UNIVERSITA' VITA-SALUTE SAN RAFFAELE

CORSO DI DOTTORATO DI RICERCA

IN MEDICINA MOLECOLARE

Curriculum in Medicina Clinica e Sperimentale

LEVERAGING REAL-WORLD DATA WITH MACHINE LEARNING TO DISENTANGLE THE COMPLEXITY OF MULTIMORBID INTERNAL MEDICINE PATIENTS

Supervisore: Patrizia Rovere Querini

Co-supervisore: Cecilia Mascolo

Tesi di DOTTORATO di RICERCA di Marco Montagna

matr. 022773

Ciclo di dottorato XXXVIII

SSD MEDS-05/A

Anno Accademico 2024/2025

CONSULTAZIONE TESI DI DOTTORATO DI RICERCA

Il/la sottoscritto	Marco Montagna
Matricola	022773
nato a	Sassocorvaro
il	08/01/1992

autore della Tesi di Dottorato di Ricerca dal titolo

Leveraging real-world data with machine learning to disentangle the complexity of multimorbid internal medicine patients

X AUTORIZZA la Consultazione della Tesi

☐ NON AUTORIZZA la Consultazione della Tesi per …… mesi a partire dalla data di sottomissione della domanda di conseguimento titolo

Poiché:

□ l'intera ricerca o parti di essa sono potenzialmente soggette a brevettabilità*;*

□ ci sono parti di Tesi che sono già state sottoposte a un editore o sono in attesa di pubblicazione*;*

□ la Tesi è finanziata da enti esterni che vantano dei diritti su di esse e sulla loro pubblicazione.

E' fatto divieto di riprodurre, in tutto o in parte, quanto in essa contenuto

Data 10/03/2026 Firma ……Marco Montagna.....

DECLARATION

This thesis has been:

- composed by myself and has not been used in any previous application for a degree. Throughout the text I use both 'I' and 'We' interchangeably.
- has been written according to the editing guidelines approved by the University.

Permission to use images and other material covered by copyright has been sought and obtained. For the following image/s (specify), it was not possible to obtain permission and is/are therefore included in thesis under the "fair use" exception (Italian legislative Decree no. 68/2003).

All the results presented here were obtained by myself.

All sources of information are acknowledged by means of reference.

Acknowledgements

This thesis reflects a multidisciplinary effort, and the research described benefited from the expertise, resources, and collegiality of many individuals and organisations. I gratefully acknowledge:

- Current and past people at Porini s.r.l. for believing in the Physician Scientist program offered by Vita-Salute San Raffaele University and deciding to fund a Project-based PhS Fellowship, which then I won: Antonello Giuseppe Bianchi, Stefano Brusamolino, Luca Malinverno, Tommaso Pozzi and Barbara Elvira Ventura. They also contributed to the development of the San Raffaele Ai CEnter (S-RACE) platform, which was instrumental in enabling this work.
- Teams at Microsoft and Sketchin, for taking part in the development of the S-RACE platform by providing the cloud resources and the interface and process designing expertise.
- The current and past data scientists and project managers of the Vita-Salute San Raffaele Artificial Intelligence Team and technicians of the Vita-Salute San Raffaele University IT4Research for their invaluable and tireless technical support: Muhammad Salaar Arslan, Simone Barbieri, Giulio Cielo, Stefano Contini, Andrea Corvaglia, Marco Denti, Alessio Dimonte, Edoardo Draetta, Bruno Fabiani, Francesca Lombardo, Patrick Scuri, Alberto Traverso and Enrico Versino.
- The Cohort Genomic Platform team for setting up the MED-Cli study electronic case report form and continuously updating it according to our requests, and for assisting me with the MySQL database creation, connection and bug-fixing: Corrado Masciullo, Cinzia Sala and Guido Scicolone.
- Emanuela Setola, lead diabetologist at IRCCS San Raffaele Hospital, for her availability in sharing her clinical knowledge and her clinical questions and in discussing the content of data and my results.
- The present and past members of the internal medicine team at Vita-Salute San Raffaele University and IRCCS San Raffaele Hospital, for supporting me with knowledge, feedback, analytical support, data management and study coordination: Marina Biganzoli, Filippo Chiabrando, Chiara Curato, Sarah Damanti, Rebecca De Lorenzo, Alessandra Lucini, Aurora Merolla, Gabriele Mogliarisi, Francesco Paciullo and Aleksandar Svilenov Rabadzhiev.

- The current and past internal medicine residents and the students of medicine of Vita-Salute San Raffaele University, for their tireless effort in managing, optimising and performing the MED-Cli study patient enrolment and data entry, in particular: Chiara Bellino, Giulia Lanzetta, Laura Leoni, Giulia Pata, Chiara Pomaranzi, Elena Rela and Clara Soggetti
- The team who hosted me at Cambridge University Hospital. They allowed and helped me to access real-world data at their hospital and use their JupyterLab, while also discussing the analytical steps with me: Olivia Bentham, Ari Ercole and Xueying Nancy Zheng.
- The Italian Ministry of University and Research for supporting my time in Cambridge, UK, through the international extra Erasmus mobility program.

Dedication

I am quite sure that as a kid I wanted to become a scientist. During elementary school, a teacher would divide the class into groups and ask each group to write a fictional story. My group wrote a zombie-story where the main characters were fictional versions of us, and we would form a team that sought to kill the zombie-boss: my role in the team was "the Scientist/the Inventor". More than 20 years later, here I am, writing the dedication of my PhD thesis in Molecular Medicine. How many people accompanied, trusted and empowered me in this journey? I owe you so much. How many crossroads have I faced? They were so challenging, and I am so lucky to say I would make the same choices again. How many times could I have turned into something else? Yet, here I am, sticking to the childish plan. This is so fascinating, intriguing, and somewhat scary at the same time. I am also quite sure that as a kid I was described by my teachers as quarrelsome. This as well has never left me: I always find something I disapprove of, in anything I come across, especially around me, not *in me*. Self-critique is something I learnt only with time, thanks again to people who accompanied me, the experiences I chose to live and some timely failures. I think this self-critique will be instrumental in pushing me toward continual learning and improvement to serve the community at my best. There are so many patients in need, so many things we can do better and a planet to protect (Did I say I am quarrelsome?). Therefore, this thesis is dedicated to all the people who trusted me and empowered me, allowing me to choose when I was at my crossroads. It is also dedicated to those who reminded me what I wanted to do and what was important, when I could have turned into something else, and to the people who taught me self-critique. I wish you all could read this and know how important you were and are and you will be for everyone you will meet, grow, teach, support, walk with, do adventures with, serve with, pray with, love with.

A heartfelt thanks to my Supervisor, Prof. Rovere Querini, for her everlasting trust in me, for her advice, for involving me in many different activities that fostered my professional growth, and for giving me complete freedom along my career choices.

Finally, I want to thank my Co-Supervisor, Prof. Cecilia Mascolo, for showing me the door, maybe unintentionally. I am the one who has to walk through it.

Abstract

The application of machine learning to healthcare real-world data is emerging as a complementary approach to clinical trials for enabling data-driven decision-making in the evolving population of complex patients managed in the internal medicine setting. To fully realise this potential, significant challenges remain, including data availability and quality limitations and the absence of standardized protocols for model development. In this thesis, we explore how to build an end-to-end machine learning pipeline for real-world data and leverage it for predictive tasks. Specifically, we focus on disease control in diabetes and hospitalisation outcomes in general medicine wards.

Our first contribution is the development of an institutional platform for on-premises real-world data collection, integration and analysis, ultimately enabling model deployment in clinical practice. This multilayer modular platform is compliant with the most recent data protection regulations, adheres to interoperability standards and medical ontologies, and facilitates the prospective validation and adoption of models through user-friendly interfaces.

We then show the feasibility of extracting data of type 2 diabetes mellitus patients stored in electronic health records and using them to train machine learning models for the prediction of improved glycated haemoglobin at three years from baseline. In this context, we systematically evaluate how different data preprocessing strategies affect model performance, and we show how model interpretation can foster clinical discussion and unveil latent insights from our datasets. We additionally demonstrate robust generalisability of these models on an external dataset.

Finally, we investigate how to generate a high-quality real-world data registry of patients managed in general medicine wards to allow improved phenotyping for a multidimensional approach to their complexity. We propose the adoption of the frailty framework to improve the risk prediction of a composite negative outcome of hospitalisation. We employ a model benchmarking pipeline followed by model interpretation to yield clinically meaningful insights.

This thesis advances the currently unmet integration of machine learning with real-world data in internal medicine. Our research enhances the understanding of optimal analytical pipelines and demonstrates the potential of routinely generated clinical data to produce meaningful and actionable insights.

L’applicazione del machine learning ai dati sanitari di mondo reale sta emergendo come un approccio complementare agli studi clinici per permettere di prendere decisioni basate sui dati nella gestione dei pazienti complessi gestiti nelle medicine interne. Per realizzare a pieno questo potenziale restano aperte sfide significative, tra cui limitazioni nella disponibilità e qualità dei dati e l’assenza di protocolli standardizzati per lo sviluppo dei modelli. In questa tesi, esploriamo come costruire una filiera per sfruttare i dati di mondo reale a scopo predittivo. In particolare, ci focalizziamo sul controllo di malattia nel diabete e sugli esiti di ricovero in medicina generale.

La nostra prima contribuzione è lo sviluppo di una piattaforma istituzionale per la raccolta, integrazione e analisi di dati di mondo reale che rende poi possibile l’implementazione dei modelli nella pratica clinica. Questa piattaforma multilivello e modulare rispetta i più recenti regolamenti sulla protezione dei dati, aderisce agli standard di interoperabilità e alle ontologie mediche, e facilita la validazione prospettica e l’adozione dei modelli attraverso una interfaccia utente facile da usare.

Successivamente mostriamo la fattibilità dell’estrazione di dati di pazienti con diabete di tipo 2 dalla cartella medica elettronica e il loro uso per allenare modelli di machine learning per la predizione di un miglioramento dell’emoglobina glicata a tre anni dal basale. In questo contesto, procediamo ad una valutazione sistematica di come diverse strategie di processamento dei dati influenzino la performance dei modelli, e mostriamo come l’interpretabilità dei modelli possa favorire il confronto clinico e far emergere dai nostri dataset conoscenze nascoste. Dimostriamo inoltre una robusta generalizzabilità di questi modelli in un dataset esterno.

Infine, studiamo come generare un registro di dati di mondo reale di alta qualità su pazienti gestiti in reparti di medicina generale per una loro migliore fenotipizzazione al fine di adottare un approccio multidimensionale alla loro complessità. Proponiamo poi l’adozione di un framework basato sulla fragilità per migliorare la predizione di rischio di esiti sfavorevoli dell’ospedalizzazione. Impieghiamo da ultimo una pipeline di valutazione seguita dall’interpretazione dei modelli per fornire intuizioni cliniche.

Questa tesi avanza l’integrazione non ancora raggiunta tra machine learning e dati di mondo reale in medicina interna. La nostra ricerca migliora la comprensione delle filiere di analisi ottimali e dimostra il potenziale che hanno i dati clinici generati quotidianamente nel produrre conoscenze significative e utilizzabili.

Contents

1. Introduction

1.1 Motivation

In recent years, medicine is encountering unprecedented challenges, despite scientific and technological advances. On the one hand, rising numbers and complexity of patients increase the demand for healthcare, fuelled by ageing populations, improved survival to once incurable diseases, evolving patient expectations, and climate change. On the other hand, soaring costs, socioeconomic crisis and political instability are hampering our systems' care delivery capacity, already burdened by providers' burnout (Almyranti *et al*, 2024). The consequence is a mismatch between healthcare demand and delivery which is reaching a tipping point that will likely mean that we fail to achieve the UN Sustainable Development Goal for Health and Wellbeing, including Universal Health Coverage (United Nations, 2015; Hunter, 2023).

Among the technologies under exploration for integration into healthcare, artificial intelligence (AI) is arguably the one that shows the greatest potential to address current challenges and alleviate some of the aforementioned pressures. As healthcare workflow is increasingly being digitalised, the amount of data it produces is steeply accumulating, making it the perfect playground for such a data-hungry technology as AI is (Jha, 2010). With the availability of ever-increasing computing power and cutting-edge architectures, such as transformers, AI promises to help us in analysing healthcare data to improve our preventive, diagnostic, prognostic and therapeutic capacity, possibly reducing costs, inequalities, mistakes, latencies, inappropriateness and many, many more. A peculiar type of data we are interested in are real-world data (RWD), "data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources" (Framework for FDA's Real-World Evidence Program, 2018). The evidence that we distil from RWD is called real-world evidence (RWE) and goes beyond that produced by traditional Randomized Controlled Trials (RCT). FDA's document continues by giving examples of RWD, which "include data derived from electronic health records (EHRs); medical claims and billing data; data from product and disease registries; patient-generated data, including from in-home-use settings; and data gathered from other sources that can inform on health status, such as mobile devices.". It also states that "RWD sources (e.g., registries, collections of EHRs, administrative [*e.g. birth and death records*] and medical claims databases) can be used for data collection and, in certain

cases, to develop analysis infrastructure to support many types of study designs to develop RWE, including, but not limited to, randomized trials (e.g., large simple trials, pragmatic clinical trials) and observational studies (prospective or retrospective).". Given the many existing sources of RWD, it appears evident how studies involving RWD requires extensive data linkage. For example, to understand inequalities like those stemming from socioeconomic deprivation or ethnicity, data on population characteristics in many countries is collected outside of EHRs and then linked within studies to clinical data coming from EHRs.

A particular challenge that practitioners are facing today is that posed by multimorbid older adult patients, individuals affected by multiple coexisting chronic diseases, giving rise to unique, complex disease trajectories that, inevitably, lead to accumulating disability and ultimately to death. These patients are also especially vulnerable due to poor socio-economic conditions and reduced function. Slowing the evolution of their diseases and increasing the life span they spend with good quality of life and preserved function is one of the goal of internal medicine physicians, the "experts in complexity", doctors "especially well trained in the diagnosis of puzzling medical problems, in the ongoing care of chronic illnesses, and in caring for patients with more than one disease" (What is an Internal Medicine Physician, or Internist? | ACP Online).

The author of this thesis is a registered, practicing physician, with a specialty in internal medicine. As such, he copes with the problematic clinical decision making inherent to the management of multimorbid elderly patients stemming from their aforementioned intrinsic clinical complexity and vulnerability, but also from further elements, including: 1) the extremely nuanced expected treatment outcomes in this population, if we consider patient versus care giver versus physician preferences, and 2) the lack of scientific evidence on which to base clinical decisions. In fact, a review published in 2007 on JAMA by Van Spall *et al* stated that "the elderly, and those with common medical conditions are frequently excluded from RCTs. […] Such exclusions may impair the generalizability of RCT results" (Van Spall *et al*, 2007). Other works describe "potential pitfalls of disease-specific guidelines for patients with multiple conditions" (Tinetti *et al*, 2004), the "underrepresentation of elderly patients in adults in COVID-19 trials" (Prendki *et al*, 2020), or the differential outcomes based on frailty in trials involving anticoagulants (Wilkinson *et al*, 2020). However, as highlighted by Van

Spall *et al*, “there are benefits of stringent eligibility criteria. The inclusion and exclusion criteria in a RCT are designed to identify a population of interest in whom an intervention has the greatest likelihood to produce a clinically important and statistically significant effect. Efficacy trials with well-defined and homogeneous populations can generally be smaller, shorter, more efficient, and less expensive. This is desirable because clinical trials are increasingly challenged by high costs, limited funding, and regulatory restrictions.”. Evidently, the matter is subject to ongoing debate, and we cannot expect RCTs to be completely revolutionised soon to include multimorbid or elderly or fragile patients, in any combination. We are therefore interested in RWE as a mean to extend and better generalise, not replace, evidence coming from RCT. A game changing solution would be that of leveraging truly representative data, i.e. RWD, with innovative approaches based on AI, to support clinical decision-making processes grounded on solid RWE.

In this thesis we explore the potential of Machine Learning (ML), the technique that underpins AI, when applied to RWD, to distil RWE in the form of disease evolution predictions in internal medicine patients and answer sample clinical questions for improved decision making.

1.2 Limitations and Challenges

1.2.1 Limitations and Challenges of Machine Learning applications to Real-World Data in Internal Medicine

ML applications to internal medicine are still in their infancy, particularly for the analysis of RWD. If radiology is where ML found it easier to first land thanks to the availability of large labelled datasets of images, narrow classification tasks and already advanced image recognition tools trained in other domains, other medical specialties, more reliant on tabular data to describe their patients and with more complex clinical questions, are still lagging behind in terms of adoption of AI (Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices | FDA). Regardless, ML applications have huge potential for widespread use to leverage RWD of internal medicine patients. However, several challenges and limitations in existing research need to be addressed.

RWD availability: RWD represent a true open challenge for AI researchers due to specific characteristics, starting with their poor availability. First, they are scattered. They are by definition stored in multiple silos, be them the registries of national or local health authorities, the data warehouses of individual hospitals, which often contain multiple compartments themselves, and the Electronic Health Records (EHR) used by various healthcare actors (hospitals, GP surgeries, pharmacies, …). Of note, such data sources are proliferating and compliance to interoperability standards, such as the Health Level Seven International Fast Healthcare Interoperability Resources (FHIR), is far from being optimal. Then, they are burdened by obvious and understandable access restrictions, due to privacy, confidentiality and sensitivity issues. Therefore, key challenges in leveraging RWD with ML appear already in the data acquisition phase. We must follow specific administrative procedures and obtain approvals from ethics review board and privacy officers, often across multiple institutions. Later, we must set up ad-hoc techniques for data retrieval from silos and strategies for their linking and harmonisation.

RWD quality: as an increasing number of alternative ML models become available for our predictive tasks, it is becoming ever clearer that their performance is ultimately constrained by the data on which they are trained. These data must be unbiased, truly representative, and aligned with the clinical questions of interest; equally crucial are the ways in which we manage and preprocess them, as well as the benchmarking strategies and evaluation metrics we adopt to assess model quality. Ultimately, no matter how many different models we test, poor-quality data will undermine our efforts and prevent meaningful progress. In other words, "The level of evidence that may be derived from RWD is ultimately a function of the quality of data and the rigor of study design and analysis" (Gregg *et al*, 2023a). This represents a fundamental bottleneck in AI's clinical adoption and research shows that 81% of companies, even outside healthcare, still struggle with AI data quality resulting in possible risks for the Return of Investments and business stability (Qlik, 2025).

The quality of RWD varies substantially depending on their source. Administrative data sources are typically considered to be of high quality, whereas data derived from EHRs are often of much lower quality. EHR-based datasets commonly contain large amounts of missing information, outliers, skewed distributions, data-entry errors, and

inconsistencies in data formats or units of measurement. Testing standards and laboratory techniques may change over time or differ across institutions, resulting in heterogeneity in the measurement scales of the same biomarker. Moreover, RWD derived from EHRs are burdened by observability issues, where observability is defined as "the degree to which patient characters are captured within EHRs"(Yan *et al*, 2025). Patients leaving the system (i.e. no longer coming to the system where the EHR is in use, therefore no longer observable), may be incorrectly assumed to have no further healthcare activity or outcome events. Of note, addressing RWD data quality issues traditionally requires tedious manual curation. Thus, developing data processing pipelines that can overcome these quality issues in a scalable way to provide ML-ready datasets is a fundamental prerequisite to any subsequent modelling effort.

A one-size-fits-all approach in this context is unfeasible, as each data preprocessing pipeline requires specific components depending on the data sources, types and complexity, and the aims of the project. Additionally, given the relatively brief experience developed by the scientific community in this setting, precise protocols are not established yet, leading to fundamental reproducibility issues (Hou *et al*, 2023). Researchers are thus looking for comprehensive solutions prioritising data integrity from the outset.

RWD informative content and interpretation: RWD are inherently not organised for research. They are often large in volumes and for the most part unstructured, especially when obtained from EHRs. In fact, EHRs are mainly based on free text fields that lack any tabular or coded format. This means that the useful clinical information is possibly dispersed and very difficult to be distilled. Efforts are ongoing both by academia and enterprise to offer methods to extract structured outputs from such sources, with tools such as Natural Language Processing and Large Language Models (Wiest *et al*, 2025; Chen *et al*, 2026).

Being generated by the everyday clinical practice, RWD might not follow international nomenclatures (e.g. ICD-10, SNOMED-CT, ATC codes..) leading to interpretation issues. They provide different follow up lengths for each subject (the already mentioned "observability issue") and may be representative of different management epochs for

specific diseases, as clinical guidelines are updated over time. Last, they may provide contradictory information on the same subject.

Current organization of research features either teams of experts who work on technologies but have limited knowledge in the specific data domain (healthcare, in this case) or, vice versa, teams of healthcare domain experts that have little or no experience in data management and analysis. Approaching research on RWD instead requires a multidisciplinary and stepwise approach in which clinicians and data scientists together evaluate and filter the informative content, understand it, univocally map and standardise the clinical variables, and assess the information in light of the evolving management scenarios. The final aim is to construct a final study cohort with the right patients and the corresponding correct phenotypes (i.e. descriptive variables) (Hou *et al*, 2023).

Clinical question generation: A further challenge arises from the intrinsic clinical complexity of internal medicine patients, which complicates the translation of real-world clinical needs into well-defined machine learning tasks. Internal medicine encompasses a heterogeneous population characterised by multimorbidity, fluctuating disease trajectories, variable functional reserve, and dynamic therapeutic priorities. For many patients, the definition of "optimal care" is neither static nor unidimensional: treatment goals often shift from curative to symptomatic, from life-prolonging to quality-preserving, or from aggressive interventions to appropriateness-driven management. This fluidity makes it difficult to establish clear, universally valid outcome definitions against which ML models can be trained.

Moreover, decisions in internal medicine frequently rely on nuanced clinical judgement, incorporating patient preferences, frailty assessments, and context-dependent risk–benefit evaluations. These elements are not consistently captured within structured datasets, leading to a mismatch between the richness of clinical reasoning and the reductionism needed to formulate a supervised learning problem. As a result, transforming real-world uncertainty into discrete labels, such as "beneficial" vs "non-beneficial" treatment or "high-risk" vs "low-risk" patients, risks oversimplifying complex scenarios and embedding subjective bias.

Addressing this challenge requires close collaboration between clinicians and data scientists to ensure that ML problem formulation reflects clinically meaningful questions,

preserves patient-centred decision-making, and supports rather than replaces expert judgement in the care of complex internal medicine populations.

1.2.2 General considerations on AI applications to healthcare

AI is widely adopted in many industries and activities, from manufacturing to services. In medicine, we are witnessing a steep rise in the scientific production around AI. Searching "medical artificial intelligence" on PubMed provides 122146 results, with 8870 papers in 2020 against 35668 papers in 2025 (+400%, assessed on 10th March 2026). Financial investments in AI for healthcare are as well on the rise: it was valued at 14.92 billion US dollars and is projected to advance at a stable compound annual growth rate of 38.6% until 2030, with a forecasted value of 110.61 billion US dollars (Artificial Intelligence (AI) in Healthcare Market Growth, Drivers, and Opportunities). Despite this hype, there are relatively few actual applications of AI across the health delivery pipeline, with the FDA AI-Enabled Medical Devices List including 1247 entries, as this dissertation is written (Artificial Intelligence-Enabled Medical Devices | FDA). This is attributed to "complex ethical, technical and human-centred challenges required for safe and effective translation" (Rajpurkar *et al*, 2022). First and foremost, we should define what an application of AI to healthcare is. If we regard AI as being a software, then we should define what does it mean to use a software in healthcare, i.e. when a software becomes a medical device (Software as a Medical Device or Software in a Medical Device). According to FDA and EU regulations, a medical device is anything intended to be used "in the diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease […] and does not do so and which does not achieve its principal intended action by pharmacological, immunological or metabolic means" (Official Journal of the European Union, 2017; Software as a Medical Device (SaMD) | FDA). Software as Medical Devices (SaMD) existed even before the diffusion of AI, and they were always required to go through review by agencies such as FDA to validate their overall safety and efficacy. Now, some of them incorporate AI systems and therefore are tagged as AI-enabled medical devices. These devices require a new framework for their identification, review, tagging, and monitoring. Open questions still arise regarding the increasingly available health apps for smartphones and other electronic devices. Most of them are not defined as SaMD and therefore do not require to go through the rigorous

evaluation by regulatory agencies. However, they are advertised as being somehow related to the health sphere and could potentially interfere with the usual care delivery pathways.

More in general, regulatory and ethical concerns, coupled to a generalized mistrust by healthcare professionals, have made the deployment of AI in real life clinical settings slower than hoped. Regulations have started to appear only recently, with the approval of the European Union's AI act and European Health Data Space (Official Journal of the European Union, 2024, 2025). This should pave the way for clearer rules for research and adoption of AI in healthcare. Ethical concerns regard mainly the perpetration and amplification of existing biases intrinsically present in our clinical practice, for example gender or ethnic bias stemming from underrepresentation of specific groups in datasets, or the worsening of health inequalities rising from different technology and competence gaps already existing across different world regions. In fact, the widespread adoption of AI-powered health technologies depends on the readiness level of the infrastructure of individual institutions, health authorities and health systems, as well as the availability of computing power and storage solutions that are sustainable from the environmental and economical point of view. Finally, for a full, safe and seamless implementation of digital technologies in healthcare, professionals should acquire the required skills, either via lifelong learning (e.g. Continuous Medical Education initiatives) or via a revision of the undergraduate and post-graduate teaching programs of universities. These topics and more were at the heart of discussions we as internal medicine physicians started in national and international scientific societies. Both the Italian Society of Internal Medicine (SIMI) and the European Federation of Internal Medicine (EFIM), now feature Working Groups dedicated to Digital Health, of which the author of this thesis is part of. To foster awareness, learning and discussion, both working groups published a Position Paper on the use of AI in internal medicine, comprising recommendations and guiding principles (see Section 1.5) (Said-Criado *et al*, 2025; Balsano *et al*, 2025).

1.3 Research Questions

We have discussed several challenges regarding research on real-world data that must be addressed to fully unlock the potential of ML applications in internal medicine and enable the development of clinical decision support systems that are grounded on real-

world evidence, beyond that generated by randomized controlled trials. Accordingly, our thesis aims to advance real-world data access, linkage, harmonisation, selection and analysis to answer specific clinical questions regarding the prototypical patients managed in the internal medicine setting. Ultimately, this work addresses the following research questions:

- Research Question 1: How can we access, harmonise and filter real-world data stored in multiple locations in the data lake of a large tertiary hospital while complying to security, privacy and ethical standards?
- Research Question 2: How can we preprocess and analyse the real-world data coming from electronic health records of type 2 diabetes mellitus patients to predict their glycated haemoglobin change over time?
- Research Question 3: How can we set up a real-world data collection effort into a structured electronic case report form for the phenotyping of complex internal medicine patients and leverage the resulting dataset to predict the risk of a composite negative hospitalisation outcome?

1.4 Contributions and Chapter Outline

This thesis will start with an overview of the background and existing works on real-world data analysis in internal medicine settings in Chapter 2 before presenting the three main contributions which address the research questions posed in the previous section:

Contribution 1: The S-RACE platform

In Chapter 3, we describe the development of a novel, secure, and trustworthy cloud-based platform for responsible AI deployment: the S-RACE (San Raffaele Ai CEntre) platform. S-RACE offers an end-to-end clinical data science pipeline: on-premises anonymisation, natural language processing-driven FHIR standardisation, a "Clinician AI Hub" for exploration, and a "Data Science Lab", based on Azure ML Studio, for ML model development and deployment, the latter through web-based dashboards. It adheres to responsible AI principles, integrating explainability and a robust governance framework (ISO 42001:2023, EU AI Act). S-RACE also supports a privacy-preserving federated learning architecture to facilitate international research collaborations. S-RACE significantly advances operationalising AI for real-world evidence. Its comprehensive,

ethically governed infrastructure accelerates clinical research and translation, enhancing patient care. Interestingly, it represents a successful example of an interdisciplinary collaboration between physicians, data scientists, information technology specialists, user interface designers and experts in legal frameworks, and of an interinstitutional collaboration between an healthcare delivery actor, IRCCS San Raffaele Scientific Institute, a university, Vita-Salute San Raffaele University, and the industry, as the support of Microsoft and its partner Porini was strategic.

Contribution 2: impact of real-world data preprocessing on the prediction of glycated haemoglobin change in time

Having presented the S-RACE platform, in Chapter 4 we then report a first example of its usage in the field of internal medicine for disease evolution prediction in the outpatient settings. Here, we leverage real-world data to train a ML model with the task of classifying type 2 diabetes mellitus (T2DM) patients as either having an increase or a decrease of glycated haemoglobin (HbA1c) at three years from their first available measurement in our records. While doing this, we assess the impact of different data preprocessing pipelines on the performance of our models. To this aim, we set up a retrospective observational clinical trial (AI-TRYDIA, NCT 06280729), to extract real-world data from the IRCCS San Raffaele Hospital diabetology EHR through an application programming interface towards the S-RACE platform, producing 10 different datasets stored in separated .csv files with records of 8591 bona-fide T2DM subjects treated between 2003 and 2023. Part of the work for this contribution comes from a period as a visiting researcher at the University of Cambridge, during which we submitted a data access application to the EHR Research and Innovation database (ERIN) of Addenbrooke's Cambridge University Hospital NHS Foundation Trust to obtain a dataset of 17355 T2DM patients. We then manipulate the two datasets to isolate two similar cohorts of interest from the respective populations and select or engineer the required features and the target variable, generating the two final datasets. Next, we train logistic regression supervised classification ML models in a series of experiments where the pipeline alternates different missing data management strategies, feature choices, and scaling approaches. We demonstrate that with proper management of collinear and missing variables, the models consistently achieve an area under the receiver operating

characteristic curve (AUC) of 0.72-0.74 in the internal validation set, and that performance drop in the external test set is minimal, with an AUC in the range of 0.67-0.69. We also show that different scaling approaches influence models' sensitivity. Last, in the model interpretation step we show how higher baseline HbA1c levels are consistent predictors of HbA1c increase at three years across models. Insights such as this, distilled by harvesting diabetology real-world data with ML, enable discussions about the management of T2DM patients to foster improved practices.

Contribution 3: negative hospitalisation outcome prediction through a multidimensional approach

Moving from the outpatient to the inpatient setting, we further test real-world data ML pipelines on the data gathered from internal medicine departments of IRCCS San Raffaele Hospital in the context of the MED-Cli study (NCT05780099). A very large dedicated electronic case report form (eCRF) designed for this prospective observational clinical trial allows for standardized collection of a wide variety of data of admitted patients, from their chronic diseases and therapies, risk factors, social and family history, to their clinical status and presenting complaint at admission, the hospitalisation course, and discharge information. We continuously refine and update the eCRF according to new inputs and needs of the research team. We then upload the so-collected data onto the S-RACE platform and integrate it with other data sources such as the hospital laboratory informatic system and the EHR storing therapy records of the hospitalisations. Next, we train two logistic regression models with the task of predicting a composite negative hospitalisation outcome using either the multidimensional frailty index (FI), built without considering variables describing patient comorbidities, or the well-known Charlson comorbidity index (CCI). We demonstrate that the model trained with the comorbidity-agnostic FI achieves a better performance compared to the model trained with the CCI (AUC = 0.72 vs 0.62). Finally, we select a set of features for further ML experiments where we benchmark 9 base supervised classification models and 3 ensemble models with the same classification task. The logistic regression comes on top and is therefore retrained, achieving a validation set AUC of 0.71 (0.68-0.75) and then tested on unseen data where the AUC increases to 0.78. When looking at the coefficients of the logistic regression, the

Braden score emerges as the most important, confirming the relevance of a multidimensional approach to complex internal medicine patients.

1.5 List of Publications

This section contains the manuscripts that have been published in peer-reviewed journals and the abstract submissions that have resulted in posters or oral presentations at national and international internal medicine conferences. Some of these works are directly related to this dissertation, such as the NPJ Digital Medicine paper describing the platform where I performed some of my experiments, or the position papers on AI in internal medicine demonstrating my interest in discussions related to the wider applicability of my work. Some demonstrate my contribution to projects applying similar techniques to fields not directly related to my thesis, such as COVID19 and lung cancer.

Works related to this dissertation

Powering Responsible Artificial Intelligence with High-Quality Real-World Data: The S-RACE platform for Scalable, Multi-Specialty Clinical Research

A. Traverso, D. Tiano, A. Corvaglia, A. Dimonte, E. Draetta, B. Fabiani, P. Scuri, S. Barbieri, M. Agazzi, M. Arslan, D. Celada, F. Chiabrando, L. Cibrario, G. Cielo, A. Colombo, S. Contini, M. Liberotti, **M. Montagna**, F. Ogliari, A. Palmisano, F. Pisu, D. Serra, D. Varani, D. Vignale, A. Vitali, A. Zambello, C. Chiapponi, M. Denti, A. Esposito, C. Tacchetti.

npj Digital Medicine, 2026

From raw data to actionable insights: preprocessing real-world data for machine learning in diabetes care

M. Montagna, A.S. Rabadzhiev, A. Traverso, E. Setola, E. Draetta, A. Dimonte, S. Barbieri, B. Fabiani, L. Piemonti, A. Esposito, C. Tacchetti, P. Rovere Querini.

Frontiers in Digital Health, 2026

Artificial Intelligence in Medicine: a position paper by the Italian Society of Internal Medicine

C. Balsano, F. Cabitza, S. Cicco, M. Gori, D. Malerba, **M. Montagna**, R. Tarquini, A. Vacca, on behalf of the Working Group on Artificial Intelligence and Digital Therapies of the Italian Society of Internal Medicine (SIMI)

Internal and Emergency Medicine, 2025

Advancing Toward P6 Medicine: Recommendations for Integrating Artificial Intelligence in Internal Medicine

I. Said-Criado, F. Pietrantonio, **M. Montagna**, F. Rosiello, O. Missikoff, C. Drago, T. Leung, A. Vinci, A. Signorini, R. Gómez-Huelgas,

Clinics and Practice, 2025

Machine learning approach for predicting non-beneficial treatments in internal medicine: real-world evidence from the med-cli study

M. Montagna, C. Pomaranzi, C. Soggetti, C. Bellino, G. Pata, E. Rela, G. Mogliarisi, G. Ramirez, S. Damanti, M. Tresoldi, P. Rovere Querini.

126th National Congress of the Italian Society of Internal Medicine, Rimini, Italy, 2025

Continuous real world data collection from internal medicine wards for modern patients' patterns description and forecasting

M. Montagna, C. Pomaranzi, G. Pata, E. Rela, M. Mallus, M. Ruggiero, T. Hill, C. Soggetti, S. Santoro, F. Sciammetta, A. Cavazzana, L. Leoni, M. Izzo, A. Merolla, F. Chiabrando, A. Lucini, M. Biganzoli, R. De Lorenzo, P. Rovere Querini.

23rd European Congress of Internal Medicine (ECIM2025), Florence, Italy, 2025

Leveraging Real World Data with Machine Learning to predict long term HbA1c changes in Type 2 Diabetes Mellitus

M. Montagna, A.S. Rabadzhiev, A. Traverso, E. Draetta, A. Dimonte, S. Barbieri, B. Fabiani, A. Esposito, C. Tacchetti, P. Rovere Querini.

23rd European Congress of Internal Medicine (ECIM2025), Florence, Italy, 2025

Real world data: challenges and opportunities to improve knowledge in internal medicine

M. Montagna, A.S. Rabadzhiev, A. Dimonte, E. Versino, B. Fabiani, S. Barbieri, E. Setola, G. Ancona, S. Santoro, M. Scavini, A. Laurenzi, E. Bosi, L. Piemonti, C. Tacchetti, P. Rovere Querini.

125th National Congress of the Italian Society of Internal Medicine, Rimini, Italy, 2024

A Diagnosis-Independent Frailty Index at Admission Improves Risk Stratification in Hospitalized Internal Medicine Patients

M. Montagna, S. Damanti, Andrea Corvaglia, Simone Barbieri, Chiara Bellino, Laura Leoni, Giulia Pata, Chiara Pomaranzi, Elena Rela, Clara Soggetti, Giulia Lanzetta, Rebecca De Lorenzo, Moreno Tresoldi, Patrizia Rovere Querini.

24th European Congress of Internal Medicine (ECIM2025), Vienna, Austria, 2025

Other works (selected)

Evaluating Listening Performance for COVID-19 Detection by Clinicians and Machine Learning: Comparative Study

J. Han, **M. Montagna**, A. Grammenos, T. Xia, E. Bondareva, C. Siegele-Brown, J. Chauhan, T. Dang, D. Spathis, R.A. Floto, P. Cicuta, C. Mascolo.

Journal of Medical Internet Research, 2023

Exploring machine learning tools in a retrospective case-study of patients with metastatic non-small cell lung cancer treated with first-line immunotherapy: A feasibility single-centre experience

F.R. Ogliari, A. Traverso, S. Barbieri, **M. Montagna**, F. Chiabrando, E. Versino, A. Bosco, A. Lin, R. Ferrara, S. Oresti, G. Damiano, M.G. Viganò, M. Ferrara, S.T. Riva, A. Nuccio, F.M. Venanzi, D. Vignale, G. Cicala, A. Palmisano, S. Cascinu, V. Gregorc, A. Bulotta, A. Esposito, C. Tacchetti, M. Reni.

Lung Cancer, 2025

Impact of Clinical Decision Support Systems on Medical Students' Case-Solving Performance: Comparison Study with a Focus Group

M. Montagna, F. Chiabrando, R. De Lorenzo, P. Rovere Querini.

JMIR Medical Education, 2025

Fostering the intersection between primary care and hospital with the Integrated Care system: the PRIME (PRIMary care-hospital Embedding) project

M. Montagna, F. Chiabrando, G.P. Vitali, A. Spanò, A. Di Resta, F. Fait, P. Rovere Querini

22rd European Congress of Internal Medicine (ECIM2024), Istanbul, Turkey, 2024

A large language model-powered digital assistant for up-to-date covid19 outpatient management

M. Montagna, F. Chiabrando, F. Baldini, D. Ficocelli, G. Faraci, P. Rovere Querini.

125th National Congress of the Italian Society of Internal Medicine, Rimini, Italy, 2024

2. Related Work

In the previous chapter, we introduced the potential of leveraging real-world data with machine learning to overcome the issue of evidence coming from randomized controlled trials being poorly generalizable to the population managed in the internal medicine setting. We highlighted the most important limitations related to data availability, access, quality, interpretation and preprocessing. In this chapter, we proceed by reviewing prior research tackling the same challenges. Specifically, we briefly discuss the relevance of real-world data for the internal medicine field in Section 2.1, followed by examples of other institutional data gathering platforms in Section 2.2. We then report on prior efforts in the two scenarios of interest in this thesis: the prediction of glycated haemoglobin change in type 2 diabetes mellitus (Section 2.3) and the prediction of hospitalisation outcomes in general/geriatric medicine wards (Section 2.4).

