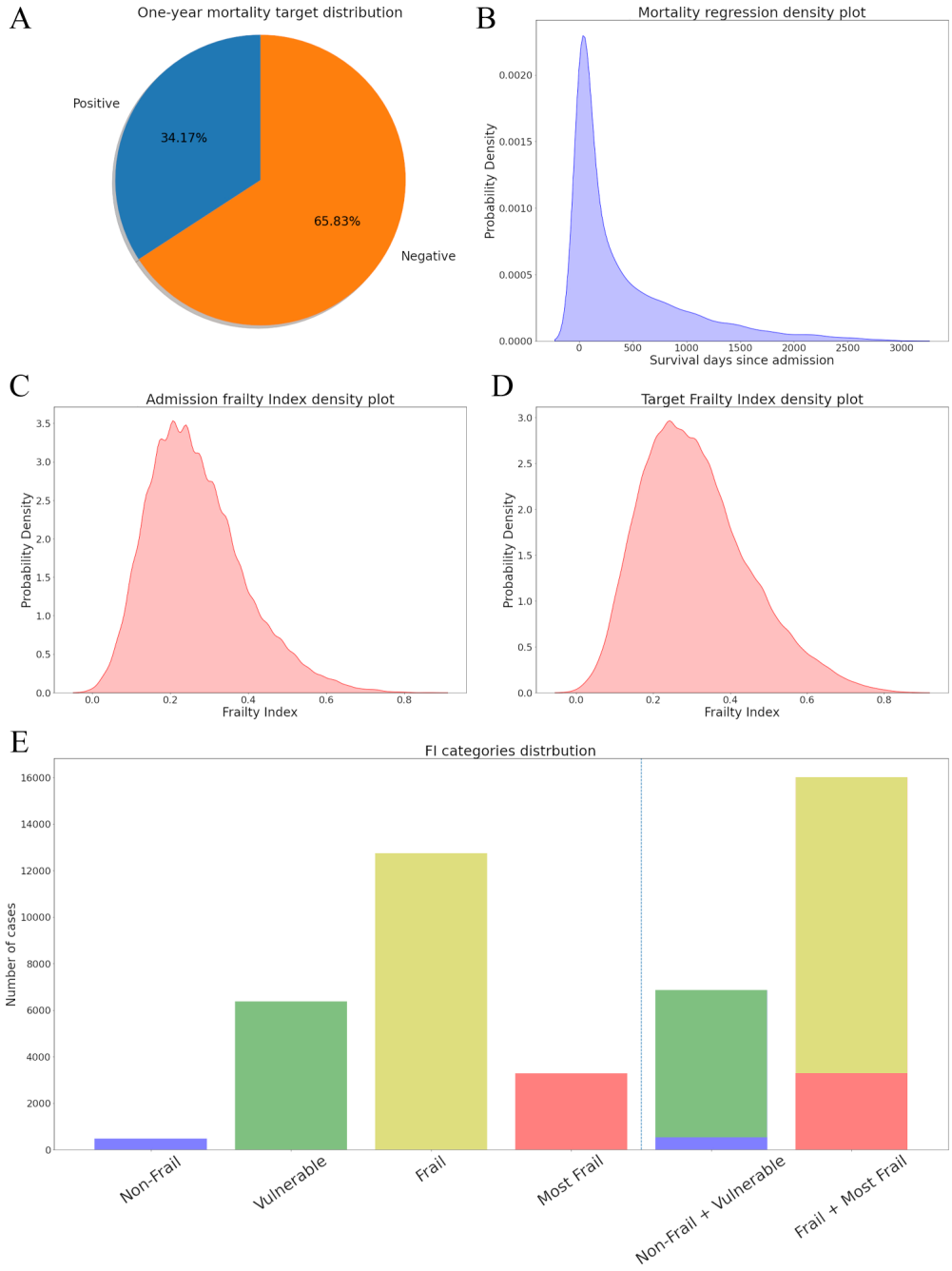


Figure 4.2: A) One-year mortality target distribution; B) Density plot from survival regression target variable; C) Density plot from the FI target variable; D) Density plot from the admission FI; E) FI categories distribution.

different problems [Touzani et al., 2018, Chen et al., 2018, Zhou et al., 2019].

Our second approximation to the predictive models is through the Deep Neural Network (DNN) [Hinton and Carbonell, 1990]. Due to the tabular nature of the data, we are using a multilayer perceptron topology, which is composed of interconnected neurons. Weights connect the neurons, and their output is a function of the sum of the inputs to the neuron, applying a non-linear activation function afterwards [Gardner

and Dorling, 1998]. Our models are using Batch Normalisation [Ioffe and Szegedy, 2015] and Dropout [Srivastava et al., 2014] as regularisation methods and the Leaky ReLU [Maas et al., 2013] function as activation function. Deep learning has been a trendy technology when dealing with the increasing amount of data, and its application to medicine is growing [Piccialli et al., 2021].

4.3.2 Hyperparameters and variable selection

To select the hyperparameters and make the selection of variables, we split the datasets (80%/20%) into a design set and an evaluation set. Then, we used a recursive feature elimination process as a filter method on the design set. This process starts with the whole set of features, trains a tree-based model, and calculates each variable's relevance using the GINI importance [Nembrini et al., 2018], which measures the average gain of purity in the tree splits. Finally, less relevant variables are eliminated. The process is repeated until the desired number of features is obtained. The number of variables was set to 20 in each task, a number of variables able to be handled by a human operator, with two variables eliminated each iteration. Table A.1 on Annex A describes each selected variable.

The selection of hyperparameters for each model was performed using the Optuna optimisation library [Akiba et al., 2019]. Using this approach, we selected the most relevant hyperparameters for the GBM and the DNN and provided feasible ranges. During the process, the method selects a value for each hyperparameter, trains the model with 80% of the design set, and evaluates it with the remaining 20% and the appropriate metric. As more iterations occur, Optuna makes a smarter selection of the hyperparameters until the algorithm reaches a selected number of iterations. The hyperparameters used in each model can be consulted in Table A.2 in Annex A.

4.3.3 Evaluation

We used the bootstrap technique [Efron and Tibshirani, 1994] to evaluate the models with 1000 resamples on the unseen evaluation set. To evaluate the performance of the one-year mortality and the frailty binary classifier, we selected the following metrics: area under ROC curve (AUC ROC), accuracy, sensitivity (or True Positive Rate) and specificity (or True Negative Rate). We selected the mean absolute error (MAE) for the survival regression model. In addition, we repeated the regression experiments using only those cases where the prediction is < 500 days. In addition, since the GBM is an explicable model, we reported the contribution of each variable in percentage.

4.3.4 Comparison with baseline models

To compare our mortality regression model with state of the art, we have performed survival analysis over the data processed with the same pipelines described above. We chose the Cox regression model [Cox, 1972], from which we obtained survival estimations for patients by calculating the survival expected time. We trained a binary Logistic Regression to compare the classification models for both mortality and frailty.

4.3.5 Software

The whole experimentation described in this work has been carried out using the python 3 programming language [VanRossum and Drake, 2010], and the following scientific libraries and packages: numpy as the main mathematical library [Harris et al., 2020], pandas' data frames to handle the data representation [McKinney et al., 2010], scikit-learn's implementation of GBM [Pedregosa et al., 2011], Pytorch's DNN implementation [Paszke et al., 2019], Optuna as hyperparameters selection [Akiba et al., 2019] and lifelines' implementation of the Cox model [Davidson-Pilon, 2019].

4.4 Results

4.4.1 Associations between distributions

The Spearman's correlation coefficient between the survival target in days and the admission FI was -0.10 while the correlation between survival and the target FI was -0.16; both correlations were statistically significant ($p < 0.001$). The similarity between the binary 1-year mortality target and the binary FI target was studied using the Chi-Squared test. However, we had to reject the null independence hypothesis ($p < 0.001$), and therefore it exists a similarity between both binary variables.

4.4.2 One-year mortality classifier

GBM and DNN performed very closely (0.87 Confidence Interval (CI) 95% [0.86, 0.87] and 0.86 CI 95% [0.85, 0.86] AUC ROC), both outperforming the logistic regression baseline, complete results and metrics on Table 4.3.

Model	AUC ROC	Sensitivity (TPR)	Specificity (TNR)	Accuracy
GBM	0.87 [0.86, 0.88]	0.78 [0.76, 0.82]	0.79 [0.75, 0.81]	0.79 [0.77, 0.80]
DNN	0.86 [0.85, 0.86]	0.79 [0.74, 0.83]	0.76 [0.71, 0.81]	0.77 [0.75, 0.79]
Log. Reg.	0.80 [0.79, 0.81]	0.75 [0.63, 0.80]	0.69 [0.64, 0.81]	0.71 [0.69, 0.75]

Table 4.3: One-year mortality classifier evaluation. Reporting the mean and the 95% CI

4.4.3 Survival regression

The cox regression produced a MAE of 444.8 days while the GBM and the DNN model achieved a MAE of 333.13 and 338.88 days, respectively. The GBM outperformed the other models when using only samples with survival < 500, complete performance for survival regression models on Table 4.4.

4.4.4 One-year frailty

The classification model based on the logistic regression achieved an AUC ROC of 0.84, while the GBM and DNN outperformed it with an AUC ROC of 0.89. Complete

Model	MAE	MAE (<500d)
GBM	333.13 [323.10, 342.49]	94.67 [92.02, 97.49]
DNN	338.88 [329.07, 349.37]	103.21 [100.47, 106.08]
Cox	444.80 [438.90, 450.90]	116.71 [115.23, 118.08]

Table 4.4: Mortality regressor evaluation. Reporting the mean and the 95% confidence interval.

metrics for the frailty classification are available in Table 4.5.

Model	AUC ROC	Sensitivity (TPR)	Specificity (TNR)	Accuracy
GBM	0.89 [0.88, 0.90]	0.77 [0.73, 0.81]	0.85 [0.81, 0.89]	0.79 [0.78, 0.81]
DNN	0.89 [0.88, 0.90]	0.76 [0.72, 0.83]	0.85 [0.78, 0.89]	0.79 [0.77, 0.82]
Log. Reg.	0.84 [0.83, 0.85]	0.74 [0.70, 0.78]	0.78 [0.73, 0.83]	0.75 [0.73, 0.77]

Table 4.5: Frailty classifier evaluation. Reporting the mean and the 95% confidence interval.

4.4.5 GINI Importances

Following the previous methodology, we have calculated the GINI importance for each of the GBM predictive models. For the one-year mortality model, the most important variables were: Number of Active Groups, Charlson Index and Age. In the regression task: Number of Active Groups, Charlson Index and Service whereas in the model version including only cases with survival < 500 days were: Leukocytes, C-reactive protein and Urea. Finally, the most relevant features in the frailty model were the Charlson Index, Number of previous Emergency Room visits and Hypertension. Complete details in Table 4.6.

4.5 Discussion

The overall aim of this study was to develop machine learning models capable of making predictions about mortality and frailty focused on older adults so that health professionals can benefit from quantitative approaches based on data-driven evidence. We have developed a ML model to predict frailty status within the year without using other problems as proxies. Regarding the mortality criterion, and despite different approximations to this task in the literature, we decided to focus on older patients to be more specific within this age group.

Our one-year mortality model ranked among the best general admission models in terms of AUC ROC (0.87 CI 95% [0.86, 0.88]). Outperforming PROFUND (0.77) [Bernabeu-Wittel et al., 2011], scoring slightly below HOMR (0.89-0.92) [van Walraven et al., 2015], mHOMR (0.89) [Wegier et al., 2019] and our previous work [Blanes-Selva et al., 2021a]. However, the results are in the same range as Avati's deep learning approach (0.93, 0.87 for admitted only patients) [Avati et al., 2018]. However, our

Variable	GINI 1YM(%)	GINI Reg.(%)	GINI Reg.<500(%)	GINI Frailty (%)
Charlson Index	**14.45**	**8.40**	1.65	**29.86**
Number Active Groups	**16.28**	**12.97**	3.24	-
Service	7.66	**10.46**	8.37	2.05
Leukocytes	5.58	6.78	**14.41**	0.60
Age	**9.36**	5.83	3.25	4.51
Barthel Index	6.23	6.09	5.43	5.10
Urea	5.28	4.62	**9.62**	0.83
Number Previous ER	1.04	4.42	2.25	**10.9**
C-Reactive Protein (CRP)	2.68	4.45	**10.31**	-
RDW-SD	4.80	3.20	4.77	2.28
DRG	3.89	4.05	4.92	1.16
Admission Diagnose Code	1.78	7.26	2.96	0.63
Glucose	2.20	2.18	5.89	1.63
RDW-CV	2.80	2.88	3.60	2.00
Creatinine	2.42	2.57	3.49	1.65
Number of previous stays	1.60	2.56	5.20	0.72
Hypertension	-	-	-	**9.67**
Haematocrit	2.15	1.56	3.61	1.85
Filtered Glomerular CKD	-	7.41	1.24	-
Psychiatric Disease	-	-	-	8.29
Atrial Fibrillation	-	-	-	7.92
Gastro. or Liver Disease	-	-	-	7.59
Potassium	1.38	1.27	4.12	0.76
Metastatic Tumour	5.35	-	-	-
Sodium	3.07	-	-	-
Number Previous ER 365d	-	1.04	1.67	-

Table 4.6: GINI importance of the GBM for mortality and frailty tasks. Variables are sorted using the sum of the GINI importances in all tasks.

model is not fully comparable since it targeted older adults (>= 65 years old); meanwhile, all the mentioned studies use inclusion criteria of >= 18, except Avati, which includes paediatric records. Yourman *et al.* [Yourman et al., 2012] reviewed prognosis indices for older patients, where the better AUC ROC for the 1-year index was 0.83, which is below our lower 95% CI bound. The authors believe that excluding younger and possibly healthier patients from the sample made the problem more difficult and negatively affected the metrics. This is the case of our previous work in [Blanes-Selva et al., 2021a], which used data from the same hospital but reported better results using the whole adult population. As expected, the GBM model performed significantly better than the Logistic Regression counterpart and slightly better than the DNN model.

Our survival regression model scored a mean absolute error of 329.97 days, outperforming the 444.8 days scored by the cox model. Despite obtaining better predictions than one of the most used models when dealing with survival time, a mean error of almost a year does not adequately meet this model's original purpose. When removing cases where survival time is longer than 500 days, the GBM performs better than the other models achieving a mean error of 94.67 days; this improves the prediction error and will be likely better accepted by the health care professionals. This improvement in the predictive power is likely due to removing the long tail in the distribution that

includes infrequent values and outliers. It would also be possible to train a model using cases where survival was less than 365 days. In this case, the model would be used only when the one-year mortality produces a positive result; a preliminary result using the GBM configuration produced an MAE of 69.89 CI 95% [67.83, 72.08]. A further study concerning health care experts' preferences is needed to know if this alternative is preferred over the standard approximation.

The 1-year frailty model scored a 0.89 AUC ROC on GBM and DNN, outperforming the logistic regression version (0.84 AUC ROC). These results demonstrate a significant predictive power for assessing a patient's frailty index category one year from admission. As far as the authors know, this is the first study where a model is used to predict a future frailty status without using proxies such as mortality or disability. These models use variables containing information about the current frailty status combined with other factors such as the previous stays in the emergency room or the age to determine the future frailty status. Since most of the variables are shared with the other two mortality models, the addition of a few extra variables means that we can obtain a prediction regarding the patient's health decay with a low extra effort. Each model was set with the 20 most relevant variables from a total of 147, a number that was arguably too high to be used by a human operator. This selection was performed using the Random Forest's GINI importance criteria with recursive feature elimination as a data-driven method. This method is known to have a favourable bias towards categorical variables with many categories and continuous variables. However, it is widely used because it is fast and straightforward to compute [Nembrini et al., 2018]. In the end, all three models share a great number of variables (Table 4.6), being only 26 different variables. The selected variables by the recursive feature elimination algorithm are coherent with the different mortality works in the literature [Bernabeu-Wittel et al., 2011, van Walraven et al., 2015]. In addition, this final set of variables can be obtained easily a few hours after admission, where the first diagnosis and laboratory tests are performed.

These results provide a complementary perspective based on an objective measure of frailty to initiate early PC. The mean admission FI was 0.27 ± 0.12, and its shape resembles a normal distribution. This is a coherent behaviour with the findings in the Mitnitski *et al.*study [Mitnitski et al., 2001], where the most impaired groups have a bigger FI mean, and the distribution is shaped like a normal distribution, as opposed to the less impaired groups, which had a smaller mean FI and can be approximated using a gamma distribution. The correlation between our admission FI and MR target in days is –0.10, lower than the one reported in [Mitnitski et al., 2001], which was –0.234. This means that the frailty index used in this work for this sample is less associated with mortality. However, the Chi-Squared test performed on both binary targets discarded the hypothesis of independence, so in our sample, we can confirm a weak association between both criteria.

The relationship between frailty and mortality have been studied previously [Shamliyan et al., 2013], pointing to the association between both. Despite the similarity in the input variables, the target variable distributions are poorly correlated and have different shapes. Both criteria have been highlighted as important for accessing PC in previous studies and are related. However, they reflect two different distributions, and

the authors think of them as two complementary criteria. Therefore, we conclude that the best approximation for taking advantage of both mortality and frailty criteria is to have different predictive models working simultaneously, increasing the information to support the decision-making process. The incorporation of the frailty criterion may represent an added value for those health professionals deciding about inclusion in PC services. This is in line with Almagro *et al.*(2017) [Almagro et al., 2017], showing that poor vital prognosis as the sole criterion for initiating PC among COPD patients should be critically appraised.

This study's clinical impact resides in the potential to predict adverse outcomes for hospital admitted patients within the following year. First, we choose one year as a horizon to make the mortality prediction; as stated elsewhere [Avati et al., 2018], longer than 12 months is not desirable due to the difficulty in the predictions and the limited resources of the programs, which are better to focus on immediate needs. Thus, referral to PC may be focused on immediate needs. Also, despite being more difficult to predict, the information provided by the survival regression model may help contextualise the one-year mortality model results. Therefore, healthcare professionals would be supported with additional information such as the magnitude of the remaining time until death in days, weeks or months. Including these models into clinical practice could help anticipate the decline in admitted patients, allowing healthcare professionals to allocate scarce resources to patients who will need them the most.

The main contribution of this work is the development of the frailty predictive model, which is a novel approach to try to identify patients in need of ACP. This frailty approach complements the more traditional mortality approach, which we also tried to enrich by adding one-year mortality classification and regression to provide more information to healthcare experts during the decision-making process without providing excessive extra information burden. The three models were implemented as an online Clinical Decision Support System [Hajioff, 1998] available to any healthcare expert for academic use until further validations at [Blanes-Selva, 2021]. Besides, we have demonstrated the complementariness of the mortality and frailty models testing the low correlation between both factors in our dataset, so we should treat them as complementary criteria.

The main limitation of this study is the use of data from only one hospital. Therefore, internal validation only assures the performance of the models with similar data. We cannot ensure the reported efficiency in other hospitals and other patient populations [Sáez et al., 2021]. Also, data from the same centres can change over time for various reasons, such as a change in protocols or external agents such as a pandemic [Sáez and García-Gómez, 2018, Sáez et al., 2020]. Additional external validations are needed for future work. Broader populations can be approached by implementing predictive models using EHR, supporting an effective identification of patients needing further specialised care [Jung et al., 2019]. Thus, besides external validation of the models, future authors' work will require significant software development and implementation project to connect these systems with hospital EHR and avoid manual input by professionals. Also, the maturity of the models and the software wrapping them needs to be field-tested before their inclusion as a standard tool to the hospital information system.

