ALMA MATER STUDIORUM – UNIVERSITA' DI BOLOGNA

DEPARTMENT OF STATISTICAL SCIENCES

Second Cycle Degree in Statistical Sciences

TITLE

Comparing explainable AI models to allow routine use in healthcare: an application using claims data to optimise expenditure of high-cost patients in Germany

Presented by:

Sebastian Benno Veuskens

Matricola: 0001082417

Supervisor:

Prof Fabrizio Carinci

II SESSION

ACADEMIC YEAR 2023 / 24

ABSTRACT

BACKGROUND

AI and machine learning are increasingly used to improve services to citizens in everyday activities through better predictions in diverse areas and tasks such as risk stratification. Growing costs in healthcare in many countries and financial constraints bring the system to the brink of collapse. A small group of patients at the upper end of the cost scale accounts for about half of the overall spending. Focusing on this high-cost patient group to optimize healthcare spending is essential to maintain a high-quality healthcare level in the future. Effective precautions instead of expensive follow-up care could prevent the escalation of such high illness-related costs. Traditional statistical models can moderately predict these high-cost patients in advance. Modern machine learning models offer even more accurate predictions. Yet, their complex, unintelligible nature prevents the widespread deployment of these powerful technologies in the regulated healthcare domain so far. The field of eXplainable AI (XAI) explores methods to explain these models.

OBJECTIVE

This thesis aims to answer the following questions: How do different models that predict high-cost patients compare? How do XAI methods help explaining these models? Are the given explanations accurate and understandable enough to support the routine use of XAI in healthcare optimisation?

METHODS

We collect claims data accessing insurance fund records from 2019 to 2022 from two German regions to predict high-cost patients yearly. Our paper utilizes the same methods and is based on similar patient data from 2015 to 2018 as a previous study. Firstly, we apply several ML models and evaluate their predictive performance. Secondly, we use the most popular model-agnostic XAI methods SHAP and LIME as well as additional global methods to explain our best-performing model. Finally, we evaluate the suitability of these explanation methods.

RESULTS

The tree ensemble models outperformed logistic regression and the deep neural network in our study. The outstanding performance from the random forest in the previous study in Langenberger,

Schulte, and Groene was confirmed and even slightly improved. On a variety of measures, our model outperformed most or even all other models from similar studies. Global explanations for the random forest model identified total costs and the current status of high-cost patients as the most important features. Explanations for these and the age variable agreed among different methods and revealed clear non-linear relationships between feature value and prediction. Local explanations partly corresponded to global explanations; others explain the prediction for a single sample through different feature contributions. The ICE plot revealed significant variability hidden by the PDP plot for the age and high-cost patient status variables.

CONCLUSIONS

Our random forest model to predict high-cost patients offers great potential to improve care in a healthcare system under cost-saving pressure due to its ability to identify high-cost patients effectively. The way we structured our model evaluation in this study facilitates the deployment in medical practice to leverage the outstanding performance of our model. Explanations allow us to overcome trust issues in healthcare. They extract general knowledge from the model in the form of feature importance and feature effects. We show how to explain the prediction of single patients. While these explanations are not always reliable, we explore different approaches to assess their stability and mitigate some of their shortcomings by demonstrating a path to more reliable explanations. In this study, we paved the way for deploying a prediction model for high-cost patients in the real world.

KEYWORDS:

high-cost patients, machine learning, prediction model, XAI, interpretability

CONTENTS

INTRODUCTION

In the past decade, much research has focused on advancing the power of model predictions [1]. There is great progress in (un-)supervised learning for both image and non-image data detection and deployment of AI systems. Many domains have already been transformed by AI and ML, including biology, energy, and game strategy. New exciting opportunities were opened by the introduction of enhanced protein structure predictions [2], [3]. Smart solutions are developed for the sustainable transformation of the energy sector [4], [5]. Special attention on AI and ML was evoked by the defeat of the world GO champion against a deep neural network model [6]. Computer Security is enhanced by ML models that identify suspicious users reliably [7]. Most recently, Large Language models entered the stage and are likely to transform many of these sectors like protein structure predictions [8] even further. These systems allow for much more accurate predictions than traditional statistical models like logistic regression (see [9], [10], [11], [12], [13], [14]).

