Christoph Eberharter, BSc

Predictive analytics in a Smart Heart Failure Registry

Master Thesis

to achieve the university degree of

Master of Science

Master's degree programme: Biomedical Engineering

submitted to

Graz University of Technology

Supervisor

Priv.-Doz. Dipl-Ing. Dr.techn. Günter Schreier, MSc

Institute of Neural Engineering
Head: Univ.-Prof. Dipl-Ing. Dr.techn. Gernot Müller-Putz

Graz, October 2024

This master thesis has been conducted
in cooperation with:

AIT Austrian Institute of Technology GmbH
Center for Health & Bioresources
Digital Health Information Systems

Supervisor

Martin Baumgartner, MSc

Reininghausstraße 13/1

8020 Graz

Austria

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

02.10.2024

Date

Christoph Eberhardter

Signature

Kurzfassung

Die täglich wachsende Menge an Gesundheitsdaten unterschiedlicher Form und Herkunft, sowie die zunehmende Entwicklung neuartiger Analysemethoden bergen das Potential für eine personalisierte Gesundheitsversorgung. Das Digital Health Information Systems Team am AIT Austrian Institute of Technology entwickelt eine Infrastruktur, die hier ansetzt: Mittels datenschutzgerechter Aggregation und Standardisierung von Gesundheitsdaten, sowie smarten Analysemethoden zur Erstellung prädiktiver Modelle sollen Services zur maßgeschneiderten Patientenbehandlung ermöglicht werden. Dies wurde in einem ersten Pilotprojekt in einem sogenannten "smarten" Register für Patienten mit chronischer Herzinsuffizienz implementiert.
Ziel dieser Arbeit war es die Entwicklung rund um dieses smarte Register zu unterstützen, indem zunächst eine Verifikationsanalyse durchgeführt wurde, bei der die Registerdaten mit den Daten einer wissenschaftlichen Publikation verglichen wurden. Dabei konnte gezeigt werden, dass basierend auf den automatisch synchronisierten Daten des smarten Registers bereits ein Großteil der im Vergleich dazu manuell gesammelten Publikationsdaten repliziert werden konnte. Durch diese Analyse konnten vor allem auch Fehler und Abweichungen in den Registerdaten identifiziert und somit die Datenqualität verbessert werden. Des Weiteren wurde ein einfaches Machine-Learning-Modell entwickelt, welches für Registerpatienten, welche auch im Telemonitoring-Programm "HerzMobil Tirol" teilnahmen, eine Verlängerung um eine zweite Telemonitoring-Periode voraussagt. Das Vorhersagemodell erzielte dabei gemischte Performanceresultate. Gleichwohl konnte damit ein erster Anwendungsfall für eine Prädiktion auf Basis von aggregierten Gesundheitsdaten im smarten Register demonstriert werden.
Weitere Forschung zur Vernetzung von Gesundheitsdaten kombiniert mit der Anwendung modernen Analysemethoden ist notwendig, um die personalisierte und datengetriebene Gesundheitsversorgung voranzutreiben.

Abstract

The daily growing amount of health data of various forms and sources, as well as the increasing development of novel analysis methods, hold the potential for personalized healthcare. The Digital Health Information Systems team of the AIT Austrian Institute of Technology is developing an infrastructure that addresses this issue: Through privacy-preserving aggregation and standardization of health data, as well as smart analysis methods to create predictive models, services for customized patient treatment are to be made possible. This was implemented in an initial pilot project in a so-called "smart" registry for patients with chronic heart failure.
The aim of this work was to support the development of this smart registry by first carrying out a verification analysis in which the smart registry data was compared with the data from a scientific publication. It was shown that, based on the automatically synchronized data of the smart registry, a large part of the manually collected publication data could already be replicated. This analysis also made it possible to identify errors and deviations in the smart registry data and thus improve the data quality. Furthermore, a basic machine learning model was developed that predicts an extension of a second telemonitoring period for smart registry patients who also participated in the "HerzMobil Tirol" telemonitoring programme. The prediction model achieved mixed performance results. Nevertheless, it was possible to demonstrate a first use case for prediction based on aggregated health data in the smart registry.
Further research on the networking of health data combined with the application of modern analytical methods is necessary to advance personalized and data-driven healthcare.

Contents

Acronyms

ACE-I	Angiotensin-Converting Enzyme Inhibitor
AHF	Acute Heart Failure
AI	Artificial Intelligence
AIT	Austrian Institute of Technology
API	Application Programming Interface
ARNI	Angiotensin Receptor-Neprilysin Inhibitor
AUROC	Area Under the Receiver Operating Characteristic Curve
BB	Beta-Blocker
CDM	Common Data Model
CHF	Chronic Heart Failure
EMR	Electronic Medical Record
HF	Heart Failure
HFrEF	Heart Failure with reduced Ejection Fraction
HIS	Hospital Information System
HMT	HerzMobil Tirol
KIT	Keep in Touch
LVEF	Left Ventricular Ejection Fraction
ML	Machine Learning
MRA	Mineralocorticoid Receptor Antagonist
NYHA	New York Heart Association
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcome Partnership
PATH	Predictive Analytics Toolset for Healthcare
QOL	Quality of Life

1 Introduction

This introductory chapter provides information about heart failure and its conventional and telehealth-supported treatment options. Furthermore, the disease management programme HerzMobil Tirol (HMT) and the recently concluded project D4Health Tirol are presented. Finally, this leads to the description of possible data-driven methods for improved healthcare, especially in the context of heart failure patients.

1.1 Heart failure

1.1.1 Defintion, classification, epidemiology

Heart failure (HF) is not defined as a single pathological diagnosis, but as a clinical syndrome. Therein, symptoms of breathlessness, ankle swelling, and fatigue go hand in hand by signs of a jugular venous pressure, pulmonary crackles, and peripheral oedema. This is due to a structural and/or functional abnormality of the heart, which results in elevated intracardiac pressures and/or inadequate cardiac output at rest and/or during exercise [1]. A more traditional and easier to understand, but thus somewhat inaccurate definition of HF is the reduced pumping ability of the heart, resulting in insufficient oxygen being supplied to the body [2].

The condition and progression of HF can be divided into several categories. Traditionally, it is subdivided based on the measurement of left ventricular ejection fraction (LVEF) because of different therapeutic approaches:

- HF with reduced LVEF (HFrEF) of $\leq$40%
- HF with a mildly reduced LVEF (HFmrEF) between 41% and 49%

- HF with a preserved LVEF (HFpEF) of $\geq$50%

The New York Heart Association (NYHA) functional classification is another commonly used and simple system to describe the severity of HF. Based on symptoms and functional capacity of the patients, it categorizes HF on a scale of I to IV, where I refers to the mildest and IV to the most severe manifestation.

Usually, HF is divided into acute heart failure (AHF) and chronic heart failure (CHF). Patients with CHF have been diagnosed with HF in the past or obtain a more gradual onset of symptoms. AHF on the other hand may be presented by a rapid onset or progressively escalating of symptoms and/or signs of HF. This can be due to an acute singular event or the deterioration of CHF. The latter, which is also described as decompensated HF, is the more common reason for patient hospitalization. [2]

Studies estimate the prevalence of known HF in developed countries to be 1-2% of the adults, increasing to >10% in those aged 70 and over. The 5-year mortality of all-type HF patients is estimated to be 57%. Of all hospital admissions, HF hospitalizations represent 1-2%. The diagnosis HF is the most common in hospitalized patients aged >65 years and associated with the highest 30-day readmission rate. [3]

1.1.2 Conventional treatment

There are several approaches to treat HF patients, which are documented as recommendations by the European Society of Cardiology in their guidelines [1]. Their common goals focus on the reduction of mortality, the prevention of recurrent hospitalization, and the improvement of the patients' clinical status, functional capacity, and quality of life (QOL).

First and foremost, pharmacotherapy is used for HF treatment. Here, the drug groups of angiotensin-converting enzyme inhibitors (ACE-I) or angiotensin receptor-neprilysin inhibitors (ARNI), beta-blockers (BB), and mineralocorticoid receptor antagonists (MRA) serve as a foundation for therapy. They show to improve survival, reduce the risk of HF hospitalization,

and reduce symptoms in patients with HFrEF. Angiotensin-receptor blockers (ARB) for patients intolerant to ACE-I or ARNI, and sodium-glucose co-transporter 2 (SGLT2) inhibitors are added to this therapy. Other recommended drugs are diuretics to handle patients' congestion, and I_f-channel inhibitors.

Cardiac rhythm management using implantable cardioverter-defibrillators (ICD) and cardiac resynchronization therapy (CRT) is another measure for the treatment of HFrEF.