2.1 The Relevance of Real-World Data in Internal Medicine

2.1.1 Population ageing

Global population is ageing. This observation, captured by national and supranational statistics, warns us that we could be facing a once unobserved patient phenotype that needs to be studied harnessing RWD. At the beginning of the 19th century, the world's population amounted to roughly one billion people. Since then, it witnessed an uncontrolled and continuous growth of around 1-2% per year (2.3% maximum rate in 1963), especially thanks to reduced mortality, especially among children, and increased life expectancy. In fact, if before 1800 life expectancy was under 30 years, today it sits at an average of 73 years. In 2022, world's population reached 8 billion people, with an annual growth rate below 1% (0.87% in 2023) (Ritchie *et al*, 2023; Population | United Nations). As humans increase in number, so does the number of them potentially becoming sick, that is, becoming patients, especially if their age distribution changes as well. Particularly burdensome is the fact that the percentage of people aged above 65 in the world increased from 5% in 1961 to 10% in 2023: doubled. If we focus on OECD countries, a change from 17.3% of 2017 to 28% in 2050 is forecasted (Berchet *et al*, 2019). As disease prevalence, especially of chronic illnesses, rises with advancing age, the overall number of patients is increasing accordingly, as their complexity also does,

especially because of multimorbidity. Therapeutical successes, as those in the field of cardiology or oncology, produced a subpopulation of patients that, even if cured, accumulate disability. Almost one third of individuals aged above 65 live with at least one chronic disease (Berchet *et al*, 2019).

2.1.2 Evidence gap for the aged population

Within the Evidence Based Medicine framework, randomized controlled trials (RCTs) are considered the most rigorous and classical methodology for evaluating the efficacy of interventions, as they minimize bias through random allocation and strict protocol adherence (Kennedy-Martin *et al*, 2015; Dreyer, 2022; Tenny & Varacallo, 2024). On the other hand, in the dynamic landscape of healthcare, critical and real-time evaluation of our management strategies along the care pathways could significantly influence the quality and effectiveness of patient care. This becomes particularly vital in internal medicine settings, where clinical decision making is particularly challenging. In fact, complex, multimorbid and elderly patients with extensive therapeutic needs and frequent diagnostic and management dilemmas, consequence of the previously described massive and fast ageing phenomenon, are often redirected and cared for in general medicine wards (Naik *et al*, 2024; Colacci *et al*, 2025). However, it is hard to transfer the knowledge generated by traditional RCTs to the everyday clinical practice. This is especially true in the internal medicine context, as RCTs often exclude patients with complex health profiles and are therefore not representative of the real-world populations cared for in internal medicine wards (Subbiah, 2023; Gregg *et al*, 2023b; Barlean *et al*, 2023; Tinetti *et al*, 2019, 2004; Tinetti & Fried, 2004; Inouye *et al*, 2007; Herrera *et al*, 2010). As stated by Michael Vassallo, from the UK National Healthcare System, “It is often difficult to practice evidence-based medicine in older people because there are not the research data to support it” (Vassallo, 2019). Additionally, those multimorbid and frail subjects often exhibit reduced physiological reserves and are highly susceptible to adverse outcomes, even with maximized medical interventions (Tinetti *et al*, 2019).

We can even describe a negative feedback loop: elderly patients have often many vulnerabilities and, as such, they are excluded from interventional research. This leads to a reduced access to new management strategies, that, in turn, causes even more vulnerability. Moreover, therapeutic goals in such patients are often multifaceted, and available resources limited, making the decision-making process even more demanding.

It becomes then evident that, if informed and context-relevant decisions are to be made, there is a need for complementary evidence generated from the everyday clinical activity to supplement that coming from RCTs. This evidence, referred to as real-world evidence (RWE), provides precise insights that not only reflect our current practices but also highlight potential areas for optimization in resources and everyday patient management (Concato & Corrigan-Curay, 2022; Batko & Ślęzak, 2022; Liu & Demosthenes, 2022; Dang, 2023; Framework for FDA's Real-World Evidence Program, 2018). The raw material from which RWE is derived are real-world data (RWD). On PubMed, RWD are identified with the MesH term "routinely collected health data", inserted in 2021 and defined as "Data collected for purposes other than research. Examples include health administrative data, EHR data, and disease or clinical registry data." (Routinely Collected Health Data - MeSH - NCBI). Those data hide information regarding precisely those patients that often don't meet the strict inclusion criteria of RCTs but that make most of the current population attending our hospitals nowadays, particularly in internal medicine settings, and that therefore are mostly eligible only for observational/epidemiological research. Moreover, RWD contain information regarding our management strategies and their outcomes, also with longer observation times compared to traditional approaches. Therefore, exploiting RWD to generate RWE that in turn can inform our management decisions is highly relevant for us internists.

2.1.2 Innovation in research through Real World Evidence

With RWE we expect to have more diverse and representative populations and settings, more insights on realistic implementation of evidence-based management, varied long-term outcomes and unanticipated benefits and adverse events (Figure 1).

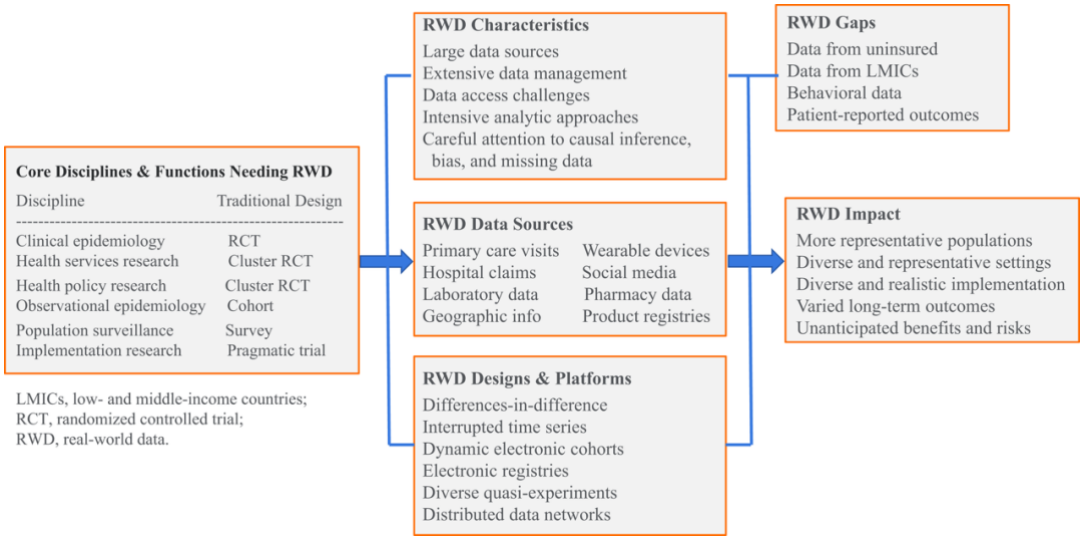


Figure 2.1.1 *– Graphical abstract of "Use of Real-World Data in Population Science to Improve the Prevention and Care of Diabetes-Related Outcomes". From Gregg, Patorno, et al. Diabetes Care 1 July 2023; 46 (7): 1316–1326. © 2023 by the American Diabetes Association, license number 6145321137929.*

RWD are in fact evolving population-based scientific disciplines that traditionally generated the evidence on which advancing current management standards. These disciplines include clinical effectiveness research, to guide the choice of interventions, health service research, to shape care delivery models, and health policy research to inform population-wide approaches for prevention and health promotion. If to date these disciplines have taken advantage of randomized controlled trials, cohort studies and regular surveys, they can now benefit from RWD (Table 1).

Table 2.1.1 *– Spectrum of applications and designs using RWD for population and public health research for diabetes. From Gregg, Patorno, et al.; Use of Real-World Data in Population Science to Improve the Prevention and Care of Diabetes-Related Outcomes. Diabetes Care 1 July 2023; 46 (7): 1316–1326. © 2023 by the American Diabetes Association, license number 6145321137929.*

Discipline/application	**Core purpose**	**Dominant design**	**New data sources being integrated**	**Dominant RWE designs**
Clinical epidemiology	Test treatments	RCT	Primary care visits Hospital claims	Parallel groups
Health services research	Test health services	Cluster RCT	Laboratory data Pharmacy data	Differences-in-difference
Health policy research	Test health policies	Cluster RCT	Health monitoring devices Behavioral dietary data	Interrupted time series
Observational epidemiology	Identify and prioritize risk factors	Cohort	Physical activity data Social media	Dynamic electronic cohorts
Population surveillance	Monitor population health	Population survey	Geographic information Product registries	Electronic registries

Implementation research	Determine reach, adoption, implementation, and sustainability	Pragmatic RCT	Disease registries	Diverse quasi-experiments

Interestingly, in research involving RWD we are forced to consider our research questions considering the data we have. As variables are not collected deliberately, they might sometimes limit the scope of our research objectives or, on the contrary, broaden it. Designing experiments requires complex decisions along the process for which guidelines are lacking, leaving room for innovation. Keeping in mind these methodological challenges, this evolution in the way we gather and harness data for research complements traditional approaches allowing to solve some of their limitation or go beyond the evidence they generate and their narrow scope. As reported in Figure 2, RWE in facts allows to: 1) measure the efficacy of evidence-based management when applied outside the research setting; 2) gain insights about populations that are underrepresented in clinical trials (children, elderly, pregnant, low socio-economic status..); 3) evaluate the effectiveness of interventions and policies where conventional experimental or randomized approaches are infeasible, prohibitively expensive, or ethically unacceptable; 4) extend our knowledge of adverse or beneficial outcomes beyond what traditional methods can achieve, especially in terms of observation time.

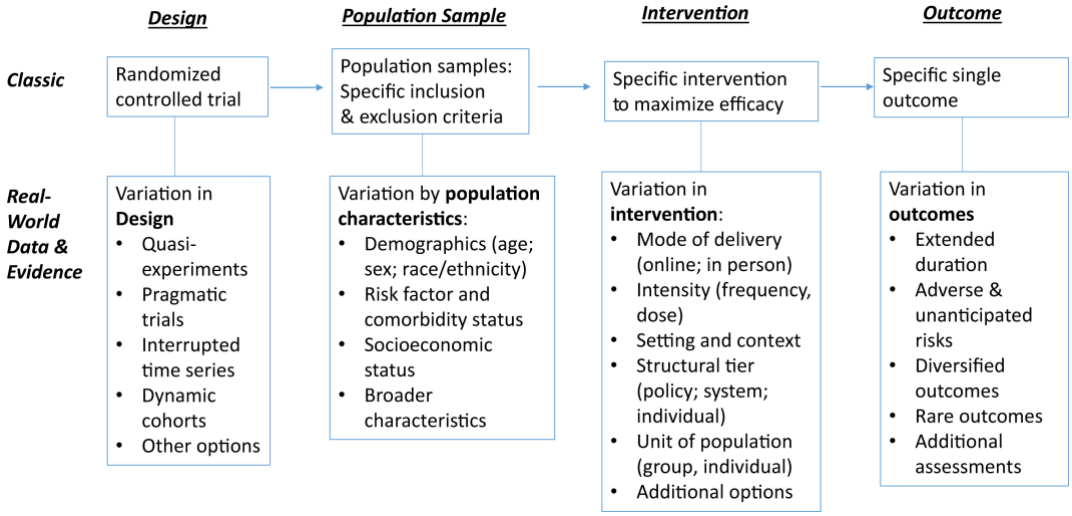


Figure 1.1.2 *– Extension and applications of RWD beyond traditional RCTs. From Gregg, Patorno, et al.; Use of Real-World Data in Population Science to Improve the Prevention and Care of Diabetes-Related Outcomes. Diabetes Care 1 July 2023; 46 (7): 1316–1326. © 2023 by the American Diabetes Association, license number 6145321137929.*

2.1.3 Artificial Intelligence for improved decision making in internal medicine

As healthcare systems increasingly seek to improve patient outcomes and optimize resource allocation, the predictive capabilities of AI, when combined with the vast and

diverse datasets generated in everyday clinical practice, offer unprecedented opportunities for enhancing decision-making processes, made so complex by the inherent complexity of the internal medicine scenario. Hospitalization outcomes are critical indicators of patient health and healthcare efficiency, making their accurate prediction essential for clinicians and healthcare administrators alike. As we delve deeper into the intersection of technology and clinical practice, it becomes imperative to explore the methodologies, challenges, and implications of such approaches. The synergy between RWD and AI could not only enhance the accuracy of hospitalization outcome predictions but also foster personalized care strategies tailored to individual patient needs.

2.2 Platforms for Real-World Data collection

Despite RWD's transformative potential, its effective utilisation for clinical translation faces inherent challenges, as already introduced. To overcome these complex data challenges and unlock the full potential of RWD, the healthcare sector is witnessing the emergence of sophisticated end-to-end clinical data science pipelines mostly hosted on cloud-based AI platforms. These infrastructures are specifically engineered to ingest, process, and analyse RWD, streamlining data harmonisation and curation, and the development of AI-driven clinical decision support systems which can be rapidly but consciously evaluated for clinical deployment (Kaur & Mann, 2018). Such platforms are essential for generating reliable RWE, but they of course depend on the existence of already fully operational and well populated EHRs. The percentage and maturity in adoption of EHRs is extremely different across different countries worldwide, and even inside each of them and among different healthcare actors. The US strategically pushed their office-based physicians and hospitals towards EHR implementation already in 2009 with the 30-billion dollars worth Health Information Technology for Economic and Clinical Health and the subsequent Meaningful Use program. Accordingly, by 2021, 76% of office-based physicians had certified EHRs and 96% of hospitals had adopted certified EHRs, with, however, differences between urban and rural areas, for example (Anzalone *et al*, 2025). The UK government paved the way to towards 100% EHR adoption by NHS actors with clear pledges, as documented in 2022 in (A plan for digital health and social care - GOV.UK). EHR adoption in the EU is undergoing a major transformation driven by the European Health Data Space Regulation, which aims to create a standardized

system for interoperable EHRs across member states by 2030. Italy is making major steps forwards thanks to the National Recovery and Resiliency Plan, whose planned investments in digital health amount to 2.8 billion euros, of which 1.45 specifically allocated for the digital modernisation of 280 first- and second-level public hospitals. According to the 2024 report of the Digital Health Observatory (Milan's Polytechnic University), some kind of EHR was adopted by 85% of Italian hospitals (+13% over previous year). However, only half of them had the EHR adopted across all wards, signifying a still fragmented and partial EHR adoption. Some Italian regions have launched initiatives to foster and standardise EHR adoption (e.g. 33 millions for 19 public hospitals in Lombardy), while the National Government has supported digital healthcare and investments with two major centralized tenders aimed at promoting the adoption of EHRs: the first worth 450 million euros and the second worth 420 million euros (Digital Innovation Observatory of Politecnico di Milano, 2025).

The development of the S-RACE platform is situated within this rapidly evolving ecosystem of RWD infrastructures in healthcare. Several other prominent platforms have emerged:

- The Memorial Sloan Kettering Cancer Center Clinicogenomic, Harmonised Oncologic Real-World Dataset (MSK-CHORD) is a large, integrated dataset developed by Memorial Sloan Kettering Cancer Center, combining NLP annotations with structured medication, patient-reported demographic, tumour registry, and tumour genomic data. Its primary purpose is to enable the discovery of clinical genomic relationships not apparent in smaller datasets and to improve cancer outcome prediction (Jee *et al*, 2024).
- The EHR Analysis in Python framework (ehrapy), from Helmholtz Munich, is an open-source Python modular framework designed for comprehensive exploratory analysis of heterogeneous epidemiology and EHR data. The foundation of ehrapy is a robust and scalable data storage backend where data are organized in matrices and encoded in consistent, reusable formats (CSV, OMOP, SQL databases). The storage is then combined with a series of pre-processing and analysis modules which allow for causal inference, trajectory inference, survival analysis, creation of patient clusters and association with diseases states. Ehrapy aims to standardise

analysis pipelines on EHR data and serve as a cornerstone for the community. (Heumos *et al*, 2024).

- Researchers at University College London and Kings College London, together with the National Health System (NHS) of England and industrial partners such as Amazon Web Services (AWS) and Databricks, recently developed Foresight. Foresight is a transformer-based pipeline that leverages structured (age, sex, and ethnicity) and unstructured (free-text) de-identified NHS data to train a generative model and predict future clinical events based on past events. The published study reports on a use case with EHR data from (1) King's College Hospital NHS Foundation Trust; (2) South London and Maudsley NHS Foundation Trust; and (3) the US Medical Information Mart for Intensive Care III. Such data was retrieved and harmonised. Free-text information was then structured using a Medical Concept Annotation Toolkit for name-entity linking to the Systematized Nomenclature of Medicine and Drug Extension concept database. The resulting dataset was used to train a Generative-Pretrained Transformer model where the token represented biomedical concepts rather than words. The pipeline ends with a web application that allows to interact with the model (Kraljevic *et al*, 2024).
- Stanford's Personal Health Dashboard (PHD) is an end-to-end solution for biomedical big data analysis which utilises state of the art security and scalability technologies, while being fully interoperable. It combines and stores large biomedical datasets ranging from wearable biosensor data and multi-omics profiles to clinical data. All structured wearable, clinical, and processed multi-omics data along with metadata are stored either in Google Cloud Platorm BigQuery, AWS Athena, and Apache Presto using SQL as a standard; unstructured data are saved on AWS S3 and Google Cloud Storage. In its dedicated back-end ML cluster, such data can then be decrypted and fed into ML models. (Bahmani *et al*, 2021).
- While not being precisely a platform for RWD collection, storage and analysis, the Italian Annals of Diabetology initiative deserves to be mentioned, if anything for being a true pioneer in the field. The initiative was launched in 2006 by the Associazione Medici Diabetologi (Diabetes Physicians Association), the largest Italian scientific society for diabetologists. It aims at continuously monitoring, benchmarking and improving current care patterns across a network of Italian

outpatient clinics (Rossi *et al*, 2008). To do so, it leverages a common EHR that was already in use in those clinics and that since then is being adopted by more centres. This EHR, distributed by Meteda ltd, went under different names across different releases: MyStar Connect, Smart Digital Clinic, and now Metaclinic. We will mention this EHR again in Sections 3 and 4. Each year, a pre-defined set of variables is extracted from the EHR in a standard format and conveyed to a team of statisticians and physicians that produce a yearly report (the "Annals"). With time, a large quantity of RWD is being amassed, opening the possibility of large studies beyond the simple benchmarking: from studies focusing on subgroups, like elderly patients, to gender medicine studies or evaluations on drug use appropriateness, up to in-depth analyses on cardiovascular/renal/hepatic outcomes. The latest publication reports that, in 2022 alone, 295 centres provided RWD for about 502747 type 2 diabetes mellitus patients (Russo *et al*, 2023). As we will see in the next section, Italian researchers are attempting to also apply ML to such a large amount of RWD to predict diabetes evolution.

2.3 Real-World Data for the Prediction of Disease Evolution in Type 2 Diabetes Mellitus

There are several examples of works leveraging RWD in the landscape of type 2 diabetes mellitus (T2DM) for the prediction of disease evolution that we identified through convenience awareness, including metabolic control without weight gain (Giorda *et al*, 2020a), accumulation of complications (Nicolucci *et al*, 2022a, 2025) or response to treatments (Bielinski *et al*, 2023; Dennis *et al*, 2025). Table 2.3.1 summarises the main characteristics of these studies. Surprisingly, we found only a couple of reviews aiming at collecting the available literature on the topic (García-Jaramillo *et al*, 2023; Kiran *et al*, 2025).

Table 2.3.1 *- Summary of key machine learning studies investigating predictors of glycaemic outcomes and treatment response in type 2 diabetes using real-world data. The list was identified based on convenience awareness of published studies in this area. The table reports essential methodological characteristics of each study, including cohort size, data source, study design, analytical approach, and performance metrics.*

Authors	*Title*	*Year*	*Journal*	*n° of patients*	*Study type*	*Study aim*	*model used*	*Main performance metrics*	*Performance*
Giorda CB, et al.	Determinants of good metabolic control without weight gain in type 2 diabetes management: a machine learning analysis.	2020	BMJ Open Diabetes Research & Care	802348	Retrospective, multicentric, national database	Identify variables associated to achievement of glycated haemoglobin <7% with no weight gain	Logic Learning Machine	AUC	0.74
Nicolucci A, et al.	Prediction of complications of type 2 Diabetes: A Machine learning approach.	2022	Diabetes Research and Clinical Practice	147664	Retrospective, multicentric, national database	Identify variables associated to having different diabetes-related complications	XGBoost	AUC	> 0.8 (depending on tasks)
Bielinski SJ, et al.	Predictors of Metformin Failure: Repurposing Electronic Health Record Data to Identify High-Risk Patients	2023	Journal of Clinical Endocrinology and Metabolism	22047	Retrospective, multicentric, hospitals EHRs	Identify demographic and clinical predictors of metformin failure across three large U.S. healthcare datasets	XGBoost (Cox proportional hazards extension)	C-index	0.731 (w/ HbA1c, sex, age, ethnicity); 0.745 (full 150-feature set)
Dennis JM, et al. (Mastermind Consortium)	A five-drug class model using routinely available clinical features to optimise prescribing in type 2 diabetes: a prediction model development and validation study.	2025	The Lancet	100107 (development cohort; validation cohorts split geographically and temporally)	Retrospective, multicentric, national database	Use routine clinical features to predict 12-month HbA1c response to five classes of glucose-lowering drugs and identify optimal therapy at the individual level	Flexible linear regression model with interaction terms (drug-class–by-predictor interactions)	Observed HbA1c difference between concordant vs. discordant treatment; clinical outcome associations (Cox regression)	15.2% of initiations were model-predicted optimal; HbA1c benefit when concordant: 5.3 mmol/mol (geographical validation), 5.0 mmol/mol (temporal validation); Concordant therapy associated with lower risk of diabetes complications

As highlighted in the previous section, Italy possesses an exceptionally rich repository of RWD on T2DM, and Italian diabetologists recognise its potential for descriptive, predictive and prescriptive analyses (Musacchio *et al*, 2020). For example, Carlo Bruno Giorda and colleagues (Giorda *et al*, 2020b) applied a Logic Learning Machine to identify and rank factors predicting metabolic control (i.e. HbA1c% < 7) without weight gain in T2DM patients. To this aim, they first isolated a proper subset of the whole population, as data of patients with any kind of diabetes and any age are uploaded in the repository. By excluding patients younger than 30 years of age and patients with either type 1 or gestational diabetes, they obtained an exploratory cohort of 1.3 million patients spanning from 2005 to 2017 (first-quarter only). Next, they generated a curated dataset for the analysis. This required a series of complex decisions on how to exclude measurements out of a reasonable range, how to select two different measurements to assess the outcomes, how to associate predictive features (past diseases, complications, medications..) to each timepoint. In this way, more than 5 million HbA1c measurements and related weight variations were consolidated in the final dataset, corresponding to 802348 patients, with 93 predictive features. This loss of numerosity (-38%) is a common trait of studies using RWD. Out of 93 features, the model identified 19 as predicting metabolic control without weight gain. The most relevant ones were low fasting plasma glucose values, low distance to the 7% target, a fast HbA1c reduction speed and no use of insulin. The authors report an accuracy of 0.75. Of interest, the study shows an attempt to manage the extreme complexity in alternative treatment regimens available for T2DM: 800 different combinations were reduced to 18 by grouping molecules in 8 classes. The study shares typical limitations of RWD, where relevant variables to control for possible biases are not available as you cannot deliberately collect all the variables of interest. In this case, information about hypoglycaemic events and medication adherence are lacking, both of which have obvious implications whenever we want to assess the attainment of goals in the management of T2DM. Furthermore, longitudinal studies from RWD repositories typically exclude patients for which the available follow up is not long enough, thereby leading to a possible selection bias, as in the case of this study.

In another Italian study performed by Antonio Nicolucci and colleagues (Nicolucci *et al*, 2022b), data from 147664 patients from 23 diabetes centres during a 15-year period were used to develop several eXtreme Grandient Boosting (XGBoost) models for two

predictive tasks: 1) appearance of a T2DM complication label in a patient EHR within 5 years; 2) early (0-2 years) vs late (3.5 years) appearance of a specific complication. Also in this case, the authors had to implement extensive data manipulation, selection and curation steps to finalise the ML-ready datasets. First, the pathology codes of interest were selected. Then, patients were filtered and classified as either being initially free of complications and either never developing one or developing one. The time between the first no-complication code and either last no-complication code or first complication code was defined as the time window of interest. The authors chose a 10-fold cross validation strategy, employing 46 features, and oversampling the minority class with the SMOTE algorithm (Synthetic Minority Oversampling Technique). Class imbalance is ubiquitous in medical datasets. For example, in screening datasets negative cases vastly outnumber positives. This imbalance is further amplified in RWD, where the absence of controlled enrolment, unlike in clinical trials, precludes artificial balancing of cohorts. As a result, RWD often reflects the true, skewed prevalence of diseases, with positive-class ratios as low as 0.1–1% for rare or asymptomatic conditions. The excess of negative cases can often make the performance metrics such as the area under the receiver operating characteristic curve (AUC) look very good without being meaningful for clinical performance. Alternative approaches to reporting performance in these settings are warranted, such as the area under the precision-recall curve where there is more transparent reporting against outcome prevalence, and of the trade-off between precision and recall (see section 4.8.3). Oversampling the minority class, as in the case of the work by Nicolucci et al., or adopting class-weighted loss, are other common solutions to this challenge. Authors report good performance metrics in both tasks and across the different diabetes complications, with AUCs consistently above 0.8. Interestingly, the study comprised an external validation on 5 different datasets (n = 3912 to 20007) from diabetes clinics that were not involved in the training. Across these unseen datasets, the performance was not uniform, a phenomenon that could not be explained by the authors, being it not associated with the dataset sizes or their completeness. The same group later applied the models prospectively by embedding them in the EHRs of 38 centres in Italy (Nicolucci *et al*, 2025). At each encounter, the system informed the physician that complication risk scores could be generated for that specific patient. Between April 2023 and December 2025, 153377 patients with T2DM were seen, of whom 138558 had

enough data to compute the scores, which were generated for 14.7% of them. Patients for which the scores were calculated achieved better disease control metrics (lower HbA1c, body mass index and low density lipoprotein cholesterol).

A group at Mayo Clinic used an XGBoost model as well to identify demographic and clinical predictors of metformin failure (Bielinski *et al*, 2023). They leveraged three data sources from 3 geographic regions in the United States: 1) the impressive Rochester Epidemiology Project (REP), that links medical records of 1.7 million people since January 2010 getting data from many healthcare providers of the area, comprising the Mayo Clinic; 2) data from the University of Mississippi Medical Center; and 3) data from the Mountain Park Health Center. Again, the first challenge is selecting the cohort of interest out of this enormous amount of data and define an index date for all enrolled subjects. To identify the diabetic patients, researchers looked for records with at least one positive diabetes screening test but excluded all inpatient point of care glucose tests, glucose measured from tissue or other nonblood fluid, glucose tolerance tests, and all glucose or HbA1c tests measured within 10 months of a pregnancy. The date of the first metformin initiation was set as the index date. Metformin failure was defined as either the failure to achieve or to maintain a target HbA1c (<7%) within 18 months of index date or the addition of other pharmacotherapeutic agents. Across the 3 sites, they identified 22047 metformin initiators with 9407 metformin failures (43%). They trained XGBoost models based on the Cox proportional hazards extension for time to metformin failure using 2 feature sets. Interestingly, a first set of predictive features was very small, including only baseline HbA1c, sex, age and race/ethnicity. With this set, the model achieved a C-index of 0.731 (95% CI 0.722, 0.740) on the leave out test set. Based on the SHapley Addittive exPlanation (SHAP) method, baseline HbA1c was the most relevant feature used by the model for the predictions, with higher levels associated with metformin failure. By extending the features to all 150 available ones, the performance improvement was modest, albeit significant (C-index of 0.745; 95% CI 0.737, 0.754, P < .0001), and baseline HbA1c remained the feature with the largest impact on model's decisions.

At metformin failure, physicians add a second-line drug, likely choosing among: 1) analogues of glucagon-like peptide-1 (GLP-1a); 2) inhibitors of dipeptidyl-peptidase (DPP-4i); 3) inhibitors of sodium-glucose linked transporter 2 (SGLT-2i); 4)

sulfonylureas and 5) thiazolidinediones. Of note, the comparative glucose-lowering efficacy across these different antidiabetic drugs is quite similar and current guidelines consider appropriate most of available treatments for most patients, aside from those with cardiorenal risk (The GRADE Study Research Group, 2022; American Diabetes Association Professional Practice Committee, 2025; Davies *et al*, 2022). Ideally, we would use clinical characteristics to individualise treatment choices. Based on these premises, several UK-based universities founded the MASTERMIND Consortium to establish whether routinely available clinical features could be used to predict the relative glycaemic effectiveness of the five previously mentioned drug classes (Dennis *et al*, 2025). The consortium accessed observational data from the Clinical Practice Research Datalink (CPRD) Aurum database (October 2021 release) to identify the training dataset. CPRD is an exceptional not-for-profit, UK government research service providing primary care data for public health research since more than 30 years, leveraging the fact that 56% of English practices use the same patient management software, EMIS Web® (Wolf *et al*, 2019). Of note, its Aurum database is already linked to other datasets such as Hospital Episode Statistics, Death Registration, Cancer data, Mental Health Services Dataset, and Small Area-Level Data (deprivation measures and rural–urban classification), allowing investigators to do studies that span multiple settings and can observe very diverse outcomes in space and time. Dennis and colleagues selected individuals aged 18-79 years with T2DM initiating for the first time one of the molecules of interest between Jan 1, 2004, and Oct 14, 2020. They excluded patients initiating them as a first-line, those with concurrent insulin treatment, those with a pre-existing end-stage kidney disease diagnosis or those with missing or extreme baseline HbA1c values (identified with quite a large cutoff time window of –183 days to +7 days from initiation date; extremes: < 53mmol/mol, i.e. patient already at target, or >110mmol/mol). They split the obtained dataset for a holdback geographical and temporal validation, and, on the remaining 100107 drug initiations, they fitted a flexible linear regression model to predict the HbA1c value after 12 months from drug initiation. Features were divided in two groups: predictive factors (current age, duration of diabetes, sex, baseline HbA1c, body mass index [BMI], estimated glomerular filtration rate [eGFR], high-density lipoprotein [HDL] cholesterol, total cholesterol, and alanine aminotransferase), interacting with a five-level categorical variable specifying the initiated drug class;

prognostic factors (number of current and previously prescribed glucose-lowering drug classes, ethnicity, smoking status, and Index of Multiple Deprivation 2015 quintile), not interacting with the initiated drug class in the regression equation, but representing independent predictors of treatment response. The trained model predicted the HbA1c value 12 months after initiation of any of the 5 drug classes for any individual patient, allowing to identify for each of them the optimal therapy, i.e. the one that achieved the lowest absolute predicted 12-month HbA1c. In the real life, obviously, each patient was administered either one of the five at a time. The counter-factual outcome, that is the 12-month HbA1c level after initiation of the other drugs, could not be observed at an individual level. Therefore, in the validation phase the consortium assessed differences in observed glycaemic effectiveness between matched (1:1) concordant and discordant groups receiving therapy that was either concordant or discordant with model-predicted optimal therapy. They were also able to evaluate the associations with long-term outcomes such as all-cause mortality, major adverse cardiovascular events or heart failure, microvascular complications using Cox proportional hazards regression. Interestingly, only 15.2% of all the drug initiations (combining development and validation cohorts) were of model-predicted optimal therapy. When there was concordance, researchers observed a HbA1c benefit of 5.3 mmol/mol in the geographical validation cohort and 5.0 mmol/mol in the temporal validation cohort with respect to matched model-discordant groups. Individuals on model-predicted optimal therapy also had a lower risk of diabetes complications.

In summary, we have seen that there is very high-quality research that leverages large amounts of RWD from various sources to train and validate models for the prediction of disease trajectories in T2DM. Of note, one of these models has already shown to benefit clinical decisions in Italian diabetology outpatient clinics (Nicolucci *et al*, 2025). We have seen how the reliance of researchers on initiatives of RWD collection and linkage is strategic, as it is the complexity and relevance of the very first steps of the whole experimental pipeline: the selection of the study cohort and the definition and calculation of the target variable. In the next section, we will discuss how different it is when moving to another setting: that of the complex multimorbid and frail patients that nowadays crowd our internal medicine wards.

2.4 Real-World Data for the Prediction of Hospitalisation Outcomes in Internal Medicine

Research on predicting clinical evolution for hospitalised patients from real-world structured data has historically prioritised high-risk events such as sepsis (Gultepe *et al*, 2014; Gupta *et al*, 2020). This focus reflects two key factors: 1) these conditions are frequently managed in Intensive Care Units (ICUs), which generate high-frequency, high-granularity data (e.g., continuous monitoring, lab results, interventions); and 2) ICUs were among the earliest adopters of EHRs, enabling large-scale data collection and analysis (Nemati *et al*, 2018; Jalilian & Khairat, 2022; Johnson *et al*, 2023). However, this emphasis on ICU-centric events has led to a relative gap in predictive tools for non-ICU hospitalizations, such as those of complex multimorbid and frail patients that are usually directed towards our internal medicine wards. RWD collected in EHRs of non-ICU wards is largely unstructured, sparse, heterogeneous, poorly standardised. Therefore, efforts in this field have been initially directed towards data mining, levering deep learning methods to extract (e.g. from textual documents), represent (e.g. for representation learning) and classify information, trying to map it to medical concepts (e.g. towards the International Classification of Diseases, or Current Procedural Terminology) and predict diseases or outcomes (Shickel *et al*, 2017).

Hospitalisations are complex events with a day-by-day evolution resulting in a trajectory that is challenging even in its description. In internal medicine, such events involve patients that are themselves complex, with multiple concomitant diseases and medications, diverse socioeconomic weaknesses, and different performance statuses (Nobili *et al*, 2011; Mannucci *et al*, 2018; Ceriani *et al*, 2024). Care teams must take management decisions to optimise outcomes of patients balancing, among the others, diagnostic uncertainty and invasiveness, expected short- and long-term benefits of therapies and their side effects, values and expectations of patients and caregivers, and available resources. Such a challenging decision making would greatly benefit from robust, and trustable clinical decision support systems that must be, given the setting and what we have discussed so far, grounded on RWE. To this aim, we must first establish whether we can manipulate RWD of such a complex and heterogeneous setting such as internal medicine to curate a research ready dataset for longitudinal ML analyses. Then, we must find the best approach to model disease evolution, focusing on two different

timeframes: the shorter one of hospitalisation events, where we want to understand activity within hospitals for people with multimorbidity (e.g. complications, adverse events, resource usage, outcomes etc..), and the longer one of chronic diseases progression, where we want to predict indicators such as accumulation of organ damage, disability, and new exacerbation events.

These two were the aims of a work published by a multidisciplinary group of clinicians, data scientists, clinical informaticians and research governance experts from Cambridge (Herrero-Zazo *et al*, 2023). They extracted de-identified data of Addenbrooke's Hospital for more than 10 thousand Admission Events (AE) of older adults admitted as an emergency and that were not discharged by the Emergency Medicine. Further inclusion criteria were aimed at regularizing the dataset whilst limiting missingness: only the first AE included per patient, a length of observation of at least 3 days (to exclude insignificant events), and less than one-third information-poor days out of the whole-length of the observation (to limit missingness). The raw data consisted in the multivariate time series (MVTS) of 23 commonly measured blood tests and vital signs recorded during each of the 11158 AE, on which a Hidden Markov Model (HMM) was applied to generate 17 different states. This allowed to brilliantly move from the complexity of a MVTS to the more interpretable sequence of daily disease states. Stratification of the distribution and overall proportion of the states by outcome of the AE (discharged alive vs inpatient mortality [IM]) showed significant differences; the sequence of disease states was as well different in patients experiencing IM or not. HMM is an unsupervised ML method that, without any knowledge of the hospitalisation outcome, uncovered hidden states in the RWD that are visually associated with risk of IM throughout the AE. Clinicians could review those states and label them as representing either disease-like states (e.g. unstable renal function), admission-like states (e.g. acute presentation, treatment response, pre-discharge..), and physiological-like states (i.e. the inflammatory, autoimmune, thrombotic.. responses to disease states), thereby making the representations interpretable. The researchers then trained a series of logistic regression (LR) and random forest (RF) models using either the MVTS or the HMM states as features, the predictive tasks for the models being 1) IM, 2) 30-day clinical outcome, 3) diagnosis at admission, and 4) diagnosis at discharge. For task 3 and 4 phenotypic information that included age, sex, Clinical Frailty Scale score, and admission diagnosis

were omitted. Going from a multivariate to univariate time-series could lead to a reduction of information. In fact, as expected, a higher performance was achieved by the models trained on MVTS data compared to HMM states. However, models based on state representations still performed similarly well, supporting the validity of the HMM-based approach. The way the group approached the data curation step is particularly interesting. For each patient, they defined 24-h bins and selected a unique observation for each variable in each bin by choosing the earliest record of blood tests (although blood tests are not often measured more than once per day in internal medicine wards) and the closest record to that test for vitals. Subsequently, they described bins as "rich-information" or "poor-information" days. A "rich-information" day was defined as a day with information for at least 14 blood tests and four vital signs. The first and last days of included AEs had to be "rich-information" days. Following this distinction, they approached data imputation using linear interpolation for "poor-information" days and multiple imputation for "rich-information" days. To sum-up, as authors note in the discussion: "Such modeling could be further developed for purposes such as the 'forecasting' of bed capacity, helping hospital managers prioritize resources, or development of tools to support identification of patients clinically fit for discharge. These tools could facilitate early discharge planning and use of early supported discharge pathways. Older adults are frequent users of inpatient services and even relatively modest gains in healthcare delivery and effectiveness could have large system wide effects." (Herrero-Zazo *et al*, 2023).

Recent advances in deep learning, particularly the introduction of attention mechanisms and transformer architectures, have reshaped the landscape of disease-trajectory modelling. These developments stem from an analogy between natural language and multimorbidity patterns: just as large language models (LLMs) learn the statistical structure of text by encoding dependencies between tokens, transformer-based models can treat diagnoses, physiological states, and lifestyle factors as sequential "tokens" within an individual's lifetime health record. This conceptual shift has enabled the emergence of large-scale models capable of capturing temporal dependencies across complex multimorbid histories. A prominent example is Delphi-2M, a transformer model trained on population-level datasets such as the UK Biobank and externally validated on Danish national registries (Shmatko *et al*, 2025). By learning from past diagnoses,

demographic characteristics, and behavioural risk factors, Delphi-2M accurately predicts incidence rates for more than 1,000 diseases simultaneously and can generate synthetic health trajectories spanning decades. Importantly, its attention layers provide interpretable insights into how early-life or past conditions influence subsequent morbidity, while also revealing biases inherent in real-world datasets. It is in fact worth noting that while the UK biobank is a very large, data-rich population-level cohort, it is not fully representative of the general UK population. It suffers from healthy volunteer selection bias, contains more white British citizens, and more educated and richer participants compared to the general UK population (Fry *et al*, 2017). The external validation in the Danish registry of RWD carried on by the group was therefore particularly important and relevant to demonstrate generalisability of Delphi-2M performance and applicability across national healthcare systems.