4.6 Conclusion

This work proposes using three different machine learning models based on hospital admission data to assess the PC needs of older adults and help healthcare professionals in the decision-making process. The authors constructed three different but complementary predictive systems: a one-year mortality model, a regression mortality model to provide more information about the first prediction, and a one-year frailty model. Previous mortality models are using machine learning methods available elsewhere, but they are not specifically focused on older populations. Also, to our knowledge, this is the first study predicting one-year frailty status based on a frailty index. As previous studies have shown, mortality and frailty could be relevant criteria to admit patients to PC programs. Therefore, health professionals could benefit from using data-driven accurate predictions of these two dimensions on patients over 65. In addition to the benefits experienced by patients and their families, the early identification of these patients' needs can help better manage the available health and social care resources and reduce costs overall. Consequently, the authors propose using predictions in both mortality and frailty as complementary predictions to help assess PC needs due to its relevance but weak correlation, reliability and great predictive power. The described models have been implemented and publicly available for the academic purpose at [Blanes-Selva, 2021].

Chapter 5

Responsive and minimalist bedside mortality calculator

Mortality and frailty models published in scientific literature have demonstrated to have a great predictive power and being able to make the correct prediction consistently. However, the models usually include administrative variables, which values depend on the hospital's idiosyncrasy. This means that some variables may not be available in every Electronic Health Record (EHR), or that the variables present a different set possible values e.g., the admission department. This difficult the implantation of models trained with other data sources. Besides this problem, models tend to incorporate a large amount of variables in order to maximize their predictive power. This trend does not represent a problem if the model is fully implemented with hospital's information systems and is able to collect most of the information automatically. However, a lot of work and maturity is needed to reach this scenario and therefore CDSSs including these models need a manual input that can interfere with the clinical workflow if the number or required inputs is too large. In this chapter we present our approach to a minimalist mortality predictive model, deployed into a website adapted for use on portable devices.

*The content of this chapter was published in the scientific journal article by Blanes-Selva et al, (2021b) - thesis contribution **P3**. The software produced during this work is accessible through: http://palliative-calculator.upv.es*

5.1 Introduction

The World Health Organization (WHO) [Roberts et al., 1998] defines healthcare sustainability as "the ability to meet the needs of the present without compromising the ability to meet future needs". This topic is increasingly important [Borgonovi et al., 2018] and possibly especially challenging in developed countries, where ageing raises healthcare costs [Kyeremanteng et al., 2018]. However, regardless of the health policies in different countries, there is an interest in cost-effective alternatives that deliver (at least) the same quality of care, especially among older populations and those with chronic conditions and morbidity. PC has gained the attention of clinicians and researchers in recent years as both an alternative to standard care and through a

combined approach for these patient groups [Wallerstedt et al., 2019].

According to the recent redefinition of PC by Radbruch *et al.*in 2020 [Radbruch et al., 2020]: "Palliative care is the active holistic care of individuals across all ages with serious health-related suffering due to severe illness and especially of those near the end of life. It aims to improve the quality of life of patients, their families and their caregivers". Several studies have shown positive clinical results in patients involved in PC programs: an improvement in the quality of life and mood [Temel et al., 2010, Bakitas et al., 2009], symptom control [Yennurajalingam et al., 2011] and the reduction of emergency department visits and hospitalisations [Quinn et al., 2020].

Besides the clinical implications, the economic impact of PC programs has also been studied. In 2010, Simoens *et al.* [Simoens et al., 2010] reviewed different studies trying to compare the cost of PC against standard care; the authors found that PC was consistently cheaper. Kyeremanteng *et al.* [Kyeremanteng et al., 2018], in 2016, reviewed how PC affected the length of stay in the intensive care unit (ICU), which is an expensive form of healthcare. The authors found that PC consultations tend to reduce the length of stay in the ICU. Similarly, Smith *et al.* [Smith et al., 2014] found in another review that PC programs are frequently less expensive than comparator groups, and, in most cases, the difference is statistically significant.

It is estimated that approximately 75% of patients nearing end-of-life may benefit from PC interventions [Etkind et al., 2017]. The same authors projected an increase of 25% to 42.4% of people requiring PC in England and Wales, and it is reasonable to expect other developed countries will also have increased PC needs. However, despite the sound evidence of PC being clinically beneficial and cost-effective, it remains under-resourced in many countries.

Clearly, it is important to identify those patients who may benefit from PC at the appropriate time, since those interventions are positive in clinical and economic healthcare areas. Different criteria have been used, also known as triggers, based on different clinical diagnoses or the detection of more personalised PC needs [Kayastha and LeBlanc, 2020]. A limited prognosis is still a widely used criterion when screening for patients who may benefit from PC interventions [Etkind et al., 2017]. The SQ has been widely used and promoted to identify patients likely to die within the next year, and therefore possibly benefit from PC [Downar et al., 2017]. However, the SQ performs poorly to modestly, and further studies are needed to develop more accurate tools [Downar et al., 2017]. In addition to the SQ, there are other tools aimed to predict all-cause mortality. Some of the more accurate ones are data-driven and based on statistics and machine learning: the PROFUND index [Bernabeu-Wittel et al., 2011], HOMR [van Walraven et al., 2015], which has been modified in further publications [van Walraven and Forster, 2017, Wegier et al., 2019], or a deep learning approximation [Avati et al., 2018]. In addition, our team presented a machine-learning-based approximation that targets adults ($\geq$ 18) [Blanes-Selva et al., 2021a].

Despite the different literature approximations to this problem and the good predictive power reported for the different models, implementing these predictive models in clinical practice is not easy. The CDSS implementation in one organisation can fail due to user acceptance, as potential users can consider the outputs irrelevant or unreliable, or that the CDSS interfere in their workflow [Mahadevaiah et al., 2020]. Leslie

et al. [Leslie et al., 2006] provide a list of key features that are very important during the CDSS design, which we can summarise as follows: the tool should meet the user needs, adapt to the clinical workflow and be flexible enough to allow the healthcare professional to manage in their own way.

In concrete, CDSS powered by machine-learning technology has been utilised successfully in several clinical applications, especially dealing with medical imaging in fields such as radiology, dermatology or ophthalmology [Yu et al., 2018]. However, other fields, such as biomarker discovery or clinical outcome prediction, have also benefitted from incorporating machine learning algorithms. In the review performed by Yu *et al.* [Yu et al., 2018], two categories of challenges are discussed: technical challenges, which includes the data quality and interpretability of the models, and social, economic and legal challenges, where the author remarks on the importance of the integration of the tool with the clinical workflows.

A particularly interesting use of CDSS is in bedside applications, where the clinical professionals do not need to spend much time working on documentation or on a desktop computer using the available tools through mobile devices [Ehrler et al., 2018]. Some studies focused on the usability of mobile clinical decision support models have proposed checklists, also known as heuristics, to determine if applications comply with usability standards. A recent example is Reeder *et al.* [Reeder et al., 2019], who compile a heuristic focused on mobile CDSS applications. When designing these bedside apps, one essential principle is to show only the necessary information needed to proceed. In other words, the application should be as minimalist as possible, whilst keeping the original functionalities and objectives. Some approximations to bedside mortality risk assessment have been studied in the past, proposing scores and indexes for different health issues, such as acute pancreatitis [Singh et al., 2009], acute ischemic stroke [O'Donnell et al., 2012], non-cardiac surgery [Glance et al., 2012] and, recently, for COVID-19 [Bertsimas et al., 2020]. Despite the increase in research around predictive models using mobile applications in the health domain, we found very little information about any apps used in practice. One of the few apps available is the COVID-19 mortality calculator [Bertsimas et al., 2020]. However, to the best of the authors' knowledge, there is no mobile app to predict 1-year general mortality at the bedside. The aim of this study is twofold. On the one hand, the aim is to create a compact version of the 1-year mortality model using common and easy-to-gather variables during admission based on the larger model developed by the authors previously [Blanes-Selva et al., 2022a]. On the other hand, the aim is to implement a web application (web app) so that health care professionals can assess PC needs during bedside examinations utilising a smartphone or a tablet.

5.2 Materials and Methods

5.2.1 Data

This study makes use of the same dataset as the one described in [Blanes-Selva et al., 2022a]. EHR information was collected from admissions for older patients (65 years old), excluding those admitted to the psychiatry department. Data comprise 1 January

2011 to 31 December 2018, containing 39,310 different admission episodes from 19,753 different patients. Patients death date was available in the EHR and was used to calculate the one-year mortality target variable.

5.2.2 Feature Selection and Modelling

To create a compacter model, we applied a strategy to reduce the number of variables, starting with the list of the 20 most important variables obtained from our previous work [Blanes-Selva et al., 2022a]. In addition, we removed the administrative variables that may be not compatible with other information systems: Diagnosis Related Group (DRG), admission code and department code. The final list of variables composing the dataset is available in Table 5.1.

Variable	Rank	Mean ± Std	Missings
Number of Active Groups (Meds)	1	2.44 ± 3.68	0%
Charlson Index	2	4.77 ± 3.34	0.2%
Barthel Index	3	51.91 ± 39.75	73.4%
Metastatic Tumour	4	-[k]	0%
Age	5	79.4 ± 8.36	0%
Urea (mg/dL)	6	61.19 ± 43.36	37.1%
RDW SD[l] (fL)	7	49.66 7.36	21.7%
Leukocyte ($10^3/\mu$L)	8	9.23 ± 6.85	21.6%
RDW CV[m] (%)	9	15.26 2.43	21.7%
Sodium (mEq/L)	10	139.89 ± 4.92	21.2%
C Reactive Protein (mg/L)	11	55.1 ± 70.54	47.3%
Creatinine (mg/dL)	12	1.23 ± 1	20.7%
Haematocrit (%)	13	36.24 ± 5.8	21.6%
Glucose (mg/dL)	14	122.5 ± 54.98	24%
Number of Previous ER	15	6.04 ± 6.56	1.4%
Number of Previous Admissions	16	7.62 ± 7.49	0%
Potassium (mEq/L)	17	4.22 ± 0.61	22.5%

Table 5.1: Variable summary and ranking. Rank represents the order of the variable according to the GINI importance. Mean *pm* Std column describe the statistical mean of the feature and its standard deviation. Missings column represent the percentage of missing values of that feature

In the next step, we ranked the variables according to their importance using an iterative method. The algorithm starts with a whole list of variables, and then an explainable model is trained using them and the importance of every variable is obtained. We used the random forest model as this explainable model, using the GINI criteria to determine the importance of each variable after the model was fitted. After this, the variable with less importance was removed from the list. The process ended

[k]Distribution for categorical variable metastatic tumour is Yes: 77 (0.2%); No: 38339 (99.8%).
[l]Red blood cell distribution width standard deviation.
[m]ed blood cell distribution width coefficient of variation.

when only one variable was left. The ranking of every variable was the iteration when it was removed from the list in reverse order.

To determine the optimum number of variables, we took the ranking obtained in the previous step and applied the following algorithm: starting with the most relevant variable, a model-based gradient boosting machine is trained and validated with the 10-fold validation method, using nine sections of the data to train the model and one to test it. The AUC ROC [Bradley, 1997] is computed for the 10 test splits, averaged and stored as a result for the first variable. The process continues, adding the following variable in the ranking to the model until the final iteration, where all variables are included in the model.

Among the selected variables (Table 5.1), there are missing values, except in the age and the number of active groups. We hypothesised that the main mechanism producing these missing values is the clinical criterion, where tests are not performed if the physicians do not consider their results important to diagnose or treat the patient. The missing values produced by this mechanism are known as Missing Not At Random (MNAR) [Haneuse et al., 2021]. To use the maximum amount of information, an imputation method was needed. In our case, we combined the inclusion of an imputation mask (a dummy variable indicating if the data are present) and an iterative imputation technique [Van Buuren and Groothuis-Oudshoorn, 2011]. We decided to use both techniques following the results of Sperrin and Martin [Sperrin and Martin, 2020] because it improves the modelling when dealing with MNAR, but has no detrimental effects for missing at random. Following the original data criteria, the only required input for the user is the age and the number of active groups prescribed to the patients during admission.

5.2.3 Explainability Layer

When a new sample introduced by the user reaches the gradient boosting machine, an explainer object is created to interpret the effect of the different variables of the sample in the prediction. This explainer object is the TreeExplainer, which is implemented in the SHAP library [Lundberg and Lee, 2017]. The output of this process is the Shapley values. Every value corresponds to one variable; positive values indicate that the value on this variable pushes the prediction to the positive class, which is mortality within the year in our case. In addition, the bigger the absolute value, the greater the effect on the prediction. We have created a bar graph to represent these values visually, showing only the most relevant ones (bigger absolute value) and adopting the green for the negative values (greater for the patient survival) and red for the positive ones.

5.2.4 APP Implementation and Software

Finally, and after determining the optimal number of variables, a model was trained using all the available cases. The model was then deployed in a publicly available Django web application. We used the Bootstrap library to make the website responsive and adapt the size and elements display in all screen sizes. The app's interface was designed to be as minimalist and functional as possible, following the heuristic from Reeder *et al.* [Reeder et al., 2019] when applicable. Figure 5.1 illustrates the whole

methodology workflow. Model creation and evaluation process were performed using the NumPy [Harris et al., 2020], pandas [McKinney et al., 2010] and scikit-learn [Pedregosa et al., 2011] libraries working with the Python programming language in its 3.8 version.

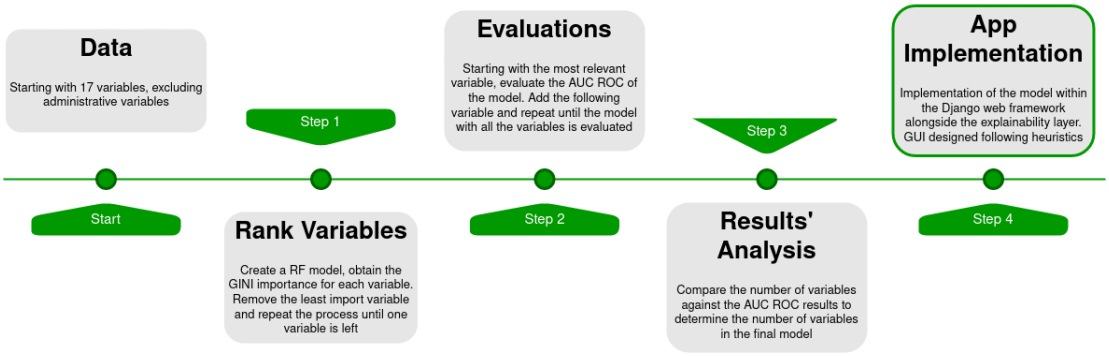


Figure 5.1: Methodology process summary from the initial data to the web application.

5.3 Results

Table 1 shows the results of the data exploration through variable parameters and the number of missing values. In addition, it includes the importance ranking obtained for each variable. The most relevant variables are the number of active groups (in the medications provided), the Charlson Index, the Barthel Index and the patient's age. Results for the second experiment can be observed in Figure 5.2. Each coloured bar represents the mean AUC ROC using the 10-fold model with the N most important variables determined by the previous experiment. Grey bars on the top represent the 95% confidence interval. The mean AUC ROC values increase as the number of variables grow, but changes are not statistically significant when the number of variables is greater than nine.

Figures 5.3a,b show the final aspect of the tool when accessing it from a smartphone. Figure 5.3a presents the form where the health care professional enters the patients' data. Units have been added to laboratory tests fields, as well as a tooltip with the reference values to help the users. The number of variables has been selected according to the heuristic principle of *Is only (and all) information essential to decision making displayed on the screen?* and *Has the need to scroll been minimized and where necessary, are navigation facilities repeated at the bottom of the screen?*. The order and position of the layout have been influenced as well by the heuristic. Figure 5.3b shows a results example, a modal panel that appears over the previous screen, including the numeric value for the prediction in percentage, and the variable importance bar plot built from the Shapley values. The missing indicators are displayed on the bar graph with the suffix "missing" if the explainer found the variable relevant.

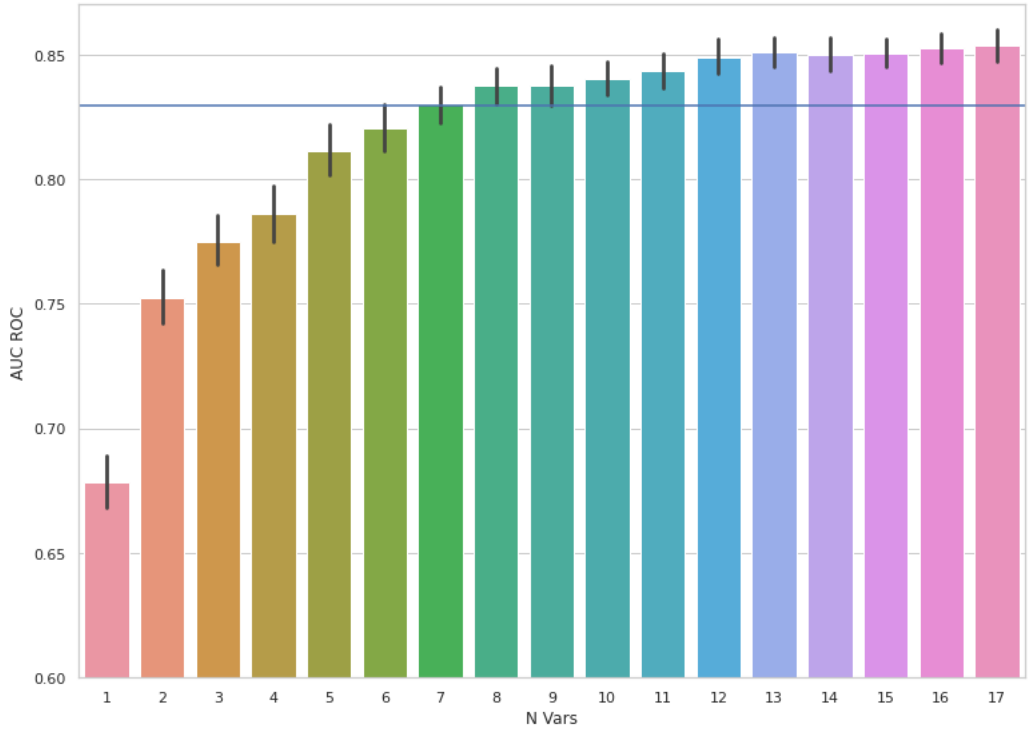
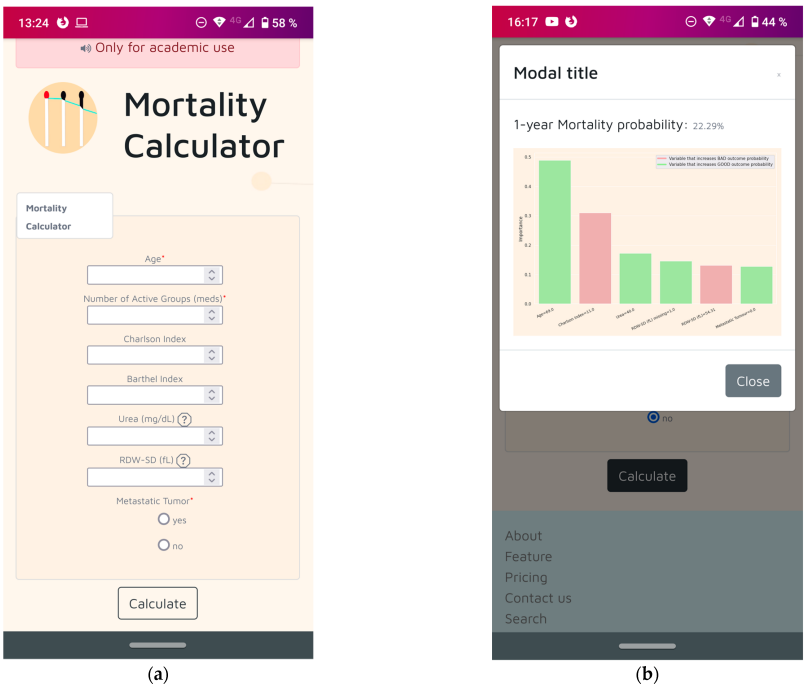


Figure 5.2: AUC ROC values when increasing the number of variables. The horizontal line marks the result for the selected number of variables in the final version. Vertical lines on top of each bar represent the width of the confidence interval at 95%



(a) (b)

Figure 5.3: Screenshots from the application running on Firefox Focus in Android device: (a) main screen where the user has to input the data; (b) results and explainability graph.