In addition to drug and device-based HF therapy, increased attention has focused on how HF care is delivered and the need for multidisciplinary management of CHF. An important area here is patient education for self-care. Patients with an improved knowledge for their HF condition benefit through better QOL, lower readmission rates, and reduced mortality [1]. Subsequently, there is consistent evidence that patients also benefit from physical conditioning through exercise rehabilitation. Further, an appropriate follow-up for HF patients discharged from the hospital is recommended to ensure continued optimal therapy and detect asymptomatic progression of HF. Lastly, telemonitoring is used to support and optimize care of HF patients. [1]

1.1.3 Telehealth-supported treatment

The high readmission rate and increased risk of death in the early period after discharge from hospital reflect the elevated vulnerability of patients in the post-discharge period after HF admission. In addition, the treatment costs caused by readmissions contribute substantially to the overall economic burden on the healthcare systems. This is despite the fact that an estimated two-thirds of triggering factors for HF readmission could potentially be avoided. These include suboptimal discharge planning, inadequate follow-up, non-adherence to heart failure medication, insufficient social support, and also delays in seeking medical attention. [4]

For these reasons, telehealth-based approaches for HF treatment have been widely investigated and compared to usual care in the past. Multiple studies demonstrate that post-discharge disease management programmes can

lead to a reduction of readmissions, mortality, and healthcare costs [4]. The remote monitoring of HF should achieve earlier identification of decompensation risk, better adherence to lifestyle changes and medication, and interventions that reduce the need for hospitalization [5].

Telehealth systems provide healthcare services using telecommunication technologies, and implementations vary widely [6]. The Keep-in-Touch (KIT) Telehealth Solutions Platform, developed by the AIT Austrian Institute of Technology (AIT), offers such a disease management programme for HF. With HMT, a first major programme has been implemented here. [7]

1.2 HerzMobil Tirol

HMT is a multidisciplinary post-discharge disease management programme for heart failure patients. It consists of a telemedical monitoring system used by a comprehensive network of specialized HF nurses, local physicians, and secondary/tertiary referral centers. Since 2012, the project has gone through a number of project phases, and is in routine operation in the Austrian state of Tyrol since 2017. One year later, HerzMobil was also implemented in the state of Styria (HerzMobil Steiermark) [8] and in 2022, also in the state of Carinthia (HerzMobil Kärnten).

For a period of 3 months, patients are cared for in an integrated care network. Therein, network physicians and HF nurses are responsible for the monitoring of telemedical patient data, face-to-face visits, and phone contacts with patients if required. In the HMT network, optimal patient treatment is ensured by regular communication between all participants. Further, a web-based telehealth software supports all stakeholders in their individual tasks. [9]

During the programme, each patient is provided with a blood pressure and heart rate monitor, a bodyweight scale, and a mobile phone capable of near-field communication. After being trained on this equipment by nursing staff, easy and secure data acquisition and transmission of blood pressure, heart rate, bodyweight, self-reported well-being, and drug intake can be performed by patients every day with the KIT technology. This data is

analyzed by a certified medical product, to offer healthcare professionals a rule-based notification system, if any values are above or below specified thresholds. This should indicate a patient's need for closer inspection and focus awareness on patients who may need early therapeutic intervention. [9]

The HMT programme has proven to be feasible and effective in clinical practice, demonstrating reduced readmissions and all-cause mortality in HF patients compared to conventional care. [8]

1.2.1 D4Health Tirol

In addition to HerzMobil Tirol, several innovative solutions for chronically ill patients in Tyrol have already been implemented in the past as a result of a collaboration between the Landesinstitut für Integrierte Versorgung (LIV) and the AIT. Based on that, personalized patient care by the networking and smart use of health data is expected to be focused on in the future. Digital & Data-Driven Decisions for Health & Care (D4Health) applications are to be increasingly used for this purpose. [10]

Therefore, the goal of the D4Health Tirol project is to establish novel methods of health data analytics and predictive modeling. Specifically, machine learning algorithms are used to generate predictive models that can provide decision support in patient care. Furthermore, a linkage of multiple data sources is established with so-called smart registries. For example, clinical data such as those from hospital information systems (HIS) are combined with telehealth data such as those from the HMT programme. The data model used for this purpose is described in more detail in the next chapter. [10]

A raw overview of the IT architecture of D4Health Tirol can be seen in figure 1.1. It depicts the lifecycle from data extraction, through data processing and modelling, to providing predictive services. After merging different data sources and the pseudonymization of the data, it is available in a standardized data store. Using application programming interfaces (APIs), raw data can be exported from the data store and used to calculate features. The features can then be stored in a central feature store and further used

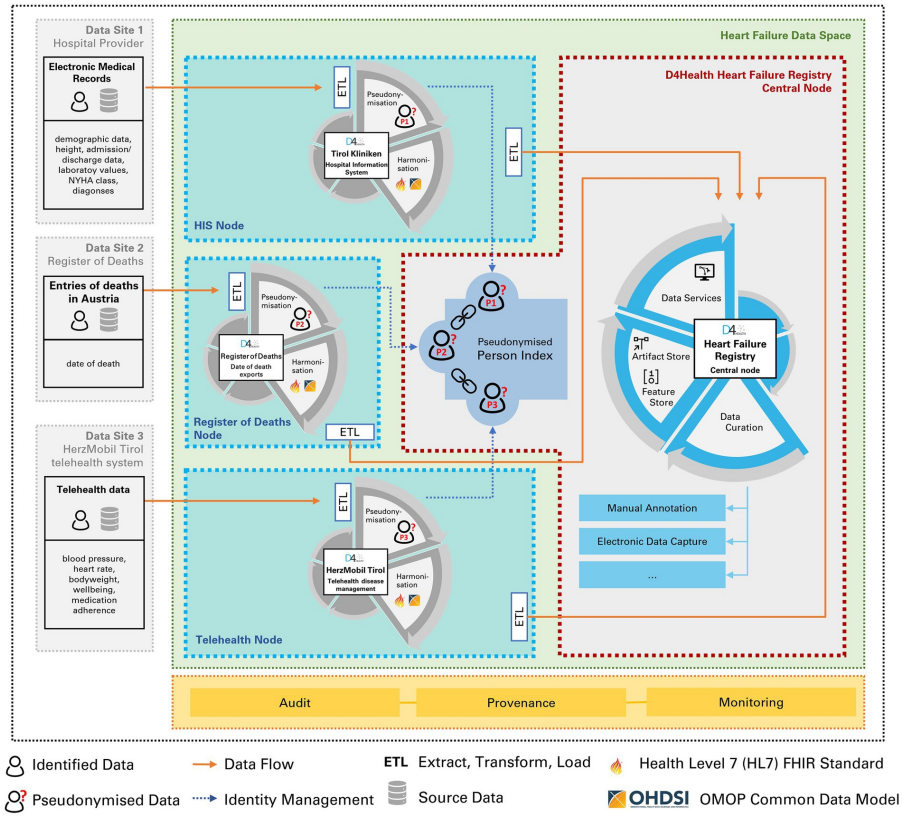

Figure 1.1: Architectural overview of the IT architecture of the D4Health Heart Failure Registry [12]

to create models for the model store. The "Predictive Analytics Toolset for Healthcare" (PATH) developed by the AIT can be used to support the development of predictive models from telemedicine data, but also any other external modelling environment can be used here [11]. For a first pilot in D4Health Tirol, the care programme HerzMobil Tirol was selected and a smart registry for CHF patients was set up. [10]

1.2.2 OMOP CDM

As medical data can vary greatly in structure and is collected for different purposes, it may be stored in different formats using different database

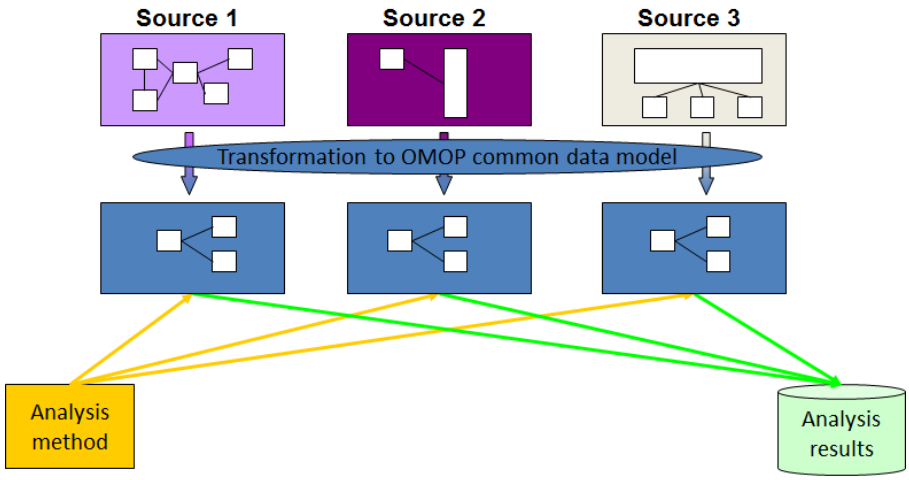

Figure 1.2: Basic schema of the OMOP CDM concept [13]

systems and information models. Data standardization is the critical process of bringing this data into a common format, allowing collaborative research and large-scale analytics. The Observational Medical Outcome Partnership (OMOP) Common Data Model (CDM) is an open community data standard, developed by the Observational Health Data Sciences and Informatics (OHDSI) initiative. It is designed to standardize the structure and content of observational data and also to enable efficient analyses that can lead to reliable findings. [13]

A basic schema of the OMOP CDM concept can be seen in figure 1.2. Systematic analysis of observational databases of diverging structure is made possible by transforming the data contained in these databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes), and then performing systematic analyses using a library of standard analysis routines written based on the common format. [13]

The AIT uses the resources of the OHDSI to perform data conversions, whereby heterogeneous data from different sources can be combined in an OMOP CDM. This was used in the smart registry for CHF to aggregate data such as from a HIS or the HMT programme into a standardized format, allowing for clean further processing.