Models such as the Delphi-2M demonstrate that transformers are well suited for the challenges of internal medicine, where heterogeneous disease patterns, recurrent hospitalisations, and complex interactions among comorbidities complicate traditional prediction approaches. The success of such population-scale architectures highlights the potential of adapting transformer-based methods to hospital-based real-world data to support forecasting, risk stratification, and personalised decision-making in internal medicine.

2.5 Summary

This chapter has reviewed the state-of-the-art in the use of ML approaches on RWD produced in the internal medicine setting, with a focus on T2DM and hospitalisation outcomes.

Significant progress has been made in ML techniques in general, for example with the introduction of transformers and attention heads, leading to the birth of large language models. Also, digitalisation of the healthcare sector is proceeding at great pace, leading to the production of massive amounts of data. Despite this, we revealed how there are still relatively few applications of ML to internal medicine settings.

In Section 2.1 we highlighted the relevance of coupling RWE to that generated by RCTs to fill the representativeness gap faced by the growing elderly multimorbid patient population, and gain insights on the reality of everyday clinical practice, especially in

internal medicine where most of such patients are cared for. This relies on a multicentric, systematic, standardised and thorough collection of RWD describing any event and actor along the healthcare delivery process. In Section 2.2 we therefore explored platforms for RWD collection, integration, storage and analysis. In Chapter 3 we present the San Raffaele Ai centre platform (S-RACE) that we developed to the same aims. S-RACE empowers our clinicians and researchers with an end-to-end data science pipeline for studies that train and validate ML models on monocentric or multicentric RWD, also featuring a model deployment module.

Next, Section 2.3 showed some efforts done by researchers to leverage RWD of patients with T2DM to generate ML models for predictive tasks such as treatment response or development of disease-related complications. Notably, in only one instance a prospective real-world deployment is reported. This relied on the long-standing strategic RWD collection initiative available for Italian diabetology outpatient clinics. Section 2.3 also highlighted the complexity of data manipulation steps for cohort identification, feature selection and model training. In Chapter 4 we propose our own strategy for leveraging RWD of T2DM generated in our diabetology outpatient clinic. We study how different feature selection and preprocessing pipelines influence performance metrics of ML models for the prediction of T2DM control over time.

Last, in Section 2.4 we discussed the complexity of predicting hospitalisation outcomes, given the unstructured and scattered nature of data sources and their scarcity. We also emphasized the need to extend studies beyond the ICU setting to the general medicine wards. There, diverse challenges arise given the inherent complexity of patients, of management decisions and of treatment goal definitions. We try to disentangle this complexity in Chapter 5, where we first discuss new approaches to holistically approach complex and frail internal medicine patients. Second, we describe our systematic and standardised hospitalisation RWD collection effort in the context of a 10-year prospective observational study. Third, we demonstrate the superiority of the multidimensional approach over a disease-centric approach for the prediction of hospitalisation outcomes. Last, we showcase a ML benchmarking experiment on a set of 15 features, followed by training, validation and testing of a logistic regression model for the same task.

3. The S-RACE Platform

3.1 Introduction

IRCCS San Raffaele Scientific Institute (OSR) is a private tertiary referral hospital in the outskirts of Milan, Italy. Founded in 1971 by an Italian priest, Luigi Maria Verzé, the next year it was appointed the "IRCCS" status, meaning a hospital with publicly funded clinical research activities, specifically in the field of diabetes. Since 1999, OSR delivers public health care services for the health authority of Lombardy Region, making it a part of the Italian "Servizio Sanitario Nazionale" (Ospedale San Raffaele - Wikipedia). Most recent yearly figures report 1.5 million visits, 50 thousand admissions and 1300 active clinical trials (STORIA, MISSION, VALORI e POLITICA per la QUALITA' OSR; OSPEDALE SAN RAFFAELE - Assolombarda Servizi).

Although not yet fully digitized, with physicians' and nurses' hospitalisation notes still recorded on paper, OSR features a very large and intricate software infrastructure. Aside from the Picture Archiving and Communication System (PACS) and Laboratory Information System (LIS, relying on DNLab, by Dedalus), the following units have different and autonomous EHRs provided by different vendors: the local emergency department (Smart Health, by Finmatica Group), the pathology unit (Pathox, by Mediko), the diabetology and organ transplant units (Metaclinic, by Meteda), and the radiotherapy unit (by Varian), just to mention a few. Therapies for admitted patients are managed via a dedicated EHR (DC4H, by Dedalus), that is however different from that through which physicians view blood tests results, reports of diagnostics and procedures, and create the discharge letters for admissions, outpatient clinic visits and day-hospital stays (Galileo, by Dedalus). As the adoption date of each of these products is different, they contain information of patients across different time windows. At the first interaction with OSR at any entry point in the system (be it a laboratory test, the review of some pathology slides, a diabetes outpatient visit, an ER admission..), a unique numerical identifier is created in the central registry for a patient, the "Master Patient Index" (MPI), that will from then on be the key linking the patient identity in the whole information technology (IT) ecosystem. Still, each of the EHRs will generate an internal unique identifier, therefore requiring extensive mapping between all the different internal identifiers and the common MPI for each patient.

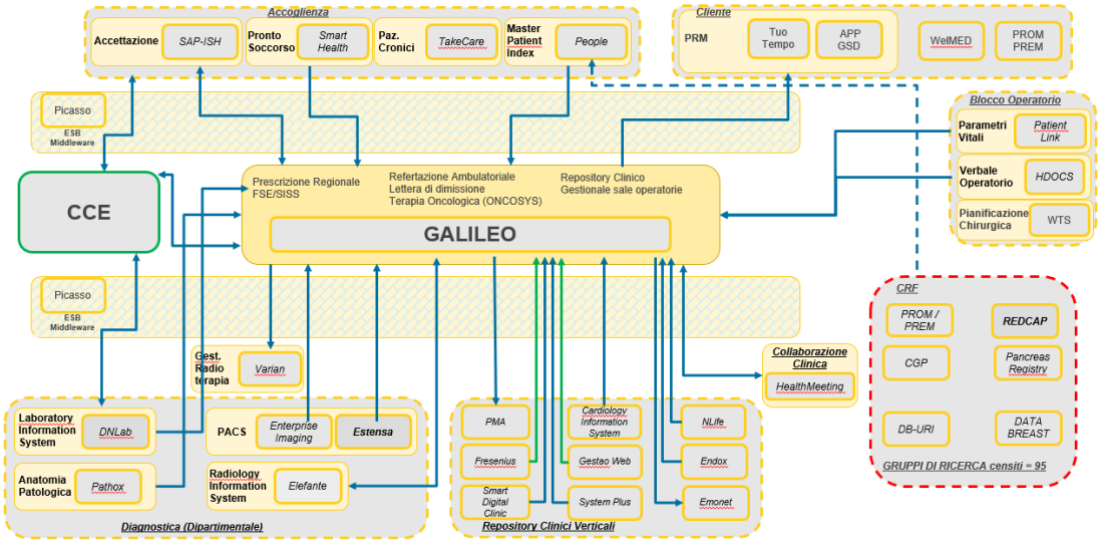


Figure 3.1.1 *– Software application map of IRCCS San Raffaele Scientific Institute as of June 2024. The image was provided by the hospital's Information Technology service in Italian and with no possibility for editing. Upper left module pertains to reception of patients. CCE stands for electronic health record (EHR). Lower left panel gathers software of diagnostic departments. Lower centre panel lists the EHRs of specific specialties (such as Smart Digital Clinic, for the diabetology unit). The lower right panel lists the research databases currently operating in the hospital ecosystem, such as those built with RedCap, or those hosted in the Cohort Genomic Platform (CGP). The right panel (Blocco Operatorio) contains the EHRs used in the operating theatres. The upper right panel (Cliente) shows the consumer-facing software, such as those for telemedicine or for report consultation. In the centre, the master EHR (Galileo, by Dedalus) used by all practitioners of the hospital is shown. All the other EHRs send their data for consultation to Galileo, which is also where discharge letters, outpatient clinics reports and consultant notes are written.*

This is the complexity that we faced when it was decided to build an institutional platform for data aggregation and analysis following a successful collaboration born during the COVID-19 pandemic. At that time, researchers of the Experimental Imaging Center of OSR met with people from the Healthcare branch of Microsoft Corporation Italy (MS). That meeting generated a multidisciplinary team willing to tackle the resource allocation problems caused by the pandemic. The team was composed by us clinicians and our statisticians at OSR, by data scientists from MS and Orobix Life ltd, and by IT infrastructure specialists from Porini ltd, a MS partner company, that had also expertise in business intelligence reporting dashboards. We aimed at generating a semi-automatic and easy to use prognostic model for 30-day mortality of COVID-19 patients admitted to the emergency department (ED) (Palmisano *et al*, 2022). To this aim, we created a pipeline that automatically extracted radiomic features from chest CT scans in approximately 30 minutes from the test leveraging pretrained neural networks, and fed them together with clinical features collected at ED admission to a voting ensemble ML

algorithm that generated a final probability presented to the attending physician on a web-based interface. In that preliminary work, we started realising the complexity of building a RWD ML pipeline connecting multiple data silos, the relevance of the selection of the right patients, features and models, and the importance of having interactive dashboards and a seamless user experience to increase usability for the end-users, that is physicians and researchers.

The COVID-19 pandemic had fortunately resolved, but the team decided to renew the collaboration to set up the San Raffaele Ai CEntre (S-RACE) platform, managed by Vita-Salute San Raffaele University, given its long-standing, bidirectional, and strategic research and teaching agreements with OSR. MS and Porini Ltd acted as technical partners, together with Sketchin, a company responsible for designing the user experience. In summary, as most of the published examples discussed in Section 2.2, our RWD pipeline as well stems from an industry-academy-hospital partnership.

3.2 Platform description

S-RACE is a multi-component platform. In this section we will go through each of the components and describe their structure and function for the whole platform (Figure 3.2.1).

3.2.1 Universal Data Platform

This component serves as the foundational entry point for RWD within the S-RACE ecosystem. The process of data ingestion commences with the selection of: 1) a specific list of patients treated at OSR, identified via the MPI within the hospital's existing IT system; 2) the data sources from which the data should be imported (i.e. clinical reports, laboratory data, pathology data), along with relevant DICOM images, when required, based on desired imaging modality and acquisition date. Then, a crucial preprocessing step is triggered, where data is pseudo-anonymised internally in the hospital by an anonymisation engine that turns the internal unique patient identifier of the specific data source into a unique Cloud Patient Index (CPI). This design choice directly addresses critical security and privacy concerns inherent in managing highly sensitive healthcare data. Once securely on the platform, AI technologies, specifically Microsoft Cognitive Health Services, parse the anonymised data. These services extract clinically relevant information using text analytics and standard medical ontologies. The transformed RWD,

now standardised through the FHIR standard, is stored in a dedicated data lake. To guarantee data segregation, investigators can only access the data assigned to their use cases.

3.2.2 Clinician AI Hub

This component is specifically designed to facilitate preliminary data analyses and exploration, making AI-driven insights accessible and interpretable for clinicians. It leverages MS PowerBI for intuitive data visualisation and interactive preliminary analyses, enabling clinicians to engage directly with the processed RWE. The component provides a structured and quantitative approach to assess a dataset's suitability for analysis before proceeding with more complex procedures, such as ML modelling. This is achieved using five principal parameters: “Data Quality, Feature Distribution, Variability & Redundancy, Information Gain, and Modelability”.

3.2.3 Data Science Lab

This component is specifically tailored for data scientists, enabling them to perform end-to-end ML modelling. The lab utilises MS Azure ML Studio, a comprehensive and integrated data science environment, to support the entire machine learning lifecycle, from data preparation and model training to evaluation and deployment. Beyond its foundational architectural principles, S-RACE's commitment to responsible AI is further solidified through practical implementation measures. These include rigorous traceability and reproducibility, achieved by using MLflow for detailed experiment logging and standard model saving, alongside virtual environments to manage software dependencies. Robust evaluation is ensured through different approaches, from conventional statistical significance testing, strict data segregation (across training, validation and test set) and cross-validation opportunities. A critical feature for clinical translation is the interpretation of generated models using standard explainability techniques, such as SHapley Additive exPlanations (SHAP). Comprehensive documentation of all project phases also underpins this commitment to responsible AI development. The S-RACE platform also leverages other components of the Microsoft Responsible AI Toolbox to ensure comprehensive responsible AI practices (GitHub - microsoft/responsible-ai-toolbox). These include “Data Explorer” for visualising dataset statistics and feature distributions to identify underrepresentation or significant differences across cohorts, and

“Model Statistics” to assess performance metrics like accuracy, precision, recall, and false positive/negative rates for various data subsets. Counterfactual 'What-if' analysis allows for exploring minimal changes to input features required to achieve a desired model outcome, aiding in understanding model behaviour and identifying potential biases. Additionally, Causal Inferencing capabilities enable the estimation of causal effects from observational data, providing insights into how changes in specific features might influence outcomes. These tools collectively support the identification and mitigation of model bias, ensuring fairness and reliability in AI-driven clinical applications.

3.2.4 Model registry

For the monitoring and release of ML models in production, S-RACE leverages the MLflow Model Registry, a centralised component for managing the full lifecycle of ML models. This functionality provides robust versioning, aliasing, and tagging capabilities. Each registered model has a unique name and can have multiple versions, allowing for detailed lineage tracking. Each registered model is further enriched with comprehensive metadata, crucial for ensuring transparency, traceability, and responsible governance throughout its lifecycle, which are based on the AIMe registry for AI in biomedical research (Matschinske *et al*, 2021). Metadata include information such as the model's unique identifier, version number, creation date, and the author responsible for its development. It also captures critical information about the training dataset used, key performance metrics achieved during validation, the hyperparameters configured, and all software dependencies. Furthermore, metadata specifies the model's intended clinical use, its known limitations, and any ethical considerations, alongside its current validation and deployment status. This rich descriptive layer is vital for maintaining a clear audit trail, facilitating reproducibility, and enabling effective oversight of AI models in a high-stakes clinical environment, aligning with the principles of trustworthy AI.

3.2.5 Model web-based deployment

The S-RACE platform allows to run a dedicated web-based application for each validated AI model. This further accelerates the adoption of developed models by the clinical and research community. These applications are specifically designed to facilitate the prospective validation of the AI models within real-world clinical practice and demonstrate the models' utility and impact in a live healthcare setting. To this aim, our

research groups will meticulously design and execute prospective clinical trials to measure the tangible impact of incorporating these AI models into clinical decision-making processes. The web-based model deployment involves leveraging a managed Azure Machine Learning Inference endpoint on the S-RACE Azure tenant, allowing predictions to be obtained via JSON input. The user-friendly interface is built with the Python library Streamlit for the frontend, while a Python backend handles data processing and communication (Raghavendra, 2023).

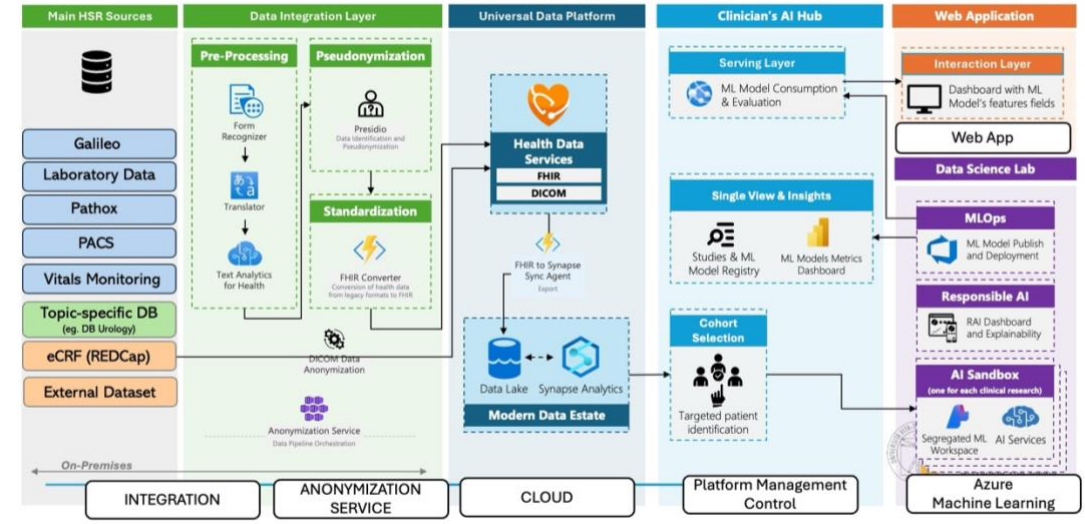


Figure 3.2.1 *– The multi-component structure of the San Raffaele Ai CEnter (S-RACE) data gathering, integration and analysis platform for RWE generation and ML model development and deployment.*

3.2.6 User Creation and Access

Access to the S-RACE platform is granted by the platform administrator, who adds users to the tenant. Following account creation, users can either initiate a new study or accept an invitation to an existing one. Each study incorporates the following defined roles: Principal Investigator (PI), Co-Principal Investigator (Co-PI), study coordinator, collaborator, and data scientist. For users assigned the data scientist role, a dedicated compute instance is automatically deployed within the workspace in the already described MS Azure ML Studio. Access to this Azure environment necessitates users to download and configure the Azure VPN application. Connectivity is then established by connecting to the Azure VPN using the account associated with the study invitation.

3.3 Author's contribution in the development of the platform

The author of this dissertation was involved in the development of the platform from the very first steps. He brought to the S-RACE interdisciplinary team his dual expertise as a clinician and as a researcher.

As a clinician, he aided in 1) identifying relevant data sources to prioritise when the team started to approach the data acquisition and linking phase; 2) explained their content, meaning and use to the non-clinical counterpart; 3) guided the first data extraction efforts, filtering and selecting relevant fields, checking data integrity, consistency and coherence with clinical practice (e.g. name of variables, units of measurements, informativeness..); and 4) supervised information extraction from free text coming from anonymised reports extracted from the hospital's EHR and helped with the adoption of medical ontologies. As a researcher and end-user of the platform, he aided in 1) delivering a user interface that ensured a friendly and intuitive user experience; 2) translating the current research procedures for clinical study application, management, amendment into a digital workflow; 3) designing a robust and usable cohort selection framework to select patients to be included in each use case; and 4) wrote and filed one of the clinical study applications to unlock data access for diabetes patients.

4. Glycated Haemoglobin Change Prediction

4.1 Introduction

T2DM is a chronic disease characterised by high levels of glucose in the blood as a consequence of our body's cells resisting to the normal effect of insulin. It is estimated that approximately 589 million adults are living with diabetes, making up almost 1 in 9 adults worldwide. In 2024 there were 3.4 million deaths attributed to T2DM and it is widely recognised that half of patients do not meet the treatment target (International Diabetes Federation, 2025). As discussed in Chapter 2, HbA1c is the major biomarker of T2DM control, being considered a three-month average of blood glucose levels. As such, it informs the management of patients and is directly associated to the development of T2DM complications, which are ultimately responsible for morbidity and mortality caused by the disease. Specifically, the higher the HbA1c level, the higher the risk of developing macrovascular (myocardial infarction, stroke) and microvascular complications (nephropathy, retinopathy and neuropathy) (Skyler, 1996). Being able to guess the risk of HbA1c increase in time would allow improved patient stratification and optimise treatment decisions, justifying more aggressive treatments in patients at high risk of having a worsening of T2DM.

We therefore leveraged two different RWD datasets of T2DM patients to perform training and external validation of classification machine learning models for the prediction of 3-year HbA1c change.

4.2 Data Structure Description

The first step of most research projects involving RWD from hospitals' EHRs is the identification and extraction of data with a specific query on the original data lake.

4.2.1 CUH dataset

Thanks to a visiting researcher position at the University of Cambridge, United Kingdom, we were granted the access to RWD of Cambridge University Hospital NHS Foundation Trust (CUH) through a Letter of Access (A097255). The de-identified data, stored in the electronic hospital record research and innovation (ERIN) database, belonged to T2DM patients who had any kind of encounter with the clinic (either because of admissions, outpatient visits or simply because their blood tests were analysed there).

ERIN is hinged on Clarity, a Structured Query Language (SQL) relational database abstracted and updated daily, that mirrors CUH's EHR (Epic) (ERIN - Information for researchers - NIHR Cambridge Biomedical Research Centre). Together with a data analysts of CUH, we explored it and identified T2DM patients. Out of all T2DM patients, we selected those that had at least two HbA1c measurements 3 years +/- 180 days apart, with the first being ≥ 48 mmol/mol. The date of the first available HbA1c measurement marked the T0, i.e. the baseline situation of each patient. We then queried for other measurements taken in a 90-day window around T0. On 18th March 2025 we finally extracted the resulting set of variables of 17355 patients in a single .csv file with 17355 rows (Table 4.2.1):

- PAT_MRN_ID is the unique and anonymous patient identifier.
- Age_death_days is available for those patients that are identified as dead in Epic thanks to its connection with the health authority database.
- Trig_T0 = triglycerides at T0
- ALT_T0 = alanine aminotransferase at T0
- TSH_T0 = thyroid stimulating hormone at T0
- SBP_T0 = systolic blood pressure at T0
- DBP_T0 = diastolic blood pressure at T0

Table 4.2.1 *– Variables extracted from the Clarity database of Cambridge University Hospital Foundation Trust for 17355 patients identified as having type 2 diabetes mellitus. The non-null count for each variable is reported.*

	Column	***Non-Null Count***
1	PAT_MRN_ID	17355
2	Age_T0	17355
3	age_death_days	4830
4	Gender	17355
5	HbA1c_T0	17355
6	HbA1c_T1	17355
7	cholesterol_T0	10803
8	creatinine_T0	13971
9	LDL_T0	9457
10	HDL_T0	9458
11	Trig_T0	9608
12	AlkalinePhosphatase_T0	11396
13	ALT_T0	11124
14	TSH_T0	6345
15	BMI_T0	4121

16	SBP_T0	3316
17	DBP_T0	3316

BMI, SBP and DBP are the variables that are less populated. This is explained by the fact that vitals and anthropometrics are measured only if the patient has an in-person encounter with the hospital (e.g. because of an admission to the emergency department) and not if CUH is only the place where his/her blood is sent for analysis by his/her general practitioners who is managing T2DM in the outpatient setting. We did not have information on the proportion of CUH patients managed by general practitioners without a specialist input. The different prevalent treatment setting marks a first important difference with the next population, that from OSR, where all patients are fully managed in the hospital by our diabetologists in secondary care diabetes services.

4.2.2 OSR dataset

OSR is part of the network of Italian diabetology clinics that join their data in the yearly Annals of Diabetology initiative (see Section 2.2). In fact, our diabetes physicians operate on the Metaclinic EHR, by Meteda, already cited in Chapter 3, where we also showed how S-RACE can pull data from any of the EHRs of OSR. Indeed, using S-RACE we extracted data of T2DM patients from the Metaclinic database on January 31st, 2025. Any record present in the database until that day and belonging to a patient labelled as having T2DM was pulled into S-RACE, and we had first to explore the dataset before continuing with the cohort selection step. As described in Chapter 3, a preprocessing step pseudo-anonymised each of the Metaclinic internal patient identifier into a Cloud Patient Index (CPI). To access the data, we prepared and filed to the local Ethics Review Board the documents to start an observational clinical study that we named "AI-Predicted Disease Trajectories in Diabetes: A Retrospective Study" (AI-TRYDIA, NCT06280729). The extraction resulted in 10 different .csv files of varying numbers of columns and rows, each containing portions of the EHR:

1. Anagrafica.csv

   Description: registry of patients.

   Columns: CPI, date of birth, sex, centre (adult or paediatric diabetology).

   Rows: 9063, with 9050 unique CPIs. The 13 repeated CPIs belong to 5 patients that moved from the paediatric to the adult diabetology clinic, thereby generating a new record, and 8 patients that are duplicated for no apparent reason.

Missing data: none.

2. AnamnesiDiabetologica.csv

   Description: label with kind of diabetes and diagnosis date.

   Columns: CPI, date, diabetes.

   Rows: 8870, with 8719 unique CPIs. 138 CPIs have at least two records in this .csv, of which 114 have different registered diabetes types. In fact, we counted 14 different diagnosis types, as there are: 1) patients that were initially labelled as having T2DM that then turned out to be actually an immune-mediated disease; 2) patients that were initially non-diabetic or pre-diabetic (impaired glucose tolerance or impaired fasting plasma glucose) and that then turned overt T2DM; 3) T2DM that had their diagnosis date updated.

   Missing data: none. However, date is 1900-01-01 00:00:00 for 227 CPIs, which likely means that the physician did not input the date while recording the T2DM label during the visit.

3. AnamnesiRemota.csv

   Description: label with diagnoses other than diabetes, and diagnosis date.

   Columns: CPI, date of diagnosis, ICD9 code, description.

   Rows: 1871, with 1019 unique CPI. Therefore, 1019 CPIs belong to patients that have at least another disease, with 1393 having more than one; the others are missing or belong to patients that no other disease. In 14 instances, all myocardial infarctions, the events are likely wrongly dated back to 1905.

   Missing data: 129 rows of the ICD9 code column (7%) have missing data but the corresponding description column contains valuable information to allow proper mapping back to the ICD9 code.

4. Esami.csv

   Description: recorded measurements, either anthropometric (height, weight, BMI and circumferences), vitals or laboratory.

   Columns: CPI, date, description, unit of measurement, value, lower limit, upper limit.

   Rows: 1007683, with 8723 unique CPIs and 219 distinct measurement types.

   Missing data:

Column	***Missing (n)***	***Missing (%)***
Upper limit	220230	21.86

Lower limit	131652	13.06
*Unit of measurement**	4978	0.49
Value	87	0.01
Description	0	0.0
CPI	0	0.0
Date	0	0.0

*notably for scores such as the fibrosis 4 score or for auto-antibody qualitative yes/no measurements

5. StiliDiVita.csv

   Description: lifestyle information of patients.

   Columns: CPI, date, smoke (no/ex/yes), cigarettes per day, alcohol (no, occasional, moderate..), alcohol per day, physical activity (absent, modest, regular..), self-glycaemic checks (no/yes), frequency of self-glycaemic check, health education (no/yes), eating, self-pressure checks (no/yes).

   Rows: 3765, with 3584 unique CPIs.

   Missing data:

Column	***Missing (n)***	***Missing (%)***
Self-pressure checks	3746	99.5
Health education	3710	98.54
Frequency of self-glycaemic check	3703	98.35
Eating	3691	98.03
Self-glycaemic check	3685	97.88
Alcohol per day	2302	61.14
Cigarettes per day	1815	48.21
Physical activity	1468	38.99
Alcohol	1336	35.48
Smoke	1143	30.36
CPI	0	0.0
Data	0	0.0

6. Terapia.csv

   Description: recording of a new prescription event with a unique identifier that grows +1 at each encounter; the CPI+id duplet allows to uniquely identify each encounter.

   Columns: CPI, id, date, note (free text).

   Rows: 43504 with 7986 unique CPIs (i.e. for some patients no prescription was ever recorded, or no note was ever wrote).

   Missing data: values are missing in 18638 (43%) rows in the note column. Those records are likely a placeholder for the CPI+id duplet allowing then to match the records of the prescribed drugs to the date of prescription (“date”, in this .csv).

7. AntiDiabetica.csv

Description: recordings for all prescriptions of glucose-lowering medications.

Columns: CPI, id (matching with the "id" column from the Terapia.csv, where the date of the prescription can be recovered), meal, Italian 'AIC' code, brand name, active principles, units.

Rows: 133375, with 7498 unique CPIs. Repeated rows for same patients identify different drugs prescribed at the same encounter. When the "id" changes, data from a new encounter is recorded.

Missing data: none.

8. AltraTerapia.csv

   Description: recordings for all long-term (beyond hospital care) prescriptions, of medications not in the glucose-lowering category.

   Columns: CPI, id (matching with the "id" column from the Terapia.csv, where the date of the prescription can be recovered), Italian 'AIC' code, brand name, active principle.

   Rows: 124691, with 5455 unique CPIs. Repeated rows for same patients identify different drugs prescribed at the same encounter. When the "id" changes, prescriptions from a new visit are recorded.

   Missing data: none.

9. EsamiStrumentali.csv

   Description: collects inputs of the diagnostics tab of the EHR, where physician record information about procedures such as eye diagnostics, EKG, heart ultrasound, and carotid ultrasound.

   Columns: CPI, date, code (from Italian procedural codes), description, result (normal vs pathologic).

   Rows: 2985, with 1314 unique CPIs and 45 distinct types of diagnostics.

   Missing data: values are missing in only 1 (<1%) row of the code column while in 1051 (35%) rows of the result column.

10. SchedaComplicanze.csv

    Description: collects inputs of the complications tab of the EHR, where physicians record information about different body districts and the relative complications.

Columns: CPI, date, tab (19 different values: diary, heart, eye, history, cerebral vessels, other, peripheral vessels..), text (unstructured), classification (structured), ICD9 code.

Rows: 40908, with 7282 unique CPIs.

Missing data:

Column	*Missing (n)*	*Missing (%)*
ICD9 code	36994	90.43
Classification	31573	77.18
Text	7118	17.4
CPI	17	0.04
Tab	17	0.04
Date	12	0.03

4.3 General Population Characteristics

4.3.1 OSR dataset

As described in the previous section, the data extraction produced registry information for 9050 unique CPIs (Anagrafica.csv) but information about diabetes type for 8719 of them (AnamnesiDiabetologica.csv). Exploring the content of AnamnesiDiabetologica.csv, we realised that patients with 13 diabetes diagnoses other than "Type 2" leaked into the extraction. Therefore, we excluded 114 CPIs of patients with multiple diagnosis labels. In the remaining dataset of patients with one diagnosis each, still there were 14 of them with diagnoses other than "Type 2". We excluded them as well, ending up with a dataset of 8615 rows and 8591 unique CPIs. 24 CPIs indeed had multiple rows, all with T2DM labels, but with an updated diagnosis date, the first being "1900-01-01 00:00:00" (i.e. value not imputed by the clinicians) and the second one being the actual diagnosis date. Overall, thus, we identified the CPIs of 8591 T2DM patients out of the 9050 CPIs available in the registry .csv (Anagrafica.csv), meaning that some patients had no information about their diabetes type and therefore had to be excluded. From then on, we filtered the content of all other datasets by keeping only rows belonging to the 8591 bona-fide T2DM patients.

5386 (63%) patients in the T2DM cohort are males, while 3205 (37%) are females. We had the birth date of patients, but no information regarding them being alive or dead at the RWD extraction date. We thus decided to calculate the age of patients at their last measurement recorded in the dataset. The obtained median (Q1-Q3) age at the last measurement recorded at OSR is 70 (61-78) years (Figure 4.3.1).

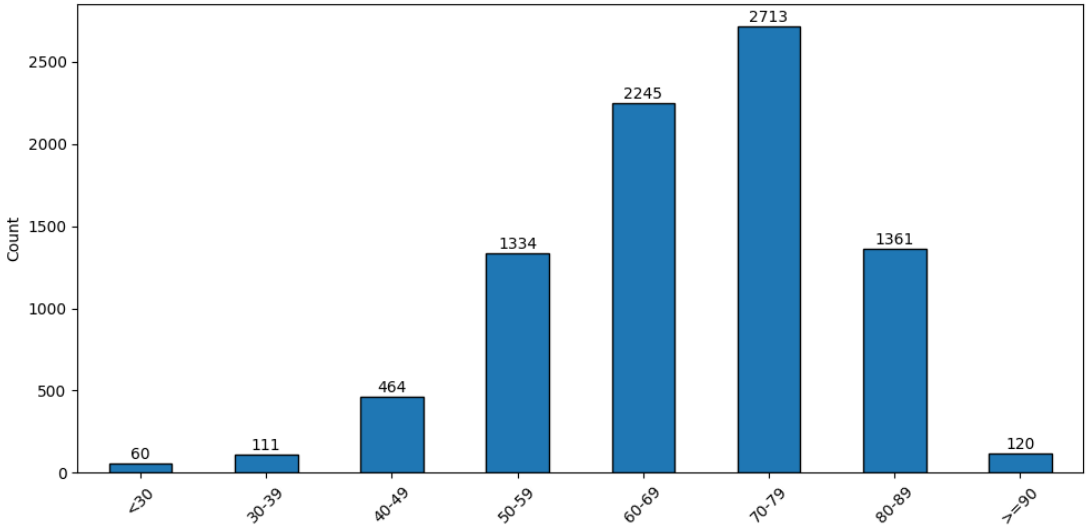
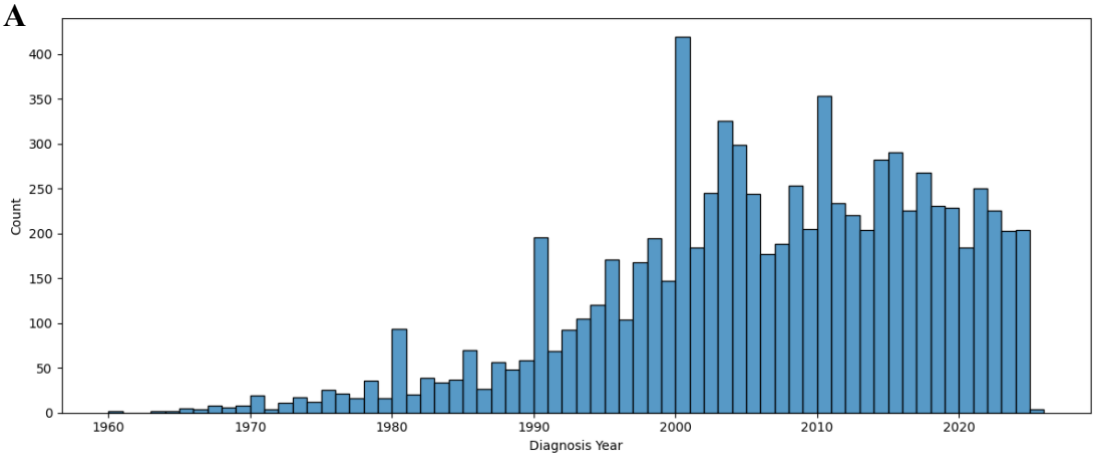


Figure 4.3.1 *– Age distribution of T2DM patients at their last measurement recorded in OSR diabetology RWD dataset.*

Panel A of Figure 4.3.2 reports the distribution of recorded diagnosis dates after 01-01-1959, that since 2001 reached a yearly plateau of roughly 225 new diagnoses per year. We then calculated the number of years with a T2DM diagnosis for each patient as the difference between the date of the last available measurement and the date of diagnosis (Figure 4.3.2, panel B). Patients with no T2DM diagnosis date (1900-01-01 00:00:00) were excluded from the calculation. The median number of years with a T2DM diagnosis is 11 (4-20). Additionally, by subtracting the birth date from the diagnosis date we calculated the age at diagnosis for each patient, whose median value is 56 (47-65) years old. Diagnoses made in more recent years have a tendency towards an older mean age at diagnosis (Figure 4.3.2, panel C)

A



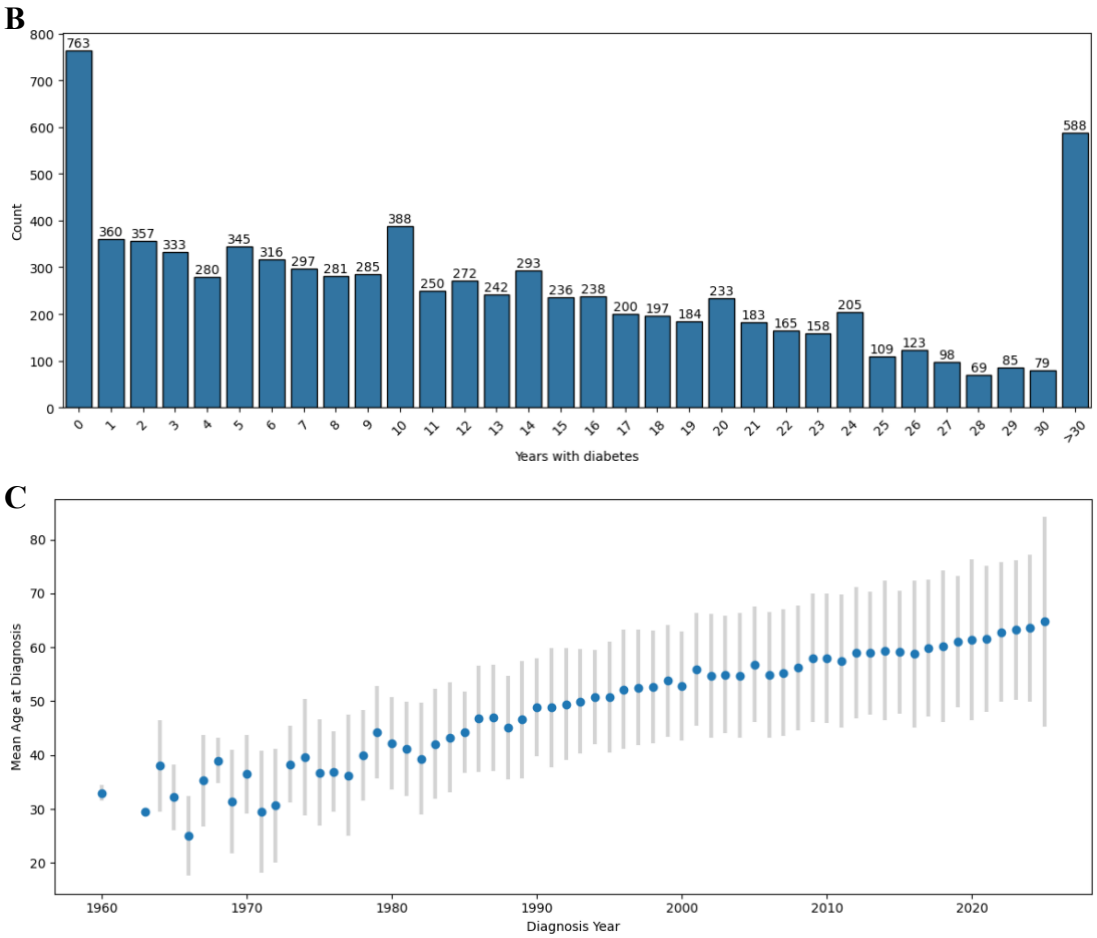


Figure 4.3.2 *– (**A**) Distribution in time of T2DM diagnosis dates. (**B**) Distribution of years elapsed from the recorded date of T2DM diagnosis and the last measurement encounter at OSR. (**C**) Mean +/- SD of age at diagnosis, by diagnosis year.*

As anticipated in the previous section, the dataset with comorbidity information for this population was particularly poorly populated. After filtering for bona-fide T2DM patients, the past medical history dataset had 1007 unique CPIs, 12% of the starting population, with 289 distinct ICD9 codes. Figure 4.3.2 shows the most common 10 ICD9 code. As the EHR field for past medical history allows physicians to pick any ICD9 code with no guideline or standardisation, the same disease is identified with different codes according to physician preference or chance (e.g. Ischemic heart diseases, Acute myocardial infarction [410] and History of myocardial infarction [412]), leading to a ballooning of alternative ICD9 codes that have relatively few counts each, with 555 "other" codes (Figure 4.3.3). The top 10 diagnoses include hypertension, dyslipidaemias, myocardial infarction, atrial fibrillation and ischemic stroke. The maximum number of comorbidities in a single patient is 8, with a median of 2 (1-2) comorbidities per patient, to which T2DM must be added.