5.4 Discussion

The experimentation reported in the first place which variables were more informative to create the predictive model. With this information, we calculated the metrics of the model using an increasing number of variables to determine the optimum number of variables. Our objective was to maximise the performance of the model while keeping the required number of inputs low. Setting a low number of required inputs is important from the design perspective [Bargas-Avila et al., 2010] to minimise the time required to use the tool, which directly affects the interference in the clinical workflow and the chances of successful implementation [Leslie et al., 2006]. Therefore, we followed the heuristic by [Reeder et al., 2019] when possible, although some checkpoints did not apply due to the app's small size.

Using the minimalist approach, the authors believe that seven variables is the most convenient number of variables to incorporate in this model. This gives a final evaluation of the model is an AUC ROC of 0.83 with a 95% confidence interval of [0.82 to 0.84]. Arguments favouring increasing or reducing the number of variables can be made, but there is no precise study about how this factor affects repetitive tasks over time. Previous mortality models proposed in the literature have achieved a greater predictive performance, such as HOMR (0.89 to 0.92 AUC ROC) in its different versions [van Walraven et al., 2015, van Walraven and Forster, 2017, Wegier et al., 2019], Avati's deep learning approach (0.87 AUC ROC on admitted patients) [Avati et al., 2018] and our previous work, which achieved 0.91 of AUC ROC [Blanes-Selva et al., 2021a].

However, this is not a fair comparison for three main reasons. First, the previously mentioned studies are aimed at adults (18), and Avatis work even includes paediatric information, whereas this work was focused on older patients (65). Since age is a determinant mortality factor recognised in the literature [Bernabeu-Wittel et al., 2011, van Walraven et al., 2015, van Walraven and Forster, 2017, Wegier et al., 2019, Avati et al., 2018, Blanes-Selva et al., 2021a], limiting the dataset to older patients removes younger and healthier patients and makes classification more difficult. There are few 1-year mortality admission predictive indices focused on older patients. Inouye *et al.* [Inouye et al., 2003] reported an AUC ROC of 0.83 in development and 0.77 in validation, whereas Fischer *et al.* [Fischer et al., 2006] reported 0.82 in development. Both studies report a similar performance to our results. Still, the indices were developed with a relatively small number of cases, 525 and 435, respectively, which may lead to weaker results than modern studies with bigger samples. Secondly, despite some of these studies using easily obtained variables, such as the HOMR-now! [van Walraven and Forster, 2017] derived from the HOMR model using data immediately available at hospital admission, none of them have been designed to be a compact and bedside clinical decision support tool. Lastly, with the design of this model, administrative variables have been avoided to predict mortality as they could be difficult to obtain or could not be exchangeable between different healthcare systems, such as the department where the patients are admitted. Other works in the literature include variables that may present difficulties implementing the system in different hospitals due to having varying protocols or administration; for example, the original HOMR study [van Walraven et al., 2015] used, as variables, if the admission has been

performed by ambulance or if the patient receives oxygen at home.

The variables that compose this predictive model are similar to those used in other studies. The Charlson Index was also used in HOMR and HOMR-now! [van Walraven et al., 2015, van Walraven and Forster, 2017]; The Barthel Index had been used in the PROFUND index [Bernabeu-Wittel et al., 2011]; BUN was the second most relevant variable in our previous study [Blanes-Selva et al., 2021a] and some studies have identified that this variable is associated with mortality in different contexts [Arihan et al., 2018, Wu et al., 2009, Cauthen et al., 2008]. Age has been used in all of the previously mentioned models, and is strongly associated with mortality. Other variables have been tested in their association with mortality. In our study, the number of active groups in medication resulted in one of the most informative predictors, despite other studies not finding a direct association [Martín-Pérez et al., 2019]. Gelder *et al.* [de Gelder et al., 2016] considered this variable in their 90-days mortality model, but it did not make the cut of the final model. Using the explainability method, we detected a strong negative interaction when the variable has a value of 0. We hypothesise that this is the case for terminal patients on non-curative treatment. Lower values for this variable have positive effects on the prediction, whereas this effect declines and turns negative as the value increases. This pattern is expected, since patients with more prescribed medications are likely to have a more complex health status and an increased number of comorbidities. RDW-SD measures the standard deviation in the variability of red cell volume/size (RDW). In 2010 Patel *et al.* [Patel et al., 2010] conducted a meta-analysis where the RDW test was a strong predictor for mortality in older adults. The metastatic tumour is a condition related to survival and has been used in previous works as well [Blanes-Selva et al., 2021a].

Accurate predictions can be helpful to health care professionals using these types of tools, but the users need to understand why a specific prediction has been made. As Shortliffe and Sepúlveda stated in [Shortliffe and Sepúlveda, 2018], black boxes are unacceptable. It is necessary to provide an explanation of why the tool has made a certain recommendation in order for the professionals to accept or override it. Carroll *et al.* [Carroll et al., 2002] also point to the lack of explanation for the proposed solution as a pitfall for CDSS. In this sense, in this work, we tried to address this situation by providing a graphic representation of the Shapley values [Lundberg and Lee, 2017]; this graphic shows the most relevant variables pushing the prediction to higher or lower mortality within the year probability.

Prognosis-based interventions are one of the most common ways to detect patients in need of PC. According to Kayastha and Leblanc [Kayastha and LeBlanc, 2020], there are two other identification types: need-based and trigger-based. On the one hand, need-based assessments use different tools, such as questionnaires and the need from the clinician's agreement on how to interpret the results, which could lead to too many referrals. On the other hand, trigger-based assessments define a set of conditions to trigger the intervention that is a mix of both prognosis and needs. The tool proposed in this work falls into the prognosis-based category. One of the downsides of this approximation is that it does not provide information about the concrete needs a patient may have. It may be interesting to complement the output provided from our tool with other criteria in order to apply an additional filter and improve the PC

needs identification. Recently, Wegier *et al.* [Wegier et al., 2021] studied this possibility in their work by assessing PC needs using symptomatology and readiness to engage in advance care planning on patients already identified using their predictive model mHOMR. Most of the patients identified presented unmet PC needs in accordance with the study criteria. This result opens the possibility to improve the effectiveness of mortality models with other PC needs assessment tools.

Resource allocation to PC programs is a relevant challenge because these programs present a list of clinical benefits to the patients but are often under-resourced [Temel et al., 2010, Bakitas et al., 2009, Yennurajalingam et al., 2011, Quinn et al., 2020]. If available, PC is usually provided to patients with a high symptom burden or in the terminal phase of illness. However, a shift to an early PC approach is advocated [Haun et al., 2017]. In addition to this, some studies have proven different PC interventions to be cost-effective compared to standard care [Kyeremanteng et al., 2018, Simoens et al., 2010, Smith et al., 2014]. Early PC has shown lower hospital costs during hospital admission [May et al., 2015] or costs savings over routine care [Lowery et al., 2013]. Combined with the anticipated increased PC demand alongside limited resources, these facts create the necessity to identify which patients could benefit more from PC interventions. A bedside tool, such as the one presented in this work, can help identify those patients in an agile way, so that the PC delivery can be improved and, therefore, bring clinical benefits and less expensive care to those in need, improving the sustainability of the healthcare system.

The main goal of this work was to create a quick-to-use tool to determine which patients may benefit from PC. Besides the identification power, the main benefit of having a compact tool is the reduced time needed for completion, which is essential to the tool's success [Shortliffe and Sepúlveda, 2018, Carroll et al., 2002]. The main strengths of this study are, first, the creation of a compact all-cause mortality model during hospital admission, obtaining a good discriminative power of 0.82 of AUC ROC. Second, this minimalist design, in addition to the lack of hospital-specific variables, allowed us to create a clinical decision tool as a web app to be used in any portable device with an internet connection and a web browser. Finally, the tool provides a numerical result and uses explanatory techniques to help the health care professionals in their final decision making, providing more context to accept or override the prediction. This simple tool presented a good predictive power and can quickly detect patients in need of PC and aid the effective allocation of resources to improve healthcare sustainability.

The main limitation of this study resides in its validation. Evaluation of the predictive model has been performed only within the same dataset using the K-fold validation strategy from a single institution. Thus, we cannot ensure the reported effectiveness in other contexts or other populations [Sáez et al., 2021]. External validation with other hospitals and populations is needed. Another important limitation of this tool is that the data input and the output calculation are disconnected from the EHR system. Simplifying the data input for the models to be used in a bedside tool was addressed during the study methodology. Still, the output prediction is not recorded anywhere, so it will require a manual introduction of the results in another system; this could prove difficult for both the case review process for healthcare workers and

for the acceptance of the tool for other stakeholders, such as hospital administrators or policymakers, who are unable to obtain a 'big picture' from the application records [Petersen et al., 2015, Ganasegeran and Abdulrahman, 2019]. However, the disconnection from the EHR and the application not storing any data solve most of the privacy and data security difficulties, which are a barrier to the large scale adoption of mHealth applications [Petersen et al., 2015].

5.5 Conclusions

Older populations with chronic conditions or multimorbidity are increasing, which may mean an increased demand and use of health care services, with PC interventions among them. Unfortunately, there are many barriers in accessing PC, such as limited resources or late referrals when a person is in their last phase of the end-of-life process. This minimalist and simple predictive model can support the early identification of patients in need of PC, requiring only a minimum investment of time by clinicians. Tools such as this can facilitate the management of complex patients and overcome decision-making difficulties in integrating PC in daily clinical practice. A better PC identification delivers clinical benefits to the patients and could help to allocate resources, improving the system's sustainability. The tool is available at: http://palliative-calculator.upv.es (accessed on July 23, 2022).

Chapter 6

Validation of a clinical decision support platform for palliative care: The Aleph

The use of ML to aid in clinical problems such as the PC referral has generated promising results. However, these solutions rarely leave the laboratories and get implemented in clinical practice. Scientific literature presents an exhaustive list of pitfalls and conditions that present the development of a CDSS. In this chapter, we are going to review the most common problems present during the CDSS implantation and evaluate our system, the Aleph, using a user-centred methodology. This methodology includes using the system in a 'near-live' scenario and the use of usability and user experience questionnaires.

*The contents of this chapter as preprint in medRxiv by Blanes-Selva et al, (2022b) and under consideration for SAGE Digital Health - thesis contribution **P4**. The software produced during this work is accessible through: https://thealeph.upv.es*

6.1 Introduction

Clinical Decision Support System (CDSS) are computer systems designed to impact clinician decision-making about individual patients at the point in time that these decisions are made [Berner, 2007]. Interest and research about CDSS are motivated by their potential benefits documented in the scientific literature: increased patient safety by reducing medical errors or avoiding advice against protocol; improved service quality due to better adherence to guidelines, and increased service time dedicated directly to the patients; cost reduction by processing faster the demands and avoiding duplicated tests; improved administrative functions by incorporating elements such as automatic documentation; diagnosis support and workflow improvement [Sutton et al., 2020, Tundjungsari et al., 2017].

However, despite the multiple virtues that CDSSs could bring, there has been a lack of adoption of these systems into the clinical practice [Belard et al., 2017, Yang et al., 2016, Devaraj and Viernes, 2014, Elwyn et al., 2013, Yang et al., 2015]. Several studies pointed to the main barriers to adoption, in which we could find two broad categories:

socio-cultural factors and usability. Socio-cultural barriers refer to the beliefs of health care professionals or their organisation regarding the CDSSs, such as the idea of loss of autonomy, the feeling of being replaced by the system, low computer literacy, lack of trust in the system, failure to fulfil a perceived clinical need, legal uncertainties and misalignment between human needs and the technical system [Khairat et al., 2018, Liberati et al., 2017, Yu et al., 2018, Carroll et al., 2002]. Liberati *et al.* [Liberati et al., 2017] proposed several strategies to deal with those barriers depending on the physicians' beliefs regarding the CDSSs. Most of them are based on communication, training, and highlighting the system's benefits.

On the other hand, usability barriers refer to the difficulties found by the user while using the software. The most common problems in this category are the difficulty to operate the software, the disruption of the workflow, the decrease of face-to-face time with the patients [Sutton et al., 2020] and the alert fatigue due to excess notifications by the system [Carroll et al., 2002, Press et al., 2015]. These challenges have been addressed previously by other authors by performing usability pilots with potential software end-users, mostly HPs, in order to identify and correct the different usability problems of their CDSSs [Press et al., 2015, Genes et al., 2016, Richardson et al., 2017, Thum et al., 2014].

Usability studies usually follow a general scheme. The participants are exposed to the software in a controlled environment and the session is usually taped and/or with the researchers taking field notes. Participants must try accomplishing tasks mimicking real scenarios, which in some studies receive the name on "near-live" simulations [Li et al., 2012]. Think-Aloud methodology [Eccles and Arsal, 2017] is commonly used during the whole study, this method consists of asking the participants to express their doubts, opinions, and in general, any thought regarding their experience with the tool. Finally, the usability of the software is quantified through a scale or an index. One of the most popular evaluation tools is the System Usability Scale (SUS) [Brooke et al., 1996, Lewis, 2018].

It is generally accepted that a positive User eXperience (UX) is essential to any software acceptance [Wallach et al., 2020]. Despite existing a close relationship between usability and UX concepts, there are some differences worth studying, primarily related to the hedonic category [Bevan, 2009], i.e., how 'pleasurable' the users find to use the software. In addition, UX also studies emotions, beliefs, preferences and perceptions. These concepts directly impact the adoption of a CDSS, concretely, they are strictly related to the previously mentioned socio-cultural barriers. Therefore, a UX study is essential to assess and improve technology adoption.

Another crucial aspect while maximizing the probability of a successful implementation of a CDSS in clinical practice is their design from the initial stages. An interdisciplinary team is highly recommended, including data scientists, programmers, usability and UX experts, the HPs as potential users of the software and other stakeholders such as representatives from hospital management to have a clear vision of the requirements [Yu et al., 2018, Mahadevaiah et al., 2020]. Planning a pleasant interface is also important since some studies reported users being more tolerant to minor usability issues if they found the interface visually appealing, which is known as the aesthetic-usability effect [Moran, 2017].

In our previous work [Blanes-Selva et al., 2022a], we developed a set of predictive models to assist the PC referral with hospital admission data on older patients using mortality and frailty predictions as main criteria. The result of that study was a demonstrator for a complete CDSS called The Aleph PC. Our study reported that these models accurately predicted which patients had a short survival time and were likely to become frail. Thus, our goal in this work is to validate the Aleph PC using user-centred techniques [Cai et al., 2019] and determine how different health professionals with PC experience envision the use of a PC CDSS in the clinical practice. First, we evaluated The Aleph PC's mortality and frailty models against the HPs predictions to obtain a baseline, and second, we assess the usability and UX of the system alongside the different insights of the HPs on how to build a useful PC CDSS.

6.2 Materials and methods

6.2.1 The Aleph CDSS Platform

The Aleph PC is an open-access ML-based CDSS implemented as a web platform. The application is divided into three main screens; in the first one, the user introduces the different data required for the PC predictions, including administrative information, Barthel [Mahoney et al., 1965] and Charlson [Charlson et al., 1994] indexes, laboratory results and a few diagnosis variables (Figure 6.1a). After completing the form, the results are calculated and displayed on another screen (Figure 6.1b); these results include a numerical result for each model and a Machine Learning (ML) explainability figure. We have used the Shapley Values (SHAP) [Lundberg and Lee, 2017] to display a graph with the relation between the input and the prediction obtained. Finally, the Files section (Figure 6.1c) allows the user to save the current case, load a different case or test the application with predefined test cases. The version tested in this study can be accessed here: https://demoiapc.upv.es/[n].

6.2.2 Recruitment process

We recruited healthcare professionals used to treat patients with bad prognostic within a wide variety of roles and possible end users of the CDSS. In concrete, we focused on: nurses, primary care physicians, hospitalist physicians, PC consultants and specialists like oncologists, neurologists or pulmonologists. This decision ensured the inclusion of different approaches working with complex patients in need of PC. The authors drafted a list of possible participants with no direct relationship with the development of The Aleph PC, and followed the snowball sampling technique [Parker et al., 2019]: once identified the first volunteers we asked them for other colleagues willing to participate until we completed our target sample size. Invitations to participate in the study were sent by email.