1.2.3 Current status

The smart registry for CHF patients has already been implemented in a real world setting and is automatically synchronized with various health data sources. It is based on an infrastructure which enables pseudonymization and privacy-preserved record linkage of cross-domain health data [14]. Further, standardization is implemented by using, for example, the OMOP CDM and the Health Level 7 (HL7) FHIR data exchange standard. Finally, by applying machine learning, supported by the feature and model store, personalized healthcare can be enabled through smart decision support services.

In addition, this infrastructure works in the spirit of the European Health Data Space proclaimed by the European Union, which aims at a secure and efficient exchange of health data within and across national health care systems. [15]

Lastly, the smart registry aggregated health data of 5004 HF patients coming from the HMT programme, the HIS of the Tirol Kliniken GmbH, and an extract of the Austrian National Register of Deaths. A more detailed description of the data of the smart registry can be found in chapter 2.1.

1.3 Data-driven methods to improve care

1.3.1 Artificial intelligence

The term artificial intelligence (AI) can be broadly described as the use of computers to model intelligent behavior with minimal human involvement [16]. It is used in engineering to solve complex tasks through novel concepts and solutions.

Machine learning (ML) is a subfield of AI and is based on enhanced learning through experience. Deep learning can be further considered as a specific form of ML with even less human preprocessing by using complex algorithms and deep neural networks to train a model. [17]

ML algorithms can be divided into three categories. Unsupervised algorithms can discover patterns or features from data. Supervised algorithms can be used to make predictions and classifications based on past data. Reinforcement learning is used to solve tasks in a specific problem space with training through rewards and punishments. [16]

More and more health data are available and at the same time the development of big data analytics is growing. This has recently led to the increased successful application of AI in healthcare. Powerful AI algorithms can extract clinically relevant information and features from vast masses of healthcare data, assisting doctors in their decision-making by predicting health status and intervening in real-time on health risk alerts. [18]

To do this, AI systems must first be trained with clinical data (demographics, records from medical devices, clinical laboratory, medical notes, images, etc.) to learn to recognize patterns in the provided features and how they relate to the desired outcome. [16]

AI can be applied to a wide variety of healthcare data, which is often divided into structured and unstructured data. Structured data such as electrophysiological data are analyzed using ML techniques such as support vector machines and neural networks to cluster patient characteristics or conclude the probability of disease outcomes. Unstructured data such as medical notes are processed with natural language processing methods to extract information that can be further analyzed as machine-readable structured data. [18]

To accelerate the application of AI on electronic medical records (EMR), data from laboratories and clinics must be aggregated and made available in real time in order to implement AI systems that can generate clinically relevant knowledge to support clinical decision-making for cost-effective personalized patient care. [16]

For a deeper insight, Eric Topol describes the multifaceted potential of AI in healthcare and the accompanying transformation towards data-driven and patient-centered care in detail in his book "Deep Medicine". [19]

1.3.2 Machine learning examples

AI is becoming more prevalent in healthcare literature, but is mostly concerned with diseases related to cancer, nervous system, and cardiovascular system, as these are among the leading causes of death [18].

Gontarska et al. investigated a ML model that predicts the risk of a patient requiring an intervention based on the patient's daily vital parameters. The data came from a telemedicine study of HF patients who were in NYHA stages II or III. Using their model, they predicted the daily per-patient risk of being in a medically critical condition and then sorted the patients by this estimated risk. Their deep neural network model reached an area under the receiver operating characteristic curve (AUROC) of 0.84 and with that outperformed a rule-based model with an AUROC of 0.73. This approach was to help medical practitioners focus their limited capacities on the most critical patients. [20]

Another example would be the EMR-based study of Schrempf et al., were they developed ML-based models for the 5-year risk prediction of major adverse cardiovascular events (MACE), such as myocardial infarction or stroke. MACE may be prevented by identifying patients at risk at an early stage, and with a random forest model an AUROC of 0.88 was achieved there. [21]

In a more recent study, Herman et al., used longitudinal patient data to predict all-cause 30, 90-, 180-, 360-, and 720-day mortality of patients with a new onset or worsened HF. Based on the combination of a wide variety of electronic health data recorded in the standard care setting, their ML-based algorithm achieved a robust AUROC performance ranging from 0.83 to 0.89, suggesting its potential in point-of-care clinical risk stratification. [22]

1.3.3 Opportunities of AI for HerzMobil Tirol and its patients

Just as described in the previous chapter, AI can also open opportunities in the context of HMT and its patients. The basis for this is the smart registry for CHF, which allows data analysts to easily work with features and models

of patient data. From this, predictions about the health status of patients can be made, supporting healthcare professionals in their decision-making.

For example, predictions of adverse events, such as death, rehospitalization, or an adjustment of medication could be made, allowing care plans to be adapted accordingly. Furthermore, similar to the examples described in the previous chapter, different types of risk stratification could be implemented to improve patient outcome and optimize resources. In this way, smart decision support services could contribute to personalized patient care and thus advance data-driven healthcare.

1.4 Aims

The objectives of this thesis have been divided into two parts. As a basis for this, a pre-processing of the existing health data from the smart registry for CHF should be made with PATH and MATLAB. Based on this, a quality assurance step of the smart registry should then be performed in the first part using a verification approach. For this purpose, the automatically synchronized data from the smart registry should be compared with the manually collected data of a recent study concerning the HMT programme [8]. In the second part, a basic prediction model should be implemented as a proof of concept. More precisely, the model should predict whether a patient within the HMT programme would need an extension, i.e., a second 3-month telemonitoring period, after his or her first 3-month period.

2 Methods

This chapter first describes the smart registry and its current data. This is followed by a brief description of the work with the smart registry data, which served as the basis for the analyses of the two core topics of this thesis. Then, the verification analysis of the smart registry is described, with its data overview, processing, and comparison. In addition, a post-analysis update is presented, which was carried out after a data update in the smart registry. Finally, the development of a simple prediction model based on the smart registry data is shown.

The analyses were approved by the Ethics Committee of the Medical University of Innsbruck (vote no. 1035/2022).

2.1 Smart Heart Failure Registry

The smart registry for CHF patients is still an ongoing project in development and at the time of writing held health data from a total of 5004 HF patients. The data came from three different sources and were aggregated in the smart registry according to the privacy preserving record linkage and in the format of the OMOP CDM.

The first data source was the HMT system and included 960 patients who participated in the HMT programme, encompassing a total data transmission period from April 2016 to November 2022. Accordingly, the smart registry included HMT patients' self-reported daily data on blood pressure, heart rate, body weight, well-being, and medication compliance. In addition, a NYHA score, which is usually determined during the initial admission in the hospital before inclusion into HMT, was available for most patients. Furthermore, in HMT, patient-specific lower and upper thresholds for blood

pressure, heart rate, and body weight are set by the patients' respective physicians. These thresholds, which are used for automatic detection of possible deterioration of health status within the HMT network, were also loaded into the smart registry. Finally, in addition to basic patient data, such as demographics, clinical free text notes are also included.

The second major data source of the smart registry consisted of an extract from the HIS of Tirol Kliniken GmbH. This contains EMR data of HF patients, which were selected as such under certain filter criteria. These included at least one diagnosis from a list of ICD-10 codes and a minimum value of the biomarker NT-proBNP in the past, a previously established HF diagnosis, or if the patient had previously been included in the HMT programme. The HIS export covered the period from January 2016 to April 2023. Through record linkage, the existing 960 HMT patients in the smart registry were assigned their data from the HIS, if available, and all remaining 4044 patients were newly created. The HIS extract consisted of patient master data (e.g., name, birthday, sex, date of death with reason), physiological measurement data (e.g., blood pressure, pulse, weight), laboratory data (e.g., sodium, troponin, NT-proBNP), as well as date and reason of outpatient and inpatient admissions.

The third data source consisted of an extract from the Austrian National Register of Deaths, which was obtained within the D4Health Tirol project for HMT patients. This contained the date of death of deceased HMT patients until June 2022 and was accordingly added to these patients in the smart registry.

As mentioned above, the smart registry uses the open community data standard OMOP CDM to aggregate the health data from these different sources in a standardized format. The data is mainly managed in the standardized clinical data tables, which can be seen in detail in the current version of the OMOP CDM in figure 2.1. The standardized vocabularies used, such as SNOMED-CT or ICD-10, are also an important component here, as they ensure a uniform representation of medical concepts/terms and thus interoperability at data level. For a brief insight, some of the most important tables used are described in the following.