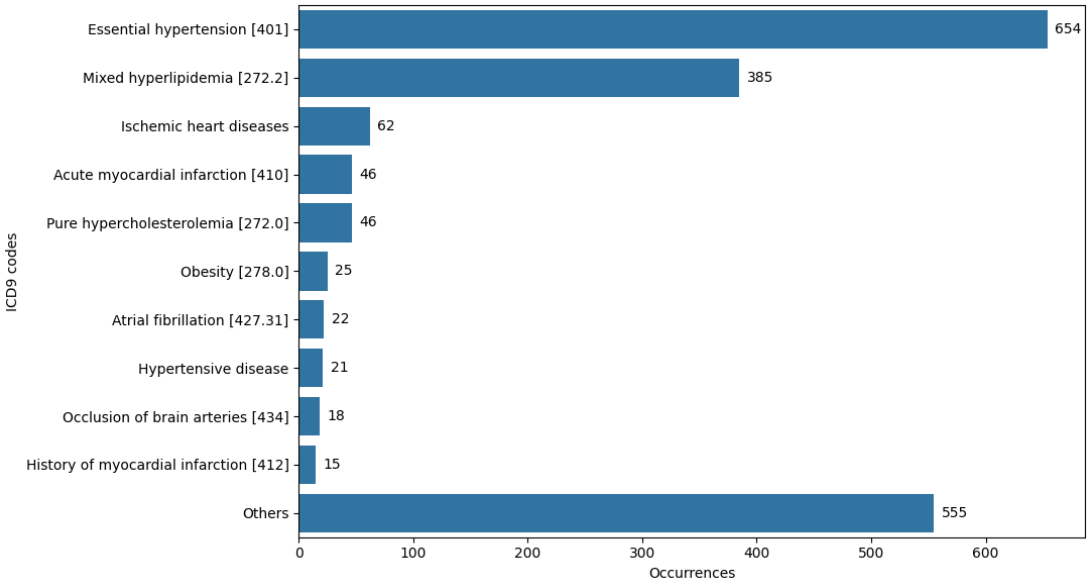


Figure 4.3.2 *– Top 10 ICD9 codes by occurrence without preprocessing of codes to merge overlapping diseases.*

4.4 Overview and Trends Over Time of Major Measurements in OSR

Blood tests, vitals and anthropometric measurements are fundamental data in the evaluation of T2DM patients, allowing to estimate disease control (HbA1c), complications, like kidney function decline (creatinine, estimated glomerular filtration rate [eGFR]), and comorbidities (hypertension via systolic and diastolic blood pressure, obesity via BMI, dyslipidaemias via low-density lipoprotein [LDL], HDL and total cholesterol). Patients are usually requested to draw blood some days before the planned outpatient visit, during which results are evaluated together with freshly recorded vitals and weight and height. Diabetologists input these data in the dedicated EHR fields, which don't have embedded plausibility checks, exposing to the risk of data entry mistakes. Only when patients had their blood tests in OSR in the days preceding the outpatient clinic visit, physicians can automatically import some of them into the EHR, reducing the risk of typos. Figure 4.4.1 shows distributions of selected measurements on the whole bona-fide T2DM population prior to data preprocessing. Evidently, extreme values, certainly due to data entry mistakes, are common.

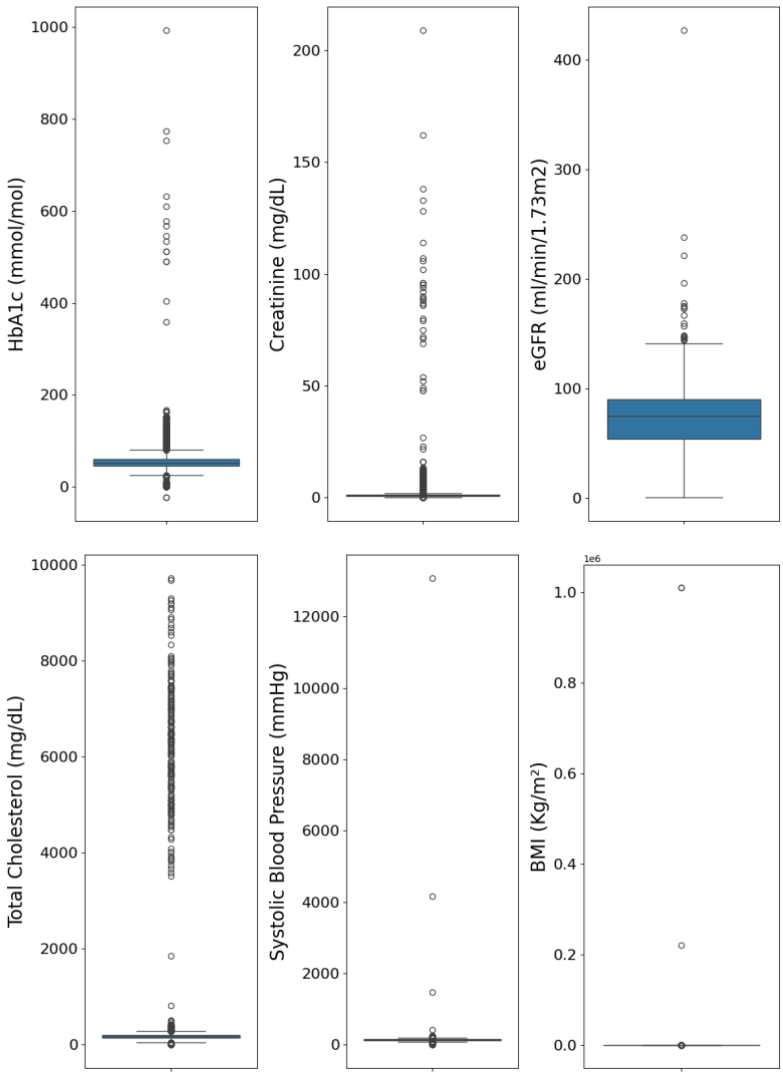


Figure 4.4.1 *– Box plots of selected blood tests, of systolic blood pressure and of body mass index. Round circles represent individual outlier recordings. BMI: body mass index; eGFR: estimated glomerular filtration rate; HbA1c: glycated haemoglobin.*

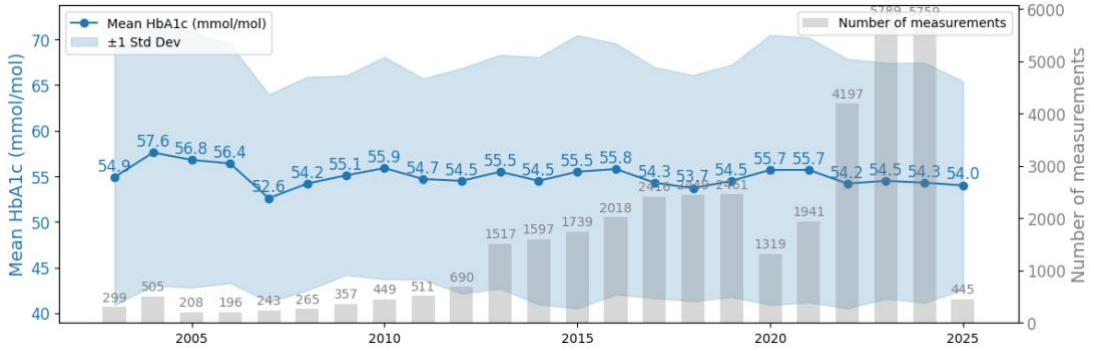
An interesting value of RWD spanning long time windows is that they allow to evaluate trends in the real-world incidence and prevalence of diseases, in the control of their biomarkers, in the prescription of medications and many, many more. After removing extreme or impossible values from the measurements of interest, we plotted their mean and standard deviation over time in the dataset (Figure 4.4.2). For the aims of this trend analysis, the lower and upper thresholds for identification of values likely due to data entry mistakes were decided together with our diabetologists with no specific statistical rule but based on clinical expertise, and are reported in Table 4.4.1.

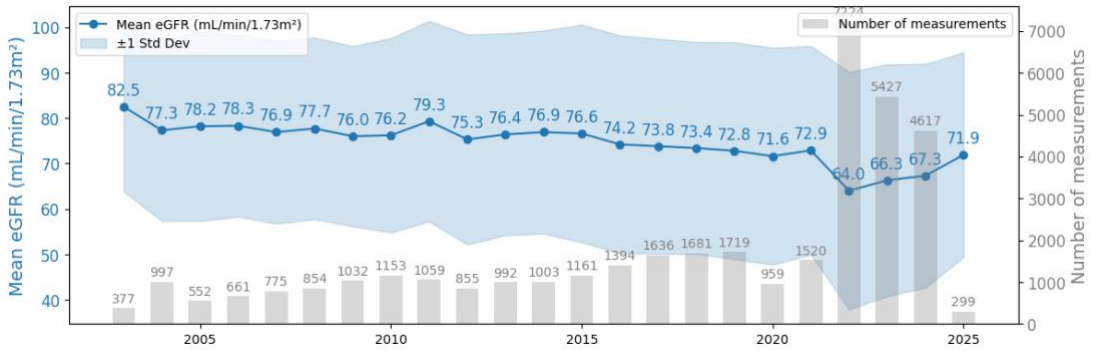
Table 4.4.1 *– Lower and upper thresholds for preprocessing selected measurements before analysing their 20-year trends.*

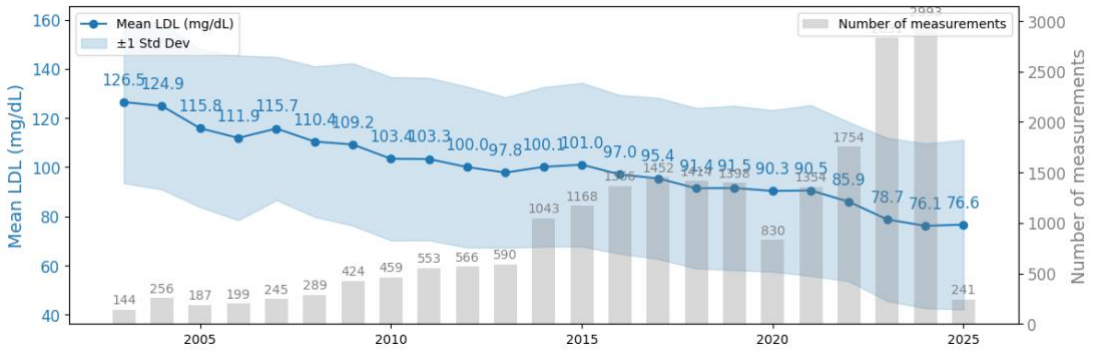
Measurement	***Lower bound***	***Upper bound***

HbA1c (mmol/mol)	20	200
eGFR (ml/min/1.73m^3)	0	200
LDL cholesterol (mg/dL)	0	400
Systolic blood pressure (mmHg)	90	180
BMI (kg/m^2)	15	60

A first striking trend is that of the number of yearly measurements, sharply increasing from 2022 on. We know that 2022 marked the beginning of a staff turnover in our diabetology clinic, due to some physicians reaching retirement age. This led to the establishment of an overall younger and more data-conscious group of diabetologists.







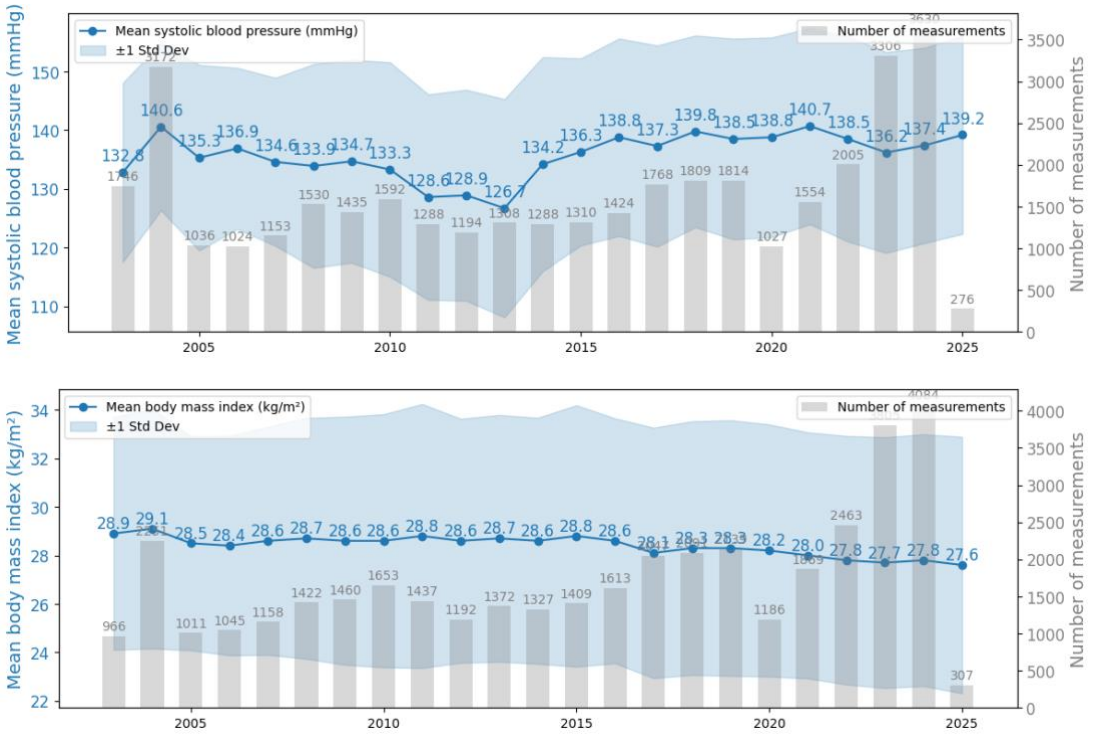


Figure 4.4.2 *– Trends over time for major measurements of interest in our T2DM dataset from 2003 to 2025 (January). In blue: annual mean and standard deviation. In grey: annual number of records. eGFR: estimated glomerular filtration rate; HbA1c: glycated haemoglobin; LDL: low-density lipoprotein cholesterol.*

The aims of the management of T2DM and its comorbidities are a decreased HbA1c, a stable eGFR, a decreased systolic blood pressure, a decreased LDL cholesterol and a decreased BMI. It is thus interesting to see that most trends of the past 20 years don't show major improvements of these endpoints, if not for a decrease in LDL cholesterol. It must be noted that these descriptions are based on averages of the whole, dynamic, population and on RWD that, as already mentioned, suffer from biases such as the survivorship or observability bias.

The trend of HbA1c shows stable values being just a little over the therapeutic target of 53 mmol/mol; however, as the mean age at diagnosis increases, as shown in Figure 4.3.2C, we expect targets to be less tight and patients being generally more complex to be treated having more resistant diseases. Mean eGFR slowly declines with time, something that can be explained by patients with some follow up in the dataset being more than new patients and showing a decline with time of their eGFR. Additionally, we can argue that given the shortage of diabetologists and recent trends in healthcare organisation in Italy, less complex patients are increasingly treated by their general practitioners. Our RWD might therefore suffer from a selection bias of "worse" patients.

Mean LDL cholesterol drops by 39% from 124.9 mg/dL in 2004 to 76.1 mg/dL in 2024 (-48.8 mg/dL). Mean systolic blood pressures are consistently in the "elevated" range according to the latest European Society of Cardiology guidelines but may suffer from the "white-coat effect" (McCarthy *et al*, 2024). Only between 2011 and 2013 we see a relative decrease in the mean systolic blood pressure values, a finding that we could not explain. Last, the trend for mean BMI shows a decline of only 1.3 kg/m^2 from 2004 (29.1 kg/m^2) to 2024 (27.8 kg/m^2), remaining well high in the overweight category. Medications with prominent weight-loss effects such as GLP-1 are starting to become prevalent only in recent times. We might therefore need to wait some more years to see their impact on such population-wise trends.

4.5 Cohort and Feature Selection

Having explored the general characteristics of the study population, we could proceed to the crucial cohort and feature selection steps.

4.5.1 Cohort selection in the CUH dataset

As we extracted from the ERIN database only patients meeting the inclusion criteria, the CUH cohort of interest was the full dataset.

4.5.2 Cohort selection in the OSR dataset

As we extracted from Meteda all T2DM patients ever recorded in the EHR from its implementation, we had to select the final OSR cohort of interest.

For each of the 8591 bona-fide T2DM patients we looked for HbA1c measurements, either in % or mmol/mol. We found at least one value for 7936 patients and converted the % values in mmol/mol with the formula: $HbA1c_{mmol/mol=}$ 10.929 * ($HbA1c_{\%}$ - 2.15). Mmol/mol was preferred as a unit of measurement as it was consistent with CUH and since 2012 it is the obligatory unit of measurement in Italy. We then removed rows with HbA1c values ≤ 20 mmol/mol and ≥ 200 mmol/mol. For each patient, we isolated the baseline value (i.e. the first recording available), which marked the T0 of our experiment, and looked for a second HbA1c recording at T1, 3 years +/- 180 d after T0. We found a T1 HbA1c value for 1946 patients but decided to include only patients with a T0 HbA1c ≥ 48 mmol/mol in the final list, which contained 1406 unique CPIs.

4.5.3 Feature selection

Table 4.5.1 shows the availability of features in the two identified cohorts.

For OSR, having extracted raw .csv files from the different portions of Meteda EHR, we had to extensively manipulate them to generate the final dataset with the predictive features and the target variable. We used the CPIs of the final OSR cohort to look for other measurements done at T0, namely: creatinine, BMI, systolic and diastolic blood pressure, fasting plasma glucose, triglycerides, total, LDL and HDL cholesterol, alanine aminotransferase, aspartate aminotransferase. We also engineered a few further features: the time from diagnosis date to T0, that we considered the "time to first visit", the age at T0, the magnitude of HbA1c change from T0 to T1 and, most importantly, the target variable, that is a binary label defined as "HbA1c decrease" true or false based on the previously calculated T1-T0 value.

For CUH, the time to first visit could not be computed, as we did not have the T2DM diagnosis dates. Additionally, the number of comorbidities was not available as we did not extract other diagnosis labels for the 17355 patients. Also, the aspartate aminotransferase (AST) is not part of the liver panel at CUH and was therefore not available. On the other hand, dead or alive status was not present in the OSR EHR and TSH at T0 was not selected in the OSR cohort.

Table 4.5.1 *– Availability of predictive features and target variable for the two cohorts of included patients. Variables with >50% missingness are marked as (X). ALT: alanine aminotransferase; AST: aspartate aminotransferase; BMI: body mass index; LDL: low-density lipoprotein cholesterol; HDL: high-density lipoprotein cholesterol; TSH: thyroid stimulating hormone.*

	Feature	OSR	CUH
1	Age	✓	✓
2	Gender	✓	✓
3	HbA1c_T0	✓	✓
4	BMI	✓	(X)
5	Systolic Blood Pressure	✓	(X)
6	Diastolic Blood Pressure	✓	(X)
7	Glycaemia	✓	(X)
8	Triglycerides	✓	✓
9	Total Cholesterol	✓	✓
10	LDL	(X)	✓
11	HDL	✓	✓
12	AST	(X)	✗
13	ALT	(X)	✓
14	Creatinine	✓	✓
15	Age at death	✗	✓
16	Time to first visit	✓	✗

17	Number of comorbidities	(✗)	?
18	TSH	?	✓
	HbA1c_T1	✓	✓
	HbA1c change	✓	✓

The final list of features selected for the exploratory data analysis and comparison between the two centres in the end comprises: 1) age; 2) sex; 3) HbA1c at T0; 4) HbA1c at T1; 5) creatinine at T0; 6) total cholesterol at T0; 7) LDL cholesterol at T0; 8) HDL cholesterol at T0; 9) triglycerides at T0; 10) alanine aminotransferase (ALT) at T0; 11) BMI at T0; 12) systolic blood pressure (SBP) at T0; 13) diastolic blood pressure (DBP) at T0; and 14) HbA1c change.

4.5.4 Outlier management

After we identified the patient cohorts and the set of features, we had to preprocess the datasets to remove outlier values before moving to the exploratory data analysis. When assessing the distribution of our variables, we found them to be strongly skewed, as expected for variables in the biomedical domain, with long tails made both of impossible values (e.g. 0 or negative values, or extremely high values) and of values belonging to patients with very extreme presentations. Such subjects are uncommon, but not implausible from the clinical point of view, and excluding them from our cohorts would have on the one hand reduced the noise and possibly improved models' performances, but on the other hand it would have made the models less fair. This made an exclusively mathematical approach to outlier identification impractical.

Consequently, we chose to apply a mixed approach, comprising Tukey's fences and clinical judgement. First, we identified the first (Q1) and third quartile (Q3) values of the distributions and calculated the interquartile range (Q3-Q1 [IQR]) for each variable in the two datasets. Second, we multiplied the IQR times a preset factor ("threshold"). Then, we calculated the lower and upper boundaries. The lower boundary was set to be the biggest between either 0 or Q1 – threshold * IQR, as the calculation sometimes yielded negative values. The upper boundary was set to be Q3 + threshold * IQR. Values being outside these boundaries were turned into Not-a-Number (NaN). The threshold was chosen by the author of this manuscript according to clinical expertise and manual revision of extreme values for each variable.

For the CUH dataset, the threshold multiplier was set at 1.5 for SBP and DBP and at 5.5 for blood tests and BMI. For the OSR dataset, the threshold multiplier was set at 2.5 for all the variables. No other approaches to outlier management (Z-score, multivariate methods) were considered. Table 4.5.2 and 4.5.3 detail the lower and upper bounds, and the number and % of identified outliers for CUH and OSR respectively.

Table 4.5.2 *– Outliers were identified in the CUH dataset by setting lower and upper bounds mathematically. The lower bound was set at 1.5 times (SBP/DBP) or 5.5 (BMI and blood tests) the interquartile range below the first quartile or at 0. The upper bound was set at 1.5 times (SBP/DBP) or 5.5 (BMI and blood tests) times the interquartile range above the third quartile. Absolute number and % of identified and discarded outliers are also reported.*

Variable	*Lower bound*	*Upper bound*	*Number of outliers*	*% of outliers*
SBP_T0	80.5	204.5	59	1.78%
DBP_T0	41.5	109.5	43	1.30%
cholesterol_T0	0	14.0	15	0.14%
LDL_T0	0	9.68	73	0.77%
HDL_T0	0	3.75	2	0.02%
triglycerides_T0	0	9.75	94	0.98%
ALT_T0	0	148.0	81	0.73%
BMI_T0	0	87.92	7	0.17%

Table 4.5.3 *– Outliers were identified in the OSR dataset by setting lower and upper bounds mathematically. The lower bound was set at 2.5 times the interquartile range below the first quartile or at 0. The upper bound was set at 2.5 times the interquartile range above the third quartile. Absolute number and % of identified and discarded outliers are also reported.*

Variable	*Lower bound*	*Upper bound*	*Number of outliers*	*% of outliers*
ALT_T0	1e-08	95.0	27	5.72%
triglycerides_T0	1e-08	417.5	34	2.90%
LDL_T0	1e-08	259.17	2	0.29%
HDL_T0	1e-08	96.0	8	0.68%
cholesterol_T0	1.0	373.0	4	0.30%
DBP_T0	45.0	105.0	8	0.55%
SBP_T0	46.5	223.5	3	0.20%
creatinine_T0	1e-08	1.9	27	2.77%
BMI_T0	9.35	47.15	8	0.53%
age_T0	16.0	112.0	9	0.47%

4.6 Exploratory Data Analysis

CUH and OSR dataset showed a different pattern in data missingness (Figure 4.5.1). In CUH, as anticipated, we measured larger portions of missing values in the anthropometric and vitals measurements, as they require the patient to physically encounter the hospital around the same period in which they have their HbA1c test. This indeed did not happen for most of CUH patients, that have their T2DM managed by their GPs outside of the hospital. In OSR, ALT, creatinine and LDL had more missing values than CUH, as they depended on clinicians to manually transcribing their values in the EHR from the lab reports that patient carried with them during the outpatient visit.

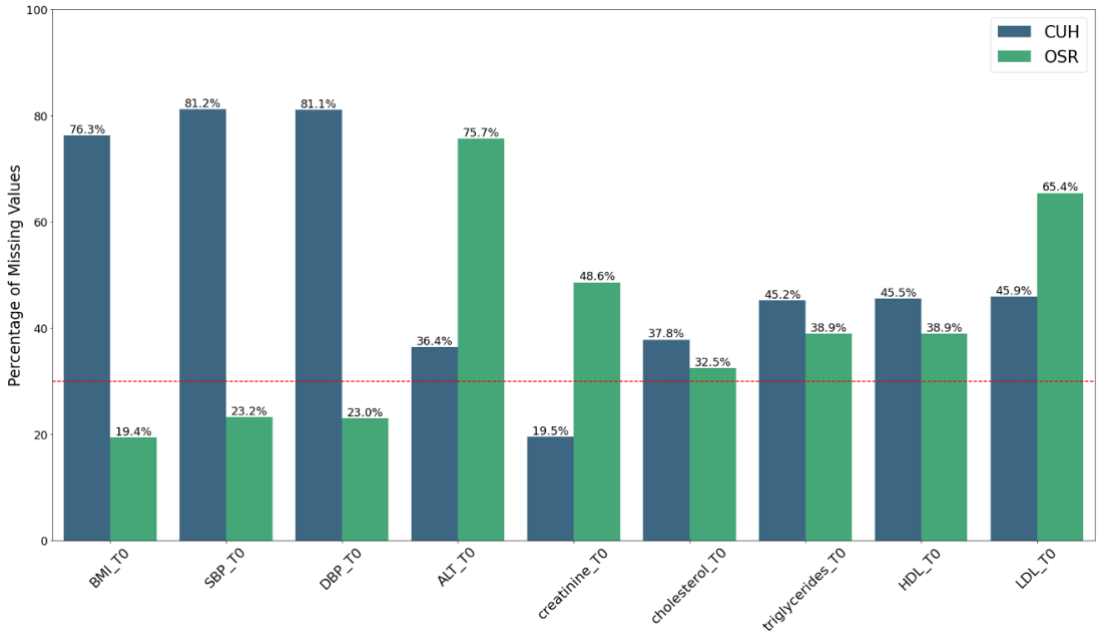


Figure 4.6.1 *– Histogram plot of the percentage of missing values for the final set of measurements that advanced to the exploratory data analysis step. In blue: CUH measurements. In green: OSR measurements. The red dashed line marks the 30% threshold, as most machine learning pipeline tend to consider 70% as an acceptable completeness threshold to include variables in models.*

Demographics are not included as they were all populated in the two datasets, as well as HbA1c and related engineered features (HbA1c change and HbA1c boolean decrease variable) that were available by definition, being necessary for patient inclusion.

Table 4.6.1 *– Distribution of variables in the overall population and statistical comparison of the distributions in the two cohorts. Values are reported as mean (SD) or median (IQR) as appropriate.*

Variable	***Overall***	***CUH (N=17355)***	***OSR (N=1406)***	***p-Value***
age_T0 (years)	66.00 (20.00)	66.00 (20.00)	64.00 (16.00)	<0.001
HbA1c_T0 (mmol/mol)	58.00 (21.00)	58.00 (20.00)	60.00 (19.00)	<0.001
HbA1c_T1 (mmol/mol)	56.00 (18.00)	56.00 (19.00)	54.00 (13.00)	<0.001
cholesterol_T0 (mg/dL)	170.15 (58.00)	166.28 (58.00)	189.00 (64.00)	<0.001
creatinine_T0 (mg/dL)	0.86 (0.34)	0.85 (0.35)	0.90 (0.30)	<0.001
LDL_T0 (mg/dL)	84.69 (48.34)	83.91 (47.18)	102.60 (51.30)	<0.001
HDL_T0 (mg/dL)	44.08 (16.36)	43.70 (16.63)	46.00 (16.00)	<0.001
Triglycerides_T0 (mg/dL)	159.48 (124.04)	159.48 (117.84)	134.00 (94.50)	<0.001
ALT_T0 (U/L)	25.00 (19.00)	25.00 (19.00)	27.00 (23.00)	<0.001
BMI_T0 (kg/m^2)	30.00 (8.79)	30.70 (9.43)	28.20 (6.30)	<0.001
SBP_T0 (mmHg)	140.00 (29.00)	142.00 (30.00)	135.00 (25.00)	<0.001
DBP_T0 (mmHg)	75.00 (15.00)	75.00 (17.00)	80.00 (10.00)	<0.001
HbA1c_change (mmol/mol)	-3.00 (17.00)	-2.00 (16.00)	-5.90 (18.00)	<0.001

As shown in Table 4.6.1, CUH and OSR cohorts are comparable in terms of age, T0 HbA1c and T1 HbA1c. In both cohorts, the HbA1c value after 3 years is lower, meaning that the management has generally a positive effect, with a more pronounced decrease in OSR, but not enough to reach the treatment target in either of the cohorts, as median HbA1c at T1 is above 53 mmol/mol in both. Cholesterol control is better in the English cohort while mean BMI and median systolic blood pressure are lower in the Italian cohort. We then decided to drop patients from CUH without a baseline BMI value, considering the CUH BMI+ subgroup to be more comparable to OSR patients. As shown in Table 4.6.2, participants with a baseline BMI are younger and with a slightly worse baseline HbA1c but more negative HbA1c change compared to those without a BMI at T0.

Table 4.6.2 *– Statistical comparison between CUH subgroup of patients with available BMI at T0 and OSR patients. Values are reported as mean (SD) or median (IQR) as appropriate. The characteristics of the CUH subgroup missing the baseline BMI are also reported for completeness.*

Characteristic	***CUH BMI- (N=13241)***	***CUH BMI+ (N=4114)***	***OSR (N=1406)***	***p-Value***
age_T0	67.00 (20.00)	63.00 (22.00)	64.00 (16.00)	0.001
HbA1c_T0	58.00 (19.00)	59.00 (26.00)	60.00 (19.00)	0.215
HbA1c_T1	57.00 (18.00)	55.00 (21.00)	54.00 (13.00)	p < 0.001
cholesterol_T0	166.28 (58.00)	166.28 (65.74)	189.00 (64.00)	p < 0.001
creatinine_T0	0.85 (0.33)	0.86 (0.43)	0.90 (0.30)	0.003
LDL_T0	83.91 (46.89)	82.75 (50.66)	102.60 (51.30)	p < 0.001
HDL_T0	43.70 (16.24)	43.70 (17.40)	46.00 (16.00)	p < 0.001
triglycerides_T0	159.48 (117.84)	159.48 (126.70)	134.00 (94.50)	p < 0.001
ALT_T0	25.00 (18.00)	26.00 (22.00)	27.00 (23.00)	0.391
BMI_T0	N/A	30.70 (9.43)	28.20 (6.30)	p < 0.001
SBP_T0	146.00 (31.00)	142.00 (30.00)	135.00 (25.00)	p < 0.001
DBP_T0	77.00 (17.00)	74.00 (17.00)	80.00 (10.00)	p < 0.001
HbA1c_change	-2.00 (15.00)	-4.00 (20.00)	-5.90 (18.00)	p < 0.001

We discussed the problem of class imbalance in RWD already in section 2.3. Before moving to the training step, we therefore wanted to check the relative distribution of the two classes of the binary outcome: HbA1c decrease true or false. As shown in Figure 4.6.2, the true outcome has a higher prevalence in both cohorts. When removing participants with baseline HbA1c below 48 mmol/mol in the OSR cohort, the true class becomes even more prevalent, indicating that for patients starting from low values of HbA1c it is more unlikely to experience an improvement at 3 years. This can be expected given the fact that these patients are more likely treated less aggressively or managed with

lifestyle adjustments only. In the subgroup of patients with available baseline BMI at CUH, we detected only a slight increase of prevalence in the true outcome.

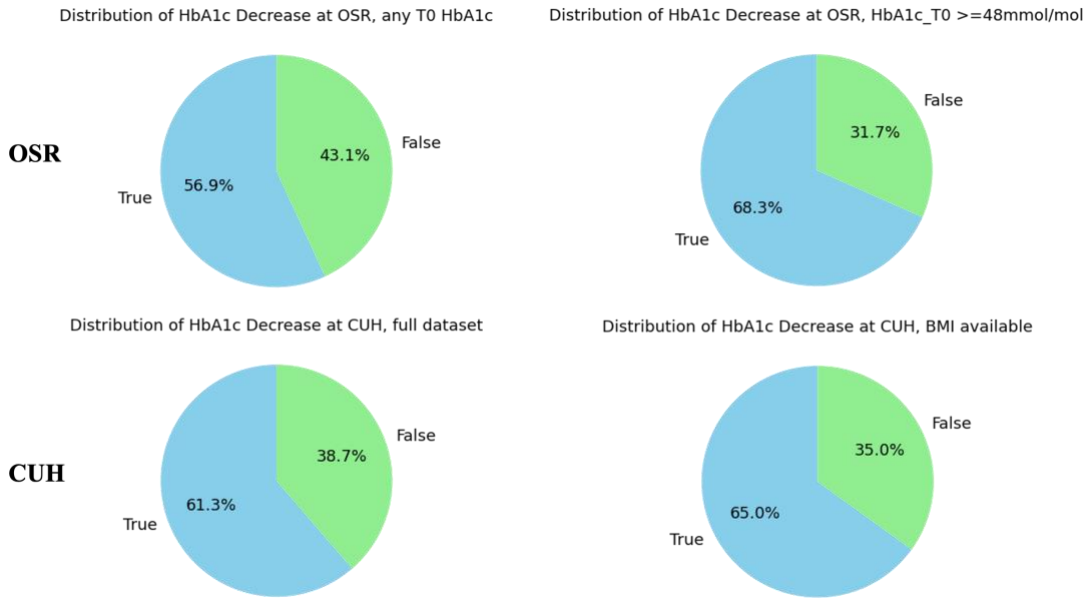


Figure 4.6.2 *– Pie charts comparing the prevalence of patients experiencing or not an HbA1c decrease in 3 years. Upper charts: OSR dataset. Lower charts: CUH dataset. In blue: true class. In green: false class.*

We next wanted to compare the subgroups of patient with and without the outcome (HbA1c decrease), in both the OSR and CUH cohorts. Table 4.6.3 summarises the results for numerical features. Although we expected it to be significantly different by definition, we left the HbA1c_change variable to measure the magnitude of change in the two outcome classes and compare relative changes in the two institutions. We also wanted to assess the HbA1c at three years (HbA1c_T1) to evaluate the change in the two subgroups. In the OSR cohort, HbA1c decrease patients have worse baseline HbA1c, LDL, total cholesterol and triglycerides values, have a slightly higher BMI and are younger. Of note, they achieve a median T1 HbA1c value below the 53 mmol/mol treatment target. Baseline HbA1c follows the same pattern in the CUH cohort as well, with identical median baseline values in the two subgroups: 64 mmol/mol for patients who then decrease in time vs 54 mmol/mol for those who worsen their diabetes control in 3 years. At CUH as well, patients who successfully reduce their HbA1c reach a median value below the treatment target. On the contrary, in both cohorts the patients what have an HbA1c increase fail to meet the treatment targets at 3 years (median HbA1c_T1 in the HbA1c increase subgroup = 60.70 mmol/mol in the OSR cohort, 66.00 mmol/mol in the CUH cohort).