[n]Last accessed July 23, 2022

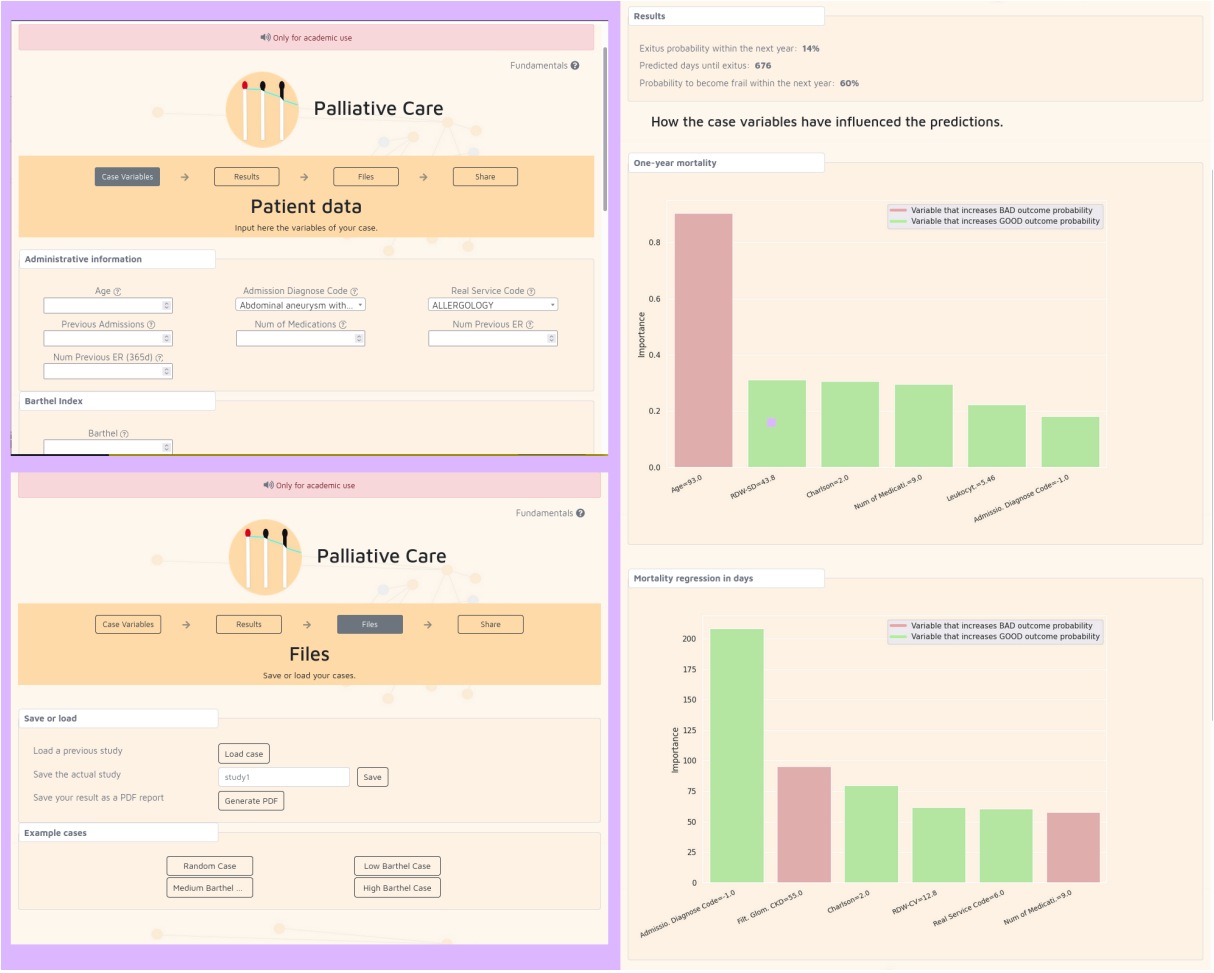


Figure 6.1: a) (left-top) Screen where the user inputs the data b) (right) Screen where the results are shown c) (left-bottom) Screen to manage the files and the predetermined examples

6.2.3 Study structure

Participation

The study was defined as an iterative user-centred validation. Participants were invited to individual evaluation sessions where one of the team members acted as session guide. On some occasions, a second member of the team spectated the evaluation session and collaborated taking field notes. Evaluation sessions were performed by video-conference, where the participants shared their screens while interacting with The Aleph PC. The think-aloud methodology [Li et al., 2012] was used during the whole session. The duration of each evaluation session was around one hour and their overall structure is displayed in Figure 6.2.

We defined two rounds of sessions, with a period of 15 days between them to adapt the software based on the feedback obtained during the first round. Sixteen participants were invited in the first round, 15 of them (93.75%) agreed to participate. For the second round, 8 participants were invited, and 6 finally responded (75%). We performed a greater number of sessions during the first round to detect as many

usability problems as possible. We settled on six respondents during the second round due to the difficulty of finding participants and the fact that we already reached a number of participants that allow us to detect most of the usability problems according to the Nielsen Normal Group [Nielsen, 2000].

The distributions between the participants in both rounds was the following: fifteen of them were physicians with the following roles: 7 general practitioners, 5 hospitalists, 1 PC consultant, 1 oncologist and 1 neurologist. The other 6 participants were nurses. The distribution between sex was: 13 males (61.9%) and 8 females (38.1%). Distribution by country was Italy (5), Brazil (4), Spain (4), Greece (4), Scotland (2) and Portugal (2).

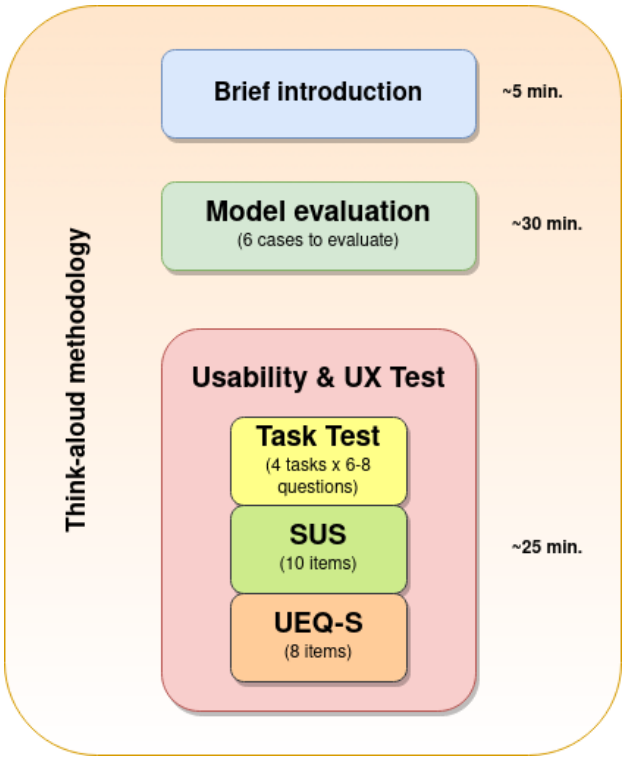


Figure 6.2: Overview structure of the sessions including three sections and the approximate time spent in each one of them: 1) Brief introduction, 2) Model evaluation and 3) Usability and UX test.

Model validation

First, we implemented a model evaluation which was performed in an unlinked section[o] of The Aleph PC. After introducing some basic information, the participants faced 6 vignettes: already filled, non-editable input forms containing real cases. Then, at the bottom of the page they were asked to fill in their own predictions regarding one-year mortality (yes/no), mortality regression (months interval) and one-year frailty. The same vignettes, in the same order, were asked for all 21 participants across both rounds. No time limit was established, and participants were asked to use any information resource (internet search, books...) they needed to make their own predictions.

[o]http://demoiapc.upv.es/validations/launch_validation Last accessed July 23, 2022

We calculated the accuracy, sensitivity and specificity for the one-year mortality (1ym) and the one-year frailty (Frailty) models. Since we asked the participants for their predictions in months to facilitate their response, we had to transform the output of the regression model from days to months. Therefore, we divided the number of days by 30, discarding the decimal part, and then transformed it into an interval. The interval was determined to have a width of 4 months based on the results of our previous study, so we used the prediction in months $\pm$ 2 months as interval bounds. We then calculated the accuracy of the participants and the model by checking if the real value of the cases in months belonged to the interval (lower bound $\leq$ real value $\leq$ higher value).

Usability and UX validation

In the usability and UX section, the participants answered a Google Form questionnaire while they were testing the The Aleph PC. Participants were asked to do a 'task test': to perform four simple tasks using The Aleph PC and answer a series of questions after each of them. The tasks covered all the implemented functionality for the CDSS: 1) input a feasible case, 2) check the results and understand the graphics, 3) save the current case and, 4) load a case previously stored. The questions after each task interrogated the participants about the difficulty, the perception of time spent, the number of errors encountered by the participant (including unexpected behaviours and elements they did not understand) and the satisfaction obtained by performing the task. All questions were mandatory.

After the task test, we tested the usability and experience with the SUS questionnaire [Brooke et al., 1996, Lewis, 2018] and the User Experience Questionnaire (short version) (UEQ-S) [Schrepp et al., 2017]. Both questionnaires were implemented into the same Google Form page. The participants were asked to stop sharing their screens during the completion of both tests.

6.3 Results

6.3.1 Model evaluation

The ML models outperformed the healthcare professionals' predictions in both mortality and frailty (see Table 6.1). The mean width for the intervals provided by the participants in the regression prediction was: 16.2 months CI 95% (13.5 to 18.9) against the fixed 4 months for the models.

6.3.2 Qualitative results

Regarding the qualitative results based on the think-aloud method and the authors' notes on the participants' behaviour, we created a list of improvements after each round, the changes were focused on interface details: removal of the Diagnosis Related Group variable because it could be inferred from the ICD9 code, replacement of the ICD9 codes by their name, improved tooltip descriptions, and added reference values

Task	Predictions	Accuracy	Sensitivity	Specificity
1ym	Participants	0.5 (0.42 - 0.58)	0.54 (0.42 - 0.56)	0.46 (0.34 - 0.58)
	The Aleph PC	0.83	0.75	1
Frailty	Participants	0.78 (0.7 - 0.85)	0.8 (0.72 - 0.88)	0.67 (0.45 - 0.89)
	The Aleph PC	1	1	1
Regression	Participants	0.45 (0.36 - 0.55)	-	-
	The Aleph PC	0.67	-	-

Table 6.1: Summary of the metrics for the participants and the ML models in the three tasks for the 6 cases evaluated. Mean and 95% confidence intervals are reported per participant. ML are deterministic so no variability on the prediction was found.

for the laboratory variables. Table 2 in supplementary materials[p] contains the complete log of changes introduced in both rounds. Most of the participants provided feedback regarding the subset of variables, suggesting other variables they are more familiar with or disregarding present variables as unimportant or unavailable in their workflow. The feeling towards the CDSS was primarily positive, and the idea of the PC identification using ML technology was well received. Few of the participants felt confused with the first interaction with the software but many of them affirmed they have learned to use it after the tasks test. Participants from hospital settings suggested picking automatically the diagnosis and laboratory results from the EHR, whereas other participants did not care about complete integration due to the lack of system integration in their respective environments. We found a participant in each round who was sceptic about the use of computers for decision making and did not believe in the benefits of the technology, rating every aspect of the system very low and providing poor opinions about the system through the think-aloud method.

6.3.3 Performance of tasks

Figure 6.3 shows the distribution of the answers for both rounds. Almost every measured feature increased their percentage of positive feedback during the second round. The most significant improvement was on task four (load a case), where despite the increased number of errors, the perceived difficulty, time spent, and satisfaction improved.

6.3.4 Usability

Responses to the 10 SUS items scores were recorded, all items were mandatory, so no missing values were present. The first round of the evaluation sessions obtained a mean of 62.7 ± 14.1, while the second round increased its score to 65 ± 26.2. The distribution of the answers for the different items can be found in Figure 6.4. We have used the adjusted scores instead of the raw scores for all items to help with visualization. Round

[p] https://www.medrxiv.org/content/medrxiv/early/2022/06/05/2022.06.03.22275904/DC1/embed/media-1.xlsx?download=true

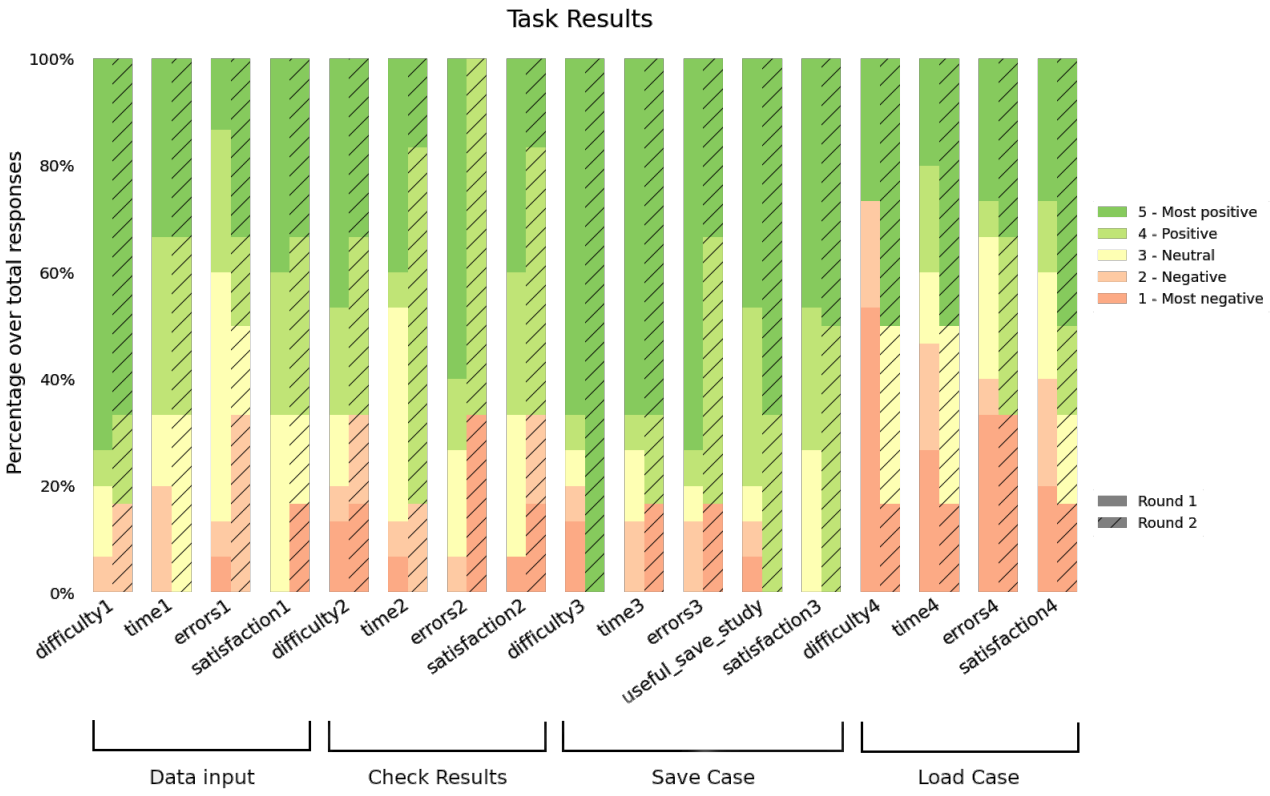


Figure 6.3: Results for the different tasks during the task test. Bars represent the distribution of the responses. A positive response means that the participant found the task: easy, short, with a low number of errors or satisfactory.

2 has a greater proportion of positive responses in 6 out of 10 items, nonetheless the first round has a lower score and lower standard deviation.

Previous studies have tried to map intervals of the score into categories such as "Poor", "OK" or "Good" or school grading scales [Bangor et al., 2009] in order to provide a better usability reference. According to these frameworks, our results for both rounds would be classified as D (lowest passing score), the first round as "OK - low marginal acceptance" and the second round as "OK - high marginal acceptance". However, if we recalculate the SUS average score excluding the sceptical participants, the average rating would be 63.9 $\pm$ 13.8 and 74.5 $\pm$ 16.8 which are D "OK - High marginal acceptance" and C "Good - Acceptable".

6.3.5 User experience

Answers to the UEQ-S questionnaire were recorded, with all items being mandatory. Figure 6.5 shows the distribution of the responses for each item in the questionnaire. The median for the second round was always greater than the first round, and the average scores were: 1.4 in the first round and 1.5 in the second one. The Pragmatic score was slightly higher in the first round (1.3 vs 1.2) and the hedonic score improved during the second round (1.5 vs 1.8).

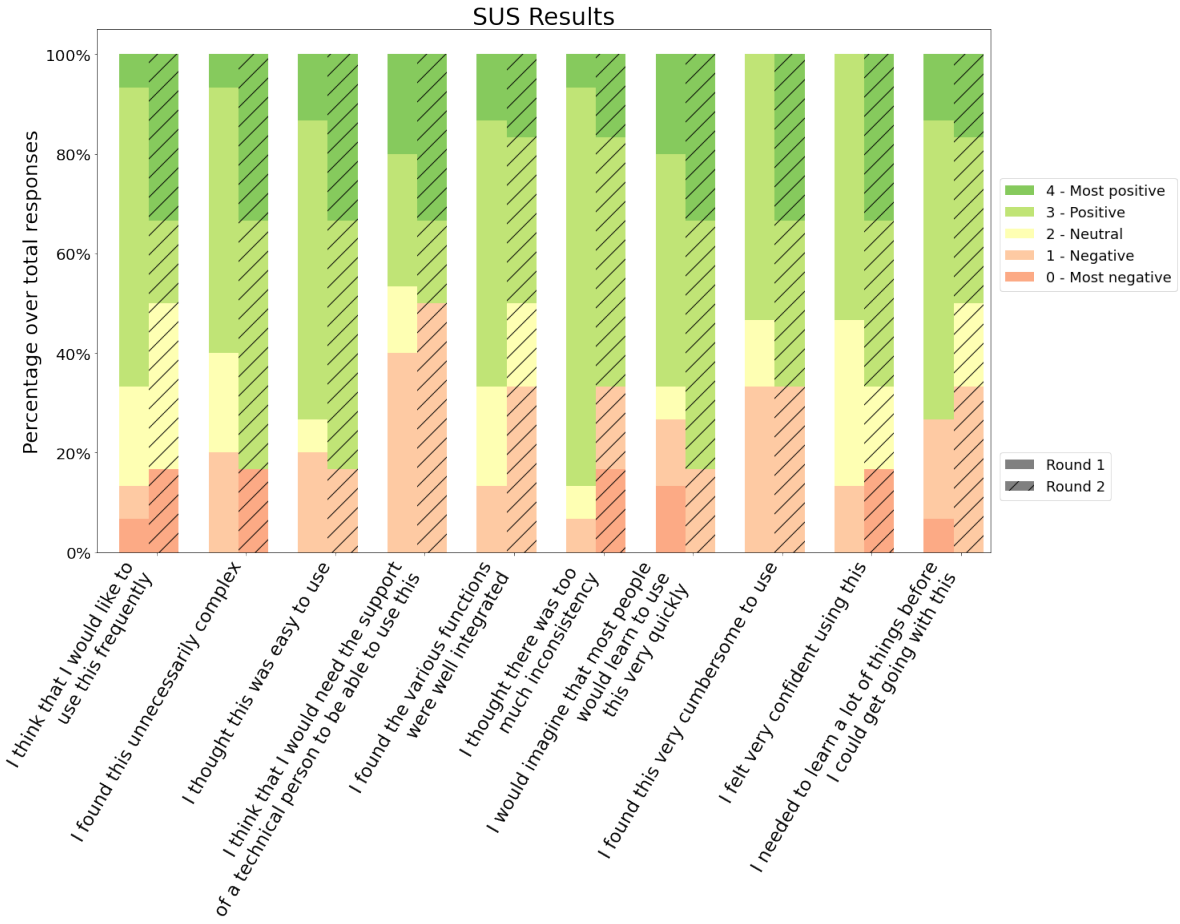


Figure 6.4: Responses to the SUS questionnaire. Bars represent the distribution of the responses using the adjusted scores: raw scores minus 1 for items in the odd position; 5 minus raw scores for items in an even position.

Authors of the UEQ-S provide a benchmark in order to compare the study results, despite this benchmark being intended for the full-size UEQ, the results may be acceptable to estimate how good the user experience is. Figure 6.6 shows the results of both rounds in the three categories and their benchmark score.