Also in healthcare, ways to incorporate AI and ML receive mounting attention [15]. Two approaches can be observed for AI systems: Applications directed at the consumer and systems aiding physicians in decision-making [16]. Many promising applications were developed: Complex ML models that outperform traditional statistical models in readmission predictions [17]. A Convolutional Neural Network showing superior recognition skills of dermoscopic melanoma over dermatologists [18]. SVMs helped to discover new biomarkers and effective drugs against cancer as well as better understand its genetic drivers [19]. A boosted random forest helped to predict COVID-19 patients based on their travel data and personal characteristics [20]. AI demonstrated its ability to reduce costs for caries detection in Germany [21]. Choosing the appropriate algorithm is not a trivial task. While in some medical applications, a random forest might perform best [22], other models like SVMs [23] or XGBoost algorithms [24] might be more suitable elsewhere. While these new technologies are developing rapidly, research is just at the beginning on how to include these technologies in healthcare applications beneficially (see [25], [26]). This emerging technology is set out to change the healthcare sector at its core [27], as it has the potential to substantially improve effectiveness in healthcare due to its superior predictive accuracy [28].

So how can ML and AI be applied to benefit patients and enhance healthcare system quality? Despite promising performances, these relatively new technologies are still in the validation phase

in healthcare [29]. Evaluations of such systems on economic health impact are scarce and are behind due to the rapid development of AI models. [30] Very few models have become part of the clinical routine, leading to a gap between the potential and current poor adaptation in clinical practice [31]. A promising approach towards reducing bias in these models are standardized data sets. They allow standardizing model outcomes and enable comparison across algorithms and research teams. A widely adapted publicly available data set is the breast cancer image collection [32]. This data set allows researchers to benchmark algorithms against each other to report the effectiveness of different ML approaches to the scientific community in a unified manner. Especially because of their central role in model comparison, these data sets need to be thoroughly monitored to avoid biases like different representations of genders that can impact patient classifications [33].

Implementation has been slowed by several obstacles. Some concern data-related issues such as data quality, privacy, availability, and security [27]. An additional concern is a possible automation bias for AI systems in healthcare [34]. However, the main challenges are almost all linked to model trust, accountability, and fairness of the AI systems [29]. Fairness and accountability are two of the three pillars that represent the minimal requirements to comply with agreed basic ethic guidelines in AI, joined by privacy [35]. Not being able to understand these intransparent systems is a major obstacle to the adaptation of these systems in healthcare [36].

Healthcare systems worldwide are under pressure. The lack or delay of medical interventions was responsible for almost every third death in the OECD countries. These deaths could have been avoided with more, effective interventions [37]. About every third general practitioner in Germany shows signs of burnout [38]. In the Saxony region in Germany, about 30% of the physicians want to emigrate, mostly due to unsatisfying work situations and high workloads [39]. This is not only problematic on an individual level, but also for the patients. Sub-optimal working conditions are shown to correlate with impaired subjective [40] and actual patient outcomes [41]. An additional financial burden is the increase in costs per patient for overworked physicians, probably due to less time assigned for paperwork execution [42].

For about a decade, healthcare expenditures in the OECD countries moved parallel to economic developments. About 8.8% of the GDP was spent on healthcare after the financial crisis from 2009 to 2019. However, in 2020 these relative expenditures saw a sharp increase. Even though they

slightly went down until the year 2022, they still remain at the relatively high level of 9.2%. Especially alarming is the situation in Germany: With 12.7% of its GDP in 2022, a concerningly high proportion of its economy is designated to healthcare. It is second in the country ranking only surpassed by the US [37]. Until 2040, the financial situation is expected to exacerbate. The latest projections of the annual growth in health expenditures in the OECD regions are on average twice as high as the percentual increase in government budgets (2.6% and 1.3%, respectively) [43]. These trends pose the risk of political unrest for the governments in charge. Cutting healthcare services might become inevitable and is likely to worsen patient outcomes and the already high pressure on physicians. Consequently, this will likely lead to a generally worse healthcare quality as reports from Spain show. In that case, the quality of mental health [44] and general life [45] decreased significantly after healthcare cuts.