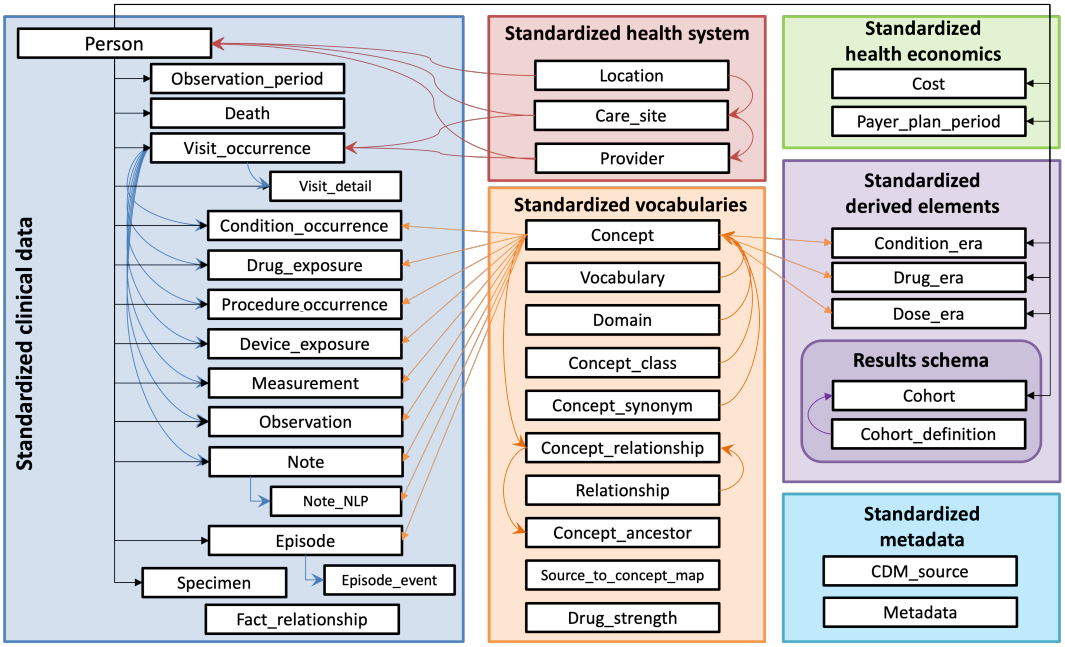

Figure 2.1: OMOP Common Data Model domains and tables [13]

The table "person" contains all basic data for unique identification of patients, such as a unique person ID, date of birth and gender. The table "measurement" contains all records of standardized measurements of persons, such as vital signs or laboratory values, in a structured form with measurement concept and associated measurement value. The "observation" table includes clinical facts about examinations or questionnaires of a person, which cannot be represented by other domains. In the case of the smart registry, for example, this corresponds to the patient's self-assessed well-being or their information on medication compliance. The tables "visit_occurrence" and "condition_occurrence" contain events about contacts with health care facilities, as well as information about diseases or diagnoses of persons. [13]

Of the 5004 patients, over 2.9 million measurements and over 570,000 observations were in the smart registry at last count. In total, over 5.2 million data points were available. A screenshot of the standard dashboard layout of the smart registry is shown in figure 2.2.

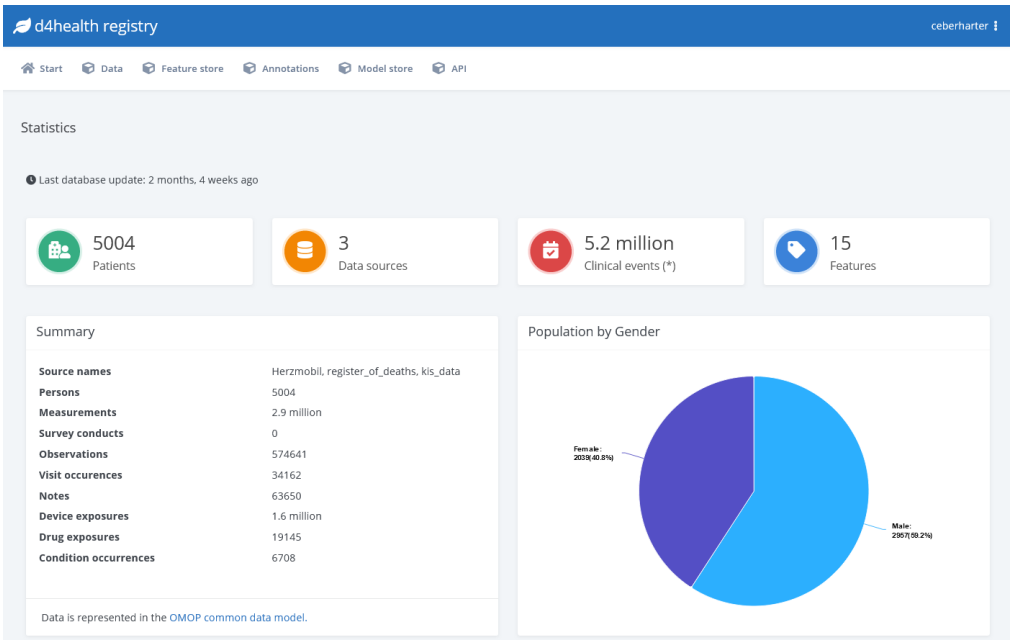

Figure 2.2: Screenshot of the default dashboard of the Smart Heart Failure Registry

2.2 Data preprocessing

The PATH software tool developed by AIT was used to start processing the health data from the smart registry data store. PATH, which is based on MATLAB R2022a (The MathWorks, Inc., Natick, Massachusetts, United States), supported data analysis and visualization. Each data processing step could be specified using definition files in Microsoft Excel and visualized in the associated PATH app with graphical user interface.

To load patient data tables from the data store, the API of the smart registry was used. SQL queries with the desired tables were defined in PATH, and the file format (JSON or CSV) in which they should be loaded from the smart registry was also specified. After all desired tables (person, measurement, observation, etc.) were exported from the smart registry in JSON format, they were further processed as MATLAB tables. In this process, all relevant information from the tables containing patient health data was aggregated into a full data table in MATLAB via several intermediate steps. Thus, this

table contained all data of the HF patients from the smart registry, and then served as the basis for all further processing.

2.3 Smart registry verification analysis

As the process of developing such a smart registry is quite complex, verification is an essential part of it. Therefore, the first task of this thesis was to analyze, to which extent the smart registry is capable of replicating data from the recent study on the feasibility and effectiveness of the HMT programme [8]. For better readability, this study will be further referred to as the "HMT study." This approach was intended to verify that the smart registry was correctly populated with data and to demonstrate a safe and efficient way to analyze health data. This first part of the present master thesis also resulted in a paper published at the dHealth 2023 conference [23].

As the development of the smart registry was an ongoing process whilst working on the thesis, the available data and with that the results of this analysis naturally changed throughout time. Therefore, a post-analysis update on how these changes have affected the analysis is provided at the end of each of the Methods, Results, and Discussion chapters of this verification analysis.

2.3.1 Data overview

For this replication attempt, a table regarding the HMT study data was available as an Excel file. This contained health data of 251 HMT patients and 257 control group (conventional care) patients over the period from April 2016 to October 2019. The data consisted of baseline characteristics and outcome data for each patient, summarized in 96 variables. These included demographic data such as age and sex, physiological measures such as blood pressure, heart rate, and body weight, or laboratory values such as creatinine, NT-proBNP, and sodium. Furthermore, previous diseases or

diagnoses at the time of HMT initiation, as well as data on hospitalizations and death during and after the HMT programme were included.

The smart registry, on the other hand, contained health data for 4680 HF patients at the time of this analysis. Of these, 506 were from the HMT programme, with a data transmission period from April 2016 to June 2021. As already described in chapter 2.1, these were HMT data such as demographics, the measurement data acquired by the patients themselves during telemonitoring, or NYHA scores. The remaining 4174 patients came from the HIS of the Tirol Kliniken GmbH with patient records between June 2022 and January 2023. These were EMR data as also described in chapter 2.1 (demographics, vital signs, diagnoses, etc.). Since the time period of these EMR data from the HIS did not overlap with that of the HMT study, no direct comparison was originally possible here for this analysis. However, a statement could be made about a possible replicability of variables regarding these data based on the existing tables in the smart registry. Finally, the extract from the Austrian National Register of Deaths with the date of death of HMT patients was linked into the smart registry.

2.3.2 Data processing

As described in chapter 2.2, the data from the smart registry had already been preprocessed in PATH and MATLAB, and were then available as a full data table for further processing. Furthermore, the 257 patients in the conventional care group of the HMT study were not considered for this analysis. Of the 251 patients in the HMT study who participated in the HMT programme, 248 patients were initially matched in the smart registry.

For the sake of clarity, the 96 HMT study variables were divided into 9 categories for analysis, which can be seen in table 3.1 in the Results chapter. For each of the 248 HMT patients, all variables were finally recalculated as far as possible using the data existing in the smart registry with PATH/MATLAB. An overview of the data processing workflow for the verification analysis is shown in figure 2.3.