6.4 Discussion

In this study, we performed an iterative user-centred validation of a CDSS aimed to support healthcare professionals in the identification of patients in need of palliative care. This two-round validation process involved decision, usability and user experience tests. During this study, the predictions provided by the models were more accurate in both sensitivity and specificity metrics for both classification models than those provided by the participants. The regression model accuracy result depends on the width of the interval; we selected 4 months as an acceptable error based on the original reported mean absolute error [Blanes-Selva et al., 2022a]. Other studies have described the low accuracy of clinicians when predicting one-year mortality through mechanisms

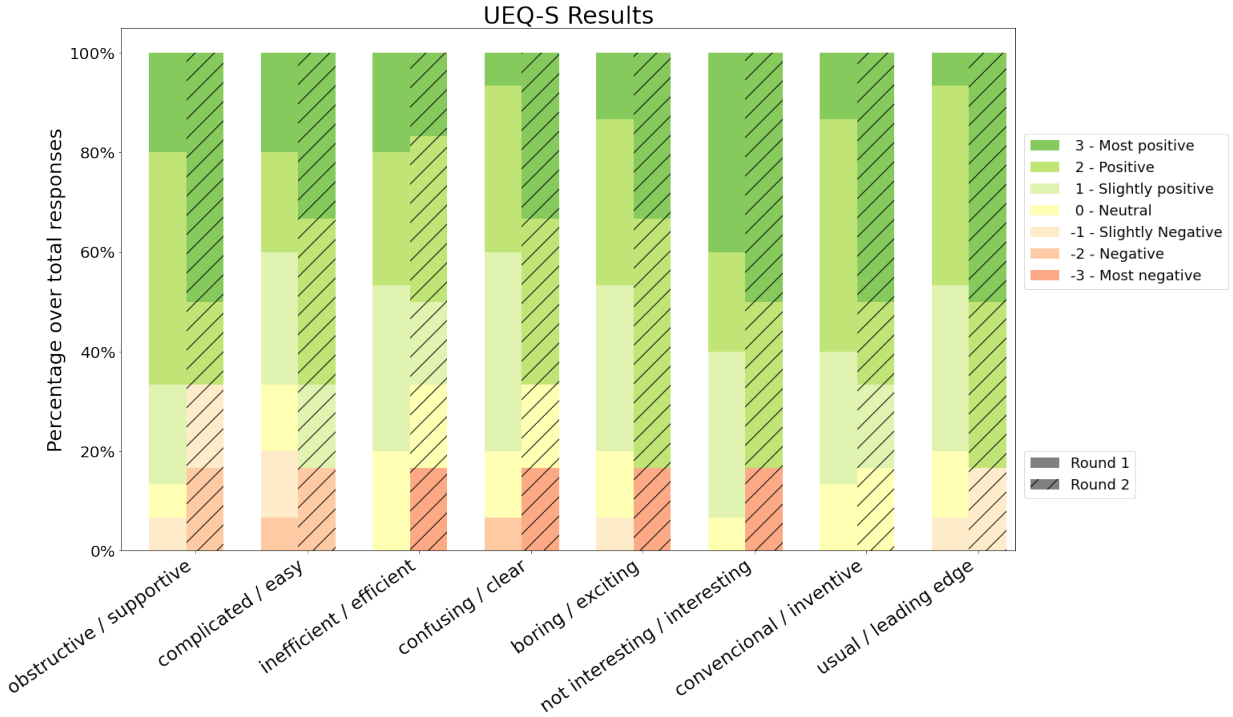


Figure 6.5: Results for the UEQ questionnaire. Bars represent the distribution of the response. Positive responses mean that the participants agreed with the positive quality of the software (supportive, easy...)

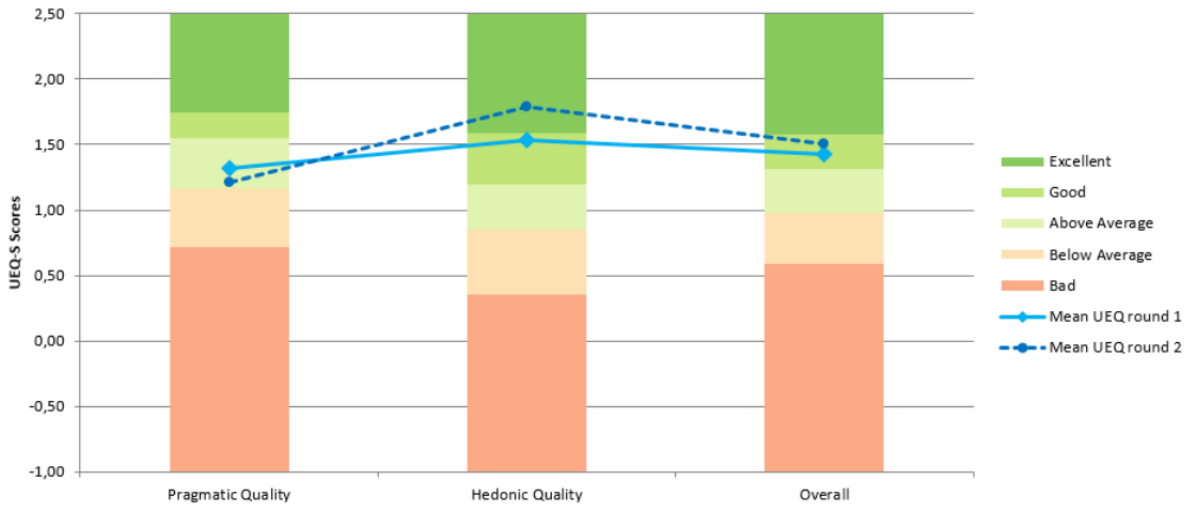


Figure 6.6: UEQ results by categories against the official benchmark. The grade assigned to each category depends on the results from different studies used to create the benchmark.

like the Surprise Question [Downar et al., 2017]. However, our intention with this comparison was to set a reasonable baseline for the predictive models. There are several factors that play against the performance of clinicians in this evaluation 1) they are not good at taking decisions over EHR data [Yang et al., 2019], 2) not having physical access to the patients affect HP's intuition [Melin-Johansson et al., 2017] and 3) the number of cases to evaluate was low. However, these results indicate that The

Aleph PC could help improve the clinicians' predictions with data.

The results of the task test indicate that the four tasks were not perceived as difficult. Task 4 (load a file in the platform) presented the highest number of negative responses regarding its difficulty during the first round, but it improved at the second round after we relocated the load button from the input variables section to the files section, as suggested by the participants. The tasks were also perceived as fast to carry on, despite the need to input all the variables manually in the first task. Levels of satisfaction were high, increasing in the last two tasks for the second round after the interface improvement. It is worth mentioning that all participants in the second round rated positively the option to store the data into a csv file so they would be able to revisit the case later.

The different SUS dimensions were considered positive (scores 3 and 4) by at least half of the participants in both rounds. The best scores were related to the perceived difficulty: *"I found the system unnecessarily complex"*, *"I thought the system was easy to use"* and *"I would imagine that most people would learn to use this system very quickly"*. Consistency was also rated among the best (*"I thought there was too much inconsistency in this system"*). The worst scores were related to the confidence of the participants (*"I found the system very cumbersome to use"*, *"I felt very confident using the system"* and *"I think that I would need the support of a technical person to be able to use this system"*). These results suggest that participants found The Aleph PC easy to use but containing some elements that they were not familiar with and/or needed further explanation. The average score was "passing" though there is room for improvement according to participants' feedback.

UEQ-S results reported in Figure 6.5 provide two main interpretations, first, the median UX scores were higher in the second round after the introduction of improvements. Second, most of the dimensions kept their values on neutral (0) and positive (1, 2, 3) values. Some outliers responses can be observed especially on the second round corresponding with the sceptical participant, which means that their experience with PC CDSS was positive, a deep look into the supportive category is needed to address possible undetected issues.

After analysing the results from SUS, the UEQ-S and the comments obtained from the think-aloud methodology we obtained a rich vision of the perception of the application. It seems that most of the participants were predisposed to use a tool such as The Aleph PC to obtain prognosis predictions that could influence PC action. Even the two participants who expressed their dislike of the CDSS and showed scepticism about the tests had their own vision about how the application should be: whereas the first-round participant did not feel that the tool was useful for their service, the second-round participant specified *"The tool is boring. The end-of-life idea should be obtained in 10-15 seconds"*. Only a few participants had experience using CDSSs, but all of them understood these kinds of applications as supporting tools instead of as a threat to their autonomy [Liberati et al., 2017, Esmaeilzadeh et al., 2015, Laka et al., 2021]. In addition, all of them understood the difficulty in identifying early PC patients so they were prone to accept a tool that may help them.

Although, the usability results were not as positive as the hedonic quality. Some hypotheses could be extracted from the sessions. Mainly, The Aleph PC did not show

a perfect fit in its current status to the different participant backgrounds. Participants from non-hospital environments commented on how some of the variables required by the models did not match with the information managed by their centres. Being unfamiliar with some required input could have a detriment on the perceived ease of use of the product. Four of the participants commented on the nature of the data introduced, they considered that introducing historical data in the application instead of a "snapshot" of a given moment could be better for the prediction. Also, a couple of the hospital physicians made suggestions about automatising the input and the output and their integration with the EHR following the schema of integrative CDSS [Yu et al., 2018]. These findings are in line with the concept of unremarkable AI [Yang et al., 2019], where the AI systems are meant to be integrated in the current workflow and did not disturb or overwork the HPs.

As stated by other works on validation, usability is a key factor of the success of the CDSS implementation. Usability tests based on the performance of tasks are sometimes described as "near-live" simulations, and the posterior usability assessment is the standard to discover usability issues and improve the final product [Press et al., 2015, Akhloufi et al., 2019]. Nonetheless, developing a perfectly usable application does not guarantee its implementation success since there are a list of socio-cultural barriers in the adoption of these technologies [Liberati et al., 2017] which are directly related to the vision and opinions of the physicians and their organisations regarding these products.

The inclusion of the UX test during the evaluation sessions allowed to detect participants' predisposition and feeling towards The Aleph and the general idea of using a CDSS as a daily tool in clinical practice. Despite the diversity of backgrounds among the participants, there was an agreement on the usefulness of the tool. Also, the whole set of participants, with two exceptions, believed in the technology and the evidence behind the predictive models. This is especially relevant since trust has been detected as one of the main issues regarding the CDSS acceptance [Liberati et al., 2017].

The inclusion of two rounds allowed us to test if the changes implemented after the first set of sessions influenced the usability and UX during the second round. The difference between SUS and UEQ-S overall scores were not significant using the T-Test ($p > 0.05$). However, we observed an improvement in certain dimensions of the metrics. We could not extract valid comparisons per role due to sample size restrictions since most of the nurses were in the second round, and most of the physicians were in the first round. At the same time, those groups were not homogeneous and contained HPs working in hospitals, primary care centres, external services and rural environments. Regarding the number of iterations, we could have set a bigger number in order to ensure that the minimum number of issues is kept in the software. However, the changes identified during the second round were either detail such as the use of abbreviations and acronyms or barriers derived from the data source from which the models were created. Therefore, we considered that most of the fixable issues were identified and we didn't need to perform any extra iterations.

The participation of professionals with different roles and backgrounds allowed us to observe the diverse needs in the highly heterogeneous PC implementation and workflows. There are significant differences between inpatient and outpatient set-

tings [Collingridge Moore et al., 2020], medicine specialities [Ferrell et al., 2020, Steigleder et al., 2019, Fricker and Serper, 2019], and urban or rural environments [Finucane et al., 2021]. In those different contexts, physicians and nurses play different roles in the PC needs identification. For example, in the work by Zemplényi *et al.* [Zemplényi et al., 2020] the authors describe how the nurses are the first to detect the needs and then the cases are discussed with the physicians. This may be different in other settings, such as rural areas where physicians visit older patients. In our study, we observed that participants working in non-hospital environments were more concerned about the availability of the variables, especially the laboratory results. Both physicians and nurses thought they could benefit from The Aleph in their workflow.

Another relevant detail in our implementation is the inclusion of ML explainability in the system (Figure 6.1b). ML explainability could defined as the human quality to understand the relation between the system input and their predictions [Reyes et al., 2020] and has been proposed many times as a solution to one of the most common CDSS adoption barriers, as stated by Shortliffe and Sepúlveda "black boxes are unacceptable" [Shortliffe and Sepúlveda, 2018]. CDSSs should be transparent to the user to allow them to accept or dismiss the prediction or recommendation. However, recent studies have highlighted possible problems when trying to create explainability mechanisms to single predictions. In its viewpoint, Ghassemi *et al.* [Ghassemi et al., 2021] discourage their implementation as patient-level systems. Since we received positive feedback from the participants on this feature, we decided to not remove the explainability graphs after the second round. However, we acknowledge the need for further study for these kinds of features and their impact on the clinical workflow.

Since the start of this work, our team has followed the design recommendations from previous studies, focusing on two main aspects: team composition and interface design. Our team included multiple roles: physicians, designers, usability and UX experts, ML researchers and programmers. This is especially relevant since a multidisciplinary team can get a better understanding of the real requisites of the project and mitigate workflow disruption [Yu et al., 2018, Mahadevaiah et al., 2020]. The interface was carefully designed, the layout was implemented focusing on usability and the colours used were extracted from the PC logo that was created previously by an artist. As described in [Moran, 2017], the aesthetic part of the application has a direct effect on the perceived usability, therefore an effort to create a visually attractive application must be set in place. The scores obtained in the UEQ-S hedonic category reflect the acceptability of the aesthetics, however, none of the participants commented explicitly on the visual aspect.

The main strength of out work was that our methodology assisted us to obtain the insight of different pitfalls identified in previous works [Sutton et al., 2020, Khairat et al., 2018, Liberati et al., 2017, Yu et al., 2018, Carroll et al., 2002, Press et al., 2015] using HPs' insights. Through the usability test we discovered that the system is *good enough* for the participants. However, concrete changes are needed depending on the context where the CDSS is deployed to maximise the usability aspect. With the predictive models being evaluated in a previous publication [Blanes-Selva et al., 2022a] and the evaluation of usability and UX in this work we followed an exhaustive user-centred validation path. We created anecdotal evidence to support the UX dimension

within the standard usability tests. In addition to this, we managed to get a very diverse sample of participants in terms of roles and countries, providing us with a richer version of the health providers' needs regarding PC in different countries.

Our work also presents some limitations. First, the model evaluation was performed with a low number of participants, and since the ML models are deterministic once trained, the evaluation on the machine side is only on the six different cases. However, with respect to the accuracy of the predictions, this is not a problem since the models were evaluated previously. Another limitation in our study is the requirement of manual input of hospital admission data because it is disconnected from the EHR. This could present a perk in roles such as primary care physicians working in rural areas but breaks the premise of automatizing the data collection as much as possible and increases the possibility of human errors. In addition, in this demo, we have not addressed some problems regarding the variability, temporal and related to different medical centres, over data distributions [Sáez et al., 2020].

As future work, we would like to adapt the tool to the different roles and clinical workflows we have identified. A further study focused on the different PC roles and their needs regarding The Aleph PC would be needed to provide a perfect fit and improve usability. Further adaptations and validations of the ML models would be needed to ensure the models keep their predictive power in other populations. In order to go further with The Aleph PC, we would need to create a pilot for potential users to incorporate the tools in their daily routine and gather long-term feedback. Further research about the ML explainability and reportability in needed to create a transparent and auditable system that could improve the acceptance of the technology by helping avoid legal problems [Yu et al., 2018, Reyes et al., 2020]. In addition, a study focused on the mortality and frailty prediction accuracy by HPs may be needed to estimate a fair baseline for predictive models to improve.

6.5 Conclusions

Our main findings indicate that the predictive models have performed better than the baseline composed of the HPs predictions. The system presents great UX hedonic qualities, i.e., participants were excited to use the tool, and they rated positively the fact of having helped to identify patients with bad prognosis. They did not feel their independence threatened by the Aleph PC. Performance regarding usability was modest but acceptable. Based on the notes taken during the think-aloud methodology, the authors hypothesise that the usability scores for the current version are maximised and would only improve if the tool was adapted for the different roles and contexts represented in the participant's sample. We have created anecdotal evidence that an iterative user-centred validation, including UX, provides a broader vision to address CDSS acceptance issues. The objective of The Aleph PC is to step further in the objective PC criteria inclusion.

Chapter 7

Concluding remarks and recommendations

This chapter summarises the work carried on during the thesis, the results and the knowledge obtained. In addition, guidelines for extending and further developing this scientific research are provided.

7.1 Concluding remarks

Research on PC is relatively recent since its first standards were introduced by the WHO in 1990 [Milligan and Potts, 2009], and their implementation on clinical practice varies strongly between regions. Despite the existing agreement in which topics should be addressed within the PC programs: symptoms and pain relief, psychological and spiritual needs and assistance to familiars and caregivers as broad categories, there is still a large number of questions to be answered. During this thesis, we have focused on providing an answer to the question: *"How can we identify patients that could benefit from PC using clinical data?"*. In order to answer this question, we proposed the adoption of criteria based on mortality and frailty, which can be modelled as predictive models. In addition, we have included these predictive models into a software stack designed to impact clinical practice.

The research carried during this thesis contributed to the SoA in the medical informatics, PC and user-interaction fields. The results of this thesis have been endorsed by publications in peer-review scientific journals specialised in these topics. Furthermore, the technological results of this thesis derived into an open-access digital platform for its use in the research community and its posterior industrialisation.

The specific Concluding Remarks (CR) of this thesis are listed as follows.

CR1 Supervised Machine Learning algorithms are adequate to predict accurately One-Year Mortality using data collected from patients few hours after their admission to the hospital, including basic demographics, administrative information, laboratory results and the presence or absence of some medical conditions. However, the internal representation of these models offered the relative importance of the variables. We concluded that administrative information such as the service

where the patient is admitted and the number of previous stays, which indicate the frequency of their health resources consumption and the blood results for BUN, Leucocytes and CRP as well as their age, are the most significant factors when predicting OYM on all-cause admitted adult patients.

*This concluding remark responds to the research questions **RQ1** and **RQ3**, covers the objectives **O1** and **O2** and was derived from the work in publication **P1**.*

CR2 Frailty indexes based on the accumulation of deficits can be used as part of the criteria for PC program referral for older patients admitted to the hospital. The frailty index on [Blanes-Selva et al., 2022a] using the recommendations from [Mitnitski et al., 2001] had a weak correlation with the mortality criteria, making it a complementary criterion. Besides, our ML model to classify future frailty status based on this index reported a great predictive power. And, as far as we know, at the time of the publication, it was the only work predicting frailty without a proxy. In addition to the frailty model, including a mortality regression model was especially relevant to complement the information provided by the OYM, significantly when this one predicted positive for mortality.

*This concluding remark responds to the research questions **RQ2** and **RQ4**, covers the objectives **O3**, **O4**, **O5** and **O6** and was derived from publication **P2**.*

CR3 We have confirmed the importance of some of the CDSS critical adoption factors: answer a relevant question, provide confidence to the final user, being intuitive and well integrated with the clinical workflow. We concluded that a software evaluation including potential users is needed to detect usability problems and possible design flaws. We have realised that one of the essential parts of the validation is the unstructured feedback provided by the participants, in our case following the *think-aloud* methodology. The feedback obtained during the validation process, specially the comments provided orally by the participants helped to improve the current version of the platform. In our experience, the inclusion of a UX evaluation allowed us to assess the participant's perception of the systems in terms of perceived value, trust and relevance of the problem.

*This concluding remark responds to the research question **RQ5** and **RQ6**, covers the objectives **O7** and **O8** and was derived from publications **P3** and **P4**.*

7.2 Recommendations

PC is a growing clinical research topic due to its high impact on the patients QoL and the expected growth in their demand. It is estimated that almost 75% of the population would require some kind of PC intervention in 2040 [Etkind et al., 2017]. Despite this high demand forecast, the clinical decisions involving PC are often taken using subjective our outdated criteria. As stated in this thesis, AI, especially data-driven models, could help in some aspects of the PC clinical pathway. However, the challenges do not end with constructing a good predictive model. Implementing a CDSS or any other software device requires a careful design. Iterative cycles involving:

implementation, evaluation and re-design are required in order to avoid the different adoption barriers.

This methods and research findings of this thesis point to the incorporation of data-driven PC referral tools and can serve as a starting point for further research. In this sense, the following recommendations are suggested.

R1 The research carried out during this thesis has offered evidence that, when using the OYM as PC referral criterion, predictive models using EHR data outperform other subjective methods such as the SQ in standard evaluation metrics [Downar et al., 2017]. Relevant variables for this problem are the service where the patient was admitted on, BUN, Leukocytes, CRP or Age. When focusing the OYM prediction problem to older patients, some of the variables that gained importance were the Charlson Index [Charlson et al., 1994], the number of prescribed medications and the Barthel Index [Mahoney et al., 1965].