It is well known that a small group of individuals is responsible for about half of the healthcare expenditures (e.g. see [46], [47]). Cost distribution among patients is extremely skewed [48]. This is widely recognized as a pressing concern for healthcare worldwide (e.g. in the US [49], Singapore [50], and Denmark [51]). Different authors identified the importance of this group and attempted to categorize these patients into a distinct group. Overlapping but not equivalent definitions include *high-need, high-cost patients* [52], *high-* or *super-utilizers* [53], or *high-resource users* [54]. Instead of expenditures, these patients are defined occasionally also based on prior healthcare utilization or clinical profiles (see [55], [56]). In this study, total expenditure within one year is the defining characteristic to be considered part of this group. We refer to the individuals in this group as *high-cost patients* (HCPs).

Most commonly, the cumulative costs for a whole year are used to identify HCPs [57]. HCPs are often defined as the top-end 5% or 10% of patients accounting for around 45%-65% of the healthcare costs (see [58], [59], [54] [47], [60]). However, the exact definitions differ. Generally, as a rule of thumb, it is assumed that 5% of the patients account for about 50% of healthcare costs [48]. Especially for cost-comparison studies it is common to regard HCPs as outliers. In that case, cost savings for HCPs are often truncated or omitted [61]. Contrarily, the main concern of this study is the characteristics and reasons why an individual becomes part of this high-cost group.

Several recent studies investigated clinical criteria to identify HCPs. Chronic conditions in all their forms (multiple, specific) and especially chronic pain were found important factors that determine

HCPs in the US, as well as behavioural and social risk factors [52]. Others found comorbidity alongside outpatient and prescription expenditures as the most relevant features [50]. In addition, mental health conditions can be profound causes of extensive costs among patients [53]. In intensive care indicators for HCPs include diagnosis of subarachnoid hemorrhage, acute respiratory failure, and complications of procedures [47].

Not only traditional clinical, but also other criteria offer valuable insights into HCP's characteristics. Very heterogeneous clinical, social, and demographic backgrounds characterise these patients. This involves the most complex and sick part of the population [62]. Old age, being white, and being female are connected to a higher percentage of becoming an HCP [63]. Others found a contrary [47] or no impact of age [52]. Lower immigrant status and non-white ethnicity were found to correlate with a greater proportion of HCPs [54]. Independent research found HCPs to form 3-10 separate subgroups [64], including cardiovascular diseases, mental disorders, and neoplasms, [62], surgical conditions, high-cost high-admission and low-cost-high-admission hospitalizations [65] and pregnancy-related complication [66].

Temporal consistency of HCPs is an ongoing topic of discussion. According to a systematic review by de Ruijter et al., 28%-51% remain HCPs in the next year [55], but only about 8% of patients in the highest-cost decile stay in this group for 3 or more consecutive years [50]. These patients are not necessarily identifiable by the high frequency of hospital admission. While 48% of the HCPs were found to have at least three admissions in [65], Lee et al. found that less than 10% of HCPs had four or more admissions [60],

Interestingly, almost two-thirds of all beneficiaries do not become HCPs in the year or years before they die [62]. Hence, being an HCP is not an inevitable part of life, but only involves a minority of the population. Consequently, a higher age does not automatically imply higher costs. Furthermore, in intensive care, a younger age even results in a higher risk of being an HCP [47]. Analogously, HCPs whose majority of costs was designated to mental health services were generally younger and more costly than other HCPs [57]. These findings underline the complexity and non-linearity of the reasons for becoming an HCP.

How can we deliver better results in healthcare for HCPs? Before we answer this, we must come up with adequate tools for this question. To assess healthcare success, indicators are essential for policymakers and governments (see [67], [68]). The official framework from 2015 for the OECD

countries focused on *quality* indicators in their official framework. Yet, it recognized access and costs/expenditures as additional dimensions for healthcare system performance [69]. Cumulative costs/expenditures are our main indicator for our HCP analysis.