Table 4.6.3 *– Statistical comparison between subgroup of patients reaching the HbA1c decrease outcome or not in both CUH and OSR cohorts. Values are reported as mean (SD) or median (IQR) as appropriate.*

	Characteristic	***Overall (N=1406, 100%)***	***HbA1c decrease (N=961, 68%)***	***HbA1c increase (N=445, 32%)***	***p-Value***
OSR	***age_T0***	**64.00 (16.00)**	**64.00 (15.00)**	**66.00 (16.00)**	**0.02**
	ALT_T0	27.00 (23.00)	27.00 (24.00)	23.00 (22.00)	0.09
	BMI_T0	**28.20 (6.30)**	**28.40 (6.60)**	**27.70 (5.50)**	**0.03**
	cholesterol_T0	**189.00 (64.00)**	**192.00 (64.00)**	**183.00 (62.25)**	**0.00**
	creatinine_T0	0.90 (0.30)	0.90 (0.29)	0.92 (0.34)	0.52
	DBP_T0	**80.00 (10.00)**	**80.00 (10.00)**	**77.00 (10.00)**	**0.01**
	HbA1c_change	**-5.90 (18.00)**	**-12.00 (17.40)**	**5.40 (7.80)**	**0.00**
	HbA1c_T0	**60.00 (19.00)**	**64.00 (21.80)**	**54.00 (9.90)**	**0.00**
	HbA1c_T1	**54.00 (13.00)**	**50.00 (11.60)**	**60.70 (13.80)**	**0.00**
	HDL_T0	46.00 (16.00)	46.00 (15.00)	48.00 (18.00)	0.05
	LDL_T0	**102.60 (51.30)**	**107.40 (52.80)**	**95.20 (52.90)**	**0.01**
	SBP_T0	135.00 (25.00)	140.00 (25.00)	135.00 (23.00)	0.20
	triglycerides_T0	**134.00 (94.50)**	**142.00 (104.25)**	**121.00 (77.50)**	**0.00**
	Characteristic	***Overall (N=4114, 100%)***	***HbA1c decrease (N=2676, 65%)***	***HbA1c increase (N=1438, 35%)***	***p-Value***
CUH	*age_T0*	63.00 (22.00)	63.00 (22.00)	64.00 (22.00)	0.19
	ALT_T0	**26.00 (22.00)**	**26.00 (22.00)**	**25.00 (23.00)**	**0.02**
	BMI_T0	**30.70 (9.43)**	**30.47 (9.51)**	**30.99 (9.21)**	**0.01**
	cholesterol_T0	166.28 (65.74)	170.15 (65.74)	166.28 (58.00)	0.28
	creatinine_T0	0.86 (0.43)	0.86 (0.43)	0.86 (0.40)	0.13
	DBP_T0	74.00 (17.00)	74.00 (17.00)	74.00 (17.00)	0.83
	HbA1c_change	**-4.00 (20.00)**	**-11.00 (21.00)**	**8.00 (14.00)**	**0.00**
	HbA1c_T0	**59.00 (26.00)**	**64.00 (33.25)**	**54.00 (12.00)**	**0.00**
	HbA1c_T1	**55.00 (21.00)**	**50.00 (16.00)**	**66.00 (23.00)**	**0.00**
	HDL_T0	43.70 (17.40)	43.70 (17.40)	43.31 (17.01)	0.69
	LDL_T0	82.75 (50.66)	83.53 (51.82)	81.98 (47.18)	0.57
	SBP_T0	142.00 (30.00)	142.00 (30.50)	141.00 (30.00)	0.70
	triglycerides_T0	159.48 (126.70)	164.80 (132.90)	159.48 (132.90)	0.06

These observations enable several clinically relevant considerations. First, the worse baseline clinical values in patients who achieved the outcome (HbA1c decrease) may reflect more aggressive management by physicians in response to severe presentations, or improved treatment adherence due to enhanced patient education or heightened patient concern. Alternatively, this pattern could arise from regression to the mean, where patients with extreme baseline values are more likely to show improvement over time (Barnett *et al*, 2005). Second, some patients with baseline HbA1c between 48 mmol/mol (diagnostic threshold) and 53 mmol/mol (treatment target) may not have initiated therapy

at the first encounter with the hospital. In fact, for T2DM patients in this HbA1c range, a trial of lifestyle management alone is reasonable if the patient is motivated and asymptomatic (Initial management of hyperglycemia in adults with type 2 diabetes mellitus - UpToDate). However, most patients will eventually require medication, and our observations possibly stem from therapeutic inertia or latency in starting a drug, a common finding in RWD datasets as this. Finally, we observed that patients who failed to meet the outcome (HbA1c worsening) not only missed the treatment target but deviated substantially from it. This finding underscores the critical importance of prompt therapeutic escalation in patients exhibiting upward HbA1c trajectories, as delayed intervention may contribute to poorer long-term outcomes (Overcoming Therapeutic Inertia | therapeuticinertia.diabetes.org).

As a last EDA step, we built correlation heatmaps for the variables of interest in the two cohorts with the aim of checking for the presence of possible collinear variables to be managed before training (Figure 4.6.3). Correlation patterns in the two cohorts are overlapping, and several interesting significant correlations can be commented. Increased BMI is correlated to increased DBP, SBP, baseline HbA1c, and triglycerides while decreased HDL and age. Obviously, DBP and SBP are positively correlated. Lipid profile shows clear correlations, with triglycerides, LDL and total cholesterol being positively correlated and HDL being negatively correlated to triglycerides and to male sex. Of note, in the OSR dataset, HbA1c decrease is positively correlated with HbA1c at T0, i.e. patients with higher baseline HbA1c are more likely in the HbA1c decrease true (= 1) class. Along the ML experiments we additionally adopted the Variance Inflation Factor approach to define collinearity among features and better select them.

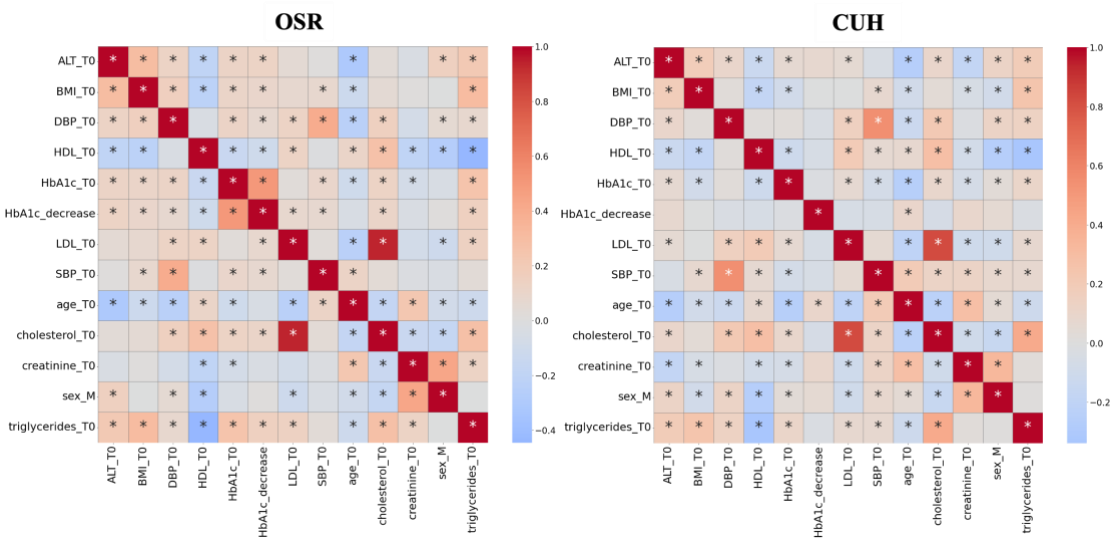


Figure 4.6.3 *– Spearman correlation heatmaps of the selected features and target variable for the two cohorts, OSR to the left, CUH to the right. Colour grading represents the value of spearman r for each couple of variables, with values closer to one being red and values closer to -1 being blue. * $p < 0.05$*

4.7 Logistic Regression Dropping Missing Values

Having completed the EDA step, we moved to modelling. Previous experience in our group with ML models on the same T2DM RWD had revealed that Logistic Regression (LR) achieved comparable results to XGBoost, a more complex, less explainable model (see Section 1.5). We therefore here chose to begin our experiments with the LR. This allowed us to focus on preprocessing decisions with a simple ML model, and evaluate their impact on performance metrics, addressing the already introduced challenge of a lack of standardised methods for ML applications to RWD. We performed all the experiments with a 10-fold stratified shuffle split approach, to estimate performances of each pipeline in the validation set with a confidence interval while making sure that no bias was caused by class imbalance in the splits. For this set of experiments, we used the OSR dataset as a train-validation set and CUH dataset as a test set and chose as a random state the value of 2. We started from the full set of features and then progressively removed collinear ones and those missing above specific thresholds. Following a trial-and-error approach, we also tested different data scaling methods, different sizes of the validation set, and different number of maximum iterations allowed for the models to converge.

4.7.1 Full set of 12 features

LR notably does not internally manage missing (NaN) values. We initially dropped all the "examples" (a patient with features and target) were there was at least one NaN value, and we set the validation set size at 20%. Resulting dataset rows were: OSR total = 155 split in 92 for train and 23 for validation; CUH test = 777. The classes of the target variable were strongly imbalanced (e.g. in the 10th split: true 71 vs false 21). At this stage, we fed the model with all 12 features: 1) age; 2) sex; 3) HbA1c at T0; 4) creatinine at T0; 5) total cholesterol at T0; 6) LDL at T0; 7) HDL at T0; 8) triglycerides at T0; 9) ALT at T0; 10) BMI at T0; 11) SBP at T0; and 12) DBP at T0. The maximum number of iterations taken for the solvers to converge was set at 100. As Table 4.7.1 shows, these settings lead to a very high sensitivity but at the expenses of a miserable specificity. Confidence intervals are also very large, given the few examples provided. However, performance is stable in the CUH test set (Figure 4.7.1).

Table 4.7.1 *– Performance metrics of a logistic regression model across the 10 OSR validation sets and in the CUH test set. Figures are reported as mean (95% c.i.). Input features: age, sex, HbA1c, creatinine, total cholesterol, LDL, HDL, triglycerides, ALT, BMI, SBP, and DBP. Missing management: drop. Validation set size: 20%. Class weight: not provided and provided. AUC: area under the receiver operating characteristic curve; MCC: Matthews correlation coefficient.*

		OSR validation set (N=23)	***CUH (N=777)***
Class weight not provided	Accuracy	0.70 (0.63 - 0.76)	0.62
	Precision	0.74 (0.67 - 0.81)	0.65
	Specificity	0.09 (0.03 - 0.16)	0.18
	Recall	0.92 (0.87 - 0.97)	0.89
	F1	0.81 (0.77 - 0.86)	0.75
	AUC	0.61 (0.47 - 0.67)	0.63
	MCC	0.02 (0.09 - 0.14)	
Class weight provided	Accuracy	0.61 (0.55-0.68)	0.59
	Precision	0.81 (0.73-0.88)	0.75
	Specificity	0.53 (0.37-0.7)	0.65
	Recall	0.63 (0.53-0.74)	0.56
	F1	0.7 (0.63-0.77)	0.63
	AUC	0.61 (0.52-0.71)	0.64
	MCC	0.14 (0.02-0.27)	

Basically, the model is optimistic, predicting improvement for most patients, including many who won't actually improve and achieving an accuracy that is like the class ratios. This happens because the model is biased towards the majority class.

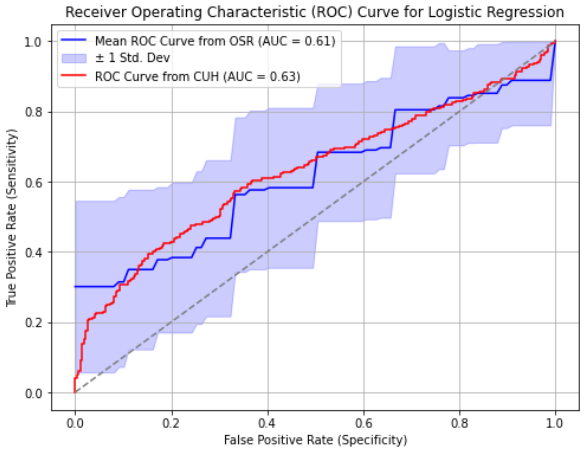


Figure 4.7.1 *– Receiver operator characteristic curves of the logistic regression model across the 10 stratified splits of the OSR dataset (blue, with shaded area for +/- 1 standard deviation) and in the CUH test set (red). Input features: age, sex, HbA1c, creatinine, total cholesterol, LDL, HDL, triglycerides, ALT, BMI, DBP and SBP. Missing management: drop. Validation set size: 20%. Class weight: not provided. AUC: area under the receiver operating characteristic curve.*

By telling the model to balance the weights according to the class distribution (class_weight = 'balanced') we achieved more reasonable sensitivity with a comparable specificity (Table 4.7.1). Of note, the area under the receiver operating characteristic curve (AUC) was not affected. This highlights the importance of evaluating models with a comprehensive set of metrics, without relying solely on the AUC.

4.7.2 Removal of collinear or largely missing variables

For the next set of experiments, we began by calculating the variance inflation factor (VIF) for each of the 12 variables contained in the two datasets, aiming at identifying strongly collinear variables to be managed. This method, available in the statsmodels library, couples the Spearman correlation heatmap concept that we showed in Figure 4.6.3, but following which we still had not removed any feature from the set. As depicted in figure 4.7.2, VIF values were either infinite or extremely high for lipid profile variables, indicating high collinearity among them. One recommendation is that if VIF is greater than 5, then the explanatory variable is highly collinear with the other explanatory variables, and the parameter estimates will have large standard errors because of this. We then tried to remove either LDL or total cholesterol and fixed the collinearity issue. As LDL was the variable with more missing values, we decided to drop it before moving to a new round of training.

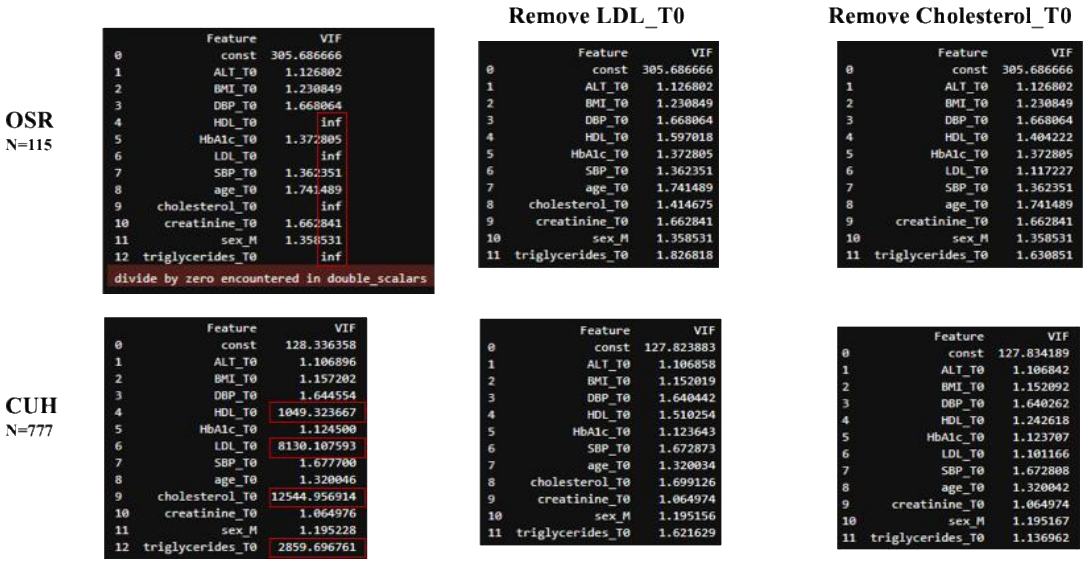


Figure 4.7.2 *– Variance inflation factor (VIF) values of the 12 features in the two datasets. To calculate the VIF we had to drop NaN values. Infinite and very high values of the lipid profile variables indicate very strong collinearity, fixed by either removing LDL (centre images) or total cholesterol (rightmost images).*

With the 11-feature dataset we achieved a major improvement in the metrics, with at least +0.1 in sensitivity, specificity, AUC and MCC. These were even more stable when we then also attempted a 10-feature model by dropping the ALT column, given its 75% missingness ratio in the OSR dataset, as already described in section 4.6. In this case, the resulting dataset rows were: OSR total = 369 split in 290 for train and 79 for validation; CUH test = 932. This led to a narrowing of the performance confidence intervals, indicating increased precision in the estimates. Table 4.7.2 summarises the metrics of the two experiments, while Figure 4.7.3 reports their receiver operator characteristic (ROC) curves.

Table 4.7.2 *– Performance metrics of two logistic regression models across the 10 OSR validation sets and in the CUH test set. Figures are reported as mean (95% c.i.). Input features: age, sex, HbA1c, creatinine, total cholesterol, HDL, triglycerides, ALT (+/-), BMI, DBP, and SBP. Missing management: drop. Validation set size: 20%. Class weight: provided. AUC: area under the receiver operating characteristic curve; MCC: Matthews correlation coefficient.*

		OSR validation set (N=39)	***CUH (N=780)***
LDL-, ALT+	Accuracy	0.65 (0.6-0.7)	0.59
	Precision	0.88 (0.83-0.93)	0.77
	Specificity	0.7 (0.58-0.82)	0.69
	Recall	0.63 (0.56-0.69)	0.54
	F1	0.73 (0.68-0.78)	0.65
	AUC	0.73 (0.67-0.8)	0.63
	MCC	0.28 (0.17-0.39)	

		OSR **validation** *set (N=79)*	*CUH (N=932)*
LDL-, ALT-	Accuracy	0.66 (0.62-0.71)	0.64
	Precision	0.81 (0.76-0.86)	0.75
	Specificity	0.65 (0.56-0.73)	0.57
	Recall	0.67 (0.62-0.72)	0.68
	F1	0.73 (0.69-0.78)	0.71
	AUC	0.7 (0.65-0.74)	0.67
	MCC	0.3 (0.21-0.38)	

By inspecting the ROC curve of the LR trained on OSR without LDL but with ALT, we noticed it does not originate from 0;0 coordinates (Figure 4.7.3A). To explore the reasons for such a behaviour, we plotted the distribution of predicted probabilities by the model and realised that the model never predicted a 0 probability of HbA1c decrease, again confirming an optimistic behaviour of the model, probably pushed by the class imbalance. As a partial solution, we tried a different scaling method from the standard one (StandardScaler), namely the QuantileTransformer, which forces features to follow a uniform distribution and works better when features are not normally distributed, as often happens with healthcare RWD (see code snippet below). In this way, we got instances of probabilities predicted between 0 and 0.2 (Figure 4.7.4).

```
scaler_cleaned = QuantileTransformer(n_quantiles=100, output_distribution='normal')
```

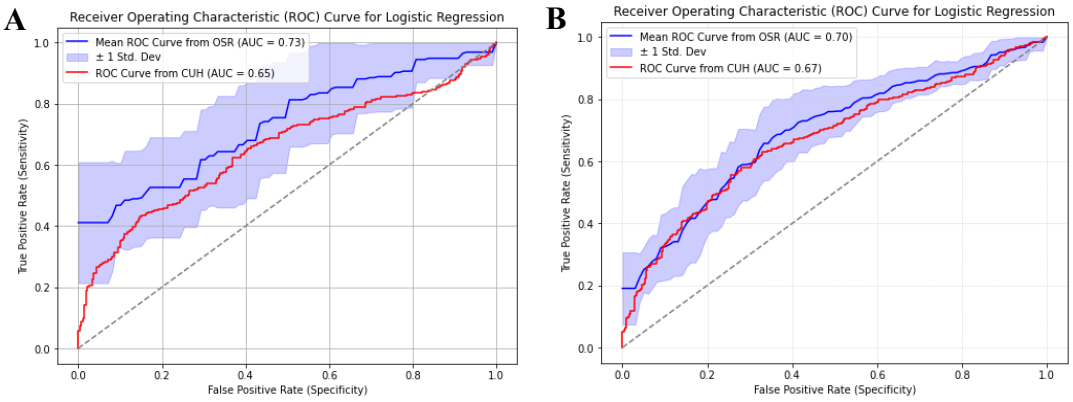


Figure 4.7.3 *– Receiver operator characteristic curves of logistic regression models with varying feature sets, across the 10 OSR validation sets (blue, with shaded area for +/- 1 standard deviation) and in the CUH test set (red). Input features (**A**): age, sex, HbA1c, creatinine, total cholesterol, HDL, triglycerides, ALT, BMI, DBP, and SBP. Input features (**B**): age, sex, HbA1c, creatinine, total cholesterol, HDL, triglycerides, BMI, DBP, and SBP. Missing management: drop. Validation set size: 20%. Class weight: balanced. AUC: area under the receiver operating characteristic curve.*

Additionally, we noticed that as we increased our train and validation size, the ROC curve origin for OSR progressively converged to 0;0 coordinates, possibly due to an increased number of negative instances learned by the model (Figure 4.7.3B).

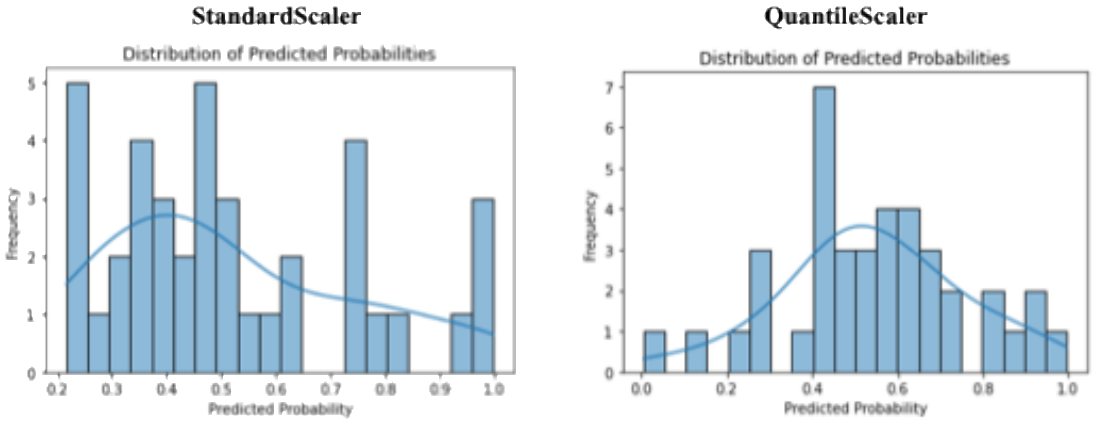


Figure 4.7.4 *– Distribution of predicted probabilities by two logistic regression models in the OSR validation set after two different scaling methods are applied. Input features: age, sex, HbA1c, creatinine, total cholesterol, HDL, triglycerides, ALT, BMI, SBP, and DBP. Missing management: drop. Validation set size: 20%. Class weight: balanced. Scaling: StandardScaler (left); QuantileScaler with 10 quantiles (right).*

4.7.3 Performance impact of split sizes and number of iterations

Figure 4.7.5 shows how by increasing the size of the validation set from 10% to 40% we obtain a progressively smoother ROC curve with a smaller standard deviation, as expected by the availability of more samples to predict. Counterintuitively, on the other hand, the mean AUC values tend to increase, despite the progressive reduction in number of examples left for the model to train on.

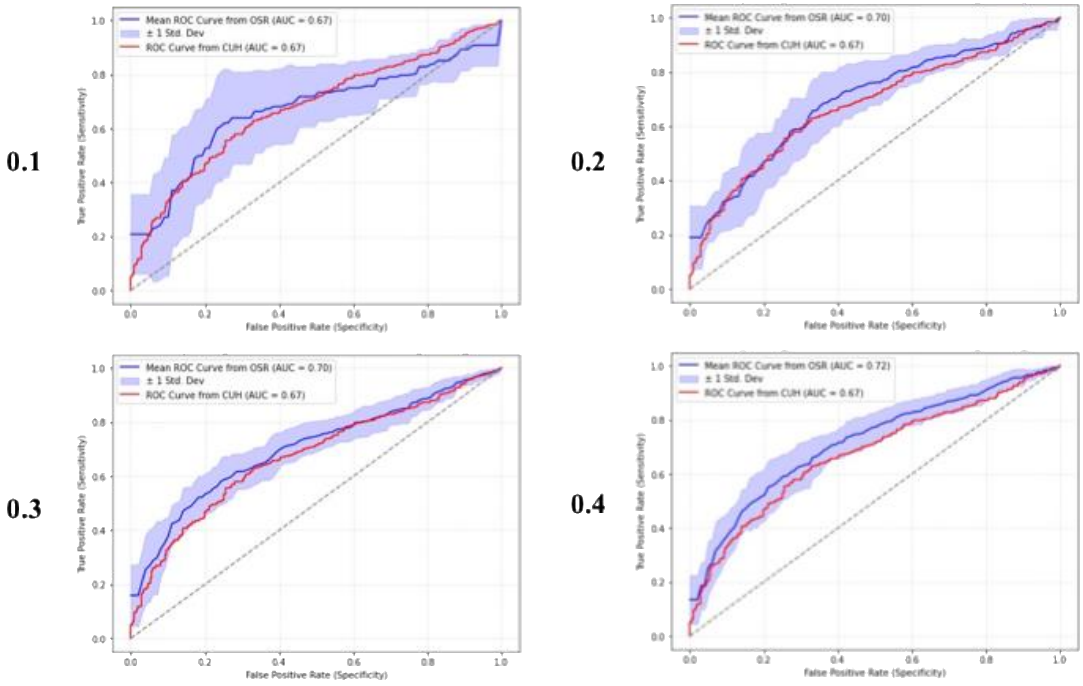


Figure 4.7.5 *– Receiver operator characteristic curves of logistic regression models at varying validation set sizes, across the 10 OSR validation sets (blue, with shaded area for +/- 1 standard deviation) and in the CUH test set (red). Input features: age, sex, HbA1c, creatinine, total cholesterol, HDL, triglycerides, ALT, BMI, SBP, and DBP. Missing management: drop. Validation set sizes are reported on the left side of the corresponding plot. Class weight: balanced. Scaling: QuantileScaler.*

After achieving strong validation performance with a larger validation set, we investigated whether increasing the maximum number of solver iterations, from 100 to 500 or 1000, would further improve model convergence or performance. However, no meaningful differences in the AUC were observed. This suggests that the model had already converged adequately within the initial 100 iterations, and additional iterations provided no further benefit.

4.7.4 Removal of all variables with >30% missingness

As a final experiment in this series, we included as input features only those with less than 30% missing values. We therefore excluded creatinine (48.6% missing), HDL cholesterol (38.9% missing), total cholesterol (32.5% missing) and triglycerides (38.9% missing), in addition to the already excluded LDL cholesterol and ALT. This allowed to get 973 rows in the OSR dataset and 2476 rows in the CUH dataset, with 6 predictive features: sex, age, baseline HbA1c, BMI, DBP and SBP. With this setting, we achieved the best performance so far, as reported in Table 4.7.3 and Figure 4.7.6.

Table 4.7.3 *– Performance metrics of a logistic regression model across the 10 OSR validation sets and in the CUH test set. Figures are reported as mean (95% c.i.). Input features: age, sex, HbA1c, BMI, DBP, and SBP. Missing management: drop. Validation set size: 40%. Class weight: provided. Max number of iterations: 500. AUC: area under the receiver operating characteristic curve; MCC: Matthews correlation coefficient.*

	OSR validation set (N=391)	***CUH (N=2476)***
Accuracy	0.68 (0.66-0.69)	0.65
Precision	0.83 (0.81-0.85)	0.77
Specificity	0.68 (0.64-0.71)	0.6
Recall	0.68 (0.67-0.69)	0.67
F1	0.75 (0.74-0.76)	0.71
AUC	0.74 (0.72-0.75)	0.69
MCC	0.33 (0.29-0.37)	

This set of experiments supports the evidence that models don't learn well if they have to deal with many variables but too few examples and that simpler models but with more data generally perform better. It also confirms the importance of managing collinear variables, commonly present in healthcare datasets, and scale the data to make it fit for the logistic regression learner, as healthcare datasets often have variables with different magnitudes and skewed distributions.

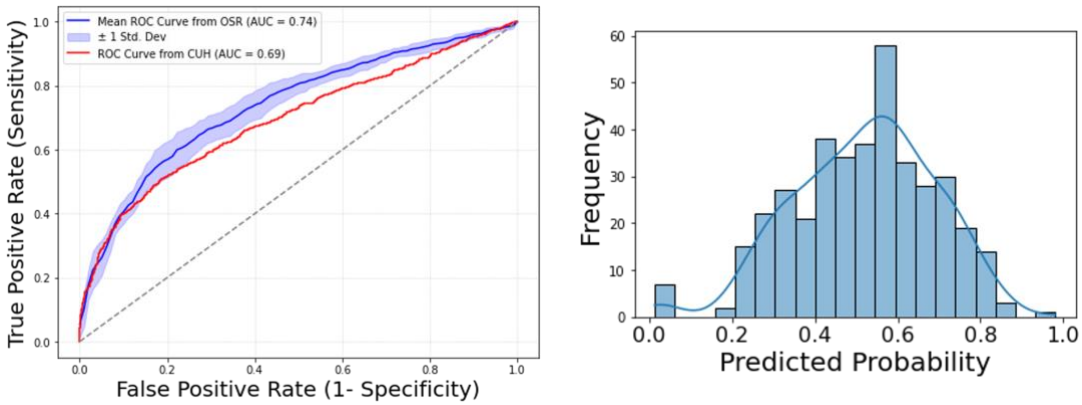


Figure 4.7.6 *– To the left: Receiver operator characteristic curves of a logistic regression model across the 10 OSR validation sets (blue, with shaded area for +/- 1 standard deviation) and in the CUH test set (red). To the right: distribution of predicted probabilities. Input features: age, sex, HbA1c, BMI, SBP, and DBP. Missing management: drop. Validation set size: 40%. Class weight: balanced. Scaling: QuantileScaler. Max number of iterations: 500.*

The code snippet below provides the final settings of the split strategy and of the LR model.

```
sss = StratifiedShuffleSplit(n_splits=10, test_size=0.4, random_state=2)
[…]
model_lr_full = LogisticRegression(max_iter=500, class_weight='balanced', solver='lbfgs')
```

4.8 Logistic Regression Imputing Missing Values

While describing challenges linked to the use of RWD in machine learning research, we mentioned that RWD often have large portions of missing values. In the previous set of experiments, we showed that if we want to use a dataset with no missing values while maintaining the full set of available variables, we often find ourselves dealing with a "large but short" dataset (few rows, many columns), which is unfit for ML. To overcome the problem of data missingness, researchers can either choose models that internally manage missing values, such as tree-based methods, or they can explore the possibility of data imputation techniques. In Chapter 2 we reported examples of both: Nicolucci and colleagues chose an XGBoost model, able to internally manage missingness, while Herrero-Zazo et al imputed the data with specific strategies. In the next set of experiments, we explored the performance of LR models after imputation of missing values for different sets of features. As a pattern of missingness, we assumed our missing values to be missing completely at random, as we attributed their missingness to random events such as the clinician not inputting the values in the EHR or patients doing blood tests at another institution and not handing the results at the visit. As an imputation method we chose the Multivariate Imputation by Chained Equations (MICE) method, a strategy for imputing missing values by modelling each feature with missing values as a function of other features in a round-robin fashion. This method is available in the scikit-learn library as IterativeImputer. We selected median as a strategy to initialize the missing values, left to right as imputation order and the Bayesian ridge regression as an estimator, which is the default (see code snippet below). We set the random state at 1. Scaling was applied after the imputation step.

```
mice_imputer = IterativeImputer(verbose=0, max_iter=70, tol=1e-10,
imputation_order='roman', min_value=0.1, random_state=1, initial_strategy='median')
```

4.8.1 Performance with and without ALT and at different validation sizes

As a first attempt, we tried to impute the full OSR dataset with all 12 variables. However, the estimator was not converging, even when we increased the maximum number of iterations. Removing the LDL variable was enough to allow the MICE to converge at 70 iterations. We trained a LR on the imputed 11-feature dataset and then did

a new round of imputation and training after dropping the ALT variable (75% missingness) to make the imputation easier and more reliable.

Table 4.8.1 *– Performance metrics of two logistic regression models across the imputed 10 OSR validation sets and in the imputed CUH test set. Figures are reported as mean (95% c.i.). Input features: age, sex, HbA1c, creatinine, total cholesterol, HDL, triglycerides, ALT (+/-), BMI, DBP, and SBP. Missing management: multivariate imputation by chained equations. Validation set size: 20%. Class weight: provided. AUC: area under the receiver operating characteristic curve; MCC: Matthews correlation coefficient.*

		OSR validation set (N=282)	***CUH (N=4114)***
LDL-, ALT+	*Accuracy*	0.67 (0.65-0.7)	0.61
	Precision	0.81 (0.78-0.83)	0.74
	Specificity	0.65 (0.6-0.7)	0.61
	Recall	0.68 (0.66-0.7)	0.61
	F1	0.74 (0.72-0.76)	0.67
	AUC	0.72 (0.7-0.75)	0.66
	MCC	0.31 (0.25-0.37)	
		OSR* validation *set (N=282)	***CUH (N*=4114*)***
LDL-, ALT-	*Accuracy*	0.67 (0.66-0.69)	0.61
	Precision	0.82 (0.81-0.84)	0.74
	Specificity	0.7 (0.67-0.73)	0.61
	Recall	0.66 (0.65-0.67)	0.61
	F1	0.73 (0.72-0.74)	0.67
	AUC	0.73 (0.72-0.75)	0.67
	MCC	0.33 (0.3-0.37)	

A few comments can be made on the metrics reported in Table 4.8.1. First, the availability of 282 patients in the validation set makes the c.i. very narrow, demonstrating a very good stability of the models. Second, the performance gain when removing the largely missing ALT variable is minimal. Third, despite the larger training set made available by the MICE, the mean metrics are comparable to those reported in Table 4.7.2. In this set of experiments too, we tried to assess the effect of varying validation set sizes. In this case the performances are overlapping across different conditions but again, interestingly, with more samples available for validation, the origin of the ROC for the 10 OSR splits tends to 0;0 coordinates (Figure 4.8.1).

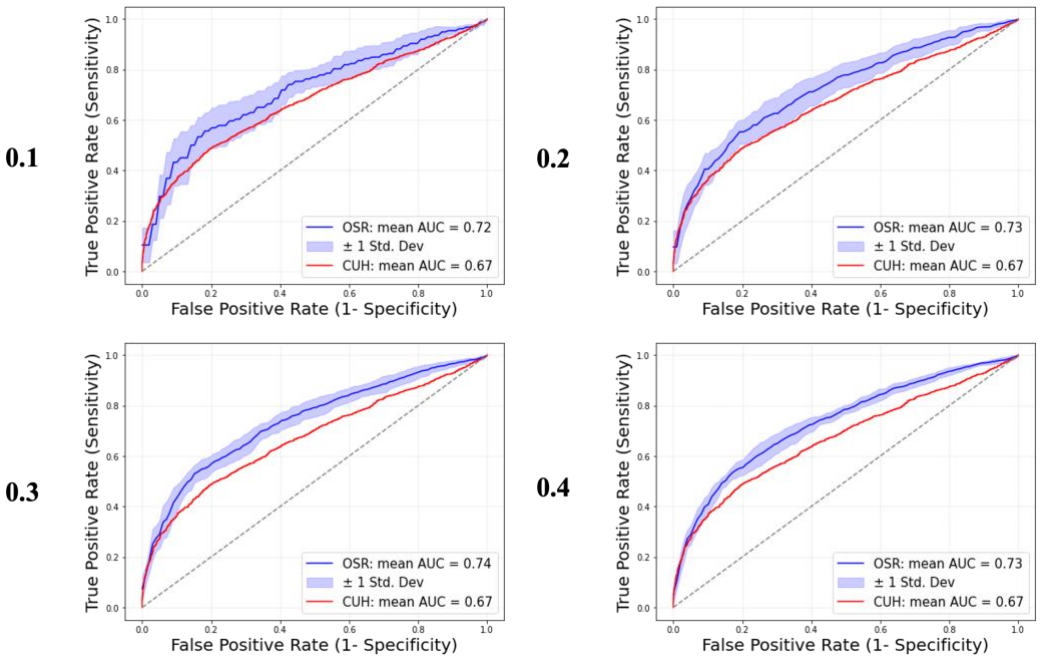


Figure 4.8.1 *– Receiver operator characteristic curves of logistic regression models at varying validation set sizes, across the imputed 10 OSR validation sets (blue, with shaded area for +/- 1 standard deviation) and in the imputed CUH test set (red). Input features: age, sex, HbA1c, creatinine, total cholesterol, HDL, triglycerides, BMI, SBP, and DBP. Missing management: multivariate imputation by chained equations. Validation set sizes are reported on the left side of the corresponding plot. Class weight: balanced. Scaling: QuantileScaler.*

4.8.2 Influence of different scaling techniques on metrics

Having fixed the validation size at 30%, we then wondered how different scaling methods after MICE could influence performance metrics. We compared them in four different settings: Quantile, Robust, Standard scalers and with no scaling. Despite achieving overlapping AUC, the models surprisingly different in terms of sensitivity, with big drops for all but the Quantile transformer which on the other hand performed worse in terms of specificity (Table 4.8.2).

Table 4.8.2 *– Performance metrics of four logistic regression models across the imputed 10 OSR validation sets and in the imputed CUH test set, with varying scaling methods. Figures are reported as mean (95% c.i.). Input features: age, sex, HbA1c, creatinine, total cholesterol, HDL, triglycerides, BMI, DBP, and SBP. Missing management: multivariate imputation by chained equations. Scaling methods: see table. Validation set size: 30%. Class weight: provided. AUC: area under the receiver operating characteristic curve; MCC: Matthews correlation coefficient.*

Scaler		***OSR validation set (N=421)***	***CUH (N=4114)***
Quantile	*Accuracy*	0.67 (0.66-0.69)	0.62
	Precision	0.82 (0.81-0.84)	0.75
	Specificity	0.7 (0.67-0.73)	0.63
	Recall	0.66 (0.65-0.67)	0.61

	F1	0.73 (0.72-0.74)	0.67
	AUC	0.73 (0.72-0.75)	0.67
	MCC	0.33 (0.3-0.37)	
Robust	*Accuracy*	0.64 (0.63-0.66)	0.61
	Precision	0.86 (0.84-0.88)	0.81
	Specificity	0.79 (0.76-0.83)	0.77
	Recall	**0.57 (0.55-0.58)**	**0.52**
	F1	0.68 (0.67-0.7)	0.63
	AUC	0.74 (0.72-0.76)	0.69
	MCC	0.34 (0.3-0.38)	
Standard	*Accuracy*	0.64 (0.63-0.66)	0.61
	Precision	0.86 (0.83-0.88)	0.81
	Specificity	0.79 (0.76-0.83)	0.77
	Recall	**0.57 (0.56-0.59)**	**0.52**
	F1	0.68 (0.67-0.7)	0.63
	AUC	0.74 (0.72-0.76)	0.69
	MCC	0.34 (0.3-0.38)	
No scaling	*Accuracy*	0.64 (0.63-0.66)	0.61
	Precision	0.86 (0.84-0.88)	0.81
	Specificity	0.79 (0.76-0.83)	0.77
	Recall	**0.57 (0.56-0.59)**	**0.52**
	F1	0.69 (0.67-0.7)	0.63
	AUC	0.74 (0.72-0.76)	0.69
	MCC	0.34 (0.31-0.38)	

4.8.3 Performance with minimal feature set and imputation

As in the previous set of experiments, we tried to reduce the feature set to those with less than 30% of missing values. Here, this allowed to use MICE to fill minimal gaps instead of imputing large portions of the dataset, something that physicians could be sceptical about. Quantile scaling was preferred. Despite using less predictive features, the model achieved overlapping performance (AUC on the validation set = 0.74 (0.72-0.75); AUC on the test set = 0.68). To evaluate the calibration of the model's predicted probabilities for HbA1c improvement at 3 years, we constructed calibration curves for both the internal validation set and the external test set (Figure 4.8.2): the model exhibits poor calibration in both sets, with predicted probabilities systematically overestimating the true likelihood of improvement.