In this sense, we recommend shifting traditional screening from ACP needs based on subjective opinions to data-driven approaches feed by EHR data. Since CDSS could predict multiple cases quickly, usually less than a second, depending on the software implementation, we also recommend their adoption as an automatised process to run upon admission and data availability.

R2 Frailty is a common clinical syndrome in older adults that carries an increased risk for poor health outcomes [Xue, 2011]. Some approaches have defined frailty as an accumulation of deficits and quantified it as an index [Mitnitski et al., 2001]. Due to its nature, it can be relevant to PC teams to assess the current frailty status of the patients. However, there are no clear criteria or thresholds for initiating PC among this type of patient [Hamaker et al., 2020]. Including a prediction of how frail the patient will become during the following year could be helpful to decide if that patient is a good fit for PC. As part of the work carried on in this thesis, we constructed a frailty index, used thresholds to create categories and created ML classification models to predict one-year frailty status. We encourage researchers to validate the use of frailty indexes as PC referral criteria using prospective data from other centres.

R3 A one-year period until *exitus* has been used as a proxy for 'bad outcome' in the PC prognosis field. We have demonstrated that a combination of relevant variables and ML models can predict this event accurately. However, during the development of this research, we found that estimating survival time was very relevant from an organisational standpoint, especially in cases where it is predicted an *exitus* within the year. A predicted survival time of two weeks is very distinct from a nine months survival period when exploring the open pathways and possibilities for an effective PC program. During this thesis, we developed a survival regression model predicting days from admission to death. We have checked that it is possible to obtain an *approximate* prediction and, that this model could help complement the OYM criterion, especially when this last one returns a positive result. Therefore, we recommend the use of both regression and OYM to obtain a more informed prediction when using the mortality criterion.

R4 Differences in the data distributions between centres are very likely to affect predictive models performances. Multiple factors can produce different data distribution over the same variables on the EHRs, for instance, the demographic differences between regions, changes in the population or the diseases over time or modifications in clinical protocols that could affect the data acquisition. Therefore, our recommendation in this point is two-fold: first, external validation is required before implementing ML models trained with data from other centres; and second, it is necessary to keep in place mechanisms that monitor the *degradation* of the predictive model (increasing unfitness of the models due to dataset shifts) and triggers a re-training and re-evaluation mechanism. This training process can focus on newer data in order to adapt the models to the new distribution. Another option is to use *continual learning* [Lee and Lee, 2020], and keep the models updated with the latest data available. This set of practices usually belongs to a methodology named Machine Learning Operations (MLOps) [John et al., 2021].

R5 Designing and developing a CDSS is a complex task. During this thesis, we implemented the Aleph, a multi-purpose clinical decision platform, in its first version. During the design phase, we gathered a team including HPs, designers, CDSS researchers and programmers. From there, we applied an iterative internal validation and designed an UX and usability evaluation with external HPs. Despite this approach, which follows the recommendations in the literature, we still identified usability issues and possible adoptions barriers.

Based on our experience building the Aleph PC, we recommend following the previously mentioned steps. The inclusion on *near-live* situations and the feedback obtained using a methodology similar to the *think-aloud* provide insightful data on how to improve the systems. In addition, incorporating UX elements to the evaluation could help to assess some socio-cultural barriers, such as the perceived importance of the problem. Therefore, we suggest paying particular attention to the external validation design.

R6 The Aleph was the result of the research carried on during this thesis, the generated knowledge, and the different scientific outputs. The Aleph offers an environment for different predictive services to be deployed with little effort, offering common services to other researchers and developers like dynamic interface creation, user management and service versioning. Alongside the platform and two predictive services for PC based on our scientific results, we provided a guide[q] on implementing a service compatible with Aleph. We, therefore, encourage other researchers of different medical fields to take advantage of The Aleph to minimise their effort regarding software implementation and, at the same time, increase Aleph's services diversity.

[q]https://thealeph.upv.es/download_integration_guide - Accessed July 23, 2022

Bibliography

Jun Liang, Ying Li, Zhongan Zhang, Dongxia Shen, Jie Xu, Xu Zheng, Tong Wang, Buzhou Tang, Jianbo Lei, and Jiajie Zhang. Adoption of electronic health records (ehrs) in china during the past 10 years: Consecutive survey data analysis and comparison of sino-american challenges and experiences. *Journal of medical Internet research*, 23(2):e24813, 2021.

Rhidian A Hughes. Clinical practice in a computer world: considering the issues. *Journal of Advanced Nursing*, 42(4):340–346, 2003.

Chris A Mack. Fifty years of moore's law. *IEEE Transactions on semiconductor manufacturing*, 24 (2):202–207, 2011.

Paul Compton and R Jansen. Knowledge in context: A strategy for expert system maintenance. In *Australian Joint Conference on Artificial Intelligence*, pages 292–306. Springer, 1988.

Reza Khajouei and MWM Jaspers. The impact of cpoe medication systems' design aspects on usability, workflow and medication orders. *Methods of information in medicine*, 49(01):03–19, 2010.

S. N. Etkind, A. E. Bone, B. Gomes, N. Lovell, C. J. Evans, I. J. Higginson, and F. E. M. Murtagh. How many people will need palliative care in 2040? past trends, future projections and implications for services. *BMC Medicine*, 15(1):102, May 2017. ISSN 1741-7015. doi: 10.1186/s12916-017-0860-2. URL https://doi.org/10.1186/s12916-017-0860-2.

Vicent Blanes-Selva, Vicente Ruiz-García, Salvador Tortajada, José-Miguel Benedí, Bernardo Valdivieso, and Juan M García-Gómez. Design of 1-year mortality forecast at hospital admission: A machine learning approach. *Health Informatics Journal*, 27(1), January 2021a.

Vicent Blanes-Selva, Ascensión Doñate-Martínez, Gordon Linklater, and Juan M García-Gómez. Complementary frailty and mortality prediction models on older patients as a tool for assessing palliative care needs. *Health Informatics Journal*, 28(2), 2022a. doi: 10.1177/14604582221092592.

Vicent Blanes-Selva, Ascensión Doñate-Martínez, Gordon Linklater, Jorge Garcés-Ferrer, and Juan M García-Gómez. Responsive and minimalist app based on explainable ai to assess palliative care needs during bedside consultations on older patients. *Sustainability*, 13(17):9844, September 2021b.

Vicent Blanes-Selva, Sabina Asensio-Cuesta, Ascensión Doñate-Martínez, Felipe Pereira Mesquita, and Juan Miguel García-Gómez. Validating a Clinical Decision Support System for Palliative Care using healthcare professionals' insights. *medRxiv*, 2022b. doi: 10.1101/2022.06.03.22275904. URL https://www.medrxiv.org/content/early/2022/06/05/2022.06.03.22275904.

Vicent Blanes-Selva, Salvador Tortajada, Ruth Vilar, Bernardo Valdivieso, and Juan M García-Gómez. Machine learning-based identification of obesity from positive and unlabelled electronic health records. In *Digital Personalized Health and Medicine*, pages 864–868. IOS Press, 2020.

Argyro Mavrogiorgou, Athanasios Kiourtis, Ilias Maglogiannis, Dimosthenis Kyriazis, Antonio De Nigro, Vicent Blanes-Selva, Juan M García-Gómez, Andreas Menychtas, Gregor Jurak, et al. Crowdhealth: An e-health big data driven platform towards public health policies. 2020.

Sabina Asensio-Cuesta, Vicent Blanes-Selva, J Alberto Conejero, Ana Frigola, Manuel G Portolés, Juan Francisco Merino-Torres, Matilde Rubio Almanza, Shabbir Syed-Abdul, Yu-Chuan Jack Li, Ruth Vilar-Mateo, et al. A user-centered chatbot (wakamola) to collect linked data in population networks to support studies of overweight and obesity causes: design and pilot study. *JMIR Medical Informatics*, 9(4):e17503, 2021a.

Sabina Asensio-Cuesta, Vicent Blanes-Selva, Manuel Portolés, J Alberto Conejero, and Juan M García-Gómez. How the wakamola chatbot studied a university community's lifestyle during the covid-19 confinement. *Health Informatics Journal*, 27(2):14604582211017944, 2021b.

Sabina Asensio-Cuesta, Vicent Blanes-Selva, Alberto Conejero, Manuel Portolés, and Miguel García-Gómez. A user-centered chatbot to identify and interconnect individual, social and environmental risk factors related to overweight and obesity. *Informatics for Health and Social Care*, pages 1–15, 2021c.

Pablo Ferri, Carlos Sáez, Antonio Félix-De Castro, Javier Juan-Albarracín, Vicent Blanes-Selva, Purificación Sánchez-Cuesta, and Juan M García-Gómez. Deep ensemble multitask classification of emergency medical call incidents combining multimodal data improves emergency medical dispatch. *Artificial Intelligence in Medicine*, 117:102088, 2021.

World Health Organization WHO. Palliative care, August 2020. URL https://www.who.int/news-room/fact-sheets/detail/palliative-care. Last accesed July 23, 2022.

Terry Fulmer, Marcus Escobedo, Amy Berman, Mary Jane Koren, Sandra Hernández, and Angela Hult. Physicians' views on advance care planning and end-of-life care conversations. *Journal of the American Geriatrics Society*, 66(6):1201–1205, 2018.

Stuart Milligan and Shirley Potts. The history of palliative care. *Palliative Nursing: Across the Spectrum of Care*, pages 1–16, 2009.

Caroline Richmond. Dame cicely saunders. *BMJ: British Medical Journal*, 331(7510):238, 2005.

Vinita Mahtani-Chugani, Inmaculada González-Castro, Amaia Sáenz de Ormijana-Hernández, Roberto Martín-Fernández, and Enrique Fernández de la Vega. How to provide care for patients suffering from terminal non-oncological diseases: barriers to a palliative care approach. *Palliative medicine*, 24(8):787–795, 2010.

Marie A Bakitas, Tor D Tosteson, Zhigang Li, Kathleen D Lyons, Jay G Hull, Zhongze Li, J Nicholas Dionne-Odom, Jennifer Frost, Konstantin H Dragnev, Mark T Hegel, et al. Early versus delayed initiation of concurrent palliative oncology care: patient outcomes in the enable iii randomized controlled trial. *Journal of clinical oncology*, 33(13):1438, 2015.

Marie Bakitas, Kathleen Doyle Lyons, Mark T Hegel, Stefan Balan, Frances C Brokaw, Janette Seville, Jay G Hull, Zhongze Li, Tor D Tosteson, Ira R Byock, et al. Effects of a palliative care intervention on clinical outcomes in patients with advanced cancer: the project enable ii randomized controlled trial. *Jama*, 302(7):741–749, 2009.

Jennifer S Temel, Joseph A Greer, Alona Muzikansky, Emily R Gallagher, Sonal Admane, Vicki A Jackson, Constance M Dahlin, Craig D Blinderman, Juliet Jacobsen, William F Pirl, et al. Early palliative care for patients with metastatic non–small-cell lung cancer. *New England Journal of Medicine*, 363(8):733–742, 2010.

Stacy M Fischer, Wendolyn S Gozansky, Angela Sauaia, Sung-Joon Min, Jean S Kutner, and Andrew Kramer. A practical tool to identify patients who may benefit from a palliative approach: the caring criteria. *Journal of pain and symptom management*, 31(4):285–292, 2006.

K Thomas, J Armstrong Wilson, GSF Team, et al. Proactive identification guidance (pig) national gold standards framework centre in end of life care. 2016, 2017. URL https://www.goldstandardsframework.org.uk/.

Nicola White, Nuriye Kupeli, Victoria Vickerstaff, and Patrick Stone. How accurate is the 'surprise question'at identifying patients at the end of life? a systematic review and meta-analysis. *BMC medicine*, 15(1):1–14, 2017.

James Downar, Russell Goldman, Ruxandra Pinto, Marina Englesakis, and Neill K.J. Adhikari. The "surprise question" for predicting death in seriously ill patients: a systematic review and meta-analysis. *CMAJ*, 189(13):E484–E493, 2017. ISSN 0820-3946. doi: 10.1503/cmaj.160775. URL https://www.cmaj.ca/content/189/13/E484.

Judith E Nelson, J Randall Curtis, Colleen Mulkerin, Margaret Campbell, Dana R Lustbader, Anne C Mosenthal, Kathleen Puntillo, Daniel E Ray, Rick Bassett, Renee D Boss, et al. Choosing and using screening criteria for palliative care consultation in the icu: a report from the improving palliative care in the icu (ipal-icu) advisory board. *Critical care medicine*, 41(10):2318–2327, 2013.

National Hospice Organization. Medical guidelines for determining prognosis in selected non-cancer diseases. *The Hospice Journal*, 11(2):47–63, 1996.

Marco Pirovano, Marco Maltoni, Oriana Nanni, Mauro Marinari, Monica Indelli, Giovanni Zaninetta, Vincenzo Petrella, Sandro Barni, Ernesto Zecca, Emanuela Scarpi, et al. A new palliative prognostic score: a first step for the staging of terminally ill cancer patients. *Journal of pain and symptom management*, 17(4):231–239, 1999.

T Morita, Junichi Tsunoda, Satoshi Inoue, and Satoshi Chihara. The palliative prognostic index: a scoring system for survival prediction of terminally ill cancer patients. *Supportive care in cancer*, 7(3):128–133, 1999.

Francis Lau, Michael Downing, Mary Lesperance, Nicholas Karlson, Craig Kuziemsky, and Ju Yang. Using the palliative performance scale to provide meaningful survival estimates. *Journal of pain and symptom management*, 38(1):134–144, 2009.

Gill Highet, Debbie Crawford, Scott A Murray, and Kirsty Boyd. Development and evaluation of the supportive and palliative care indicators tool (spict): a mixed-methods study. *BMJ supportive & palliative care*, 4(3):285–290, 2014.

M Bernabeu-Wittel, M Ollero-Baturone, L Moreno-Gaviño, B Barón-Franco, A Fuertes, J Murcia-Zaragoza, C Ramos-Cantos, A Alemán, and A Fernández-Moyano. Development of a new predictive model for polypathological patients. the profund index. *European journal of internal medicine*, 22 (3):311–317, 2011.

Carl van Walraven, Finlay A McAlister, Jeffrey A Bakal, Steven Hawken, and Jacques Donzé. External validation of the hospital-patient one-year mortality risk (homr) model for predicting death within 1 year after hospital admission. *Cmaj*, 187(10):725–733, 2015.

Carl van Walraven and Alan J. Forster. The homr-now! model accurately predicts 1-year death risk for hospitalized patients on admission. *The American Journal of Medicine*, 130(8):991.e9–991.e16, 2017. ISSN 0002-9343. doi: https://doi.org/10.1016/j.amjmed.2017.03.008. URL https://www.sciencedirect.com/science/article/pii/S0002934317303200.

Pete Wegier, Allison Kurahashi, Stephanie Saunders, Bhadra Lokuge, Leah Steinberg, Jeff Myers, Ellen Koo, Carl van Walraven, and James Downar. mhomr: a prospective observational study of an automated mortality prediction model to identify patients with unmet palliative needs. *BMJ Supportive & Palliative Care*, 2021.

Anand Avati, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H Shah. Improving palliative care with deep learning. *BMC medical informatics and decision making*, 18 (4):55–64, 2018.

J Matthew Helm, Andrew M Swiergosz, Heather S Haeberle, Jaret M Karnuta, Jonathan L Schaffer, Viktor E Krebs, Andrew I Spitzer, and Prem N Ramkumar. Machine learning and artificial intelligence: definitions, applications, and future directions. *Current reviews in musculoskeletal medicine*, 13(1):69–76, 2020.

Frances Y Kuo and Ian H Sloan. Lifting the curse of dimensionality. *Notices of the AMS*, 52(11): 1320–1328, 2005.

Arthur Szlam, Yuval Kluger, and Mark Tygert. An implementation of a randomized algorithm for principal component analysis. *arXiv preprint arXiv:1412.3510*, 2014.

Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18(1998):1–8, 1998.

Yasi Wang, Hongxun Yao, and Sicheng Zhao. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242, 2016.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Stefano Nembrini, Inke R König, and Marvin N Wright. The revival of the gini importance? *Bioinformatics*, 34(21):3711–3718, 2018.

Donato Hernández Fusilier, Manuel Montes-y Gómez, Paolo Rosso, and Rafael Guzmán Cabrera. Detecting positive and negative deceptive opinions using pu-learning. *Information processing & management*, 51(4):433–443, 2015.

Lluis Marquez, Lluis Padro, and Horacio Rodriguez. A machine learning approach to pos tagging. *Machine Learning*, 39(1):59–91, 2000.

Isabelle Guyon, B Boser, and Vladimir Vapnik. Automatic capacity tuning of very large vc-dimension classifiers. In *Advances in neural information processing systems*, pages 147–155, 1993.

William A Belson. Matching and prediction on the principle of biological classification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 8(2):65–75, 1959.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

Bjoern H Menze, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, 10(1):1–16, 2009.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4): 367–378, 2002.

Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, December 1943. ISSN 1522-9602. doi: 10.1007/BF02478259.

F. Rosenblatt. The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain. *Psychological Review*, pages 65–386, 1958.

Alekseĭ Grigorevich Ivakhnenko, A G Ivakhnenko, Valentin Grigorevich Lapa, and Valentin Grigorevich Lapa. *Cybernetics and forecasting techniques*, volume 8. American Elsevier Publishing Company, 1967.

David A Winkler and Tu C Le. Performance of deep and shallow neural networks, the universal approximation theorem, activity cliffs, and qsar. *Molecular informatics*, 36(1-2):1600118, 2017.

Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02): 107–116, 1998.

Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc. event-place: Lake Tahoe, Nevada.

Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer, 2013.

Richard D Riley, Joie Ensor, Kym IE Snell, Thomas PA Debray, Doug G Altman, Karel GM Moons, and Gary S Collins. External validation of clinical prediction models using big datasets from e-health records or ipd meta-analysis: opportunities and challenges. *bmj*, 353, 2016.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Thomas CW Landgrebe and Robert PW Duin. Approximating the multiclass roc by pairwise analysis. *Pattern recognition letters*, 28(13):1747–1758, 2007.

Edward H Shortliffe and Bruce G Buchanan. A model of inexact reasoning in medicine. *Mathematical biosciences*, 23(3-4):351–379, 1975.

Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas, and David F Lobach. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj*, 330(7494):765, 2005.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):1–10, 2020.

Tiffani J Bright, Anthony Wong, Ravi Dhurjati, Erin Bristow, Lori Bastian, Remy R Coeytaux, Gregory Samsa, Vic Hasselblad, John W Williams, Michael D Musty, et al. Effect of clinical decision-support systems: a systematic review. *Annals of internal medicine*, 157(1):29–43, 2012.

Vitri Tundjungsari, Abdul Salam Mudzakir Sofro, Ahmad Sabiq, and Aan Kardiana. Investigating clinical decision support systems success factors with usability testing. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 8(11), 2017.

Kun-Hsing Yu, Andrew L. Beam, and Isaac S. Kohane. Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10):719–731, Oct 2018. ISSN 2157-846X. doi: 10.1038/s41551-018-0305-z. URL https://doi.org/10.1038/s41551-018-0305-z.

Ida Sim and Amy Berlin. A framework for classifying decision support systems. In *AMIA Annual Symposium Proceedings*, volume 2003, page 599. American Medical Informatics Association, 2003.