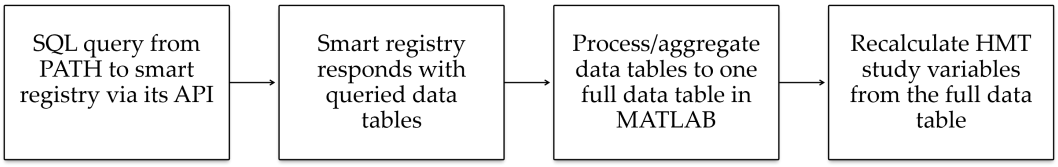

Figure 2.3: Overview of the data processing workflow of the smart registry verification approach

2.3.3 Data comparison

For the comparison of the two datasets, the Excel table of the HMT study was also converted to a MATLAB table. Furthermore, the quality of replicability of the 96 variables from the smart registry was classified into the following 4 classes:

- Not replicable variables: Variables that could not be replicated from the smart registry, as no tables in the smart registry with information regarding these variables existed.
- Theoretically replicable variables: Variables that would be replicable from the existing smart registry tables if these tables were appropriately filled with data.
- Time-deviating replicable variables: Variables that could actually be recalculated from the smart registry tables, but the time range of the data did not match that of the HMT study.
- Correctly replicable variables: Variables that could be replicated from the smart registry tables, where the time range of the data also matched that of the HMT study.

2.3.4 Post-analysis update

Since the completion of the described analyses, the smart registry had evolved and received additional data (see Chapter 2.1). This increased the number of time-deviating and correctly replicable variables, and all 251 HMT patients of the HMT study were also to be found in the smart registry at last. Of particular interest to the verification analysis was the extension of the HIS extract, back to 2016. Thus, the time period of the EMR data of

126 of the 251 patients in the smart registry overlapped with that from the HMT study, allowing a first direct comparison regarding the variables of these data.

Subsequently, a PATH-internal function was used, which made it possible to compare any number of variables from two tables with each other. The values of all variables that could be correctly replicated in MATLAB from the smart registry were compared with the values of the corresponding variables from the table of the HMT study. With this compare function, congruence, correlation, and the percentage of finite elements between the respective variables of the two tables could be calculated. The congruence describes to what extent the variable values of the respective patients in the comparison of the two tables exactly match (degree of similarity). The correlation compares the variables of both tables using linear regression. The percentage of finite elements reflects the proportion of finite values in the respective variable category, i.e. values which were not NaN (not a number) when comparing the two tables. Some variables for which it did not make sense to apply a certain comparison metric, such as correlation between categorical variables like patients' assigned hospital, were excluded in the process. The results of the three metrics were then averaged for the respective variable categories for clarity.

2.4 HMT extension necessity prediction model

The second goal of this master's thesis was to try to develop a simple prediction model based on the existing data in the smart registry. The aim was not to design an elaborate complex model, but to implement a basic machine learning model within the smart registry infrastructure as a proof-of-concept, and in turn to support the development of the smart registry with a use case. One question within the HMT environment that was of interest in this regard was whether an extension of a further 3 months is recommended for a patient after their initial 3-month stabilization phase. This can be decided individually by the healthcare professionals caring for each HMT patient based on the development and current state of his or her health before the end of the initial telemonitoring period. An automated tool

could support and improve this decision-making process on the necessity of an extension by means of objective criteria. Therefore, an attempt was made to design a model for predicting an extension based on the aggregated health data of HMT patients in the smart registry. Data pre-processing for this was done within MATLAB and PATH. Further processing and actual modeling were carried out in SPYDER as a programming environment using the Python programming language.

2.4.1 Data overview and processing

MATLAB
Based on the pre-processed full data table with the smart registry data of all 5004 patients, the further processing steps were carried out in this second part of the PATH experiment. For the given question, all HMT smart registry patients were included who had been hospitalized at least once before or during their initial HMT period and thus had HIS data in the smart registry. The health data of this patient group from the full data table was then mapped onto daily level to form "daily reports", which contained all available information for one patient on a specific date. Relevant variables were derived and calculated from the raw health data in the OMOP format of the full data table as so-called features. These were basic features such as age, gender or blood pressure values, as well as more complex ones such as the Charlson Comorbidity disease categories derived from ICD-10 diagnoses or the adherence of patients and physicians on the prescribed heart failure medication (for the complete feature list see Appendix B). Missing feature values (NaNs) were filled with the last observation carried forward method between the daily reports [24]. As a decision is made in the last days of the initial HMT period as to whether a patient is extended or not, the daily reports of all patients were logically removed after the last telemonitoring day of the stabilization phase. The information as to whether an HMT patient had an extension was previously generated and stored in the smart registry as a feature in the feature store and could therefore be easily imported via PATH and then assigned to the patients in the daily reports as a target with a binary label (i.e. "0" for no extension, "1" for extension).

In the end, this resulted in 298 patients, of which 69 patients were extended, with a total of 32650 daily reports and 56 features over a data period from January 2016 to July 2022. This 32650 x 56 matrix was then exported as a csv file for subsequent modeling in Python.

Python
For a functioning modeling, features were first filtered, which consisted of at least 99% missing values (NaNs). This applied to some features relating to Charlson Comorbidities, as the “condition_occurrence” table of the large HIS export was not available in the smart registry at the time of the analyses. All other missing values were filled in using an initially patient-wide and then dataset-wide median. All features of the resulting feature matrix were then normalized to the value range [0, 1] using dataset-wide min-max scaling. The daily reports of each patient were then sampled into windows of 12 daily reports with an overlap of 3 reports. The 12-day windows were then assigned the target regarding extension "No" or "Yes" (binary label "0" or "1") according to the daily reports. These labels were additionally smoothed with a Gaussian noise and shortened to avoid "hard" labels [25]. The principle of this label smoothing is visualized in figure 2.4. Since only 69 of 298 patients were included in an extension period, the data set was naturally unbalanced, which is why basic data augmentation was also used to increase this balance. Thereby, a twin patient was created from each extension patient, in which all variables and parameters were modified with either random noise (for continuous values such as blood pressure), random choice (for categorical categories such as gender) or a random constant offset (for constant values such as age).

2.4.2 Predictive model

A residual neural network was used to predict the necessity of an extension. The Tensorflow implementation of ResNet50V2, which is a convolutional neural network with 50 layers, was trained for 50 epochs with a batch size of 32 samples and the Adam optimizer. Furthermore, the Keras implementation of binary cross entropy was used as a loss function. The evaluation of the

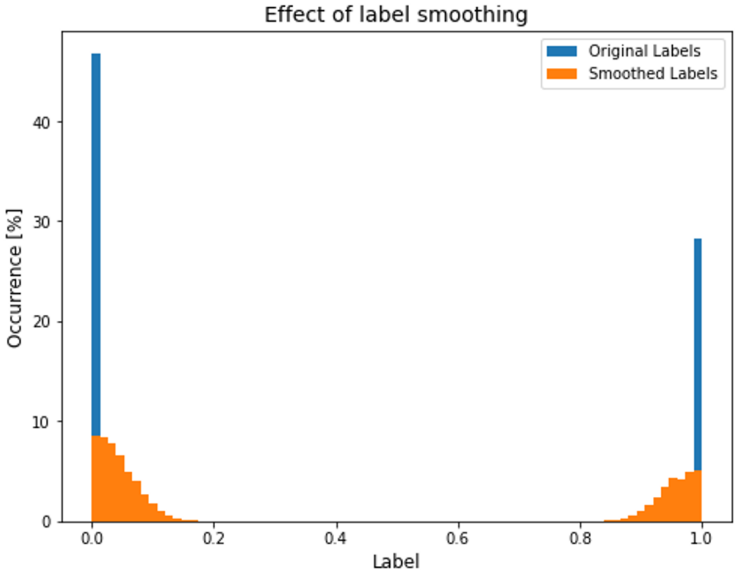

Figure 2.4: Label smoothing effect visualized. Instead of "hard" values of "0" and "1" (blue), target values are smoothed (orange).

model was based on a 5-fold cross-validation split with a test:train ratio of 20:80 in each fold.

3 Results

This chapter presents the results of the smart registry verification analysis and the HMT extension necessity prediction model.

3.1 Smart registry verification analysis

Table 3.1 shows the 96 HMT study variables divided into the 9 categories on the left, and the corresponding number of variables replicated from the smart registry in the 4 replicability classes on the right. In total, 80 of the 96 HMT study variables could be theoretically replicated from the smart registry. For the remaining 16 variables, no information existed in the smart registry from which they could have been derived. 52 variables could be replicated with a different time range compared to the HMT study, and 17 variables could be replicated correctly.

3.1.1 Post-analysis update

The results of the analysis with the latest smart registry data are shown in table 3.2. With that, 60 variables could be recalculated with a deviating time-period. Also, due to the extension of the HIS extract back to 2016, 47 variables could be correctly replicated.