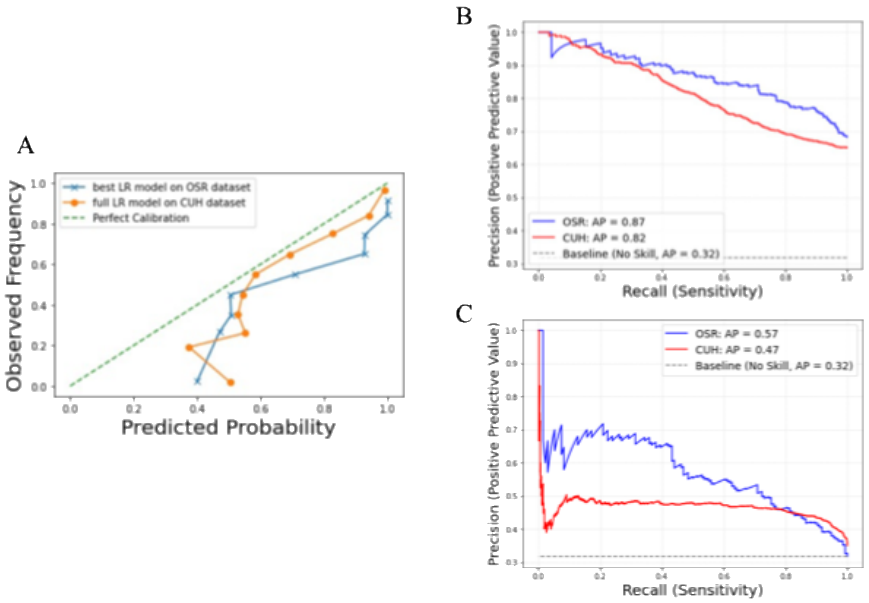
The calibration curves lie below the ideal diagonal line, indicating that the model's predictions are overly optimistic. Notably, no predicted probabilities fall below 0.4 in either set, as evidenced by the curves bending sharply toward zero at this threshold. This

suggests that the model fails to assign low probabilities to any patients, likely due to the class imbalance (67% true in training) and the model's tendency to favour the majority class (improvement). As a result, the model struggles to discriminate between patients with lower probabilities of HbA1c improvement, limiting its clinical utility for identifying cases at high risk of T2DM worsening.

Given the class imbalance in our dataset (68% improvement vs. 32% no improvement in the OSR dataset used for training, see Figure 4.6.2), in this experiment we additionally evaluated model performance using precision-recall (PR) curves. The PR curve is particularly informative for imbalanced data, as it focuses on the minority class ('no improvement') and directly reflects the trade-off between precision and recall.

For the majority class ('improvement'), the model achieved strong performance, with Average Precision (AP) scores of 0.87 (OSR validation set) and 0.82 (CUH test set), indicating high accuracy in identifying patients likely to improve. However, for the minority class ('no improvement'), the model's performance was modest, with AP scores of 0.57 (internal validation) and 0.47 (external test set), compared to a no-skill baseline of 0.32 (reflecting the 31.7% prevalence of 'no improvement' in the training/validation sets).

These results (Figure 4.8.2) highlight a disparity in model performance between the two classes. While the model excels at predicting improvement, its ability to identify patients at risk of HbA1c worsening is limited, as evidenced by the lower AP scores for the minority class. This suggests a need for further optimization, such as class rebalancing, alternative algorithms, or threshold adjustment, to improve sensitivity for high-risk patients.



. ***Figure 4.8.2*** *– **(A)** Calibration curves, **(B)** precision-recall curves for majority and **(C)** minority classes of the logistic regression model on the OSR validation set and CUH test set. Input features: age, sex, HbA1c, creatinine, total cholesterol, HDL, triglycerides, BMI, DBP, and SBP. Missing management: multivariate imputation by chained equations. Scaling method: QuantileTransformer. Validation set size: 30%. Class weight: provided. AP: average precision.*

4.8.3 Feature importance analysis

A crucial step after model training is represented by model explanation. In the case of LR models, given their simplicity, this step is basically represented by the evaluation of coefficients. We evaluated the coefficients of the 6- and 12-variable models trained on the full OSR dataset after MICE and QuantileTransformer scaling.

	Feature	***Coefficient***
6-variable model	*HbA1c_T0*	0.73
	BMI_T0	0.12
	DBP_T0	0.10
	SBP_T0	-0.06
	age_T0	0.03
	sex_M	0.02
	Feature	***Coefficient***
12-variable model	*HbA1c_T0*	0.71
	triglycerides_T0	0.14
	cholesterol_T0	0.09
	DBP_T0	0.08
	HDL_T0	0.07
	creatinine_T0	-0.07
	BMI_T0	0.05
	age_T0	0.03
	sex_M	0.02
	SBP_T0	-0.02

As reported in Table 4.8.3, we observed that baseline HbA1c was the strongest predictor of HbA1c improvement in both models, aligning with clinical expectations that patients with higher baseline values have greater potential for reduction. Other features, including lipid profile, BMI, blood pressure, age, and sex, contributed minimally to predictions (coefficients < 0.14 in magnitude), suggesting limited additive value in this linear model. These findings overlap with those of the exploratory data analysis, where the subgroup of patients achieving the outcome had a higher median HbA1c_T0. Triglycerides_T0 was the other variable with an evident difference in the two subgroups and here it comes in second place in the magnitude of coefficients.

4.9 Conclusion

In this chapter, we demonstrated how two distinct datasets, one derived from two decades of routine clinical activity within our diabetology clinic at OSR and the other curated from the Clarity database mirroring the Epic EHR at CUH, can be leveraged to investigate long-term glycaemic outcomes in patients with T2DM. The construction of these datasets required markedly different processes. For the OSR data, we relied on the S-RACE platform to retrieve and subsequently curate heterogeneous EHR tables, whereas the CUH dataset was assembled in situ via structured SQL queries and pre-analytical exploration within the local research environment. Despite their differing origins, the two datasets jointly enabled a robust, multi-institutional modelling effort.

A considerable portion of the work focused on the rigorous identification of the target cohort. This required extensive inspection of the raw extracts, harmonisation of heterogeneous variables, and the implementation of principled inclusion criteria to ensure analytical consistency. These preparatory steps not only facilitated the subsequent modelling tasks but also revealed the complementary value of RWD in characterising historical trends in clinical practice. By analysing longitudinal trajectories of key biomarkers in the OSR dataset, we provided an internal audit of the evolving management of T2DM within our institution. Such analyses highlighted how EHR-based surveillance can uncover temporal patterns in metabolic control and inform updates in clinical strategies.

Building on these foundations, we conducted a series of machine learning experiments aimed at predicting 3-year HbA1c outcome, a clinically meaningful endpoint given the

centrality of HbA1c in monitoring disease progression and risk of complications. Using logistic regression as an interpretable modelling framework, we systematically evaluated the impact of alternative preprocessing strategies. These included not only the comparison between dropping missing values and imputing them via MICE, but also the selection of missingness thresholds for feature exclusion, the choice of scaling methods, the allocation ratio for the validation set, and the maximum number of iterations permitted during model training. Our results indicated that, once collinearity was addressed and only sufficiently populated variables were retained, model performance was largely comparable between the two pipelines. This finding carries practical implications: in healthcare applications, where imputed values risk introducing artefactual patterns or eroding clinical trust, the marginal gains obtained through sophisticated imputation may not justify their complexity.

Finally, we gained substantial experience in external validation, a crucial but often underreported component of predictive modelling. Aligning the OSR and CUH datasets required a careful reconciliation of variable definitions, completeness thresholds, and clinical context. Differences in local care pathways, such as the fact that many CUH patients are managed primarily by general practitioners, resulting in fewer anthropometric measurements, directly influenced feature availability, data quality, and the interpretation of results. Nonetheless, the models demonstrated stable performance across sites, supporting the generalisability of the approach and underscoring the translational potential of RWD-based prediction in heterogeneous clinical settings.

5. Hospitalization Outcome Prediction

5.1 Introduction

In chapter 2, we discussed the appearance of a new patient phenotype originating from global population ageing. We also highlighted the potential impact of studying RWD in internal medicine, as that is the specialty taking care of such new complex and vulnerable patients for whom poor evidence generated by randomized controlled trials exists. To address this lack of evidence, at our hospital we launched a RWD gathering initiative from general medicine wards in the form of a 10-year prospective observational study named "MED-Cli" ("Prospective Observational Study to Characterize Patients Treated at Internal Medicine Clinics", NCT05780099). In fact, understanding the relevance of doing research on the evolving population of patients admitted to our wards, we wanted to set up a patient registry to generate RWE and answer relevant clinical questions. In this chapter, we describe the study, its data collection pipeline, the real-time dashboards we designed with the available data and the experiments for the prediction of hospitalisation outcomes. The very first fundamental and still open question is how to transform a complex biological entity such as a multimorbid elderly patient into a set of high-fidelity, high-quality and standardised numbers, categories and labels. We tried to answer this question with an extensive, multidimensional and standardised electronic case report form (eCRF). Once data is collected, it should be easily interrogated to enable data-driven medicine. Therefore, we inserted this eCRF in a RWD pipeline to augment it with other sources and couple it with real-time data analysis dashboards borrowed from the manufacturing and finance businesses. It is in those fields that the Gartner Analytic Continuum was conceived, a data exploitation maturity ladder where the steps beyond descriptive (what happened?) and diagnostic (why did it happen?) analytics are predictive (what will happen) and prescriptive (how can we make it happen?) analytics. Therefore, we then manipulated the RWD collected in the MED-Cli study to make it ML ready and allow us to investigate on hospitalisation outcome prediction models.

5.2 Paradigm Shift to Multidimensionality

5.2.1 Prognostic uncertainty

As introduced in Chapter 2, the global demographic shift is causing a profound transformation of the profile of patients admitted for acute conditions to internal medicine wards (Naik *et al*, 2024). What we are witnessing is an increase in the proportion of multimorbid and frail patients who often exhibit reduced physiological reserves, making them highly susceptible to adverse outcomes, even with maximized medical interventions (Ceriani *et al*, 2024; Colacci *et al*, 2025). Care for these patients is complex, often involving uncertain prognoses and difficult decisions. Clinicians must weigh the risks and benefits of initiating potentially non-beneficial treatments, interventions that may prolong suffering without improving survival or quality of life. Cultural, religious, and emotional factors further complicate shared decision-making with patients and caregivers (Lo *et al*, 2022). Uncertainty is a major contributor to conflicts between clinical teams and families, frequently resulting in the continuation of treatments perceived as inappropriate. A 2019 survey about futile or potentially inappropriate care (futile/PIC) shows that 91.3% of clinicians continue treatments despite recognizing their futility, primarily due to disagreement with patients or families (61%) (Chamberlin *et al*, 2019).

One key reason for this uncertainty is the lack of validated tools to predict short-term disease trajectories in complex, multimorbid patients. Unlike oncology, where trial-derived models guide decisions and palliative care referral, internal medicine lacks robust, generalizable predictive instruments. Existing scores often focus on one-year mortality, too long for acute decision-making, and are usually developed in disease-specific or administrative datasets with limited relevance to real-world inpatient populations. Several factors (e.g., age, malnutrition, multimorbidity, acute kidney injury, and high dependency) have been identified as being associated with higher in-hospital and post-discharge mortality, but their validation and large-scale implementation are still lacking (Lenti *et al*, 2023). Additionally, standardized algorithms capable of predicting adverse outcomes have not been implemented in clinical practice. Such algorithms, stemming from the analysis of RWD with emerging techniques such as Machine Learning (ML), would help clinicians in identifying the actual needs of patients, avoiding unnecessary treatment escalation and enabling the appropriate allocation of resources.

5.2.2 The Frailty Index

As a potential solution to the prognostic uncertainty, over the past two decades the concept of frailty has emerged as a central framework for understanding and managing the clinical complexity of older patients admitted to Internal Medicine wards. Unlike traditional comorbidity indices, which merely quantify the burden of chronic diseases, the frailty approach adopts a multidimensional and dynamic perspective, capturing the progressive accumulation of physical, functional, cognitive, and social deficits (Clegg *et al*, 2013; Kim & Rockwood, 2024). Thanks to this multidimensionality, the concept of frailty can overcome the limitations of tools focused exclusively on disease diagnosis, offering instead a more holistic representation of the older patient. It is now widely recognized that comorbidity alone is insufficient to account for the substantial interindividual variability in clinical outcomes observed in advanced age (Tinetti & Fried, 2004).

Frailty is a clinical syndrome characterized by reduced physiological reserve and impaired homeostatic capacity, which render individuals more vulnerable to even minor stressors. This condition is associated with an increased risk of falls, disability, recurrent hospitalizations, and mortality (Fried *et al*, 2004). One widely adopted international definition is the Fried frailty, which conceptualizes frailty as a phenotype based on five clinical criteria: unintentional weight loss, exhaustion, muscle weakness, slow gait speed, and low physical activity (Fried *et al*, 2001, 2021).

An alternative perspective is provided by Rockwood's deficit accumulation model, which conceptualizes frailty as the cumulative burden of impairments across functional, cognitive, and social domains, leading to the construction of a Frailty Index (FI) (Searle *et al*, 2008; Theou *et al*, 2023). This framework enables a quantitative and continuous assessment of frailty, making it highly valuable both in clinical practice and epidemiological research. The FI can be retrospectively applied to datasets if you can build at least 30 items. One famous Italian dataset of hospitalized older adults is the REPOSI (*Registro POliterapie SImi*), the result of a data collection effort launched by the Italian Society of Internal Medicine (SIMI) and the Pharmacological Research Institute "Mario Negri", which completed its first pilot phase in 2008 (1332 patients from 38 hospital wards). The study group subsequently applied the FI to the REPOSI dataset, and showed an association with overall mortality (HR 1.417, 95%CI 1.316–1.525), with

Kaplan-Meyer curves for mortality according to FI tertiles showing a clear trend towards decreased survival at the 33rd and 67th percentiles (Cesari *et al*, 2018).

Frailty, operationalized through the FI, serves as a more sensitive and earlier marker of clinical vulnerability, enabling the identification of at-risk individuals before clinically significant events occur. Moreover, the value of the FI lies in its broad clinical applicability: it can be used 1) at hospital admission as a screening tool or to avoid attributing symptoms of an acute illness to frailty, 2) as a decision-support instrument in therapeutic planning, 3) as a criterion for designing personalized care pathways, and 4) as a tool that can reveal an older person's health trajectory if used in annual review. The systematic integration of the FI into hospital information systems could represent a meaningful advancement in the organization of care for the geriatric population, ultimately improving both efficiency and quality of healthcare delivery. For example, frailty measures could be used to appropriately target the Comprehensive Geriatric Assessment (CGA), an intervention delivered either by a caring team or a specialized physician that as well represents a holistic approach considering chronic diseases, drugs, and socioeconomical variables, but which is often a limited resource (Kim & Rockwood, 2024). However, there is the need to gather further evidence to support the use of frailty index in routine care or screening and how to manage each stage of the fit-to-frail evolution.

5.3 Real-World Data Pipeline

5.3.1 Production: the MED-Cli study

Any patient admitted to three internal medicine wards at IRCCS San Raffaele Scientific Institute is offered to sign an informed consent for enrolment in the MED-Cli study. Upon signature, we are allowed to collect clinical data of that admission and any other future clinical event of that specific individual. The study started in the second half of 2022, but data gathering by healthcare staff became fully operational only in early 2024.

Data reflects routine clinical practice, with no additional test or evaluation performed on purpose. It is basically a "continuous recording" of what is going on in the study wards, of course limited to those who signed the consent, with an approximate coverage of 75% of admission episodes. Consultants, residents and research staff take care of data entry,

and this ensures a domain-informed data collection. Data quality is maximised through a series of strategies: 1) a written, constantly refined and updated guideline; 2) periodic coordination meetings to align definitions and collection strategies; 3) recurrent data-entry training sessions to reduce inter-operator variability; 4) retrospective data quality checks performed by data managers.

5.3.2 Collection: the Cohort Genomic Platform

Data is collected through a very large and comprehensive eCRF, hosted on an institutional platform built by the Center for Omics Sciences at OSR. The platform is called Cohort Genomic Platform (CGP) and can be accessed only through institutional email and password upon permission by its administrators. Among the most important features are:

- Live, multi-layered, pseudonymizations which allow users with different privileges (i.e., clinicians versus users that must not have access to sensitive information) to visualize different IDs for the same patients, thus reducing the risk of potential identity associations. This makes CGP GDPR-compliant.
- Integration with hospital data sources: APIs were designed to capture personal data, and the lab test results for the patients included in each cohort.
- Use of SurveyJS as a rendering engine for eCRFs (in CGP called collectors) thanks to which extremely powerful and highly customizable forms (collectors) can be created (Figure 5.3.1, top). Moreover, being JSON-based, they are also easily editable once data collection has begun and suitable for the collection of longitudinal data whose cardinality cannot be predicted in advance. The data are collected in a structured manner and hooked into standard dictionaries (e.g., ICD9, ICD10, AJCC staging...) that reduce the possibility of data entry errors and allow comparison of results with data from the scientific community. Fields include dynamic logic and input constraints to minimize entry errors.
- Versatile data export using an interface that allows capillary selection of fields, transformation of raw data into an entity-relationship representation, various export formats, creation of views that can be shared with other users, and the ability to create a MySQL database on the fly based on what is displayed (Figure 5.3.1, bottom).

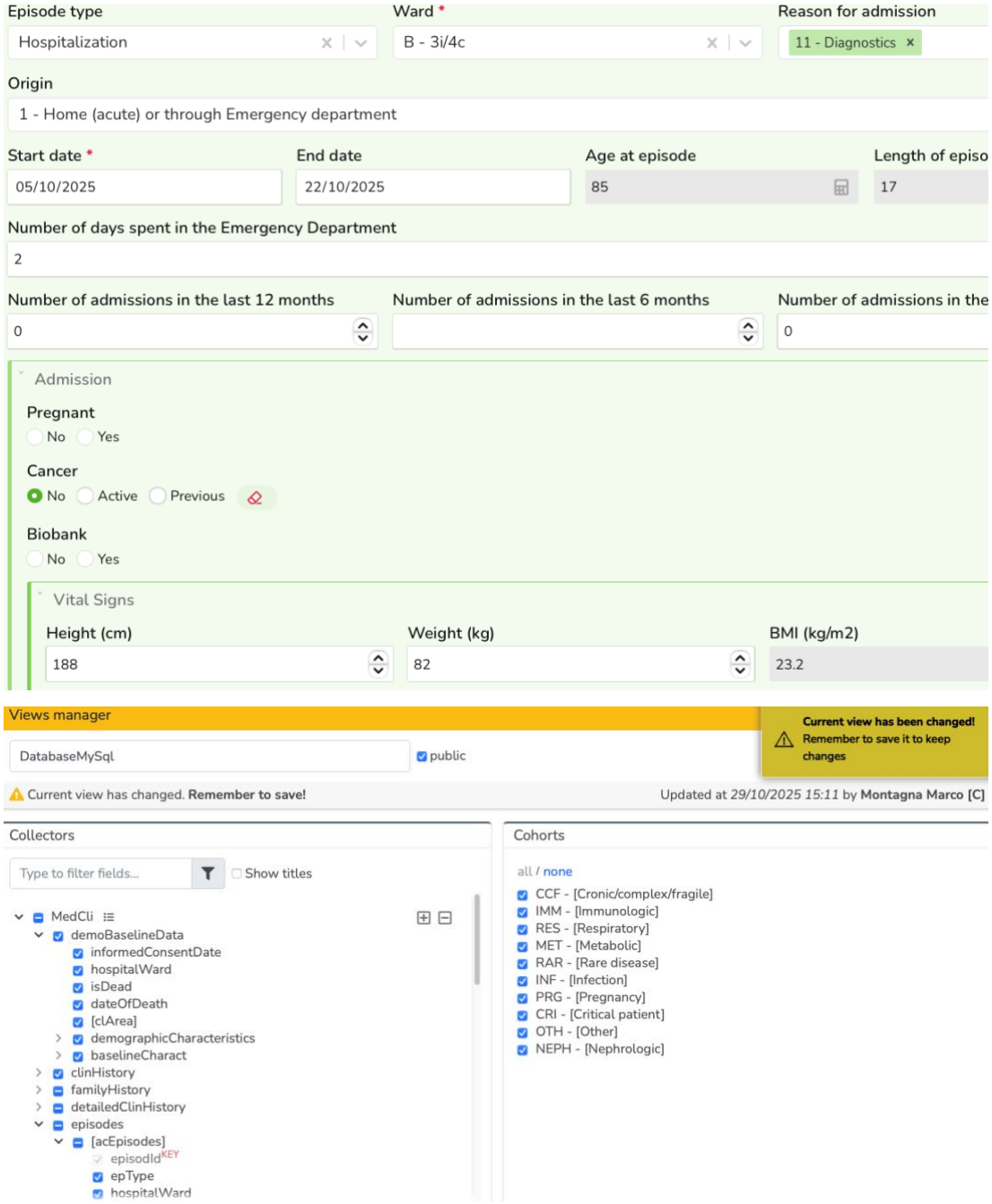


Figure 5.3.1 *– **Top**: Screenshot of one tab of the electronic case report form for the MED-Cli study hosted on the Cohort Genomic Platform. Exemplary types of fields are shown: number, date, drop-down list, calculated… As an example of an internal check: fields marked with the asterisks prevent from saving the tab if not filled. **Bottom**: Screenshot of the data export interface. "Data views" can be created by selecting the desired fields, saved, updated and shared. The tool also provides a preview of the resulting dataset. Finally, the dataset can be exported in .csv, .xlsx or MySQL, in which case the connection settings (host, port, database, username, password) to the database are provided.*

5.3.3 Integration: the data journey

Figure 5.3.2 provides a schematic representation of the multi-step pipeline which allows for the collection, integration and, finally, analysis of RWD generated by the everyday interaction between patients and practitioners. As detailed earlier, CGP features an in-built data export interface, which we use to generate a MySQL database of all the available fields and participants. The resulting database has 86 tables. We then query it with SQLAlchemy in python to load the relevant tables for the aims of each specific study.

```
# Create a MySQL connection using SQLAlchemy
engine =
create_engine('mysql+mysqlconnector://username:password@host:port/name_of_database')

# List tables to load
tables = ['main', 'chronicDiseases', 'acEpisodes', 'acEpisodes_admissReason',
'acEpisodes_disch_ICD10', 'therapies', 'hemaExames', 'fluidCultures',
'fluidCultures_bcSpecies', 'fluidCultures_urSpecies', 'acEpisodes_admc_diagnostics']
dfs = []  # list to hold dataframes

# For each table in the database, load all records
for table in tables:
    query = f"SELECT * FROM {table}"
    df = pd.read_sql(query, con=engine)
    dfs.append(df)  # append the dataframe to the list
```

We linked the same database to a MS PowerBI model to produce interactive dashboards for live data analysis. The database can be periodically refreshed to include newer patients or updated data.

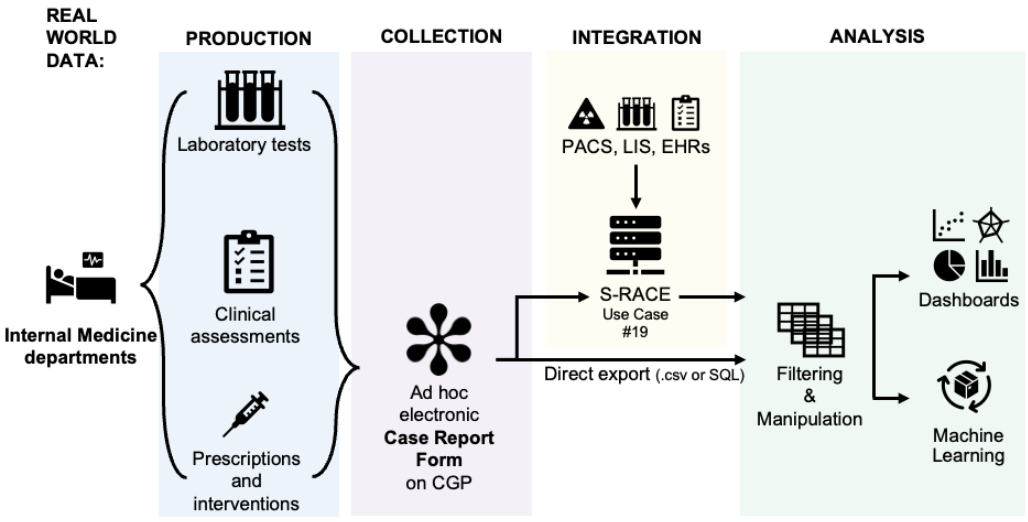


Figure 5.3.2 *– Schematic representation of pipeline from production to analysis for Real World Data (RWD) within the MED-Cli study. RWD generated during routine care in internal medicine wards include clinical assessments, prescriptions, laboratory results, and administrative records. These data are collected via a dedicated electronic case report form*

hosted on CGP, exported in structured formats, and analysed using machine learning models and interactive dashboards for clinical decision support and research. Prior to analysis, CGP exports can be uploaded onto S-RACE for further expansion via integration with data coming from imaging, laboratory or EHR systems. CGP: Cohort Genomic Platform; LIS: Laboratory Information System; EHRs: Electronic Health Records; PACS: Picture Archiving and Communication System; S-RACE: San Raffaele Ai CEntre.

Additionally, we can uploaded exported data onto the S-RACE platform for expansion and integration with data originating from silos in the hospital. Once on the platform, RWD can also be leveraged with machine learning.

5.3.4 Analysis: dashboards for data-driven medicine

Business intelligence (BI) is a set of technologies and methodologies used by enterprises to enhance their decision making and gain a competitive advantage by enabling a data-driven environment (Dedić & Stanier, 2016). As we previously stressed, decision making in medicine has never been so complex, but we now have large amounts of data that could empower physicians with data-driven medicine. Therefore, we tested as a proof of concept the Microsoft PowerBI environment leveraging the MySQL database to generate a set of interactive dashboards. These analytic tools allow to quickly explore the content of the database and gain insights on the RWD generated by the MED-Cli study. As an example, we used the dashboards to show the number of enrolled patients per month, the distribution of the length of admission events, the average number of days spent in the emergency department, the most prevalent chronic diseases in our population and more (Figure 5.2.3). Insights as this can be used both by hospital administrators and by practitioners to take decisions based on continuously updated RWE generated directly by local wards.

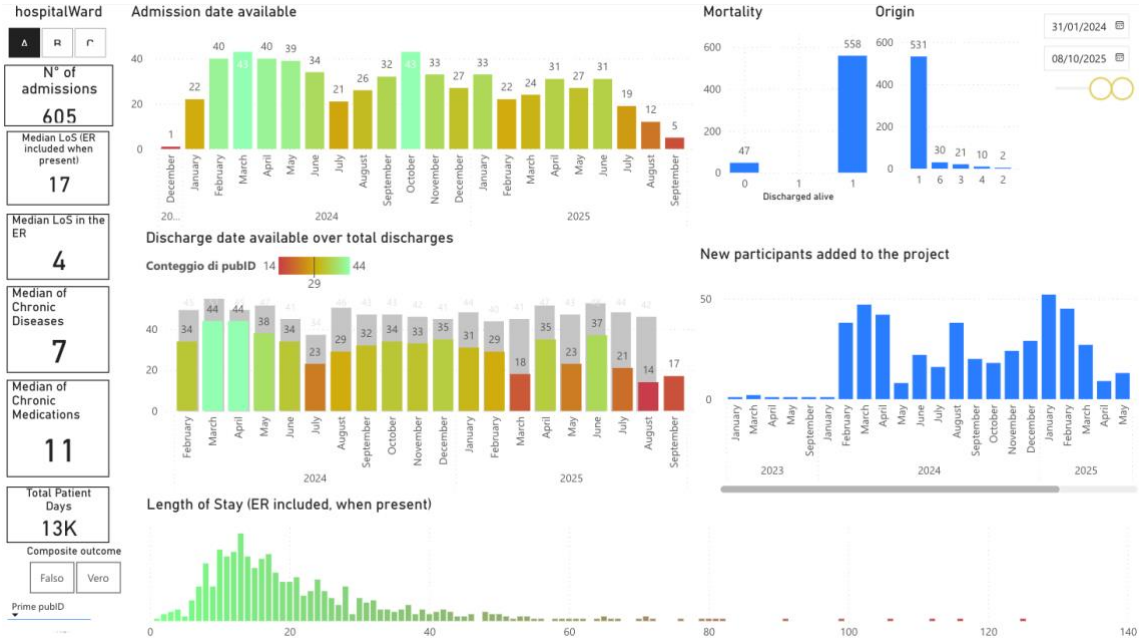


Figure 5.3.3 *– Screenshot of one of the dashboards created in Microsoft PowerBI. Bar plots can be used to explore distribution of continuous variables, count categories or dates. Boxes can be used to report statistics such as median number of chronic diseases or medications. Filters can be applied to slice the data in any desired way. Displayed data are contained in the MySQL database linked to the data model behind the dashboards. When the MySQL database is updated, so are the dashboards, allowing for live monitoring of the variables of interest.*

5.4 Comparison of Frailty and Comorbidity for Risk Stratification

Despite the demographic evolution and the development of the frailty framework, most European and other high-income healthcare systems remain largely organized around disease-specific protocols, which often prove inadequate when applied to frail patients whose needs extend beyond traditional diagnostic categories (Tinetti *et al*, 2019; Yu *et al*, 2023). Given these gaps, there is a pressing need for prognostic tools applicable at hospital admission that are not centred around the past medical history of patients. In this section we describe how we developed and evaluated an admission-based, comorbidity-independent FI in a cohort of Internal Medicine patients, comparing its prognostic performance with the Charlson comorbidity index (CCI), a well-validated score for predicting hospitalization outcomes (Frenkel *et al*, 2014; Charlson *et al*, 2022).

5.4.1 Cohort selection

For this study, we analysed patients managed in one of the Internal Medicine departments involved in the MED-Cli project: the Unit of General Medicine with a focus on Metabolic and Ageing Medicine. We focused on a 12-month time window between

February 1, 2024, and January 31, 2025. From internal registries, we knew that 506 patients were discharged in that period. Among them, 395 provided written informed consent (78%). We therefore queried CGP's MySQL database and filtered admission episodes (AE) of the ward of interest (the eCRF has a specific field to record the managing ward) ending in the selected time window. We found that 16 patients were admitted twice during the study period, yielding a total of 411 AE.

5.4.2 Composite hospitalisation outcome definition

As mentioned in Chapter 2, the therapeutic goals in this population are multifaceted, being the result of the complex interplay between the expectation of patients, caregivers and practitioners, that must also be confronted to the acute disease presentation in the context of patients' individual fitness. Defining a widely accepted "adverse outcome" of an AE is therefore challenging. In this study, we defined the hospital stay outcome as unfavourable according to a composite endpoint comprising: 1) in-hospital mortality after a length of stay exceeding the median of the study population (14 days); 2) early mortality (<8 days) in patients who had been hospitalized at least once in the previous 6 months; 3) failure to be discharged home (i.e., discharge to palliative care or to a lower-intensity care facility); and 4) discharge alive after a length of stay in the ward exceeding the mean of the study population (18 days). This definition tries to capture: 1) "meaningful deaths", excluding those happening too early in the AE, possibly as an effect of the acute presentation, unless the patient is a returning to the hospital after a recent admission; 2) AE that were not enough to restore a patient's function; and 3) very long hospital stays, that cannot be explained by logistics problems but rather reflect poor patient's response to our management.

As reported in Figure 5.4.1, of 411 analysed episodes, 186 (45%) met the criteria for the unfavourable composite outcome. In most cases (59%), this was driven by a length of stay in the ward longer than the mean. In 31% of cases, the outcome was due to discharge to a destination other than home, and in 10% it was the result of in-hospital death.

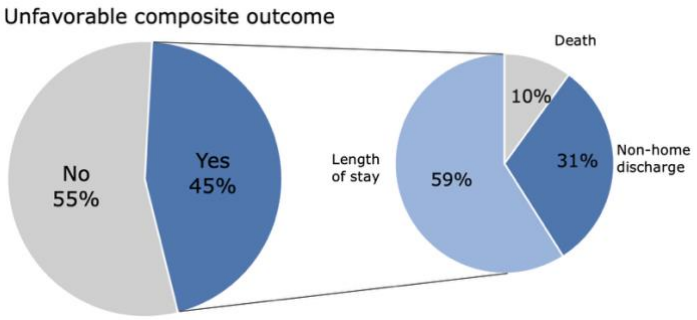


Figure 5.4.1 *– Proportion and composition of the composite negative outcome. Proportion of hospital episodes meeting the unfavourable composite outcome (Yes) or not (No) and relative contribution of each component.*

5.4.3 Construction and distribution of frailty indices

Variables covering distinct domains of patients' health were selected to construct two frailty indices according to the method proposed by Kenneth Rockwood and colleagues (see 5.2.2). Both indices contained information available at the arrival to the hospital, including a panel of 10 admission blood tests and 5 vitals, and the 10 items of the Barthel index, used to assess patients' ability to perform the Activities of Daily Living referring to the state of patients at home before the presenting acute illness (Mahoney & Barthel, 1965). The 'comorbidity FI', 41 variables, additionally incorporated patients' past medical history, whereas the 'admission-only FI', 29 variables, did not. In both indices, the cause of hospitalization was excluded. Table 5.4.1 presents the variables included in the indices along with the corresponding prevalence of deficits. Episodes with more than 20% missing variables (124, 30%) required for the calculation of the index were excluded from the analysis. Correlation between the selected variables was assessed with Spearman's correlation matrix (Figure 5.4.2A).

Table 5.4.1 *– List of variables included in the calculation of the frailty indices and their corresponding deficit prevalence. The table reports the variables used to construct the two frailty indices, with and without information on patients' past medical history. For each variable, the prevalence of the deficit is shown. COPD: Chronic Obstructive Pulmonary Disease; SBP: systolic blood pressure; DBP: diastolic blood pressure; SpO2: peripheral blood oxygen saturation; WBC: white blood cells; AST: aspartate amino-transferase; ALT: alanine amino-transferase; LDH: lactic dehydrogenase. *not used in the calculation of the admission-only FI.*

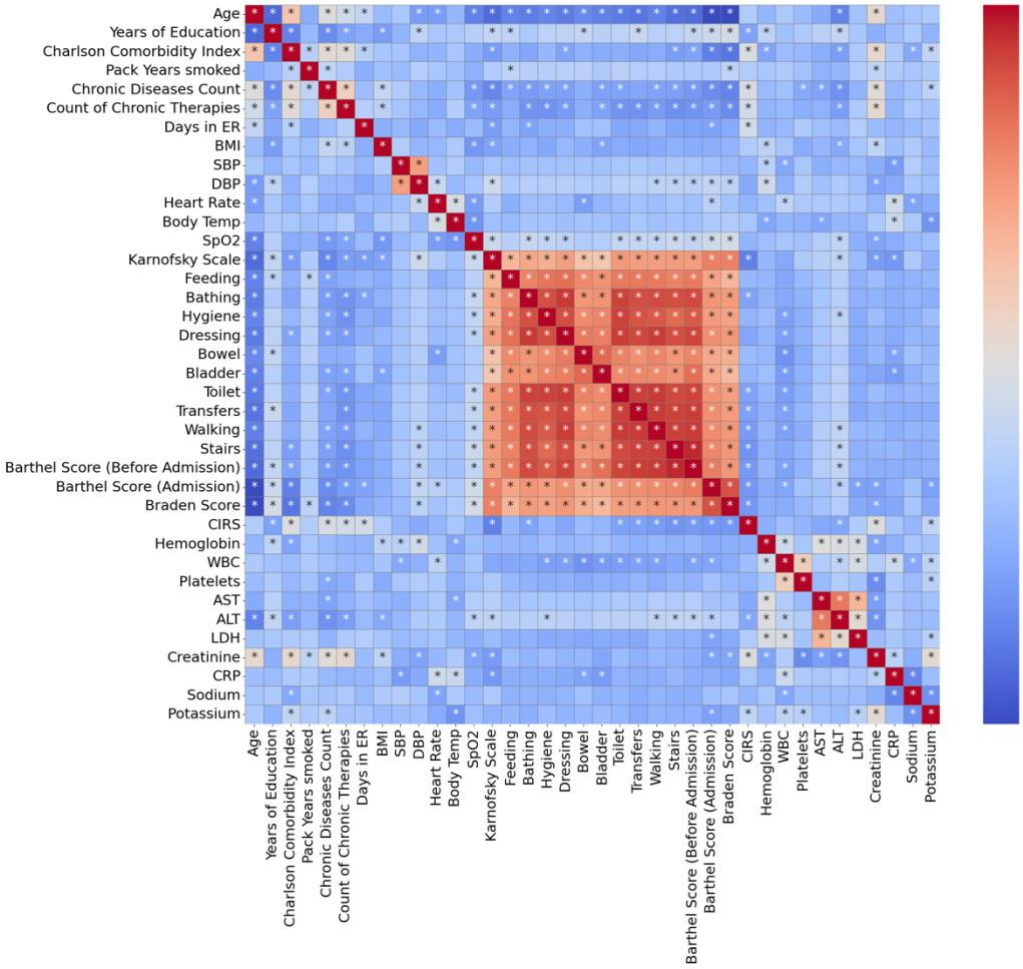
Variable	***% with deficit***	***Variable***	***% with deficit***
Past Medical History*		***Laboratory values***	
Hypertension	50.9	*Haemoglobin*	68.1
Myocardial Infarction	10.9	*White Blood Cells*	53.0
Congestive Heart Failure	15.4	*Platelets*	37.3

Cerebrovascular event	10.5
Peripheral Vascular Disease	9.8
Diabetes	27.2
Dementia	7.7
Chronic Kidney Disease	20.3
Liver Disease	8.4
COPD	16.1
Cancer	38.8
Polypharmacy	75.1
Vitals and baseline info	
Body Mass Index	55.5
Systolic Blood Pressure	39.2
Diastolic Blood Pressure	27.8
Heart Rate	24.5
Dependency	28.2
Home Assistance	23.3
Body Temperature	16.5
SpO2	20.0
Karnofsky Scale	88.6
C-reactive Protein	83.3
Creatinine	34.2
Sodium	28.4
Potassium	21.5
AST	17.4
ALT	11.7
LDH	24.9
Barthel score items	
Feeding	18.5
Bathing	28.2
Hygiene	24.0
Dressing	28.0
Bowel	16.4
Bladder	22.6
Toilet	28.0
Transfers	30.0
Walking	30.4
Stairs	35.0

No variables were strongly correlated with each other (Spearman's $r > 0.95$), indicating that each provided distinct information, avoiding redundancy and supporting the stability and reliability of the statistical and machine learning analyses. The exploration of the dataset through the correlation matrix allows to get interesting preliminary insights on the behaviour of variables, to be linked to what is expected by the domain knowledge. For example, we can see how with advancing age the Charlson comorbidity index increases, as creatinine does (possibly an expression of growing prevalence of chronic kidney disease with age). On the other hand, the blue squares of performance (Karnofsky), functional (Barthel) and pressure ulcer (Braden) scores when correlated to age, signify how much elderly patients have a reduced performance, a reduced independence in the activities of daily living and an increased risk of pressure ulcers. As expected, serum creatinine is positively correlated to serum potassium, and transaminases and lactic dehydrogenase are as well all positively correlated.

As shown in 5.4.2B, the distributions of the two indices were similar, suggesting that the inclusion of medical history did not substantially influence frailty as assessed by our FIs.