Modern healthcare must address HCPs not only from the financial perspective of society but also in view of the individual humanitarian tragedies that arise for HCPs. Focusing on improving the outcomes for HCPs is humanitarian because of their increased need for healthcare services. In addition, they are likely to benefit the most from preventive and safety improvements [48], [70].

Porter introduced a way to account for improved outcomes in [71]. He proposes *value* as the best way to enhance progress in the healthcare sector, which is defined as outcomes per cost. He argues that a sole focus on cost reduction is undesirable due to its potentially dangerous consequences and limiting impact on good care. In contrast to value-based systems, the common fee-for-service payment systems reduce incentives for clinicians to improve care for HCPs [72]. So, should one try to reduce costs by cutting seemingly necessary 'wasted' services targeted at HCPs? This is not effective because it is unclear whether healthcare service 'waste' is always proportionally larger with higher costs. Low-value or even harmful services should be targeted at the system level rather than at single patients only [49].

When outcomes are integrated into healthcare quality assessment, an effective way for cost reduction and value increase in HCPs is to improve certain services, so that other services can be reduced. A promising example is interdisciplinary primary care programs. They provide vital services such as helping HCPs to transition between care institutions. These programs require accurate targeting of HCPs to treat them effectively [48]. De Carvalho et al. [73] state that more than 20% of HCP expenditures are avoidable. Interdisciplinary care with a focus on these individuals may be an additional cost factor. Despite these costs, the total value for HCPs that are part of these programs might improve value drastically. Although a significant amount of HCPs remains in the high-cost group for the consecutive year, Crowley et al. [70] point out that the majority of HCPs that are caused by heart failures have not been hospitalized in the previous year. Relying on historical HCP records leads to a too-narrow cohort focus and a heterogeneous target group. Therefore, additional techniques are needed to target HCPs accurately.

This inability to identify HCPs reliably is a central dilemma that prevents effective cost reduction of HCPs [52]. Machine learning (ML) models have the potential to identify them effectively and

hence dramatically improve the *value* of HCP care. Even though not all agree on the improved power of AI [74] or call for more detailed cost-effectiveness analyses in the future [75], many applications in healthcare report the superior power of these new technologies. AI already demonstrated its ability to reduce costs in diverse areas of healthcare (see [21], [76], [77]). One example where AI predictions could help policymakers decide on the best treatment for patients with depression was presented by Bremer et al. [78]. Even though the outcomes slightly worsened with the use of AI by 1.98% compared to traditional treatment decisions, a significant cost reduction of 5.42% was achieved. Overall, this allowed for an improvement in *value*. However, researchers should strive to improve the outcome when introducing new interventions. In past medical practice, procedures that impaired the outcomes were the absolute exception and made up only 0.4% of the investigated studies [79].

Much research has been dedicated in recent years to identifying HCPs more reliably. Complex ML algorithms consistently outperform traditional statistical models for HCP predictions (see [58], [73], [46], [51], [50]). Discriminative capabilities are superior for these complex models. It is common practice to predict potential HCPs for a period of 12 months. Such predictions are mostly based on claims data, while a significant minority utilizes electronic health records (EHR) data instead [55]. Accurate predictions from these models can help to tailor preventive measures or care management strategies to HCPs more effectively. Especially the significant HCP subgroup of high-admission high-cost patients is expected to benefit from improved discharge care and chronic disease management [65].

The recently approved ‘Gesundheitsdatennutzungsgesetz’ (GDNG) in March 2024 opens the door for novel treatment strategies for HCPs based on data-generated insights in Germany. Its commencement encourages health insurances to conduct analyses on individual claims data. Before this act, health insurances could not address individuals even when their analysis indicated great health hazards. With the GDNG, they now can communicate their findings to the patients when their analyses detect rare diseases, cancer, or other health hazards in the beneficiaries. These insurances are now able to approach potential HCPs based on their claims data. Additionally, the act pronounces the right of beneficiaries to request information about the basis which led the health insurance to their conclusion. Hence, it demands transparency regarding the underlying analysis of the beneficiaries.