PK Anooj. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University-Computer and Information Sciences*, 24(1):27–40, 2012.

Ravi K Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A Helvie, Jun Wei, and Kenny Cha. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Medical physics*, 43(12):6654–6666, 2016.

Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158–164, 2018.

M Bikromjit Khumancha, Aarti Barai, and CB Rama Rao. Lung cancer detection from computed tomography (ct) scans using convolutional neural network. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, 2019.

Maggie Makar, Marzyeh Ghassemi, David M Cutler, and Ziad Obermeyer. Short-term mortality prediction for elderly patients using medicare claims data. *International journal of machine learning and computing*, 5(3):192, 2015.

Matthew M Churpek, Trevor C Yuen, Christopher Winslow, Ari A Robicsek, David O Meltzer, Robert D Gibbons, and Dana P Edelson. Multicenter development and validation of a risk stratification tool for ward patients. *American journal of respiratory and critical care medicine*, 190(6):649–655, 2014.

Amy S Kelley and R Sean Morrison. Palliative care for the seriously ill. *New England Journal of Medicine*, 373(8):747–755, 2015.

Camilla Zimmermann, Nadia Swami, Monika Krzyzanowska, Breffni Hannon, Natasha Leighl, Amit Oza, Malcolm Moore, Anne Rydall, Gary Rodin, Ian Tannock, et al. Early palliative care for patients with advanced cancer: a cluster-randomised controlled trial. *The Lancet*, 383(9930):1721–1730, 2014.

Alvin H Moss, Jesse Ganjoo, Sanjay Sharma, Julie Gansor, Sharon Senft, Barbara Weaner, Cheryl Dalton, Karen MacKay, Beth Pellegrino, Priya Anantharaman, et al. Utility of the "surprise" question to identify dialysis patients with high mortality. *Clinical Journal of the American Society of Nephrology*, 3(5):1379–1384, 2008.

Bianca M Buurman, Barbara C Van Munster, Johanna C Korevaar, Ameen Abu-Hanna, Marcel Levi, and Sophia E De Rooij. Prognostication in acutely admitted older patients by nurses and physicians. *Journal of general internal medicine*, 23(11):1883–1889, 2008.

Florence I Mahoney et al. Functional evaluation: the barthel index. *Maryland state medical journal*, 14(2):61–65, 1965.

Mary Charlson, Ted P Szatrowski, Janey Peterson, and Jeffrey Gold. Validation of a combined comorbidity index. *Journal of clinical epidemiology*, 47(11):1245–1251, 1994.

Giampiero Mazzaglia, Carlo Piccinni, Alessandro Filippi, Giovanna Sini, Francesco Lapi, Emiliano Sessa, Iacopo Cricelli, Paola Cutroneo, Gianluca Trifirò, Claudio Cricelli, et al. Effects of a computerized decision support system in improving pharmacological management in high-risk cardiovascular patients: A cluster-randomized open-label controlled trial. *Health informatics journal*, 22 (2):232–247, 2016.

Margaret O'Connor, Trudy Erwin, and Linda Dawson. A means to an end: a web-based client management system in palliative care. *Health informatics journal*, 15(1):41–54, 2009.

Sydney M Dy, Jayashree Roy, Geoffrey E Ott, Michael McHale, Christine Kennedy, Jean S Kutner, and Allen Tien. Tell us™: a web-based tool for improving communication among patients, families, and providers in hospice and palliative care through systematic data specification, collection, and use. *Journal of pain and symptom management*, 42(4):526–534, 2011.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

GE Hinton and J Carbonell. Connectionist learning procedures machine learning, 1990.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

M Abadi, P Barham, Jea Chen, Z Chen, A Davis, J Dean, M Devin, S Ghemawat, G Irving, M Isard, et al. Osdi'16: Proceedings of the 12th usenix conference on operating systems design and implementation. *Berkeley: USENIX Association*, pages 265–283, 2016.

Randal S Olson, Ryan J Urbanowicz, Peter C Andrews, Nicole A Lavender, La Creis Kidd, and Jason H Moore. Automating biomedical data science through tree-based pipeline optimization. In *European conference on the applications of evolutionary computation*, pages 123–137. Springer, 2016.

Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Rachel Klinedinst, Z Noah Kornfield, and Rachel A Hadler. Palliative care for patients with advanced heart disease. *J. Cardiothorac. Vasc. Anesth.*, 33(3):833–843, mar 2019.

Doran Bostwick, Steven Wolf, Greg Samsa, Janet Bull, Donald H Taylor, Jr, Kimberly S Johnson, and Arif H Kamal. Comparing the palliative care needs of those with cancer to those with common non-cancer serious illness. *J. Pain Symptom Manage.*, 53(6):1079–1084.e1, jun 2017.

Deborah Birch and Jan Draper. A critical literature review exploring the challenges of delivering effective palliative care to older people with dementia. *J. Clin. Nurs.*, 17(9):1144–1163, may 2008.

Scott A Murray, Marilyn Kendall, Kirsty Boyd, and Aziz Sheikh. Illness trajectories and palliative care. *BMJ*, 330(7498):1007–1011, April 2005.

Mary V Callaway, Stephen R Connor, and Kathleen M Foley. World health organization public health model: a roadmap for palliative care development. *Journal of pain and symptom management*, 55 (2):S6–S13, 2018.

Sriram Yennurajalingam, Diana L. Urbauer, Katie L.B. Casper, Cielito C. Reyes-Gibby, Ray Chacko, Valerie Poulter, and Eduardo Bruera. Impact of a palliative care consultation team on cancer-related symptoms in advanced cancer patients referred to an outpatient supportive care clinic. *Journal of Pain and Symptom Management*, 41(1):49–56, 2011. ISSN 0885-3924. doi: https://doi.org/10.1016/j.jpainsymman.2010.03.017. URL https://www.sciencedirect.com/science/article/pii/S0885392410005026.

Kieran L Quinn, Therese Stukel, Nathan M Stall, Anjie Huang, Sarina Isenberg, Peter Tanuseputro, Russell Goldman, Peter Cram, Dio Kavalieratos, Allan S Detsky, and Chaim M Bell. Association between palliative care and healthcare outcomes among adults with terminal non-cancer illness: population based matched cohort study. *BMJ*, 370, 2020. doi: 10.1136/bmj.m2257. URL https://www.bmj.com/content/370/bmj.m2257.

Sonja McIlfatrick. Assessing palliative care needs: views of patients, informal carers and healthcare professionals. *Journal of advanced nursing*, 57(1):77–86, 2007.

Irene Higginson and Julia M Addington-Hall. *Palliative care for non-cancer patients*. Oxford University Press, 2001.

Amy EH Kingston, Jennifer Kirkland, and Alexandra Hadjimichalis. Palliative care in non-malignant disease. *Medicine*, 48(1):37–42, 2020.

Scott A Murray, Adam Firth, Nils Schneider, Bart Van den Eynden, Xavier Gomez-Batiste, Trine Brogaard, Tiago Villanueva, Jurgen Abela, Steffen Eychmuller, Geoffrey Mitchell, et al. Promoting palliative care in the community: production of the primary palliative care toolkit by the european association of palliative care taskforce in primary palliative care. *Palliative medicine*, 29(2):101–111, 2015.

Andrew Clegg, John Young, Steve Iliffe, Marcel Olde Rikkert, and Kenneth Rockwood. Frailty in elderly people. *The lancet*, 381(9868):752–762, 2013.

Linda P Fried, Catherine M Tangen, Jeremy Walston, Anne B Newman, Calvin Hirsch, John Gottdiener, Teresa Seeman, Russell Tracy, Willem J Kop, Gregory Burke, et al. Frailty in older adults: evidence for a phenotype. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 56(3):M146–M157, 2001.

Arnold B Mitnitski, Alexander J Mogilner, and Kenneth Rockwood. Accumulation of deficits as a proxy measure of aging. *TheScientificWorldJournal*, 1:323–336, 2001.

Robbert JJ Gobbens, Katrien G Luijkx, Maria Th Wijnen-Sponselee, and Jos MGA Schols. In search of an integral conceptual definition of frailty: opinions of experts. *Journal of the American Medical Directors Association*, 11(5):338–343, 2010.

Barbara M Raudonis and Kathryn Daniel. Frailty: an indication for palliative care. *Geriatric Nursing*, 31(5):379–384, 2010.

Katalin Koller and Kenneth Rockwood. Frailty in older adults: implications for end-of-life care. *Cleveland Clinic journal of medicine*, 80(3):168–174, 2013.

Gordon Linklater, Sally Lawton, Shona Fielding, Lisa Macaulay, David Carroll, and Dong Pang. Introducing the palliative performance scale to clinicians: the grampian experience. *BMJ supportive & palliative care*, 2(2):121–126, 2012.

Anne Woolfield, Geoffrey Mitchell, Srinivas Kondalsamy-Chennakesavan, and Hugh Senior. Predicting those who are at risk of dying within six to twelve months in primary care: A retrospective case–control general practice chart analysis. *Journal of palliative medicine*, 22(11):1417–1424, 2019.

Pete Wegier, Ellen Koo, Shahin Ansari, Daniel Kobewka, Erin O'Connor, Peter Wu, Leah Steinberg, Chaim Bell, Tara Walton, Carl van Walraven, Gayathri Embuldeniya, Judy Costello, and James Downar. mhomr: a feasibility study of an automated system for identifying inpatients having an elevated risk of 1-year mortality. *BMJ Quality & Safety*, 28(12):971–979, 2019. ISSN 2044-5415. doi: 10.1136/bmjqs-2018-009285. URL https://qualitysafety.bmj.com/content/28/12/971.

Amy S Porter, Stephanie Harman, and Joshua R Lakin. Power and perils of prediction in palliative care. *Lancet (London, England)*, 395(10225):680–681, 2020.

Tatyana Shamliyan, Kristine MC Talley, Rema Ramakrishnan, and Robert L Kane. Association of frailty with survival: a systematic literature review. *Ageing research reviews*, 12(2):719–736, 2013.

František Babič, Ljiljana Trtica Majnarić, Sanja Bekić, and Andreas Holzinger. Machine learning for family doctors: a case of cluster analysis for studying aging associated comorbidities and frailty. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 178–194. Springer, 2019.

Shelley A Sternberg, Netta Bentur, Chad Abrams, Tal Spalter, Tomas Karpati, John Lemberger, and Anthony D Heymann. Identifying frail older people using predictive modeling. *The American journal of managed care*, 18(10):e392–7, 2012.

Debra Saliba, Marc Elliott, Laurence Z Rubenstein, David H Solomon, Roy T Young, Caren J Kamberg, RN Carol Roth, Catherine H MacLean, Paul G Shekelle, Elizabeth M Sloss, et al. The vulnerable elders survey: a tool for identifying vulnerable older people in the community. *Journal of the American Geriatrics Society*, 49(12):1691–1699, 2001.

Flavio Bertini, Giacomo Bergami, Danilo Montesi, Giacomo Veronese, Giulio Marchesini, and Paolo Pandolfi. Predicting frailty condition in elderly using multidimensional socioclinical databases. *Proceedings of the IEEE*, 106(4):723–737, 2018.

Samuel D Searle, Arnold Mitnitski, Evelyne A Gahbauer, Thomas M Gill, and Kenneth Rockwood. A standard procedure for creating a frailty index. *BMC geriatrics*, 8(1):1–10, 2008.

Samir Touzani, Jessica Granderson, and Samuel Fernandes. Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, 158:1533–1543, 2018.

Xing Chen, Li Huang, Di Xie, and Qi Zhao. Egbmmda: extreme gradient boosting machine for mirna-disease association prediction. *Cell death & disease*, 9(1):1–16, 2018.

Jian Zhou, Enming Li, Shan Yang, Mingzheng Wang, Xiuzhi Shi, Shu Yao, and Hani S Mitri. Slope stability prediction for circular mode failure using gradient boosting machine approach based on an updated database of case histories. *Safety Science*, 118:505–518, 2019.

Matt W Gardner and SR Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Francesco Piccialli, Vittorio Di Somma, Fabio Giampaolo, Salvatore Cuomo, and Giancarlo Fortino. A survey on deep learning in medicine: Why, how and when? *Information Fusion*, 66:111–137, 2021.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Guido VanRossum and Fred L Drake. *The python language reference*. Python Software Foundation Amsterdam, Netherlands, 2010.

Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32: 8026–8037, 2019.

Cameron Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4 (40):1317, 2019.

Lindsey C Yourman, Sei J Lee, Mara A Schonberg, Eric W Widera, and Alexander K Smith. Prognostic indices for older adults: a systematic review. *Jama*, 307(2):182–192, 2012.

Pere Almagro, Sergi Yun, Ana Sangil, Mónica Rodríguez-Carballeira, Meritxell Marine, Pedro Landete, Juan José Soler-Cataluña, Joan B Soriano, and Marc Miravitlles. Palliative care and prognosis in copd: a systematic review with a validation cohort. *International journal of chronic obstructive pulmonary disease*, 12:1721, 2017.

S Hajioff. Computerized decision support systems: an overview. *Health Informatics Journal*, 4(1): 23–28, 1998.

Vicent Blanes-Selva. Palliative care models webapp Demo Aleph, 2021. URL http://demoiapc.upv.es/.

Carlos Sáez, Nekane Romero, J Alberto Conejero, and Juan M García-Gómez. Potential limitations in covid-19 machine learning due to data source variability: A case study in the ncov2019 dataset. *Journal of the American Medical Informatics Association*, 28(2):360–364, 2021.

Carlos Sáez and Juan M García-Gómez. Kinematics of big biomedical data to characterize temporal variability and seasonality of data repositories: functional data analysis of data temporal evolution over non-parametric statistical manifolds. *International journal of medical informatics*, 119: 109–124, 2018.

Carlos Sáez, Alba Gutiérrez-Sacristán, Isaac Kohane, Juan M García-Gómez, and Paul Avillach. Ehrtemporalvariability: delineating temporal data-set shifts in electronic health records. *GigaScience*, 9(8):giaa079, 2020.

Kenneth Jung, Sylvia EK Sudat, Nicole Kwon, Walter F Stewart, and Nigam H Shah. Predicting need for advanced illness or palliative care in a primary care population using electronic health record data. *Journal of biomedical informatics*, 92:103115, 2019.

JL Roberts, World Health Organization, et al. Terminology: A glossary of technical terms on the economics and finance of health services. Technical report, Copenhagen: WHO Regional Office for Europe, 1998.

Elio Borgonovi, Paola Adinolfi, Rocco Palumbo, and Gabriella Piscopo. Framing the shades of sustainability in health care: Pitfalls and perspectives from western eu countries. *Sustainability*, 10 (12), 2018. ISSN 2071-1050. doi: 10.3390/su10124439. URL https://www.mdpi.com/2071-1050/10/12/4439.

Kwadwo Kyeremanteng, Louis-Philippe Gagnon, Kednapa Thavorn, Daren Heyland, and Gianni D'Egidio. The impact of palliative care consultation in the icu on length of stay: A systematic review and cost evaluation. *Journal of Intensive Care Medicine*, 33(6):346–353, 2018. doi: 10.1177/0885066616664329. URL https://doi.org/10.1177/0885066616664329. PMID: 27582396.

Birgitta Wallerstedt, Eva Benzein, Kristina Schildmeijer, and Anna Sandgren. What is palliative care? perceptions of healthcare professionals. *Scandinavian Journal of Caring Sciences*, 33(1): 77–84, 2019. doi: https://doi.org/10.1111/scs.12603. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/scs.12603.

Lukas Radbruch, Liliana De Lima, Felicia Knaul, Roberto Wenk, Zipporah Ali, Sushma Bhatnaghar, Charmaine Blanchard, Eduardo Bruera, Rosa Buitrago, Claudia Burla, Mary Callaway, Esther Cege Munyoro, Carlos Centeno, Jim Cleary, Stephen Connor, Odontuya Davaasuren, Julia Downing, Kathleen Foley, Cynthia Goh, Wendy Gomez-Garcia, Richard Harding, Quach T. Khan, Phillippe Larkin, Mhoira Leng, Emmanuel Luyirika, Joan Marston, Sebastien Moine, Hibah Osman, Katherine Pettus, Christina Puchalski, M.R. Rajagopal, Dingle Spence, Odette Spruijt, Chitra Venkateswaran, Bee Wee, Roger Woodruff, Jinsun Yong, and Tania Pastrana. Redefining palliative care—a new consensus-based definition. *Journal of Pain and Symptom Management*, 60(4):754–764, 2020. ISSN 0885-3924. doi: https://doi.org/10.1016/j.jpainsymman.2020.04.027. URL https://www.sciencedirect.com/science/article/pii/S0885392420302475.

Steven Simoens, Betty Kutten, Emmanuel Keirse, Paul Vanden Berghe, Claire Beguin, Marianne Desmedt, Myriam Deveugele, Christian Léonard, Dominique Paulus, and Johan Menten. The costs of treating terminal patients. *Journal of Pain and Symptom Management*, 40(3):436–448, 2010. ISSN 0885-3924. doi: https://doi.org/10.1016/j.jpainsymman.2009.12.022. URL https://www.sciencedirect.com/science/article/pii/S0885392410003635.

Samantha Smith, Aoife Brick, Sinéad O'Hara, and Charles Normand. Evidence on the cost and cost-effectiveness of palliative care: A literature review. *Palliative Medicine*, 28(2):130–150, 2014. doi: 10.1177/0269216313493466. URL https://doi.org/10.1177/0269216313493466. PMID: 23838378.

Neha Kayastha and Thomas W. LeBlanc. When to integrate palliative care in the trajectory of cancer care. *Current Treatment Options in Oncology*, 21(5):41, Apr 2020. doi: 10.1007/s11864-020-00743-x. URL https://doi.org/10.1007/s11864-020-00743-x.

Geetha Mahadevaiah, Prasad RV, Inigo Bermejo, David Jaffray, Andre Dekker, and Leonard Wee. Artificial intelligence-based clinical decision support in modern medical physics: Selection, acceptance, commissioning, and quality assurance. *Medical Physics*, 47(5):e228–e235, 2020. doi: https://doi.org/10.1002/mp.13562. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13562.

Stephen J Leslie, Mark Hartswood, Catrin Meurig, Sinead P McKee, Roger Slack, Rob Procter, and Martin A Denvir. Clinical decision support software for management of chronic heart failure: Development and evaluation. *Computers in Biology and Medicine*, 36(5):495–506, 2006. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2005.02.002. URL https://www.sciencedirect.com/science/article/pii/S0010482505000478.

Frederic Ehrler, Thomas Weinhold, Jonathan Joe, Christian Lovis, and Katherine Blondon. A mobile app (bedside mobility) to support nurses' tasks at the patient's bedside: Usability study. *JMIR Mhealth Uhealth*, 6(3):e57, Mar 2018. ISSN 2291-5222. doi: 10.2196/mhealth.9079. URL http://mhealth.jmir.org/2018/3/e57/.

Blaine Reeder, Cynthia Drake, Mustafa Ozkaynak, Wallace Jones, David Mack, Alexandria David, Raven Starr, Barbara Trautner, and Heidi L. Wald. Usability inspection of a mobile clinical decision support app and a short form heuristic evaluation checklist. In Dylan D. Schmorrow and Cali M. Fidopiastis, editors, *Augmented Cognition*, pages 331–344, Cham, 2019. Springer International Publishing. ISBN 978-3-030-22419-6.