A



B

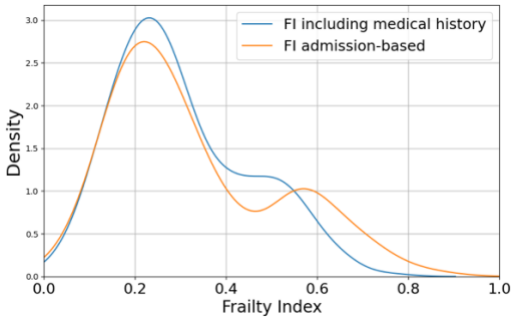


Figure 5.4.2 *– Spearman correlation matrix of numerical variables and density plots of the Frailty Index. (**A**) Demographic, clinical, functional, and laboratory variables used for frailty index FI construction were tested for correlation using the non-parametric Spearman coefficient, appropriate for non-normally distributed variables. Significance levels are indicated in each cell (*p < 0.05). Variables differing between the "comorbidity FI" and "admission-only FI" are highlighted. (**B**) Density plots showing the distribution overlap between "comorbidity FI" (blue) and "admission-only FI" (orange). BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; SpO_2, peripheral oxygen saturation; WBC: white blood cells; AST: aspartate amino-transferase; ALT: alanine amino-transferase; LDH: lactic dehydrogenase.*

Having verified an almost complete overlap between the two frailty indices, we continued our analysis focusing on the one calculated without considering the

comorbidities ('admission-only FI'). We used it to classify patients into predefined frailty categories: 0–0.09 fit, 0.10–0.19 pre-frail, 0.20–0.54 frail, and 0.55–1.00 end-stage frail (Kim & Rockwood, 2024). Overall, 3.1% of patients were classified as fit, 23.0% as pre-frail, 54.7% as frail, and 19.2% as end-stage frail. Interestingly, with advancing age, the proportion of frail and end-stage frail increased (Figure 5.4.3A), and the overall distribution in the 4 classes changed if we compared patients with and without the outcome (Figure 5.4.3B).

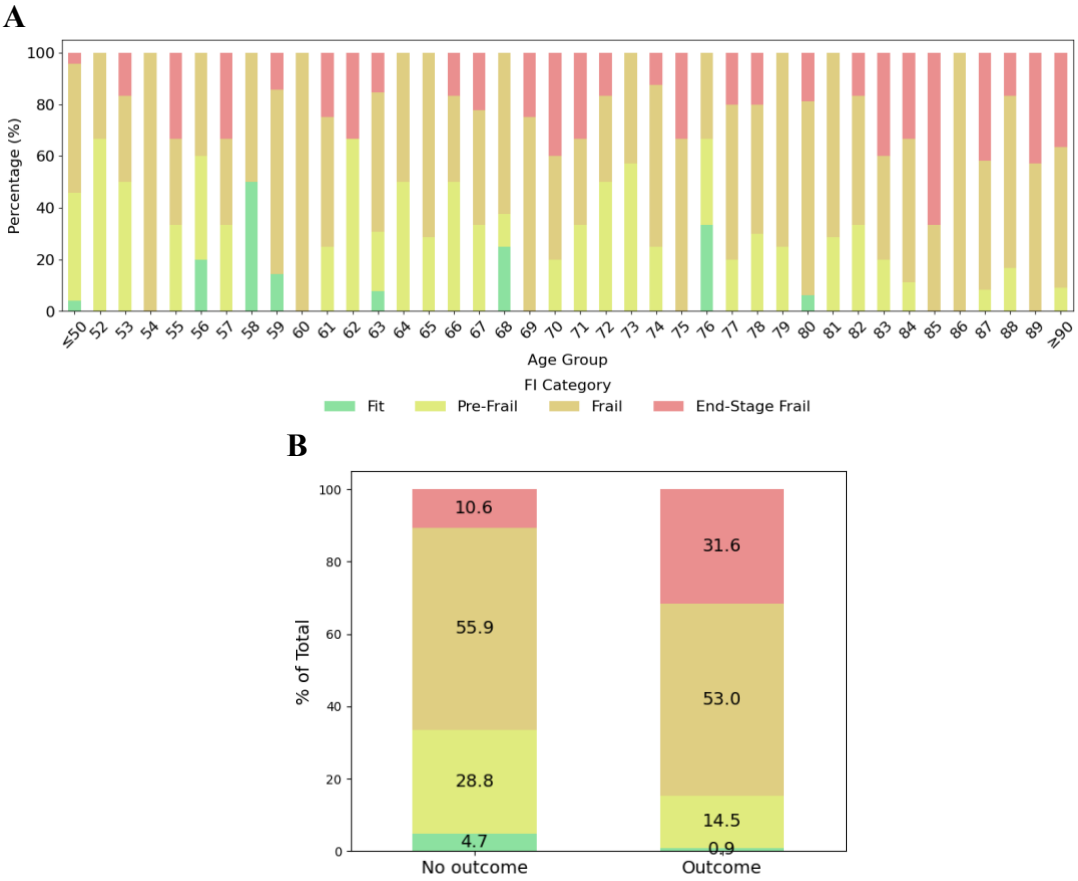


Figure 5.4.3 *– Proportion of patients in the four frailty categories by age (**A**) and by outcome (**B**). Fit = 0–0.09, pre-frail = 0.10–0.19, frail = 0.20–0.54, and end-stage frail = 0.55–1.00. Patients aged up to 50 and from 90 onwards are grouped in overflow bins.*

5.4.4 Statistical comparison by composite outcome

Table 5.4.2 provides a comparative analysis of baseline characteristics between the groups of patients meeting or not meeting the composite negative outcome. No statistically significant differences were observed in terms of sex, BMI, number of chronic diseases or vital parameters at ER admission, while the median age of patients meeting the negative outcome was significantly higher. Among laboratory values only CRP was significantly higher in patients meeting the negative outcome. These patients

had a longer stay in the ER and a higher number of admissions in the previous 6 and 12 months. Functional scores (Barthel score before admission, Barthel score at admission, eastern cooperative oncology group performance status [ECOG], Karnofsky performance status, Braden score, cumulative illness rating scale [CIRS]) were all significantly worse in patients meeting the negative outcome.

Both FIs versions showed a strong, graded association with the unfavourable composite outcome. Median values for patients experiencing the outcome were 0.34 (IQR 0.24–0.57) for the admission-only FI and 0.33 (IQR 0.23–0.47) for the FI including medical history, compared with 0.24 (IQR 0.17–0.34) and 0.25 (IQR 0.18–0.34), respectively, in those without the outcome ($p < 0.001$ for both). The magnitude of association and discriminative performance were similar, indicating that the exclusion of past medical history did not meaningfully reduce the FI's prognostic ability. Clinically, this suggests that frailty can be reliably assessed at the point of admission, using only information immediately available, while still retaining strong predictive value for adverse hospital outcomes.

Table 5.4.2 *– Baseline characteristics according to composite outcome occurrence. Sociodemographic, anthropometric, functional, and laboratory variables for the overall study population and by composite outcome status, highlighting significant differences. Data are shown as mean (SD) or median (Q1-Q3) as appropriate. CCI: Charlson comorbidity index; FI: frailty index; ED: emergency department; BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; SpO_2, peripheral oxygen saturation; GCS: Glasgow coma scale; ECOG: eastern cooperative oncology group performance status; CIRS: cumulative illness rating scale; WBC: white blood cells; AST: aspartate amino-transferase; ALT: alanine amino-transferase; LDH: lactic dehydrogenase; CRP: C-reactive protein.*

Characteristic	***Overall (N=411)***	***No outcome (N=225)***	***Outcome (N=186)***	***p***
CCI	6.00 (4.00–8.00)	6.00 (4.00–8.00)	7.00 (4.00–9.00)	< 0.001
Normalised CCI	0.16 (0.11–0.22)	0.16 (0.11–0.22)	0.19 (0.11–0.24)	< 0.001
admission-only FI	0.28 (0.19–0.45)	0.24 (0.17–0.34)	0.34 (0.24–0.57)	< 0.001
comorbidity FI	0.27 (0.20–0.40)	0.25 (0.18–0.34)	0.33 (0.23-0.47)	< 0.001
Sociodemographic				
Age	76.00 (65.00–84.00)	73.00 (63.00–82.00)	78.00 (67.00–85.00)	< 0.01
Female gender	46%	41%	59%	0.11
Years of Education	10.00 (5.75–13.00)	10.00 (6.50–13.00)	10.00 (5.75–13.00)	0.89
Pack Years smoked	0.00 (0.00–31.50)	0.00 (0.00–36.00)	0.00 (0.00–30.00)	0.92
Chronic Diseases Count	7.00 (5.00–10.00)	7.00 (5.00–9.25)	7.00 (5.00–10.00)	0.27
Days in the ED	4.00 (3.00–6.00)	4.00 (3.00–5.00)	5.00 (3.00–6.00)	< 0.05
Admissions, last 12 months	1.00 (0.00–1.00)	1.00 (0.00–1.00)	1.00 (0.00–2.00)	< 0.05
Admissions, last 6 months	1.00 (0.00–1.00)	1.00 (0.00–1.00)	1.00 (0.00–1.00)	< 0.05
Admissions in the last month	0.00 (0.00–0.00)	0.00 (0.00–0.00)	0.00 (0.00–0.25)	0.70
Admission Reasons Count	1.00 (1.00–2.00)	1.00 (1.00–2.00)	1.00 (1.00–2.00)	< 0.01
Therapy Count	7.00 (5.00–10.00)	7.00 (4.00–10.00)	8.00 (5.00–11.00)	0.08
Anthropometric				
BMI	24.34 (21.48–27.68)	24.55 (21.72–28.18)	24.08 (21.12–27.14)	0.11
SBP	126.54 (23.56)	125.53 (23.88)	127.37 (23.31)	0.43
DBP	70.00 (63.50–80.00)	73.00 (64.75–80.00)	70.00 (62.50–80.50)	0.80

Heart Rate	85.00 (76.00–100.00)	85.00 (77.00–100.00)	85.00 (75.00–100.00)	0.45
Body Temp	36.40 (36.00–37.00)	36.40 (36.00–37.00)	36.50 (36.00–37.00)	0.80
SpO2	96.00 (94.00–98.00)	97.00 (94.70–98.00)	96.00 (94.00–98.00)	0.10
GCS Score	15.00 (15.00–15.00)	15.00 (15.00–15.00)	15.00 (15.00–15.00)	0.10
Functional				
Barthel Score (Before Admission)	100.00 (65.00–100.00)	100.00 (95.00–100.00)	90.00 (50.00–100.00)	< 0.001
Barthel Score (Admission)	55.00 (25.00–100.00)	90.00 (50.00–100.00)	45.00 (10.00–60.00)	< 0.001
ECOG	2.00 (1.00–3.00)	1.00 (0.00–3.00)	3.00 (1.00–3.00)	< 0.001
Karnofsky	70.00 (50.00–90.00)	80.00 (50.00–90.00)	60.00 (50.00–80.00)	< 0.001
Braden Score	18.00 (14.00–22.00)	20.00 (17.00–22.25)	16.00 (13.00–18.00)	< 0.001
CIRS	6.00 (0.00–9.00)	4.00 (0.00–9.00)	6.50 (0.00–10.00)	< 0.05
Laboratory				
Haemoglobin	11.45 (2.51)	11.26 (2.53)	11.61 (2.48)	0.17
WBC	9.85 (7.17–13.22)	9.55 (7.12–12.82)	10.40 (7.20–14.90)	0.19
Platelets	213.00 (149.75–309.25)	213.00 (160.50–303.25)	210.50 (144.25–316.75)	0.82
AST	28.00 (20.25–47.00)	28.00 (20.50–45.00)	29.00 (20.50–50.00)	0.76
ALT	22.00 (14.00–40.00)	22.00 (15.00–40.00)	22.00 (13.00–39.50)	0.84
LDH	279.50 (227.00–370.25)	268.00 (224.25–358.00)	290.00 (236.25–422.50)	0.07
Creatinine	1.06 (0.78–1.64)	1.01 (0.78–1.62)	1.15 (0.78–1.67)	0.48
CRP	48.80 (13.10–134.00)	40.40 (9.30–121.20)	60.60 (15.28–146.57)	< 0.05
Sodium	137.75 (134.50–140.53)	138.00 (135.20–140.50)	137.25 (132.90–140.60)	0.06
Potassium	4.30 (3.92–4.75)	4.25 (3.91–4.67)	4.36 (3.96–4.79)	0.18

5.4.5 Frailty and comorbidity in relation to the outcome

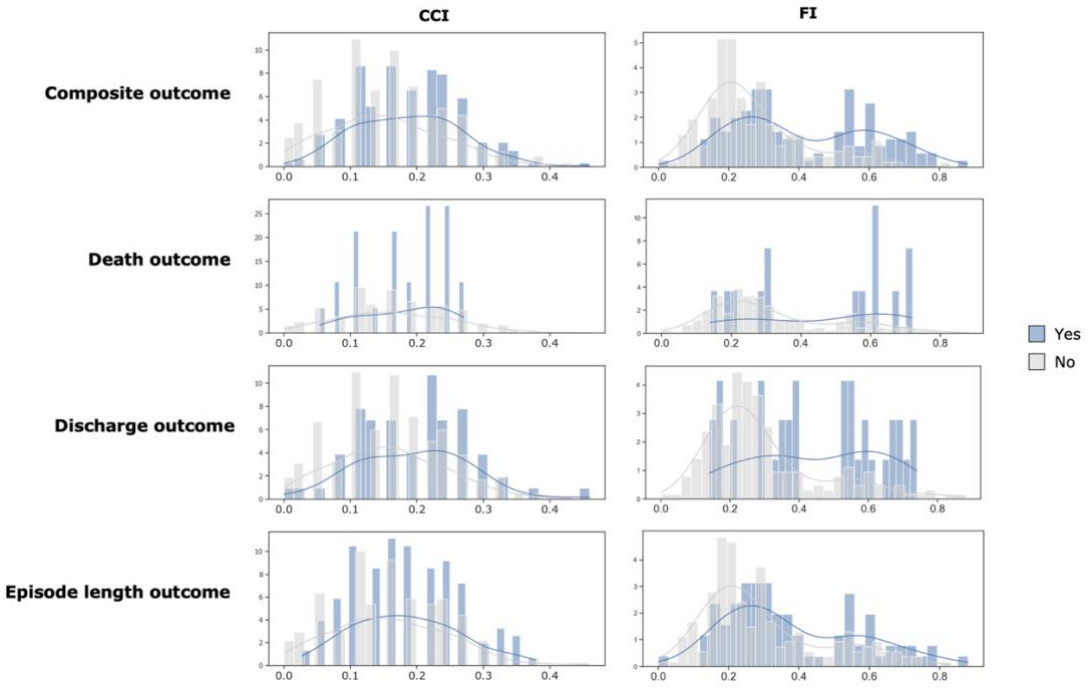
The CCI is a widely validated tool for estimating short- and long-term mortality risk in hospitalized patients, particularly older adults (Frenkel *et al*, 2014), and was therefore used as a benchmark to contextualize the prognostic performance of the admission-only FI in our cohort. Figure 5.4.4A shows that admission-only FI distributions for patients meeting (blue) or not meeting (grey) the negative outcome overlap less than those for the CCI, indicating better between groups discrimination. Moreover, CCI values clustered below 0.5, whereas FI values spanned up to 0.76 and 0.88 for the comorbidity FI and admission-only FI respectively, reflecting superior resolution and granularity. As a comparison, the FI calculated on the REPOSI register by Cesari and colleagues showed a positively skewed distribution with the median equal to 0.27 (IQR 0.21–0.37) with no patient presenting with a FI value equal to or higher than 0.8 (Cesari *et al*, 2018).

We then trained two single-feature LR models for patient classification in the two outcome groups. The model trained with admission-only FI achieved a mean AUC of 0.71 (95% CI = 0.66-0.76), while that trained with the CCI achieved a mean AUC of 0.62 (95% CI = 0.60-0.63) (Figure 5.4.4B). We then compared the across-splits paired AUC from the two models with the Wilcoxon signed-rank test and found a statistically significant difference (Wilcoxon statistic: 2.0000, two-sided p-value 0.006). We also performed a DeLong test on paired out of fold predictions for both single-feature logistic

models with a 10-fold cross validation approach in training. We obtained a pooled out of fold AUC of 0.69 for the model trained with the admission-only FI and of 0.60 for the model trained with the CCI, with a DeLong z of 1.96 (two-sided p-value 0.049). This confirms the greater predictive performance of a multidimensional index compared to the use of comorbidities alone. To estimate the odds of composite outcome at clinically interpretable increments of frailty and comorbidity, we obtained the Odds Ratios from a Logistic Regression with the Logit component of the Statsmodels package. After adjusting for age and sex, per 0.1 increase in admission-only FI, odds of composite outcome increased 1.43-fold (95% CI: 1.24-1.67, p<0.001), while per 3 points increase in CCI, odds of composite outcome increased 1.27-fold (95% CI: 0.97-1.66, p=0.08). Again, the admission-only FI appears to be a stronger predictor of the composite outcome, with a significant p-value even after adjusting for age and sex. Both of these covariates strongly influence the comorbidity burden of patients while apparently being less impactful on our FI built without taking into account patients' past medical history.

We noticed a subset of patients with the outcome but forming a peak of FI values below the median, especially driven by lengthy AE. Detailed analysis showed that these were predominantly oncologic cases: 51% had localized or metastatic tumours compared with 24% among patients with the outcome but FI above the median (p < 0.01).

A



B

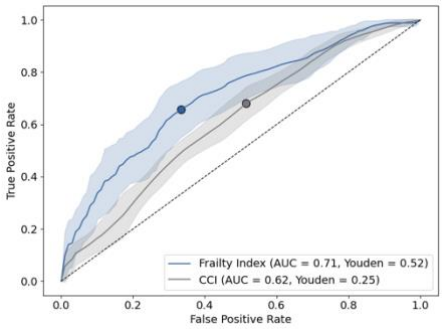


Figure 5.4.4 *– Discriminative performance of the Charlson Comorbidity Index (CCI) and Frailty Index (FI). (**A**) Density plots of CCI and FI values according to composite negative outcome status (blue = meeting outcome; grey = not meeting outcome). (**B**) Receiver operating characteristic curves and corresponding area under the curve (AUC) for two machine learning models using either CCI or FI as the sole predictive feature. Shaded areas represent the standard deviation of the performance metrics.*

5.5 Machine Learning Experiments

Given the good performance of the two very simple single-feature logistic regression ML models obtained in the previous study, we wanted to deepen the exploration of ML models for the prediction of hospitalisation outcomes. We aimed at testing a wider range of features and leveraging the fact that the MED-Cli study allows to build cohorts of patients from different wards or different time windows.

5.5.1 Cohort selection

Having curated and consolidated a first cohort of 411 AE, we preprocessed in the same way an additional cohort of 275 AE from the same time window (01/02/2024-31/01/2025) but from a distinct ward of our hospital, the Unit of General Medicine and Advanced Care, featuring also a high-intensity of care medicine section. High intensity of care wards in Italy are general medicine wards managed by internists that, despite not being intensive-care units, feature 24/24 monitoring stations, non-invasive ventilation and the use of vasopressors. As the hospital building where the Unit of General Medicine and Advanced Care operates is called “Iceberg”, we will refer to this second cohort of AE as “Ice”. We will refer to the first cohort as “4B”. The final number of AE available for training, validation and testing was therefore 686. We decided to use the larger cohort as a training and validation set (411 AE), while the smaller cohort as test cohort (275 AE). While not being truly an “external” test set, as it comes from the same hospital and treats similar patients, as said it still represents the activity of a different team of treating nurses

and physicians with different organisation and, possibly, different attitudes. We were therefore interested in understanding the behaviour of the models in the two cohorts.

5.5.2 Feature selection

Given the insight gained in the previous study, we focused these experiments on the subset of variables that were significantly different in the outcome subgroups or that we deemed clinically relevant:

- Age: we expect higher age to be associated to an increased multimorbidity, frailty and chance of having the negative outcome.
- BMI: it acts as a bimodal risk factor, where both underweight and overweight/obese patients have poorer outcomes. In both cases we expect patients to be sarcopenic, either because of malnutrition or because of detrimental effects of adipose tissues. Additionally, both patients will have high risk of being bedridden and having pressure ulcers.
- CCI, CIRS and count of chronic therapies: we expect them to reflect the burden of chronic diseases of patients and be associated to higher chance of negative outcome.
- Peripheral blood oxygen saturation (SpO2): it was interestingly the only vital parameter approaching significance, with slightly lower values in the outcome class, possibly reflecting the severity of the presenting illness (pneumonia, exacerbation of chronic obstructive pulmonary disease, decompensated heart failure), of chronic diseases or frailty (e.g. poorer muscle function).
- Functional scores (Karnofsky, Barthel before admission, Braden): they all capture the fitness/performance domain of patient in their daily living, like their nutrition, their mobility and their activity, adding fundamental and global information that could be missed by more single-organ variables.
- Lab tests: we kept those regarding the inflammatory status of patients (white blood cells counts [WBC], C-reactive protein [CRP], lactic dehydrogenase [LDH]), anaemia (Haemoglobin [Hb]), which is a biomarker of chronic diseases, malnutrition and decreased performance, and sodium as it surprisingly approximated significance, something for which we could not find an explanation.

Table 5.5.1 *– Variance inflation factors calculated for each of the variables chosen for further machine learning exploration. BMI: body mass index; CCI: Charlson comorbidity index; SpO2: peripheral blood oxygen saturation; CIRS: cumulative illness rating scale; WBC: white blood cells; LDH: lactic dehydrogenase; CRP: C-reactive protein.*

Feature	***VIF***
Constant	1433.68
Age	1.52
BMI	1.06
CCI	1.67
Count of Chronic Therapies	1.30
Admissions in the last 12 months	1.13
SpO2	1.13
Karnofsky Scale	2.39
Barthel Score (Before Admission)	2.09
Braden Score	2.70
CIRS	1.47
Admission Reasons Count	1.20
Haemoglobin	1.09
WBC	1.30
LDH	1.14
CRP	1.18
Sodium	1.14

We first calculated the variance inflation factor (VIF) for the selected variables, and we obtained values between 1 and 5, suggesting moderate correlation and no need to drop any variable for multicollinearity (Table 5.5.1). Figure 5.5.1 reports the missing fractions for the selected variables, and the only variable with >30% missingness that we needed to drop was the Karnofsky performance status (36% missing values). Differences in missingness patterns can also be noted, with functional variables being more missing in the 4B cohort while laboratory ones being more missing in the Ice cohort. This may reflect areas of improvement in data gathering standardisation and thoroughness.

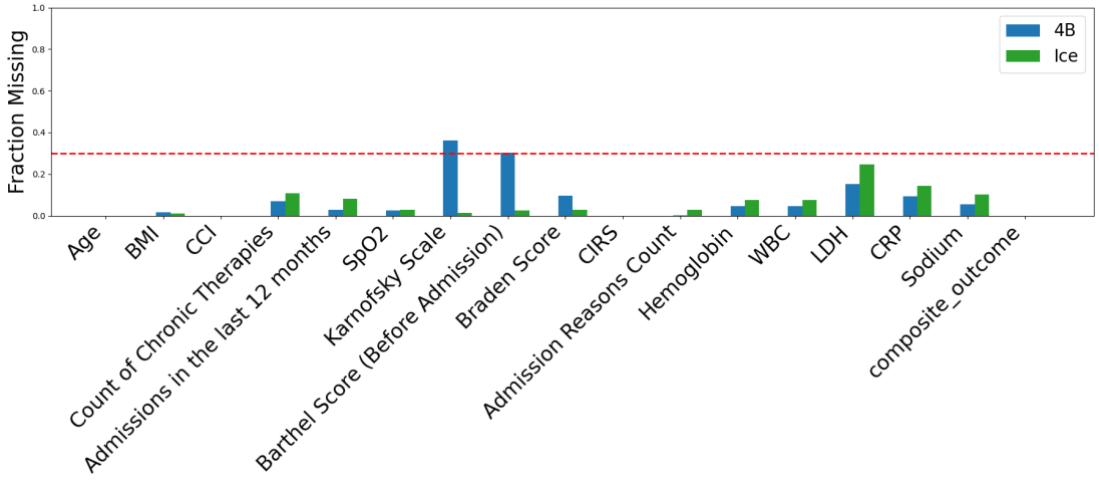


Figure 5.5.1 *– Histogram plot of the percentage of missing values for the final set of measurements that advanced to the exploratory data analysis step. In blue: 4B cohort. In green: Ice cohort. The red dashed line marks the 30% threshold.*

5.5.3 Exploratory data analysis

As detailed in Table 5.5.2, the two cohorts of patients show no significant differences for the most part, however a few considerations can be done, in line with the slightly diverse patient targets of the two wards:

1- Ice patients show a tendency towards being younger.
2- 4B patients are more prone to being readmitted to hospitals in time, possibly displaying increased frailty.
3- Ice patients have a slightly more pronounced inflammatory profile at hospital arrival, with higher average values of WBC, CRP and LDH.
4- The CIRS score is higher for Ice patients, indicating a higher clinical and functional severity of patients managed in the Ice ward, possibly in the high-intensity medicine section (see 5.5.1).

Table 5.5.2 *– Baseline characteristics of the overall patient population and the two different cohorts used for training-validation and testing. BMI: body mass index; CCI: Charlson comorbidity index; SpO2: peripheral blood oxygen saturation; CIRS: cumulative illness rating scale; WBC: white blood cells; LDH: lactic dehydrogenase; CRP: C-reactive protein.*

Characteristic	***All (N=686)***	***4B (N=411)***	***Ice (N=275)***	***p***
Age	75.00 (63.00–84.00)	76.00 (65.00–84.00)	75.00 (61.00–84.00)	0.09
BMI	24.22 (21.36–27.68)	24.28 (21.35–27.68)	24.08 (21.39–27.42)	0.60
CCI	6.00 (4.00–9.00)	6.00 (4.00–8.00)	6.00 (4.00–9.00)	0.61
Count of Chronic Therapies	7.00 (4.00–10.00)	7.00 (4.00–10.00)	6.00 (4.00–11.00)	0.91
Admissions, last 12 months	**1.00 (0.00–1.00)**	**1.00 (0.00–1.00)**	**0.00 (0.00–1.00)**	**0.00**
SpO2	96.00 (94.00–98.00)	96.00 (94.00–98.00)	96.00 (93.00–98.00)	0.71
Barthel Score (Before Admission)	100.00 (65.00–100.00)	100.00 (65.0–100.0)	100.00 (65.0–100.0)	0.49
Braden Score	18.00 (15.00–22.00)	18.00 (14.50–22.00)	18.00 (15.00–21.50)	0.98
CIRS	**7.00 (2.00–10.00)**	**6.00 (0.00–9.00)**	**9.00 (5.00–12.00)**	**0.00**
Admission Reasons Count	1.00 (1.00–2.00)	1.00 (1.00–2.00)	1.00 (1.00–2.00)	0.42
Haemoglobin	11.50 (9.80–13.30)	11.50 (9.85–13.15)	11.50 (9.70–13.35)	0.59
WBC	10.30 (7.20–14.38)	9.80 (7.15–13.40)	10.70 (7.50–15.50)	0.10
LDH	282.00 (233.00–377.50)	280.50 (227–371)	289.00 (246–390)	0.11
CRP	54.50 (14.30–134.00)	47.70 (13.05–130.53)	60.00 (17.00–140.30)	0.08
Sodium	137.70 (134.6–140.6)	137.90 (134.5–140.6)	137.40 (134.7–140.7)	0.99

We also checked the outcome labels counts and their relative percentages in the two cohorts, detecting a different prevalence for the outcome. As previously shown, in the 4B cohort the counts were 186 true and 225 false, with a 45% prevalence of patients with the outcome, while in the Ice cohort the counts were 145 true and 130 false, with a 53% of patients with the outcome.

5.5.4 Benchmarking of machine learning models

As mentioned in the introduction, one of the challenges of applying ML to RWD is there is not a one-size-fit all approach. The experimental settings must adapt to the clinical question, the task, the source data, and the computing power only to mention a few. Here, we tested a benchmarking pipeline to quickly assess the most promising model. The pipeline featured:

- A median imputer for missing data management
- A robust scaler for data scaling
- A k-fold data splitter with shuffling for cross validation (k=10)
- A set of 9 base supervised classification models and 3 ensemble models (see code snippet below)

```
base_models = [
    ("LR", LogisticRegression(max_iter = 10000, random_state=42)),
    ("RF", RandomForestClassifier(random_state=42)),
    ("XGB", XGBClassifier(use_label_encoder=False, eval_metric='logloss',
random_state=42)),
    ("SVC", SVC(probability=True, random_state=42)),
    ("KNN", KNeighborsClassifier()),
    ("NB", GaussianNB()),
    ("BNB", BernoulliNB()),
    ("MLP",  MLPClassifier(random_state=42)),
    ("HGB", HistGradientBoostingClassifier(random_state=42))
    ]

ensemble_models = [
    ("soft voting", VotingClassifier(estimators=base_models, voting='soft')),
    ("stacking LR", StackingClassifier(estimators=base_models,
final_estimator=LogisticRegression(max_iter = 10000, random_state=42))),
    ("stacking RF", StackingClassifier(estimators=base_models,
final_estimator=RandomForestClassifier(random_state=42)))
]
```

Figure 5.5.2 provides the AUCs of the 12 models trained with the described pipeline, ranked by median AUC value. The LR model achieves the best performances, followed by the stacking ensemble with LR as final estimator, and then the Bernoulli naïve Bayes.

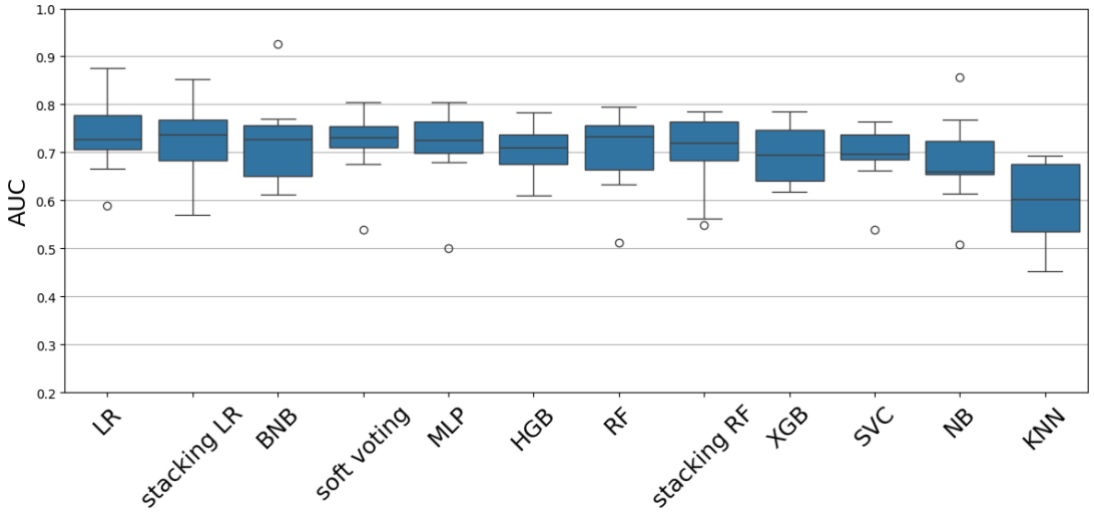


Figure 5.5.2 *– Overview of models' performance in the benchmarking experiment. Boxplots are ranked by median AUC from left to right. Metrics for both base models and ensemble models are reported. AUC: area under the receiver operating characteristic curve; LR: logistic regression; BNB: Bernoulli naive Bayes; MLP: Multi-layer Perceptron classifier; HGB: histogram-based gradient boosting classification tree; RF: random forest; XGB: XGBoost; SVC: C-Support Vector Classification; NB: gaussian native Bernoulli; KNN: K-neighbours classifier.*

5.5.5 Training, validation and testing of a logistic regression model

Having benchmarked a few models, we selected the LR with balanced class weights as the best candidate for further experiments, as we saw no advantage in picking more complex, less explainable and more computationally expensive models. Our pipeline featured MICE to manage missing data and a 10-fold stratified shuffle split (validation set size = 30%) to obtain validation metrics with confidence intervals. Following previous evidence on the impact of different scaling methods, we tested both a robust scaler and a quantile scaler and, again, the second proved to be better. Table 5.5.3 reports on the metrics on the 4B validation set and Ice test set.

Table 5.5.3 *– Performance metrics of two logistic regression models across the 10 4B validation sets and in the Ice test set. Figures are reported as mean (95% c.i.). Input features: Age, BMI, CCI, Count of Chronic Therapies Admissions in the last 12 months, SpO2, Barthel Score (Before Admission), Braden Score, CIRS, Admission Reasons Count, Haemoglobin, WBC, LDH, CRP, Sodium. Missing management: MICE. Validation set size: 30%. Class weight: provided. AUC: area under the receiver operating characteristic curve; MCC: Matthews correlation coefficient.*

Metric	*Validation set*	*Test set*
Accuracy	0.65 (0.62-0.69)	0.74
Precision	0.6 (0.57-0.64)	0.75
Specificity	0.63 (0.57-0.69)	0.72
Recall	0.69 (0.66-0.71)	0.76
F1	0.64 (0.62-0.66)	0.76
ROC AUC	0.71 (0.68-0.75)	0.78
MCC	0.31 (0.25-0.37)	

Surprisingly, performance increases in the test set, as also shown by the receiver operating characteristic curves plots (Figure 5.5.3A). To assess whether the deployed prediction model demonstrated consistent discrimination across the two clinical cohorts, we applied the same final model, trained on the full 4B dataset, to both the same 4B training set and Ice test set. The AUC was first computed directly for each cohort using the predicted probabilities generated by the deployed model. In Ward A, the model achieved an apparent AUC of 0.76, while in Ward B the AUC was 0.78. To quantify the uncertainty around these estimates and to formally compare them, we performed an unpaired stratified bootstrap with 10000 resamples independently within each cohort, thereby obtaining bootstrap distributions for the two AUCs and their difference. The resulting estimate of the discrimination difference (ΔAUC = 0.021) was accompanied by a 95% confidence interval of –0.052 to 0.094, which includes zero. Likewise, the one-sided hypothesis test evaluating whether discrimination was superior in Ice test set did not reach statistical significance ($p = 0.289$). These findings suggest that the deployed model exhibits broadly comparable performance across the two wards, with no statistically detectable improvement or degradation in the external cohort. Nevertheless, the relatively wide confidence interval highlights the role of sampling variability and the potential need for larger external samples to more precisely characterize between-cohort differences.

We also plotted precision-recall curves for both true and false class, displaying similar performances for the two classes (Figure 5.5.3 B-D).

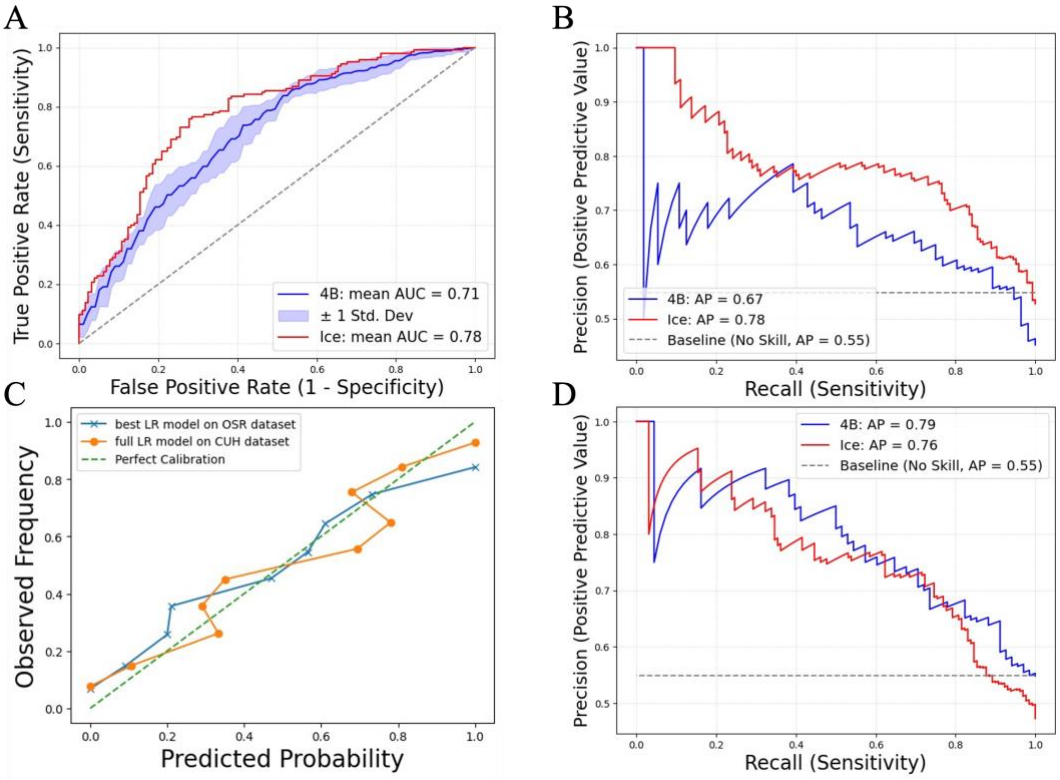


Figure 5.5.3 *– (**A**) Receiver operating characteristic curves, (**B**) precision-recall curves for true (negative hospitalisation outcome) and (**D**) false (no outcome) classes, and (**C**) calibration curves of the logistic regression model on the validation cohort and Ice test cohort. Areas under the curve (AUC) and average precision (AP) are also reported. The shaded area represents the standard deviation of the metric.*

The model appeared to be decently calibrated for both the cohorts, as shown by calibration curves following quite consistently the diagonal of the plot (Figure 5.5.3C).