The three comparisons of the 47 variables correctly replicated from the smart registry with those from the HMT study are entered in table 3.3. For each category, the three comparison metrics used were entered as an averaged value. The mean of all six replicable categories was a congruence of 0.67, a

Table 3.1: HMT study variables in each category with the associated four classes of replicated variables from the smart registry with initial data prior to the data update

	HMT study	**Smart registry**			
Category	**Variables**	**Not replicable variables**	**Replicable variables (theoretical)**	**Replicable variables (deviant)**	**Replicable variables (correct)**
Administrative	4	0	4	3	3
Demographic	4	0	4	4	4
Physiologic	9	0	9	9	0
Laboratory	13	0	13	10	0
Hospitalization	12	0	12	6	0
Diagnosis	22	0	22	8	0
Doctor's letter	6	6	0	0	0
Other sources	13	10	3	0	0
Death	13	0	13	12	10
Total	**96**	**16**	**80**	**52**	**17**

Table 3.2: Verification analysis variables with updated smart registry data

	HMT study	**Smart registry**			
Category	**Variables**	**Not replicable variables**	**Replicable variables (theoretical)**	**Replicable variables (deviant)**	**Replicable variables (correct)**
Administrative	4	0	4	4	4
Demographic	4	0	4	4	4
Physiologic	9	0	9	9	6
Laboratory	13	0	13	10	10
Hospitalization	12	0	12	12	12
Diagnosis	22	0	22	8	0
Doctor's letter	6	6	0	0	0
Other sources	13	10	3	0	0
Death	13	0	13	13	11
Total	**96**	**16**	**80**	**60**	**47**

correlation of 0.83, and a percentage of finite elements of 0.70. A detailed overview of all 96 variables can be found in Appendix A.

Table 3.3: Comparison metrics of the correctly replicable variables averaged in each category

Category	Replicable variables (correct)	Congruence	Correlation	Percentage finite
Administrative	4	0.81	1.00	0.87
Demographic	4	0.99	1.00	1.00
Physiologic	6	0.29	0.49	0.24
Laboratory	10	0.43	0.77	0.47
Hospitalization	12	0.67	0.71	0.75
Death	11	0.85	0.99	0.86
Total	**47**	**0.67**	**0.83**	**0.70**

3.2 HMT extension necessity prediction model

The results of the 5-fold cross-validation performance of the prediction model are shown in table 3.4. The metrics used to evaluate the performance of the model were the accuracy, the AUROC, as well as sensitivity and specificity. The metrics mean result plus/minus standard deviation, as well as the selected best result across all 5 folds were documented.

Table 3.4: Achieved model performance, shown as mean value plus/minus standard deviation and the performance range across all 5 folds

Evaluation metric	Mean result $\pm$ standard deviation	[Min - Max]
Accuracy	0.6194 $\pm$ 0.1501	[0.3280 - 0.7551]
AUROC	0.7020 $\pm$ 0.0615	[0.6333 - 0.8144]
Sensitivity	0.6967 $\pm$ 0.1455	[0.4573 - 0.9038]
Specificity	0.5950 $\pm$ 0.2307	[0.1444 - 0.7651]

4 Discussion

During the work within this thesis two main goals were pursued. The basis for these was the work with the data from the smart registry for CHF, and the further pre-processing.

4.1 Smart registry verification analysis

The first goal of this thesis was a verification analysis of the smart registry to ensure its data quality. To this end, the extent to which the smart registry could replicate the data of a scientific publication (HMT study) was analyzed.

This analysis showed that theoretically a large part, more precisely 80 of the 96 variables of the HMT study, could be replicated via the smart registry route. Of those variables, 52 could be replicated, but with a deviating time period, thus with no overlap, and 17 could be replicated with a matching time period. The latter were the variables from the administrative and demographic categories, which were replicated from the basic patient data of the HMT data source in the smart registry. In addition, a large proportion of the variables in the death category could be replicated using the extract from the Austrian National Register of Deaths available in the smart registry.

The 16 variables that could not be replicated often contained information that was taken manually from doctors' letters in the HMT study. These were variables such as smoking behavior, which were derived from individual patient histories. Automated annotation of such free-text documents from the EMR to make unstructured information available in the smart registry would be a potential solution here but is a very difficult task that will require further research in the future. Furthermore, non-replicable variables also

came from other sources, such as LVEF, which was taken from a subsystem of Tirol Kliniken GmbH. A connection of such additional data sources with the smart registry is planned for the future.

Because the smart registry did not yet have data from the HIS with the same time period of the HMT study at the beginning of this analysis, a direct comparison of most replicated variables was not yet possible. Overall, however, the infrastructure of the smart registry was able to theoretically replicate most of the information in the HMT publication. This was especially true for variables in the hospitalization and death categories, which were crucial for the primary and secondary outcomes of the HMT study [8].

Above all, this first analysis confirmed the importance of verification to confirm or improve the data quality of the smart registry. By comparison with a quality-controlled data source from the HMT study, relevant missing information and potential data discrepancies could be identified. It should be noted that the verification was based on the HMT study dataset mentioned above, which itself was derived from primary data sources. A verification of the source data as such was not carried out.

4.1.1 Post-analysis update

With the extension of the HIS data extract back to 2016, 60 variables could be practically recalculated at last, of which 47 variables were recalculated with a correct time period. After this update of the smart registry, all 251 HMT patients of the HMT study were present in the smart registry, and of these, EMR data was found in 126 patients. This resulted in a congruence of 100% in a number of variables for the first time, as can be seen in the table in Appendix A.

In a direct comparison of all values of the 47 correctly replicated variables with those of the HMT study, some matches were found in six of the nine categories as a first step. The administrative and demographic category variables could be replicated identically from the smart registry to a large extent. Only the variable for the HMT programme start date differed minimally for many patients, resulting in a lower congruence with simultaneously high correlation in the administrative category. In addition, some patients

in this category could not be assigned to a participating hospital because of unavailable data, which explains the imperfect percentage of finite elements. Variables of the physiological and laboratory category had a slightly lower congruence, because also here, similar to the HMT start date, the exact time of the measured values could not be derived from the manually generated HMT study data according to any fixed scheme, and therefore often deviated in the automated recalculation. The variables of the hospitalization category could again be replicated with a higher agreement, but often also depended on the deviating HMT start date. The same was the case for the follow-up variables of the death category, otherwise the replicated death data were identical.

4.2 HMT extension necessity prediction model

In the second part of this master thesis, an attempt was made to design a basic ML model based on the smart registry data, which predicts whether a subsequent second period is necessary for an HMT patient after his or her initial 90 days in the HMT telemonitoring programme. As presented in table 4, the prediction models showed a mixed performance. On the one hand, the selected model with the best fold achieved satisfactory results, on the other hand, the performance varied greatly between the folds. A common problem in training was that models tended to collapse into a version with high specificity (> 0.9) and low sensitivity (> 0.2), or vice versa. This trend is reflected in the high standard deviation values in the results. This could be due to a common problem of models collapsing into local optima rather than a global optimum, which our best-case model shows would likely exist. These results suggest that the model architecture may be sufficient, but the training methods and data preprocessing could be improved.

It should also be said that the data set of 298 patients and 69 positive targets is probably small enough for the given issue. For events that are relatively complex in their development and are influenced by many factors, it is important to have a considerable number of observations. Our question of telemonitoring extension is relatively specific and the outcome is also influenced by many individual and personalized factors. It can therefore

be assumed that a data set with a higher number of patients and features would have a positive effect on the performance and results of our model.

Of course, there are a large number of model hyperparameters that have not yet been examined in detail here. For example, the window size and its overlap could be further optimized using grid search to obtain better predictions. Likewise, noise generation during label smoothing or data augmentation could be investigated in more detail. Furthermore, other different approaches could also be tried out during data preprocessing, for example when imputing and scaling the data. Although different loss functions/configurations and batch sizes were also experimented with during model training, a systematic search could also reduce the model collapse. Other optimizing functions, early stopping and better regularization could be explored as well.

With regard to the prediction itself, it should be noted that the methods described basically predict retrospectively which feature patterns are associated with extended inclusion in a disease management programme. However, this prediction does not correlate with clinical outcomes or the future health status of patients. Thus, the model gives an indication of how likely it is that a patient with these given data will be included in an extension, but not how much an extension would help that patient, or how much a patient actually needs it to avoid adverse events.

In general, it is of course difficult to make a truly qualitative statement about whether an extension would be indicated based on the available data, considering that this decision-making process is influenced in real life by a variety of other factors and information that cannot be taken into account or are not available in this prediction setting. Extending the model to include future clinical outcomes would certainly be a more convincing use case that could help healthcare professionals in their practice as a decision support tool. This would have benefits for patient care, health economics and quality of life, but more data and research is needed to make this a reality.

4.3 Conclusion

The basic scope of this master's thesis was the work around the smart registry for CHF patients in order to support its development with quality assurance steps. The working environment was based primarily on the PATH software developed for such purposes within MATLAB, which enabled a data communication interface with the smart registry and effective further processing. The aims were to improve the data quality of the smart registry through a verification analysis and to develop a basic ML model as a use case in the smart registry, which performed a prediction based on the health data of HMT patients automatically aggregated from various sources.

The verification analysis showed that a manually collected data set of a scientific publication could largely be replicated via the smart registry. The work carried out during this analysis also made it possible to identify deviations in the smart registry data and thus improve its quality. The ML model developed was able to achieve relatively satisfactory performance results for the given data situation. Of course, one has to be careful when making a statement about whether the health of a patient would benefit from the prediction results, but it was possible to run through an initial use case with the smart registry data and thus gain some insights and knowledge for further approaches.

In summary, it can be said that the development of the smart registry infrastructure presented here can make an important contribution to the future of the European health data economy. This master's thesis has made a small contribution to its quality assurance through a data verification analysis and a basic prediction model. Further aggregation of health data and the generation of features and models from it is necessary for the further advancement of the smart registry. This could improve resource management in the healthcare system and bring us closer to the goal of assisting healthcare professionals at the point of care with smart decision support services.