Vikesh K. Singh, Bechien U. Wu, Thomas L. Bollen, Kathryn Repas, Rie Maurer, Richard S. Johannes, Koenraad J. Mortele, Darwin L. Conwell, and Peter A. Banks. A prospective evaluation of the bedside index for severity in acute pancreatitis score in assessing mortality and intermediate markers of severity in acute pancreatitis. *Official journal of the American College of Gastroenterology | ACG*, 104(4), 2009. ISSN 0002-9270. URL https://journals.lww.com/ajg/Fulltext/2009/04000/A_Prospective_Evaluation_of_the_Bedside_Index_for.26.aspx.

Martin J. O'Donnell, Jiming Fang, Cami D'Uva, Gustavo Saposnik, Linda Gould, Emer McGrath, Moira K. Kapral, and for the Investigators of the Registry of the Canadian Stroke Network. The PLAN Score: A Bedside Prediction Rule for Death and Severe Disability Following Acute Ischemic Stroke. *Archives of Internal Medicine*, 172(20):1548–1556, 11 2012. ISSN 0003-9926. doi: 10.1001/2013.jamainternmed.30. URL https://doi.org/10.1001/2013.jamainternmed.30.

Laurent G. Glance, Stewart J. Lustik, Edward L. Hannan, Turner M. Osler, Dana B. Mukamel, Feng Qian, and Andrew W. Dick. The surgical mortality probability model: Derivation and validation of a simple risk prediction rule for noncardiac surgery. *Annals of Surgery*, 255(4), 2012. ISSN 0003-4932. URL https://journals.lww.com/annalsofsurgery/Fulltext/2012/04000/The_Surgical_Mortality_Probability_Model_.13.aspx.

Dimitris Bertsimas, Galit Lukin, Luca Mingardi, Omid Nohadani, Agni Orfanoudaki, Bartolomeo Stellato, Holly Wiberg, Sara Gonzalez-Garcia, Carlos Luis Parra-Calderón, Kenneth Robinson, Michelle Schneider, Barry Stein, Alberto Estirado, Lia a Beccara, Rosario Canino, Martina Dal Bello, Federica Pezzetti, Angelo Pan, and The Hellenic COVID-19 Study Group. Covid-19 mortality risk assessment: An international multi-center study. *PLOS ONE*, 15(12):1–13, 12 2020. doi: 10.1371/journal.pone.0243262. URL https://doi.org/10.1371/journal.pone.0243262.

Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(96)00142-2. URL https://www.sciencedirect.com/science/article/pii/S0031320396001422.

Sebastien Haneuse, David Arterburn, and Michael J Daniels. Assessing missing data assumptions in ehr-based studies: A complex and underappreciated task. *JAMA Network Open*, 4(2): e210184–e210184, 2021.

Matthew Sperrin and Glen P. Martin. Multiple imputation with missing indicators as proxies for unmeasured variables: simulation study. *BMC Medical Research Methodology*, 20(1), Jul 2020. ISSN 1471-2288. doi: 10.1186/s12874-020-01068-x. URL https://doi.org/10.1186/s12874-020-01068-x.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

Javier A Bargas-Avila, O Brenzikofer, SP Roth, AN Tuch, S Orsini, and K Opwis. Simple but crucial user interfaces in the world wide web: introducing 20 guidelines for usable web form design, user interfaces. 2010.

Sharon K Inouye, Sidney T Bogardus Jr, Gail Vitagliano, Mayur M Desai, Christianna S Williams, Jacqueline N Grady, and Jeanne D Scinto. Burden of illness score for elderly persons: risk adjustment incorporating the cumulative impact of diseases, physiologic abnormalities, and functional impairments. *Medical care*, pages 70–83, 2003.

Okan Arihan, Bernhard Wernly, Michael Lichtenauer, Marcus Franz, Bjoern Kabisch, Johanna Muessig, Maryna Masyuk, Alexander Lauten, Paul Christian Schulze, Uta C Hoppe, et al. Blood urea nitrogen (bun) is independently associated with mortality in critically ill patients admitted to icu. *PloS one*, 13(1):e0191697, 2018.

Bechien U Wu, Richard S Johannes, Xiaowu Sun, Darwin L Conwell, and Peter A Banks. Early changes in blood urea nitrogen predict mortality in acute pancreatitis. *Gastroenterology*, 137(1): 129–135, 2009.

Clay A Cauthen, Michael J Lipinski, Antonio Abbate, Darryn Appleton, Annunziata Nusca, Amit Varma, Evelyne Goudreau, Michael J Cowley, and George W Vetrovec. Relation of blood urea nitrogen to long-term mortality in patients with heart failure. *The American journal of cardiology*, 101(11):1643–1647, 2008.

Mar Martín-Pérez, Ana Ruigómez, Roberto Pastor-Barriuso, Fernando J García López, Ana Villaverde-Hueso, and Javier Damián. Number of medications and mortality among residents in nursing homes. *Journal of the American Medical Directors Association*, 20(5):643–645, 2019.

Jelle de Gelder, Jacinta A Lucke, Noor Heim, Antonius JM de Craen, Shantaily D Lourens, Ewout W Steyerberg, Bas de Groot, Anne J Fogteloo, Gerard J Blauw, and Simon P Mooijaart. Predicting mortality in acutely hospitalized older patients: a retrospective cohort study. *Internal and emergency medicine*, 11(4):587–594, 2016.

Kushang V Patel, Richard D Semba, Luigi Ferrucci, Anne B Newman, Linda P Fried, Robert B Wallace, Stefania Bandinelli, Caroline S Phillips, Binbing Yu, Stephanie Connelly, et al. Red cell distribution width and mortality in older adults: a meta-analysis. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 65(3):258–265, 2010.

Edward H Shortliffe and Martin J Sepúlveda. Clinical decision support in the era of artificial intelligence. *Jama*, 320(21):2199–2200, 2018.

Carmen Carroll, Phil Marsden, Pat Soden, Emma Naylor, John New, and Tim Dornan. Involving users in the design and usability evaluation of a clinical decision support system. *Computer methods and programs in biomedicine*, 69(2):123–135, 2002.

Markus W Haun, Stephanie Estel, Gerta Ruecker, Hans-Christoph Friederich, Matthias Villalobos, Michael Thomas, and Mechthild Hartmann. Early palliative care for adults with advanced cancer. *Cochrane Database of Systematic Reviews*, (6), 2017.

Peter May, Melissa M Garrido, J Brian Cassel, Amy S Kelley, Diane E Meier, Charles Normand, Thomas J Smith, Lee Stefanis, and R Sean Morrison. Prospective cohort study of hospital palliative care teams for inpatients with advanced cancer: earlier consultation is associated with larger cost-saving effect. *Journal of Clinical Oncology*, 33(25):2745, 2015.

William J Lowery, Ashlei W Lowery, Jason C Barnett, Micael Lopez-Acevedo, Paula S Lee, Angeles Alvarez Secord, and Laura Havrilesky. Cost-effectiveness of early palliative care intervention in recurrent platinum-resistant ovarian cancer. *Gynecologic oncology*, 130(3):426–430, 2013.

Carolyn Petersen, Samantha A Adams, and Paul R DeMuro. mhealth: don't forget all the stakeholders in the business case. *Medicine 2.0*, 4(2), 2015.

Kurubaran Ganasegeran and Surajudeen Abiola Abdulrahman. Adopting m-health in clinical practice: A boon or a bane? In *Telemedicine Technologies*, pages 31–41. Elsevier, 2019.

Eta S Berner. *Clinical decision support systems*, volume 233. Springer, 2007.

Arnaud Belard, Timothy Buchman, Jonathan Forsberg, Benjamin K Potter, Christopher J Dente, Allan Kirk, and Eric Elster. Precision diagnosis: a view of the clinical decision support systems (cdss) landscape through the lens of critical care. *Journal of clinical monitoring and computing*, 31(2):261–271, 2017.

Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. Investigating the heart pump implant decision process: opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4477–4488, 2016.

Srikant Devaraj and Sara Viernes. Barriers and facilitators to clinical decision support systems adoption: A systematic review. *International journal of trends in business administration*, 3(2), 2014.

Glyn Elwyn, Isabelle Scholl, Caroline Tietbohl, Mala Mann, Adrian GK Edwards, Catharine Clay, France Légaré, Trudy van der Weijden, Carmen L Lewis, Richard M Wexler, et al. "many miles to go…": a systematic review of the implementation of patient decision support interventions into routine clinical practice. *BMC medical informatics and decision making*, 13(2):1–10, 2013.

Qian Yang, John Zimmerman, and Aaron Steinfeld. Review of medical decision support tools: Emerging opportunity for interaction design. *IASDR 2015 Interplay Proceedings*, 2015.

Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR medical informatics*, 6(2):e24, 2018.

Elisa G Liberati, Francesca Ruggiero, Laura Galuppo, Mara Gorli, Marien González-Lorenzo, Marco Maraldi, Pietro Ruggieri, Hernan Polo Friz, Giuseppe Scaratti, Koren H Kwag, et al. What hinders the uptake of computerized decision support systems in hospitals? a qualitative study and framework for implementation. *Implementation Science*, 12(1):1–13, 2017.

Anne Press, Lauren McCullagh, Sundas Khan, Andy Schachter, Salvatore Pardo, and Thomas McGinn. Usability testing of a complex clinical decision support tool in the emergency department: lessons learned. *JMIR human factors*, 2(2):e4537, 2015.

Nicholas Genes, Min Soon Kim, Frederick L Thum, Laura Rivera, Rosemary Beato, Carolyn Song, Jared Soriano, Joseph Kannry, Kevin Baumlin, and Ula Hwang. Usability evaluation of a clinical decision support system for geriatric ed pain treatment. *Applied clinical informatics*, 7(01):128–142, 2016.

Safiya Richardson, Rebecca Mishuris, Alexander O'Connell, David Feldstein, Rachel Hess, Paul Smith, Lauren McCullagh, Thomas McGinn, and Devin Mann. "think aloud" and "near live" usability testing of two complex clinical decision support tools. *International journal of medical informatics*, 106:1–8, 2017.

Frederick Thum, Min Soon Kim, Nicholas Genes, Laura Rivera, Rosemary Beato, Jared Soriano, Joseph Kannry, Kevin Baumlin, and Ula Hwang. Usability improvement of a clinical decision support system. In *International Conference of Design, User Experience, and Usability*, pages 125–131. Springer, 2014.

Alice C Li, Joseph L Kannry, Andre Kushniruk, Dillon Chrimes, Thomas G McGinn, Daniel Edonyabo, and Devin M Mann. Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support. *International journal of medical informatics*, 81(11):761–772, 2012.

David W Eccles and Güler Arsal. The think aloud method: what is it and how do i use it? *Qualitative Research in Sport, Exercise and Health*, 9(4):514–531, 2017.

John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194): 4–7, 1996.

James R Lewis. The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction*, 34(7):577–590, 2018.

Dieter P Wallach, Lukas A Flohr, and Annika Kaltenhauser. Beyond the buzzwords: On the perspective of ai in ux and vice versa. In *International Conference on Human-Computer Interaction*, pages 146–166. Springer, 2020.

Nigel Bevan. What is the difference between the purpose of usability and user experience evaluation methods. In *Proceedings of the Workshop UXEM*, volume 9, pages 1–4. Citeseer, 2009.

Kate Moran. The aesthetic-usability effect, 2017. URL https://www.nngroup.com/articles/aesthetic-usability-effect. Accessed 2022-03-02.

Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–14, 2019.

Charlie Parker, Sam Scott, and Alistair Geddes. Snowball sampling. *SAGE research methods foundations*, 2019.

Jacob Nielsen. Why you only need to test with 5 users, 2000. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/, Last accesed July 23, 2022.

Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. Design and evaluation of a short version of the user experience questionnaire (ueq-s). *International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108.*, 2017.

Aaron Bangor, Philip Kortum, and James Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.

Qian Yang, Aaron Steinfeld, and John Zimmerman. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

Christina Melin-Johansson, Rebecca Palmqvist, and Linda Rönnberg. Clinical intuition in the nursing process and decision-making—a mixed-studies review. *Journal of Clinical Nursing*, 26(23-24): 3936–3949, 2017.

Pouyan Esmaeilzadeh, Murali Sambasivan, Naresh Kumar, and Hossein Nezakati. Adoption of clinical decision support systems in a developing country: Antecedents and outcomes of physician's threat to perceived professional autonomy. *International journal of medical informatics*, 84(8):548–560, 2015.

Mah Laka, Adriana Milazzo, and Tracy Merlin. Factors that impact the adoption of clinical decision support systems (cdss) for antibiotic management. *International journal of environmental research and public health*, 18(4):1901, 2021.

Hassna Akhloufi, SJC Verhaegh, MWM Jaspers, DC Melles, Heleen van der Sijs, and Annelies Verbon. A usability study to improve a clinical decision support system for the prescription of antibiotic drugs. *PloS one*, 14(9):e0223073, 2019.

Danni Collingridge Moore, Sheila Payne, Lieve Van den Block, Julie Ling, and Katherine Froggatt. Strategies for the implementation of palliative care education and organizational interventions in long-term care facilities: A scoping review. *Palliative medicine*, 34(5):558–570, 2020.

Betty R Ferrell, Vincent Chung, Marianna Koczywas, and Thomas J Smith. Dissemination and implementation of palliative care in oncology. *Journal of Clinical Oncology*, 38(9):995, 2020.

Tobias Steigleder, Rainer Kollmar, and Christoph Ostgathe. Palliative care for stroke patients and their families: barriers for implementation. *Frontiers in neurology*, 10:164, 2019.

Zachary P Fricker and Marina Serper. Current knowledge, barriers to implementation, and future directions in palliative care for end-stage liver disease. *Liver Transplantation*, 25(5):787–796, 2019.

Anne M Finucane, Hannah O'Donnell, Jean Lugton, Tilly Gibson-Watt, Connie Swenson, and Claudia Pagliari. Digital health interventions in palliative care: a systematic meta-review. *NPJ digital medicine*, 4(1):1–10, 2021.

Antal T Zemplényi, Ágnes Csikós, Marcell Csanádi, Maureen Rutten-van Mölken, Carmen Hernandez, János G Pitter, Thomas Czypionka, Markus Kraus, and Zoltán Kaló. Implementation of palliative care consult service in hungary–integration barriers and facilitators. *BMC Palliative Care*, 19(1): 1–12, 2020.

Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: Artificial Intelligence*, 2(3):e190043, 2020.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

Qian-Li Xue. The frailty syndrome: definition and natural history. *Clinics in geriatric medicine*, 27 (1):1–15, 2011.

ME Hamaker, Frederiek van den Bos, and Siri Rostoft. Frailty and palliative care, 2020.

Cecilia S Lee and Aaron Y Lee. Clinical applications of continual learning machine learning. *The Lancet Digital Health*, 2(6):e279–e281, 2020.

Meenu Mary John, Helena Holmström Olsson, and Jan Bosch. Towards mlops: A framework and maturity model. In *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 1–8. IEEE, 2021.

Appendix A

Appendix: variables and hyperparameters

Variable	Description
Admission Diagnose Code	ICD9 code representing the main reason for the admission.
Age	Patient's age.
Atrial Fibrillation	ICD9 Diagnosis code: Atrial fibrillation (no/yes).
Barthel index	Barthel Index is an ordinal scale used to measure performance in activities of daily living (ADL). Ten variables describing ADL and mobility are scored, a higher number being a reflection of greater ability to function independently following hospital discharge.
Charlson index	The Charlson comorbidity index predicts the one-year mortality for a patient who may have a range of comorbid conditions, such as heart disease, AIDS, or cancer (a total of 17 conditions: Acute myocardial infarction, Congestive heart failure, Peripheral vascular disease, Cerebrovascular disease, Dementia, Chronic lung disease, Mild liver disease, Mild to moderate diabetes, Diabetes with chronic complications, Hemiparaplegia or paraplegia, Kidney disease, Malignant tumours, Moderate to serious liver disease, solid, metastatic tumour and AIDS). Each condition is assigned a score of 1, 2, 3, or 6, depending on the risk of dying associated with each one.
Creatine	Lab result expressed mg/dL.
DRG	Diagnosis-related group (DRG) is a system to classify hospital cases into one of originally 467 groups.
Filtered Glomerular CKD	Filtered Glomerular CKD lab result in ml/min/1,73 m^2.

Gastrointestinal or Liver Disease	ICD9 Diagnosis code: Gastrointestinal or Liver Disease (no/yes).
Glucose	Lab result expressed in mg/dL.
Haematocrit	Lab result expressed in %.
Hypertension	ICD9 Diagnosis code: Hypertension (no/yes).
Leukocyte	10^3/microL.
Number Active groups	Number of active groups (medications) in each episode.
Number of previous stays	Number of previous hospital admissions.
Number Previous ER 365d	Number of previous Emergency Room visits (last 365 days).
Number Previous ER	Number of previous Emergency Room visits.
Metastatic Tumour	ICD9 Diagnosis code: Metastatic tumour (no/yes).
PCR	C-Reactive protein lab result expressed in mg/L.
Potassium	Lab result expressed in mEq/L.
Psychiatric Disease	ICD9 Diagnosis code: Psychiatric disease (No/yes).
RDW-CV	The red cell distribution width (RDW) blood test measures the amount of red blood cell variation in volume and size. This values is the coefficient of variation of RDW.
RDW-SD	Standard deviation of RDW measure.
Service	Last Service updated during the stay.
Sodium	Lab result expressed in mEq/L.
Urea	Lab result expressed in mg/dL.

Table A.1: Variables used in the predictive models and their descriptions

Task	Model	Parameter	Value
OYM	GBM	Criterion	Friedman MSE
		Max depth	5
		Max features	Auto
		n estimators	291
	DNN	Criterion	MSE
		Learning Rate	0.01732471628757128
		Epochs	50
		Activation Function(s)	Leaky ReLU
		Final function	Softmax
		Batch norm	Yes, every layer
		Layer 1 size	512
		Layer 1 dropout	0.45
		Layer 2 size	256
		Layer 2 dropout	0.40
		Layer 3 size	512
		Layer 3 dropout	0.25
		Layer 4 size	512

		Layer 4 dropout	0.34
		Layer 5 size	256
		Layer 5 dropout	0.3
Regression	GBM	Criterion	MSE
		Max depth	5
		Max features	Auto
		n estimators	286
	DNN	Criterion	MSE
		Learning Rate	0.01732471628757128
		Epochs	30
		Activation Function(s)	Leaky ReLU
		Final function	ReLU
		Batch norm	Yes, every layer
		Layer 1 size	256
		Layer 1 dropout	0.23
		Layer 2 size	64
		Layer 2 dropout	0.35
		Layer 3 size	256
		Layer 3 dropout	0.29
		Layer 4 size	64
		Layer 4 dropout	0.44
		Layer 5 size	128
		Layer 5 dropout	0.49
Frailty	GBM	Criterion	MSE
		Max depth	4
		Max features	SQRT
		n estimators	149
	DNN	Criterion	MSE
		Learning Rate	1.301440136399707e-05
		Epochs	100
		Activation Function(s)	Leaky ReLU
		Final function	Softmax
		Batch norm	Yes, every layer
		Layer 1 size	512
		Layer 1 dropout	0.50
		Layer 2 size	128
		Layer 2 dropout	0.284
		Layer 3 size	64
		Layer 3 dropout	0.21
		Layer 4 size	16
		Layer 4 dropout	0.44

Table A.2: Hyperparameters selected by Optuna. The non-specified hyper-parameters have the default value defined in their libraries: scikit-learn v1.0 for the GBM and Pytorch v1.9.1 for the DNN