Table 5.5.3 *– Values of the coefficients for the 15-variable logistic regression model. BMI: body mass index; CCI: Charlson comorbidity index; SpO2: peripheral blood oxygen saturation; CIRS: cumulative illness rating scale; WBC: white blood cells; LDH: lactic dehydrogenase; CRP: C-reactive protein.*

Feature	***Coefficient***
Braden Score	-0.77
BMI	-0.27
CCI	0.19
CRP	0.16
Sodium	-0.16
Haemoglobin	-0.15
Barthel Score (Before Admission)	-0.14
Age	-0.13
LDH	0.11
Admission Reasons Count	0.08
SpO2	-0.05
CIRS	0.04
Count of Chronic Therapies	-0.03
WBC	0.01
Admissions in the last 12 months	0.00

We found it intriguing that the most relevant coefficient for the LR model was the Braden score, which is an evaluation of the pressure ulcer risk usually done by our nurses to decide if the patient requires a pressure-relieving mattress. The negative value of the coefficient means that patients with lower values of the Braden score, i.e. higher pressure ulcer risk, have a higher risk of negative hospitalisation outcome. This stresses again the importance of a multidimensional evaluation of complex and frail internal medicine patients, for whom information regarding nutrition, mobility, activity and sensory perception are as important, if not more, than the biochemical evaluation of the status of their organs. As expected, patients with lower BMI are at higher risk of the outcome, which seems counterintuitive, but it is compatible with a heightened frailty of malnourished and sarcopenic patients. In fact, the BMI normality thresholds for elderly patients are increased at 22-27 kg/m^2. This adjustment is because a slightly higher BMI in seniors may be associated with lower mortality risk, better functional capacity, and protection against conditions like malnutrition (Winter *et al*, 2017). Still, the burden of disease comes at third place, where higher CCI values increase the risk of negative hospitalisation outcome. CRP, LDH and haemoglobin also behave as expected, with higher level of CRP and LDH, and lower levels of haemoglobin being linked to increased risk of the outcome. Finally, lower sodium increases the risk of the outcome. Hyponatraemia is the most common electrolyte abnormality encountered in hospitalised patients and previous studies have evaluated its association to all-cause mortality and hospitalisation outcomes: a 2015 Danish cohort study of 279508 first-time acute admissions to Departments of Internal Medicine found a consistent 30-day and 1-year mortality risk increase across different hyponatremia levels (Holland-Bill *et al*, 2015); an earlier cohort studies from Massachusetts reported an association of community-acquired hyponatraemia with in hospital mortality and increased length of stay (Wald *et al*, 2010).

5.6 Conclusion

To address the challenge of characterising the complex and evolving phenotype of internal medicine patients, we introduced the MED-Cli study, which employs a sophisticated eCRF and a rigorous data collection strategy to generate a high-quality, standardised, and representative patient registry. By integrating this registry into the existing real-world data pipeline provided by the S-RACE platform, we assembled a

comprehensive dataset that reflects the everyday clinical reality of patients admitted to our general medicine wards. In fact, through S-RACE we could enrich the eCRF-derived registry with laboratory results retrieved from the Laboratory Information System and with detailed therapeutic information extracted from the EHR, thereby expanding both the breadth and temporal resolution of the dataset.

One immediate outcome is the ability to closely monitor both patient characteristics and clinical management practices through interactive business-intelligence dashboards, thereby fostering a data-driven approach to decision-making in healthcare. A key strength of this strategy is its foundation on locally generated data, which maximises relevance and representativeness for the future patients we will treat, while reducing reliance on external evidence. As discussed in Chapter 2, this evidence is often highly robust and methodologically rigorous, yet not always fully aligned with the specific operational and epidemiological context of patients treated outside clinical trials.

The second outcome of this RWD pipeline is that it enables us to run cohort studies to answer specific clinical questions. As an example, this chapter demonstrated the importance of adopting a multidimensional evaluation such as the frailty framework when assessing complex internal medicine patients, for which there is high prognostic uncertainty, as discussed earlier in Section 5.2.1. One question we had was in fact: “shall we move beyond the disease-centric paradigm when doing risk-stratification for modern internal medicine patients?”. Indeed, we found that the new framework of frailty, measured at hospital admission using a multidimensional, diagnosis-independent FI that excluded both comorbidities and the acute admitting diagnosis, was a better predictor of short-term adverse outcomes (defined as a composite of in-hospital mortality, prolonged length of stay, and discharge to a non-home destination) than comorbidities. In particular, we directly compared the admission-only FI with the CCI, a well-validated prognostic tool for hospitalized patients (Frenkel *et al*, 2014), and demonstrated superior discriminative performance (AUC 0.70 vs. 0.62) and finer risk resolution. Importantly, including information about medical history did not enhance predictive power, underscoring the robustness of this streamlined approach. While frailty is traditionally conceptualized as a predictor of long-term outcomes such as disability, institutionalization, and mortality (Cesari *et al*, 2018; Lv *et al*, 2022; Damanti *et al*, 2024; Kim & Rockwood, 2024), our findings show that it also identifies acute vulnerability that

translates into immediate risk during hospitalization. Of note, while prior studies have shown that frailty predicts outcomes in mixed or elective hospital populations (Cilla *et al*, 2023), our cohort consisted predominantly of acutely admitted internal medicine patients, a high-complexity group where swift prognostication is critical. However, it must be highlighted that our FI was not only assessing the chronic state of patients (via the pre-admission Barthel index), but, as it included vitals and laboratory data on hospital admission, it also measured the severity of the acute illness.

The relevance of a multidimensional evaluation of patients is confirmed by the Braden score and BMI being at the top places as a coefficient of the LR ML model that we trained and validated in Section 5.5.5. In the context of increasingly constrained healthcare systems, integrating targeted prognostic tools such as the admission-only FI for in-hospital risk could support a more personalized approach to care planning, ensuring that resources are directed toward the most relevant and modifiable risks for each patient. Such a strategy aligns with the principles of complexity-aware, patient-centred medicine, where precision in risk identification underpins both improved clinical outcomes and more efficient use of hospital capacity. By addressing the limitations of diagnosis-dependent tools and validating a rapid, point-of-admission method in an acute, high-risk population, our work provides a pragmatic pathway for integrating frailty assessment into routine early decision-making in internal medicine wards.

The third outcome of the availability of such a wealth of highly curated RWD is that it creates an ideal, controlled environment in which to conduct methodologically rigorous ML experiments. The MED-Cli dataset, by virtue of its standardised data collection, continuous quality checks, and rich clinical granularity, supports the implementation of benchmarking pipelines. In Section 5.5.5, we demonstrated this by evaluating multiple algorithms on a predefined set of 15 clinically and statistically relevant features, followed by a complete training–validation–testing procedure using LR as the preferred model. Importantly, because the MED-Cli study collects data from multiple general medicine wards, these distinct clinical settings can be leveraged as quasi-external datasets, enabling meaningful assessments of generalisability and mitigating the risk of overfitting to a single ward's case mix or organisational practices. This constitutes a substantial advantage, particularly given the broader challenges of RWD research introduced in Section 1.2.1, where poor data quality and heterogeneous collection standards often

hinder model development and introduce biases. In contrast, the curated, interoperable, and information-rich MED-Cli dataset provides a dependable substrate on which to "stress-test" analytical pipelines, explore modelling strategies, and generate actionable insights with a high degree of methodological reliability.

6. Conclusions, Limitations and Future Research Directions

In the introduction of this work, we highlighted the growing relevance gained by healthcare real-world data in filling the representativeness gap that randomized controlled trials have especially when we face the problematic management of a new phenotype of patients in the internal medicine setting: the complex, multimorbid and fragile elderly patient. We described the potential which lies in leveraging such data with advanced analysis techniques such as machine learning to produce real-world evidence and enable data-driven medicine or produce systems to support decision making, thanks for example to the prediction of disease evolution. However, we listed the challenges that still need to be addressed, such as real-world data availability, quality, informative content, and interpretation. We stressed how difficult it is in this field to define clear, universally valid outcome definitions for machine learning predictive tasks, and the absence of standardized protocols for model development, especially regarding the data preprocessing step. Through this thesis, we address these fundamental challenges by describing: 1) the San Raffaele Ai CEnter (S-RACE) end-to-end data science pipeline for real-world data extraction, integration, curation, labelling and analysis which establishes the technical foundations required for trustworthy clinical artificial intelligence; 2) the assembly and benchmarking of different data preprocessing strategies demonstrating how careful optimisation of preprocessing choices substantially influences downstream model performance; 3) the MED-Cli study and the Cohort Genomic Platform for the systematic collection of real-world data produced by general medicine wards, thereby enabling high-resolution characterisation of both patient profiles and clinical practices; 4) the adoption of a multidimensional frailty-based framework for assessing complex patients and the resulting enhanced risk stratification beyond conventional comorbidity measures. Ultimately, this thesis demonstrates the wealth of clinically meaningful insights that can be unveiled by analysing real-world data to foster improvements in our management strategies and decision-making, reinforcing modern, sustainable and patient-centric healthcare systems for the benefit of the whole community.

Several limitations can be highlighted together with potential solutions that represent areas for future development and work.

1- **Platform infrastructure**. S-RACE is inherently dependent on agreements with industrial partners such as Microsoft. Its core functionalities rely on Azure services, their associated interfaces, and substantial cloud computing resources. While the partnership stemming from these industrial collaborations provide important advantages like a robust infrastructure, regulatory compliance, and high scalability, it also represents a potential limitation. Specifically, it may introduce a condition of partial technological and operational dependence for research teams and for the institution as a whole. A possible mitigation strategy would consist in developing internal expertise aimed at progressively enabling migration toward open-source or cost-free alternatives, and in establishing academic consortia or national/European platforms capable of jointly maintaining shared, public, not-for-profit health data gathering and analysis infrastructures. Such approaches could strengthen long-term autonomy and reduce the risk of being forced to negotiate in case of misalignment between our scientific and common good goals versus the strategic interests of commercial providers. In future work we will better investigate opportunities stemming from the European Health Data Space framework, from the HealthData@EU Central Platform, and from European grant calls.
2- **Framing of machine learning problems**. All machine learning experiments in this thesis were framed as static classification tasks (e.g., predicting a decrease in HbA1c or identifying admissions with length of stay above the mean). Although this formulation is practical and interpretable, alternative problem definitions could yield richer and more clinically relevant information. In particular, both clinical questions could be reformulated as time-to-event analyses, enabling the use of survival modelling approaches such as Cox proportional hazards models, random survival forests, or neural survival models (Katzman *et al*, 2018). These methods would allow us to model not only the occurrence of an event but also its timing, which is often of primary importance in clinical decision-making. Furthermore, regression-based formulations could be explored, like predicting the value of HbA1c three years after baseline or number of days spent in hospital, thus providing more granular quantitative estimates than binary outcomes.

3- **Sizes of datasets**. Although the cohorts used in our studies were adequate for exploratory modelling, they remain considerably smaller than those reported in many state-of-the-art studies discussed in Chapter 2.

   For the T2DM cohort, the most direct strategy to overcome this limitation would be to access and leverage the Annali AMD (Italian Annals of Diabetes) database, which represents one of the largest national registries for diabetes care and would provide the statistical power required for more ambitious modelling efforts.

   For the MED-Cli study, the size of the cohort will naturally expand over the planned 10-year enrolment period. Nevertheless, the dataset will remain monocentric, inherently limiting the possibility of rigorous external validation. Although we demonstrated that internal interdepartmental validation can partially mitigate this issue, it cannot substitute a true multicentre assessment. Encouragingly, the Italian Society of Internal Medicine has recently initiated a nationwide data-collection programme across general medicine wards, with four dedicated collection weeks per year. This initiative is expected to generate a large multicentric real-world dataset well suited for machine learning and for robust external validation. In the interim, an alternative immediately accessible resource is the REPOSI registry, which could provide a valuable multicentric dataset for complementary analyses and further model validation. (Nobili *et al*, 2011; Mannucci *et al*, 2018).

4- **Target definition**. While HbA1c is the fundamental biomarker in T2DM, it is not the only relevant one. We left unexplored others such as BMI and kidney function both of which represent important dimensions of diabetes progression and would be strong candidates for future machine learning studies. Furthermore, recent developments in the definition of metabolic dysfunction–associated steatotic liver disease underscore the need to broaden analyses to include markers of liver health, which is increasingly recognised as a central component of metabolic disease.

   The definition of our composite negative outcome of hospitalisation could rise some criticism. First, discharge to a lower-intensity care facility is not invariably negative: in some cases, such transfers are undertaken for rehabilitation purposes and may indicate a favourable evolution in selected, functionally fit patients. Second, not all prolonged admission events represent a failure: some diseases such

as endocarditis require extended duration of intravenous antibiotic therapy, making a lengthy hospital stay clinically appropriate. Looking at the discharge diagnoses could allow more accurate classification of these cases. Third, the current definition of the outcome does not investigate important events such as death after discharge or new hospitalisation events, that are highly relevant to assess the true effectiveness of our interventions in this setting, which should aim at having a more sustained effect beyond averting in-hospital death or prolonged hospital stay. Indeed, in the context of the MED-Cli study, our group recently started a manual collection of post-discharge events (ER or hospital readmissions, deaths) via phone calls as we lack data links with registries of the local health authority. The availability of these new timepoints of patients' medical history will allow new modelling studies.

5- **Feature selection**. A substantial portion of this work was devoted to developing and optimising the data extraction pipelines and gaining familiarity with the available datasets. Several data sources became accessible only during the latter half of the PhD programme, and some data collection efforts such as the MED-Cli study reached full operational capacity progressively over time. As a result, our exploratory analyses and modelling experiments primarily focused on the most readily available variables, such as clinical measurements. However, a considerable amount of information still requires systematic curation before it can be considered suitable for machine learning. For instance, in the HbA1c prediction task, we did not incorporate baseline therapeutic information, despite its clear clinical relevance. Future research will therefore need to include additional data harmonisation and preprocessing steps to fully exploit the richness of the real-world data currently at our disposal. Also, the generation of structured data from unstructured medical reports with the potential of adding new features for analysis or filling currently missing data was only briefly attempted during the PhD. Lately, our team has resumed efforts in this direction, with the aim of comparing extraction performance of Microsofts' NLP- and FHIR-based Text Analytics 4 Healthcare with more recent LLM-based approaches.

6- **Clinical translation**. In section 1.2.2 we examined the issue of the slower than hoped adoption of Artificial Intelligence by healthcare systems. Then, when

describing the S-RACE platform, we mentioned how it is natively meant to go all the way from data acquisition to web-based model deployment. While this PhD project's main focus was on the steps happening before model adoption, a natural future direction would be that of an attempt to create and test clinical decision support systems based on the models we trained and validated, enabled by the availability of a platform with the potential for that. The next steps would therefore be: 1. The registration of our models and related pipelines 2. The production of a webapp to allow the generation of predictions for new patients based on inputted parameters; 3. The design and approval of prospective clinical trials to evaluate the adoption of the tools and their clinical utility. Furthermore, registered models can be evaluated in the Responsible Artificial Intelligence dashboard of the Azure Machine Learning environment. This module is meant for in-depth model debugging, to identify errors, biases and fairness issues and to allow the diagnosis of the reason for such mistakes. The findings generated by the dashboard can be leveraged in multiple ways. First, they generate new clinical insights, for example by identifying how the importance of specific features changes when varying their value. One may find that as age increases, it becomes less able to discriminate patients as having or not the negative outcome, meaning that being old is not sufficient to develop the negative outcome. Second, they allow for a more informed clinical decision, as practitioners using the model know how well it performs in different subgroups of patients. Third, they improve future modelling efforts, for instance by suggesting to add new data to better represent specific patient subgroups, or to change some of the models' parameters.

All the results presented in this dissertation are the outcome of a concrete and sustained multidisciplinary effort, one that, in my view, should represent the standard for any future research carried out at the intersection of healthcare and data science. Over the past 36 months, I have consistently questioned how far I should expand beyond my "original" domain as a physician and immerse myself in the methods, language, and mindset of data science. I learned to code, to structure and approach machine learning problems, and to evolve into a genuinely "data-informed" clinician. The deeper I went, the clearer it became that progress in this field relies not on isolated expertise, but on

teams capable of integrating diverse perspectives. My dual background in medicine and data science places me in a position to help bridge the existing gap between healthcare practitioners and "machine learners", maximising the translational potential of research for the benefit of patients and society at large. This is the direction I will continue to pursue as I move forward in my research journey.

References

A plan for digital health and social care - GOV.UK

Almyranti M, Sutherland E, Ash N & Eiszele S (2024) Artificial Intelligence and the health workforce. 28

American Diabetes Association Professional Practice Committee (2025) 9. Pharmacologic Approaches to Glycemic Treatment: Standards of Care in Diabetes—2025. *Diabetes Care* 48: S181–S206

Anzalone AJ, Geary CR, Dai R, Watanabe-Galloway S, McClay JC & Campbell JR (2025) Lower electronic health record adoption and interoperability in rural versus urban physician participants: a cross-sectional analysis from the CMS quality payment program. *BMC Health Serv Res* 25: 128-

Artificial Intelligence (AI) in Healthcare Market Growth, Drivers, and Opportunities

Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices | FDA

Artificial Intelligence-Enabled Medical Devices | FDA

Bahmani A, Alavi A, Buergel T, Upadhyayula S, Wang Q, Ananthakrishnan SK, Alavi A, Celis D, Gillespie D, Young G, *et al* (2021) A scalable, secure, and interoperable platform for deep data-driven health management. *Nature Communications 2021 12:1* 12: 1–11

Balsano C, Cabitza F, Cicco S, Gori M, Malerba D, Montagna M, Tarquini R & Vacca A (2025) Artificial intelligence in medicine: a position paper by the Italian Society of Internal Medicine. *Intern Emerg Med*

Barlean B, Merali K, Smith M, Yang P, Mistry S, Gouripeddi R, Facelli JC, Lantz AM, Staley C & Nicol MR (2023) 297 Identifying Opportunities and Challenges for Translational Informatics Approaches to Real-World Data: A Diabetes Case Study. *J Clin Transl Sci* 7: 89–89

Barnett AG, van der Pols JC & Dobson AJ (2005) Regression to the mean: what it is and how to deal with it. *Int J Epidemiol* 34: 215–220

Batko K & Ślęzak A (2022) The use of Big Data Analytics in healthcare. *J Big Data* 9: 1–24

Berchet C, Guanais F & Colombo F (2019) Realising the Full Potential of Primary Health Care Policy brief

Bielinski SJ, Yanes Cardozo LL, Takahashi PY, Larson NB, Castillo A, Podwika A, De Filippis E, Hernandez V, Mahajan GJ, Gonzalez C, *et al* (2023) Predictors of Metformin Failure: Repurposing Electronic Health Record Data to Identify High-Risk Patients. *Journal of Clinical Endocrinology and Metabolism* 108: 1740–1746

Ceriani E, Milani O, Donadoni M, Benetti A, Berra SA, Canetta C, Colombo F, Dentali F, Magnani L, Mazzone A, *et al* (2024) COmplexity of CARE and Discharge barriers: the 'modern internal medicine patient'. Results from the CO-CARED Study. *Intern Emerg Med* 20: 471–479

Cesari M, Franchi C, Cortesi L, Nobili A, Ardoino I, Mannucci PM, Tettamanti M, Pasina L, Perticone F, Salerno F, *et al* (2018) Implementation of the Frailty Index in hospitalized older patients: Results from the REPOSI register. *Eur J Intern Med* 56: 11–18

Chamberlin P, Lambden J, Kozlov E, Maciejewski R, Lief L, Berlin DA, Pelissier L, Yushuvayev E, Pan CX & Prigerson HG (2019) Clinicians' Perceptions of Futile or Potentially Inappropriate Care and Associations with Avoidant Behaviors and Burnout. *J Palliat Med* 22: 1039

Charlson ME, Carrozzino D, Guidi J & Patierno C (2022) Charlson Comorbidity Index: A Critical Review of Clinimetric Properties. *Psychother Psychosom* 91 doi:10.1159/000521288 [PREPRINT]

Chen L, He R, Lu P, Jin Y, Zhou L, Li N, Wu P & Hu B (2026) Operationalizing Large Language Models for Clinical Research Data Extraction: Methods, Quality Control, and Governance. *Journal of Medical Systems 2026 50:1* 50: 25-

Cilla F, Sabione I & D'Amelio P (2023) Risk Factors for Early Hospital Readmission in Geriatric Patients: A Systematic Review. *Int J Environ Res Public Health* 20

Clegg A, Young J, Iliffe S, Rikkert MO & Rockwood K (2013) Frailty in elderly people. *The Lancet* 381: 752–762

Colacci M, Loffler A, Roberts SB, Straus S, Verma AA & Razak F (2025) Patient Complexity, Social Factors, and Hospitalization Outcomes at Academic and Community Hospitals. *JAMA Netw Open* 8

Concato J & Corrigan-Curay J (2022) Real-World Evidence - Where Are We Now? *N Engl J Med* 386: 1680–1682

Damanti S, De Lorenzo R, Citterio L, Zagato L, Brioni E, Magnaghi C, Simonini M, Ruggiero MP, Santoro S, Senini E, *et al* (2024) Frailty index, frailty phenotype and 6-year mortality trends in the FRASNET cohort. *Front Med (Lausanne)* 11: 1465066

Dang A (2023) Real-World Evidence: A Primer. *Pharmaceut Med* 37: 25

Davies MJ, Aroda VR, Collins BS, Gabbay RA, Green J, Maruthur NM, Rosas SE, Del Prato S, Mathieu C, Mingrone G, *et al* (2022) Management of Hyperglycemia in Type 2 Diabetes, 2022. A Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetes Care* 45

Dedić N & Stanier C (2016) Measuring the success of changes to existing business intelligence solutions to improve business intelligence reporting. *Lecture Notes in Business Information Processing* 268: 225–236

Dennis JM, Young KG, Cardoso P, Güdemann LM, McGovern AP, Farmer A, Holman RR, Sattar N, McKinley TJ, Pearson ER, *et al* (2025) A five-drug class model using routinely available clinical features to optimise prescribing in type 2 diabetes: a prediction model development and validation study. *The Lancet* 405: 701–714

Digital Innovation Observatory of Politecnico di Milano (2025) La Sanità Digitale in Italia e i principali ambiti di innovazione.

Dreyer NA (2022) Strengthening evidence-based medicine with real-world evidence. *Lancet Healthy Longev* 3: e641–e642

ERIN - Information for researchers - NIHR Cambridge Biomedical Research Centre

Framework for FDA's Real-World Evidence Program (2018)

Frenkel WJ, Jongerius EJ, Mandjes-Van Uitert MJ, Van Munster BC & De Rooij SE (2014) Validation of the Charlson Comorbidity Index in acutely hospitalized elderly adults: A prospective cohort study. *J Am Geriatr Soc* 62

Fried LP, Cohen AA, Xue QL, Walston J, Bandeen-Roche K & Varadhan R (2021) The physical frailty syndrome as a transition from homeostatic symphony to cacophony. *Nat Aging* 1

Fried LP, Ferrucci L, Darer J, Williamson JD & Anderson G (2004) Untangling the Concepts of Disability, Frailty, and Comorbidity: Implications for Improved Targeting and Care. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences* 59 doi:10.1093/gerona/59.3.m255 [PREPRINT]

Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, Seeman T, Tracy R, Kop WJ, Burke G, *et al* (2001) Frailty in older adults: Evidence for a phenotype. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences* 56

Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, Collins R & Allen NE (2017) Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 186: 1026–1034

García-Jaramillo M, Luque C & León-Vargas F (2023) Machine Learning and Deep Learning Techniques Applied to Diabetes Research: A Bibliometric Analysis. *J Diabetes Sci Technol* 18: 287

Giorda CB, Pisani F, De Micheli A, Ponzani P, Russo G, Guaita G, Zilich R & Musacchio N (2020a) Determinants of good metabolic control without weight gain in type 2 diabetes management: a machine learning analysis. *BMJ Open Diabetes Res Care* 8

Giorda CB, Pisani F, De Micheli A, Ponzani P, Russo G, Guaita G, Zilich R & Musacchio N (2020b) Determinants of good metabolic control without weight gain in type 2 diabetes management: a machine learning analysis. *BMJ Open Diabetes Res Care* 8

GitHub - microsoft/responsible-ai-toolbox

Gregg EW, Patorno E, Karter AJ, Mehta R, Huang ES, White M, Patel CJ, McElvaine AT, Cefalu WT, Selby J, *et al* (2023a) Use of real-world data in population science to improve the prevention and care of diabetes-related outcomes. *Diabetes Care* 46: 1316–1326

Gregg EW, Patorno E, Karter AJ, Mehta R, Huang ES, White M, Patel CJ, McElvaine AT, Cefalu WT, Selby J, *et al* (2023b) Use of Real-World Data in Population Science to Improve the Prevention and Care of Diabetes-Related Outcomes. *Diabetes Care* 46: 1316–1326

Gultepe E, Green JP, Nguyen H, Adams J, Albertson T & Tagkopoulos I (2014) From vital signs to clinical outcomes for patients with sepsis: A machine learning basis for a clinical decision support system. *Journal of the American Medical Informatics Association* 21

Gupta A, Liu T & Crick C (2020) Utilizing time series data embedded in electronic health records to develop continuous mortality risk prediction models using hidden Markov models: A sepsis case study. *Stat Methods Med Res* 29

Herrera AP, Snipes SA, King DW, Torres-Vigil I, Goldberg DS & Wenberg AD (2010) Disparate inclusion of older adults in clinical trials: priorities and opportunities for policy and practice change. *Am J Public Health* 100

Herrero-Zazo M, Fitzgerald T, Taylor V, Street H, Chaudhry AN, Bradley JR, Birney E & Keevil VL (2023) Using machine learning to model older adult inpatient trajectories from electronic health records data. *iScience* 26

Heumos L, Ehmele P, Treis T, Upmeier zu Belzen J, Roellin E, May L, Namsaraeva A, Horlava N, Shitov VA, Zhang X, *et al* (2024) An open-source framework for end-to-end analysis of electronic health record data. *Nat Med* 30: 3369–3380

Holland-Bill L, Christiansen CF, Heide-Jørgensen U, Ulrichsen SP, Ring T, Jørgensen JOL & Sørensen HT (2015) Hyponatremia and mortality risk: a Danish cohort study of 279 508 acutely hospitalized patients. *Eur J Endocrinol* 173: 71–81

Hou J, Zhao R, Gronsbell J, Lin Y, Bonzel CL, Zeng Q, Zhang S, Beaulieu-Jones BK, Weber GM, Jemielita T, *et al* (2023) Generate Analysis-Ready Data for Real-world Evidence: Tutorial for Harnessing Electronic Health Records With Advanced Informatic Technologies. *J Med Internet Res* 25

Hunter DJ (2023) At Breaking Point or Already Broken? The National Health Service in the United Kingdom. *New England Journal of Medicine* 389: 100–103

Initial management of hyperglycemia in adults with type 2 diabetes mellitus - UpToDate

Inouye SK, Studenski S, Tinetti ME & Kuchel GA (2007) Geriatric syndromes: clinical, research, and policy implications of a core geriatric concept. *J Am Geriatr Soc* 55: 780–791

International Diabetes Federation (2025) IDF Diabetes Atlas, 11th edn.

Jalilian L & Khairat S (2022) The Next-Generation Electronic Health Record in the ICU: A Focus on User-Technology Interface to Optimize Patient Safety and Quality. *Perspect Health Inf Manag* 19: 1g

Jee J, Fong C, Pichotta K, Tran TN, Luthra A, Waters M, Fu C, Altoe M, Liu SY, Maron SB, *et al* (2024) Automated real-world data integration improves cancer outcome prediction. *Nature* 636: 728–736

Jha AK (2010) Meaningful Use of Electronic Health Records: The Road Ahead. *JAMA* 304: 1709–1710

Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Moody B, Gow B, Lehman L wei H, *et al* (2023) MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data 2023 10:1* 10: 1-

Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T & Kluger Y (2018) DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol* 18: 24-

Kaur J & Mann KS (2018) AI based HealthCare Platform for Real Time, Predictive and Prescriptive Analytics using Reactive Programming. In *Journal of Physics: Conference Series*

Kennedy-Martin T, Curtis S, Faries D, Robinson S & Johnston J (2015) A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* 16

Kim DH & Rockwood K (2024) Frailty in Older Adults. *New England Journal of Medicine* 391: 538–548

Kiran M, Xie Y, Anjum N, Ball G, Pierscionek B & Russell D (2025) Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis. *Front Digit Health* 7: 1557467

Kraljevic Z, Bean D, Shek A, Bendayan R, Hemingway H, Yeung JA, Deng A, Baston A, Ross J, Idowu E, *et al* (2024) Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit Health* 6: e281–e290

Lenti MV, Croce G, Brera AS, Ballesio A, Padovini L, Bertolino G, Di Sabatino A, Klersy C & Corazza GR (2023) Rate and risk factors of in-hospital and early postdischarge mortality in patients admitted to an internal medicine ward. *Clinical Medicine, Journal of the Royal College of Physicians of London* 23

Liu F & Demosthenes P (2022) Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Medical Research Methodology 2022 22:1* 22: 1–10

Lo JJM, Graves N, Chee JH & Hildon ZJL (2022) A systematic review defining non-beneficial and inappropriate end-of-life treatment in patients with non-cancer diagnoses: theoretical development for multi-stakeholder intervention design in acute care settings. *BMC Palliat Care* 21: 1–14

Lv J, Li R, Yuan L, Yang X ling, Wang Y, Ye ZW & Huang FM (2022) Research on the frailty status and adverse outcomes of elderly patients with multimorbidity. *BMC Geriatr* 22

Mahoney FI & Barthel DW (1965) FUNCTIONAL EVALUATION: THE BARTHEL INDEX. *Md State Med J* 14

Mannucci PM, Nobili A, Pasina L, Tettamanti M, Franchi C, Corrao S, Marengoni A, Salerno F, Cesari M, Perticone F, *et al* (2018) Polypharmacy in older people: Lessons from 10 years of experience with the REPOSI register. *Intern Emerg Med* 13: 1191–1200

Matschinske J, Alcaraz N, Benis A, Golebiewski M, Grimm DG, Heumos L, Kacprowski T, Lazareva O, List M, Louadi Z, *et al* (2021) The AIMe registry for artificial intelligence in biomedical research. *Nat Methods* 18 doi:10.1038/s41592-021-01241-0 [PREPRINT]

McCarthy CP, Bruno RM, Brouwers S, Canavan MD, Ceconi C, Christodorescu RM, Daskalopoulou SS, Ferro CJ, Gerdts E, Hanssen H, *et al* (2024) 2024 ESC Guidelines for the management of elevated blood pressure and hypertension: Developed by the task force on the management of elevated blood pressure and hypertension of the European Society of Cardiology (ESC) and endorsed by the European Society of Endocrinology (ESE) and the European Stroke Organisation (ESO). *Eur Heart J* 45: 3912–4018

Musacchio N, Giancaterini A, Guaita G, Ozzello A, Pellegrini MA, Ponzani P, Russo GT, Zilich R & de Micheli A (2020) Artificial intelligence and big data in diabetes care: A position statement of the Italian association of medical diabetologists. *J Med Internet Res* 22 doi:10.2196/16922 [PREPRINT]

Naik H, Murray TM, Khan M, Daly-Grafstein D, Liu G, Kassen BO, Onrot J, Sutherland JM & Staples JA (2024) Population-Based Trends in Complexity of Hospital Inpatients. *JAMA Intern Med* 184

Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD & Buchman TG (2018) An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Crit Care Med* 46: 547–553

Nicolucci A, Romeo L, Bernardini M, Vespasiani M, Rossi MC, Petrelli M, Ceriello A, Di Bartolo P, Frontoni E & Vespasiani G (2022a) Prediction of complications of type 2 Diabetes: A Machine learning approach. *Diabetes Res Clin Pract* 190

Nicolucci A, Romeo L, Bernardini M, Vespasiani M, Rossi MC, Petrelli M, Ceriello A, Di Bartolo P, Frontoni E & Vespasiani G (2022b) Prediction of complications of type 2 Diabetes: A Machine learning approach. *Diabetes Res Clin Pract* 190

Nicolucci A, Vespasiani G, Mannino D, Russo GT, Lucisano G, Rossi MC, Ponzani P, De Cosmo S, Di Cianni G, Lencioni C, *et al* (2025) A machine learning algorithm for the prediction of complications incorporated in electronic medical records improves type 2 diabetes care. *Diabetes Res Clin Pract* 229: 112900

Nobili A, Licata G, Salerno F, Pasina L, Tettamanti M, Franchi C, De Vittorio L, Marengoni A, Corrao S, Iorio A, *et al* (2011) Polypharmacy, length of hospital stay, and in-hospital mortality among elderly patients in internal medicine wards. The REPOSI study. *Eur J Clin Pharmacol* 67: 507–519

Official Journal of the European Union (2017) REGULATION (EU) 2017/ 745 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL - of 5 April 2017 - on medical devices, amending Di rective 2001/ 83/ EC, Regulation (EC) No 178/ 2002 and Regulation (EC) No 1223/ 2009 and repealing Council Directives 90/ 385/ EEC and 93/ 42/ EEC

Official Journal of the European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)Text with EEA relevance.

Official Journal of the European Union (2025) REGULATION (EU) 2025/327 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847 (Text with EEA relevance) setting out criteria for establishing and evaluating European Reference Networks and their Members and for facilitating the exchange of information and expertise on establishing and evaluating such

OSPEDALE SAN RAFFAELE - Assolombarda Servizi

Ospedale San Raffaele - Wikipedia

Overcoming Therapeutic Inertia | therapeuticinertia.diabetes.org

Palmisano A, Vignale D, Boccia E, Nonis A, Gnasso C, Leone R, Montagna M, Nicoletti V, Bianchi AG, Brusamolino S, *et al* (2022) AI-SCoRE (artificial intelligence-SARS CoV2 risk evaluation): a fast, objective and fully automated platform to predict the outcome in COVID-19 patients. *Radiol Med* 127: 960–972

Population | United Nations

Prendki V, Tau N, Avni T, Falcone M, Huttner A, Kaiser L, Paul M, Leibovici-Weissmann Y & Yahav D (2020) A systematic review assessing the under-representation of elderly adults in COVID-19 trials. *BMC Geriatrics 2020 20:1* 20: 538-

Qlik (2025) Data Quality is Not Being Prioritized on AI Projects, a Trend that 96% of U.S. Data Professionals Say Could Lead to Widespread Crises.

Raghavendra S (2023) Beginner's Guide to Streamlit with Python

Rajpurkar P, Chen E, Banerjee O & Topol EJ (2022) AI in health and medicine. *Nature Medicine 2022 28:1* 28: 31–38

Ritchie H, Rodés-Guirao L, Mathieu E, Gerber M, Ortiz-Ospina E, Hasell J & Roser M (2023) Population Growth. *Our World in Data*

Rossi MCE, Nicolucci A, Arcangeli A, Cimino A, De Bigontina G, Giorda C, Meloncelli I, Pellegrini F, Valentini U & Vespasiani G (2008) Baseline Quality-of-Care Data From a Quality-Improvement Program Implemented by a Network of Diabetes Outpatient Clinics. *Diabetes Care* 31: 2166–2168

Routinely Collected Health Data - MeSH - NCBI

Russo G, Di Bartolo P, Candido R, Lucisano G, Manicardi V, Giandalia A, Nicolucci A, Rocca A, Rossi MC & Di Cianni G (2023) The AMD ANNALS: A continuous initiative for the improvement of type 2 diabetes care. *Diabetes Res Clin Pract* 199: 110672

Said-Criado I, Pietrantonio F, Montagna M, Rosiello F, Missikoff O, Drago C, Leung TI, Vinci A, Signorini A & Gómez-Huelgas R (2025) Advancing Toward P6 Medicine: Recommendations for Integrating Artificial Intelligence in Internal Medicine. *Clin Pract* 15: 200

Searle SD, Mitnitski A, Gahbauer EA, Gill TM & Rockwood K (2008) A standard procedure for creating a frailty index. *BMC Geriatrics 2008 8:1* 8: 24-

Shickel B, Tighe PJ, Bihorac A & Rashidi P (2017) Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J Biomed Health Inform* 22: 1589

Shmatko A, Jung AW, Gaurav K, Brunak S, Mortensen LH, Birney E, Fitzgerald T & Gerstung M (2025) Learning the natural history of human disease with generative transformers. *Nature*

Skyler JS (1996) CHRONIC COMPLICATIONS OF DIABETES The Importance of Glucose Control. *VOLUME* 25

Software as a Medical Device (SaMD) | FDA

Van Spall HGC, Toren A, Kiss A & Fowler RA (2007) Eligibility Criteria of Randomized Controlled Trials Published in High-Impact General Medical Journals: A Systematic Sampling Review. *JAMA* 297: 1233–1240

STORIA, MISSION, VALORI e POLITICA per la QUALITA' OSR

Subbiah V (2023) The next generation of evidence-based medicine. *Nature Medicine 2023 29:1* 29: 49–58

Tenny S & Varacallo MA (2024) Evidence-Based Medicine. *StatPearls*

The GRADE Study Research Group (2022) Glycemia Reduction in Type 2 Diabetes — Glycemic Outcomes. *New England Journal of Medicine* 387: 1063–1074

Theou O, Haviva C, Wallace L, Searle SD & Rockwood K (2023) How to construct a frailty index from an existing dataset in 10 steps. *Age Ageing* 52: 1–7

Tinetti ME, Bogardus STJr & Agostini J V. (2004) Potential Pitfalls of Disease-Specific Guidelines for Patients with Multiple Conditions. *New England Journal of Medicine* 351: 2870–2874

Tinetti ME & Fried T (2004) The end of the disease era. *American Journal of Medicine* 116: 179–185

Tinetti ME, Green AR, Ouellet J, Rich MW & Boyd C (2019) Caring for patients with multiple chronic conditions. *Ann Intern Med* 170 doi:10.7326/M18-3269 [PREPRINT]

United Nations (2015) 70/1. Transforming our world: the 2030 Agenda for Sustainable Development Transforming our world: the 2030 Agenda for Sustainable Development Preamble.

Vassallo M (2019) Research and reducing inequity in healthcare. *Age Ageing* 48 doi:10.1093/ageing/afz051 [PREPRINT]

Wald R, Jaber BL, Price LL, Upadhyay A & Madias NE (2010) Impact of Hospital-Associated Hyponatremia on Selected Outcomes. *Arch Intern Med* 170: 294–302

What is an Internal Medicine Physician, or Internist? | ACP Online

Wiest IC, Wolf F, Leßmann ME, van Treeck M, Ferber D, Zhu J, Boehme H, Bressem KK, Ulrich H, Ebert MP, *et al* (2025) A software pipeline for medical information extraction with large language models, open source and suitable for oncology. *npj Precision Oncology 2025 9:1* 9: 313-

Wilkinson C, Wu J, Searle SD, Todd O, Hall M, Kunadian V, Clegg A, Rockwood K & Gale CP (2020) Clinical outcomes in patients with atrial fibrillation and frailty: insights from the ENGAGE AF-TIMI 48 trial. *BMC Medicine 2020 18:1* 18: 401-

Winter JE, MacInnis RJ & Nowson CA (2017) The influence of age on the BMI and all-cause mortality association: A meta-analysis. *Journal of Nutrition, Health and Aging* 21

Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J & Myles P (2019) Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol* 48: 1740–1740g

Yan M, Hong H, Wilson J & Goldstein BA (2025) Estimating the observability of an outcome from an electronic health record data set using external data. *Am J Epidemiol* 194: 3224–3432

Yu C, Xian Y, Jing T, Bai M, Li X, Li J, Liang H, Yu G & Zhang Z (2023) More patient-centered care, better healthcare: the association between patient-centered care and healthcare outcomes in inpatients. *Front Public Health* 11