Bibliography

[1] Theresa A McDonagh et al. "2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC." In: *European Heart Journal* 42.36 (Aug. 2021), pp. 3599–3726 (cit. on pp. 1–3).

[2] Biykem Bozkurt et al. "Universal definition and classification of heart failure: a report of the Heart Failure Society of America, Heart Failure Association of the European Society of Cardiology, Japanese Heart Failure Society and Writing Committee of the Universal Definition of Heart Failure." In: *European Journal of Heart Failure* 23.3 (2021), pp. 352–380 (cit. on pp. 1, 2).

[3] Amy Groenewegen et al. "Epidemiology of heart failure." In: *European Journal of Heart Failure* 22.8 (2020), pp. 1342–1356 (cit. on p. 2).

[4] Deddo Moertl et al. "Disease management programs in chronic heart failure." In: *Wiener Klinische Wochenschrift* 129 (2017), pp. 869–878 (cit. on pp. 3, 4).

[5] Darshan H Brahmbhatt and Martin R Cowie. "Remote Management of Heart Failure: An Overview of Telemonitoring Technologies." In: *Cardiac Failure Review* 5.2 (2019), pp. 86–92 (cit. on p. 4).

[6] Trisha Greenhalgh, Christine A'Court, and Sara Shaw. "Understanding heart failure; explaining telehealth - A hermeneutic systematic review." In: *BMC Cardiovascular Disorders* 17 (2017), pp. 1–16 (cit. on p. 4).

[7] AIT Austrian Institute of Technology GmbH. *KIT Telehealth Solutions information portal*. [Online; accessed 02-October-2024]. URL: https://kit.ait.ac.at/ (cit. on p. 4).

[8] G. Poelzl et al. "Feasibility and effectiveness of a multidimensional post-discharge disease management programme for heart failure patients in clinical practice: the HerzMobil Tirol programme." In: *Clinical Research in Cardiology* 111 (2022), pp. 294–307 (cit. on pp. 4, 5, 11, 17, 30).

[9] Elske Ammenwerth et al. "HerzMobil, an Integrated and Collaborative Telemonitoring-Based Disease Management Program for Patients With Heart Failure: A Feasibility Study Paving the Way to Routine Care." In: *JMIR Cardio* 2.1 (2018), e11 (cit. on pp. 4, 5).

[10] Landesinstitut für Integrierte Versorgung Tirol. "D4Health Tirol project description." In: 3 (2020) (cit. on pp. 5, 6).

[11] Dieter Hayn et al. "Predictive analytics for data driven decision support in health and care." In: *it - Information Technology* 60.4 (2018), pp. 183–194 (cit. on p. 6).

[12] Martin Baumgartner et al. "Health data space nodes for privacy-preserving linkage of medical data to support collaborative secondary analyses." In: *Frontiers in Medicine* 11 (2024) (cit. on p. 6).

[13] OHDSI Observational Health Data Sciences and Informatics. *The Observational Medical Outcome Partnership (OMOP) Common Data Model (CDM)*. [Online; accessed 02-October-2024]. URL: https://www.ohdsi.org/data-standardization/ (cit. on pp. 7, 15).

[14] Michael Nitzlnader and Günter Schreier. "Patient identity management for secondary use of biomedical research data in a distributed computing environment." In: *Studies in health technology and informatics* 198 (2014), pp. 211–8 (cit. on p. 8).

[15] European Commission. *European Health Data Space*. [Online; accessed 02-October-2024]. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en (cit. on p. 8).

[16] Pavel Hamet and Johanne Tremblay. "Artificial intelligence in medicine." In: *Metabolism: Clinical and Experimental* 69 (2017), S36–S40 (cit. on pp. 8, 9).

[17] Yann LeCun, Y. Bengio, and Geoffrey Hinton. "Deep Learning." In: *Nature* 521 (2015), pp. 436–44 (cit. on p. 8).

[18] Fei Jiang et al. "Artificial intelligence in healthcare: past, present and future." In: *Stroke and Vascular Neurology* 2.4 (2017), pp. 230–243 (cit. on pp. 9, 10).

[19] Eric Topol. *Deep medicine: how artificial intelligence can make healthcare human again*. Basic Books, New York, 2019 (cit. on p. 9).

[20] Kain Kordian Gontarska et al. "Predicting Medical Interventions from Vital Parameters: Towards a Decision Support System for Remote Patient Monitoring." In: *CoRR* abs/2104.10085 (2021) (cit. on p. 10).

[21] Michael Schrempf et al. "Machine Learning Based Risk Prediction for Major Adverse Cardiovascular Events." In: *Studies in health technology and informatics* 279 (2021) (cit. on p. 10).

[22] Robert Herman et al. "Utilizing longitudinal data in assessing all-cause mortality in patients hospitalized with heart failure." In: *ESC Heart Failure* 9 (2022) (cit. on p. 10).

[23] Christoph Eberharter et al. "Towards a Verification Approach of a Smart Registry for Chronic Heart Failure Patients." In: *Studies in health technology and informatics* 301 (2023), pp. 242–247 (cit. on p. 17).

[24] Michael Kenward. "The handling of missing data in clinical trials." In: *Clinical Investigation* 3 (2013), pp. 241–250 (cit. on p. 21).

[25] Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. "When Does Label Smoothing Help?" In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 22).

List of Figures

List of Tables

Appendices

Appendix A

Variable list of the smart registry verification analysis

Category	Variable name	Replicable (theoretical)	Replicable (deviant)	Replicable (correct)	Congruence	Correlation	Percentage finite
Administrative	Patient_ID	1	1	1	1,00	NaN	1,00
Administrative	HMT_yes_no	1	1	1	1,00	NaN	1,00
Administrative	Beginning_date	1	1	1	0,42	1,00	1,00
Administrative	Hospital_center	1	1	1	0,81	NaN	0,48
Demographic	Date_of_birth	1	1	1	0,99	1,00	1,00
Demographic	Gender	1	1	1	1,00	0,99	1,00
Demographic	Age_at_HMT_start	1	1	1	0,98	1,00	1,00
Demographic	Group_Age_Median_73	1	1	1	1,00	1,00	1,00
Physiologic	Height	1	1	1	0,75	0,56	0,49
Physiologic	Heart_rate_0	1	1	1	0,21	0,16	0,19
Physiologic	Systolic_BP_0	1	1	1	0,18	0,59	0,20
Physiologic	Diastolic_BP_0	1	1	1	0,22	0,50	0,20
Physiologic	Weight_0	1	1	1	0,33	0,64	0,18
Physiologic	BMI_0	1	1	1	0,07	0,50	0,18
Physiologic	NYHA_0	1	1	0			
Physiologic	Obesity_BMIgt30	1	1	0			
Physiologic	Group_NYHA	1	1	0			
Laboratory	Creatinin_0	1	1	1	0,46	0,84	0,49
Laboratory	GFR_0	1	1	1	0,01	0,75	0,49
Laboratory	Sodium_mmoll_0	1	1	1	0,42	0,67	0,49
Laboratory	Potassium_mmoll_0	1	1	1	0,35	0,41	0,49
Laboratory	GOT_ASAT_UI_0	1	1	1	0,62	0,81	0,47
Laboratory	GPT_ALAT_UI_02	1	1	1	0,60	0,88	0,45
Laboratory	GGT_UI_0	1	1	1	0,59	0,90	0,45
Laboratory	Troponin_T_0	1	1	1	0,46	0,69	0,41
Laboratory	NTproBNP_0	1	1	1	0,42	0,92	0,48
Laboratory	Hb_0	1	1	1	0,36	0,84	0,49
Laboratory	Group_NTproBNP	1	0	0			
Laboratory	Group_GFR	1	0	0			
Laboratory	lnNTproBNP	1	0	0			
Hospitalisation	Index_Admission_Hospitalisation	1	1	1	0,70	0,88	0,47
Hospitalisation	Index_Discharge_Hospitalisation	1	1	1	0,71	0,87	0,47
Hospitalisation	Index_Hospitalisation_Duration	1	1	1	0,64	0,77	0,49
Hospitalisation	Index_ICU_stay	1	1	1	0,86	NaN	1,00
Hospitalisation	Index_ICU_stay_duration	1	1	1	0,80	NaN	0,50
Hospitalisation	Hospitalisation_yes_no	1	1	1	0,71	NaN	1,00
Hospitalisation	Date_of_admission	1	1	1	0,68	0,99	0,09
Hospitalisation	Follow_up_Date_Hospitalisation	1	1	1	0,34	0,99	1,00
Hospitalisation	Follow_up_Duration_Months_Hospitalisation	1	1	1	0,62	0,29	1,00
Hospitalisation	Count_Hospitalisation	1	1	1	0,65	0,59	1,00
Hospitalisation	cum_Sum_Days_lost_due_Hospitalisation	1	1	1	0,66	0,55	1,00
Hospitalisation	Group_Count_Hospitalisation	1	1	1	0,66	0,50	1,00
Diagnosis	LBBB_Left_Bundle_Branch_Block	1	1	0			
Diagnosis	KHK	1	0	0			
Diagnosis	Arterial_hypertension	1	0	0			
Diagnosis	COPD_or_asthma	1	1	0			
Diagnosis	OSAS	1	0	0			
Diagnosis	pAVK	1	0	0			
Diagnosis	AS	1	0	0			
Diagnosis	MS	1	0	0			
Diagnosis	TS	1	0	0			
Diagnosis	AI	1	0	0			
Diagnosis	MI	1	0	0			
Diagnosis	TI	1	0	0			
Diagnosis	Valvular_heart_disease	1	0	0			
Diagnosis	DM	1	0	0			
Diagnosis	Depressio	1	0	0			
Diagnosis	Malignom	1	0	0			
Diagnosis	ICD10Code	1	1	0			
Diagnosis	Initialdiagnosisdate	1	1	0			
Diagnosis	Renal_insufficiency_Stage	1	1	0			
Diagnosis	Anemia	1	1	0			
Diagnosis	CharlsonKomorbiditätsIndex	1	1	0			
Diagnosis	Group_Artrial_fibrillation	1	1	0			
Physician´s letter	Cause_of_Decompensation	0	0	0			
Physician´s letter	Cause_of_HF	0	0	0			
Physician´s letter	Ejection_fraction	0	0	0			
Physician´s letter	Classification_EF_0	0	0	0			
Physician´s letter	EF_Grouping_45	0	0	0			
Physician´s letter	Cause_of_HF_groups	0	0	0			
Other sources	Date_of_diagnosis	0	0	0			
Other sources	Heart_rhythm	1	0	0			
Other sources	Device	1	0	0			
Other sources	St,p,_MCI	0	0	0			
Other sources	St,p,_ACBG	0	0	0			
Other sources	St,p,_Valve_OP	0	0	0			
Other sources	St,p,_Stroke	0	0	0			
Other sources	Smoking	0	0	0			
Other sources	Smoking_codiert	0	0	0			

Other sources	Pack_years	0	0	0			
Other sources	St,p,malignom__under_therapy	0	0	0			
Other sources	filter_$	0	0	0			
Other sources	Group_CRT_ICD	1	0	0			
Death	Death	1	1	1	0,99	0,98	1,00
Death	Dateofdeath	1	1	1	1,00	1,00	0,22
Death	CauseofDeath	1	1	0			
Death	Follow_up_Death_Date	1	1	1	0,99	1,00	1,00
Death	Follow_up_Duration_in_months	1	1	1	0,03	1,00	1,00
Death	Death_12_mo	1	1	1	1,00	1,00	1,00
Death	Followup_Duration_12_mo	1	1	1	0,84	0,99	1,00
Death	Days_lost_due_death	1	1	1	0,94	1,00	1,00
Death	Death_6_mo	1	1	1	1,00	1,00	1,00
Death	Days_lost_due_death_6_mo	1	1	1	0,98	1,00	1,00
Death	Distribution_mortality	1	1	1	0,95	1,00	0,22
Death	Time_alive_and_out_of_hospital	1	1	1	0,60	0,88	1,00
Death	death_hospital	1	1	0			
Total	**96**	**80**	**60**	**47**	**0,65**	**0,80**	**0,68**

Appendix B

Feature list of the HMT extension necessity prediction model

Parameter (Feature)	Category	Parameter source	OMOP table	Parameter type	Pre-processing (Units)	Remarks
Age	Demographics	HMT, HIS	Person	Continuous	(years)	
Is_female	Demographics	HMT, HIS	Person	Binary	(1 – True; 0 – False)	
Age_at_hmt_start	Demographics	HMT, HIS	Person	Continuous	(years)	
Outpatient_visit	Hospitalisation	HIS	Visit_occurrence	Binary	(1 – True; 0 – False)	
Inpatient_visit	Hospitalisation	HIS	Visit_occurrence	Binary	(1 – True; 0 – False)	
Intensive_care_unit	Hospitalisation	HIS	Visit_occurrence	Binary	(1 – True; 0 – False)	
Patients_medication_adherence	Physiologic	HMT	Observation	Continuous	(ratio)	Patients relative adherence to prescribted medications (drugs) per day: number_of_medications_taken/number_of_medications
Wellbeing	Physiologic	HMT	Observation	Discrete	(1 - bad; 2 - medium; 3 - good)	
NYHA	Physiologic	HMT	Observation	Discrete	(1 - Class I; 2 - Class II, 3 - Class III, 4 - Class IV)	New York Heart Association Classification
HR	Physiologic	HMT, HIS	Measurement	Continuous	(/min)	Heart rate
BP_sys	Physiologic	HMT, HIS	Measurement	Continuous	(mmHg)	Systolic blood pressure
BP_dia	Physiologic	HMT, HIS	Measurement	Continuous	(mmHg)	Diatolic blood pressure
BP_diff	Physiologic	HMT, HIS	Measurement	Continuous	(mmHg)	Pulse amplitude of Blood Pressure measurements: BP_sys - BP_dia
MAP	Physiologic	HMT, HIS	Measurement	Continuous	(mmHg)	Mean Arterial Pressure: (2*BP_dia + BP_sys)/3
Bodyweight	Physiologic	HMT, HIS	Measurement	Continuous	(kg)	
Body_height	Physiologic	HIS	Measurement	Continuous	(cm)	
BMI	Physiologic	HMT, HIS	Measurement	Continuous	(kg/m²)	Received values from HIS and calculated values from HMT data
Creatinine	Laboratory	HIS	Measurement	Continuous	(mg/dl)	
GFR	Laboratory	HIS	Measurement	Continuous	(ml/min/1,73m²)	
GGT	Laboratory	HIS	Measurement	Continuous	(U/l)	
GOT_ASAT	Laboratory	HIS	Measurement	Continuous	(U/l)	
GPT_ALAT	Laboratory	HIS	Measurement	Continuous	(U/l)	
Hb	Laboratory	HIS	Measurement	Continuous	(g/l)	
NTproBNP	Laboratory	HIS	Measurement	Continuous	(ng/l)	
Potassium	Laboratory	HIS	Measurement	Continuous	(mmol/l)	
Sodium	Laboratory	HIS	Measurement	Continuous	(mmol/l)	
Troponin	Laboratory	HIS	Measurement	Continuous	(ng/l)	
Time_of_value	Various	HMT, HIS	Measurement	Continuous	s	Time of first measured value per day
Season	Various	HMT, HIS	Measurement	Discrete	(1 - quarter 1; 2 - quarter 2; 3 - quarter 3; 4 - quarter 4)	Season of daily report
Number_of_daily_reports	Various	HMT, HIS	Measurement	Continuous	(days)	
Days_since_first_report	Various	HMT, HIS	Measurement	Continuous	(days)	
CANCER	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Any malignancy, including lymphoma and leukemia, except malignant neoplasm of skin (Charlson Comorbidity)
CEVD	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Cerebrovascular disease (Charlson Comorbidity)
CHF	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Congestive heart failure (Charlson Comorbidity)
COPD	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Chronic pulmonary disease (Charlson Comorbidity)
DEM	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Dementia (Charlson Comorbidity)
DIAB_C	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Diabetes with chronic complication (Charlson Comorbidity)
DIAB_NC	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Diabetes without chronic complication (Charlson Comorbidity)
HIV	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	AIDS/HIV (Charlson Comorbidity)
METS	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Metastatic solid tumor (Charlson Comorbidity)
MI	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Myocardial infarction (Charlson Comorbidity)
MILDLD	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Mild liver disease (Charlson Comorbidity)
MSLD	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Moderate or severe liver disease (Charlson Comorbidity)
PARA	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Hemiplegia or paraplegia (Charlson Comorbidity)
PUD	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Peptic ulcer disease (Charlson Comorbidity)
PVD	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Peripheral vascular disease (Charlson Comorbidity)
RD	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Renal disease (Charlson Comorbidity)
RHEUM	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Rheumatic disease (Charlson Comorbidity)
Anemia	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	
Asthma	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	
Atrial_fibrillation	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	
LBBB	Diagnosis	HIS	Condition_occurrence	Binary	(1 – True; 0 – False)	Left Bundle Branch Block
BB	Medication	HMT	Drug_exposure	Continuous	(ratio)	Relative drug dosis of HF relevant category (beta blockers) by comparing the daily prescribted dose with the ESC guideline recommended target dose (doctor´s adherence)
MRA	Medication	HMT	Drug_exposure	Continuous	(ratio)	Relative drug dosis of HF relevant category (mineralocorticoid receptor antagonists) by comparing the daily prescribted dose with the ESC guideline recommended target dose (doctor´s adherence)
RAASi	Medication	HMT	Drug_exposure	Continuous	(ratio)	Relative drug dosis of HF relevant category (renin-angiotensin-aldosterone system inhibitors) by comparing the daily prescribted dose with the ESC guideline recommended target dose (doctor´s adherence)
SGLT2i	Medication	HMT	Drug_exposure	Continuous	(ratio)	Relative drug dosis of HF relevant category (sodium-glucose co-transporter 2 inhibitors) by comparing the daily prescribted dose with the ESC guideline recommended target dose (doctor´s adherence)