However, none of the existing models for HCP predictions focuses on transparent model reasoning. Transparency and interpretability of risk prediction models are closely related concepts. They both facilitate trust and acceptance by patients not only for HCPs but for healthcare in general [80]. While ML models may offer superior prediction capabilities, their current black-box nature hinders the interpretation of these models. This poses a serious challenge to their application in healthcare [29]. Simpler models like logistic regression or decision trees can be explained directly.

Despite these convenient properties, many articles in the field of interpretable ML and AI associate these simple models with generally less accurate predictions. They assume the presence of a general trade-off between predictive performance vs. explainability and interpretability in statistical models [81], [82], [83], [84]. Consequently, the choice of a model depends on the application setting. In high-stakes decisions, interpretable models should be preferred. Again, when performance is of capital importance, black-box models should be considered [85]. Others argue that this trade-off is not inevitable. Prior processing of the data into a small number of meaningful features can improve the performance of simple models significantly [86]. All authors agree that simple models are less flexible and thus often less accurate with not or minimally pre-processed data [82]. Their simplicity originates from the constraints like monotonicity or linearity being imposed on them. They allow humans to understand the model behaviour [87]. The interpretable nature of simple models facilitates their troubleshooting and eventually helps to improve them [81]. On the other hand, Rudin et al. [87] acknowledged that troubleshooting slows down the deployment of the model.

Especially in recent years, growing attention has been given to understanding complex black box models. The breadth of applications where explanations are demanded is as broad as that of AI and ML itself. It stretches from autonomous drones [88] and vehicle systems [89], [90], strategy analysis for arcade games [91], drought predictions [92], administrative law for predictive algorithms used in courts [93] to financial predictions [94] and medical image analysis [95]. Such explanations have been united within the field of *eXplainable AI* (XAI) and have the potential to enhance the transparency of HCP prediction models substantially.

But what exactly is XAI? The answer to this question depends strongly on the researcher who is answering the question. Due to the lack of consistent XAI definitions, different concepts and taxonomies are common in scientific literature. For some authors, XAI describes means in general

that make models understandable [84], others emphasize the audience of the XAI explanations [82] or solely consider approximations of complex models as XAI methods [87]. XAI is also interchangeably described sometimes by the term Interpretable Machine Learning (IML) [96], even though for other authors IML only refers to simple, interpretable models [87]. Mohseni et al. on the other hand describe an XAI system as the combination of IML with an explainable user interface [97].

Several authors attempted to create a roadmap for the XAI field to avoid getting lost in the jungle of explanation methods. Three main dimensions were distilled in XAI meta-analyses that constitute the underlying structure for *post-hoc* XAI methods (see [97], [82], [98]). As post-hoc, we consider all methods that are deployed to the model after it was trained.

1. Model-agnostic vs. model-specific
2. Local vs. global
3. Type of explanation (visual vs. textual vs. feature relevance vs. simplification)

Additionally, some methods only work on a specific data input type (tabular, image, audio, etc.) [99].

Different systems like unsupervised and reinforcement learning sometimes also apply explanation methods. Analogously to most XAI research though [81], we limit our focus to classification systems and *post-hoc explainability (or post-hoc interpretability* [83] *or post-modelling explainability* [85]). Other types of explainability include *data explainability*, which refers to explanations of the data and not the model. It covers inter alia standard data summarization techniques. We also do not include *model-interpretability* [82] (or *model-explainability* [81], *transparent* [100], or *interpretable models* [85]), which is a different sub-category of XAI. Instead of explaining a black box model, model interpretability aims to construct simple models that are already understandable to humans.

But what do we need XAI for? Most prominently represented is the aim to convey the knowledge learned by the model to humans, followed by the ability to allow ML models being transferred to other settings and problems second [82]. XAI can help to empower individuals, enforce existing laws [81], avoid bias, and ensure fairness [100], [85]. Furthermore, XAI has the potential to improve debugging for statistical models [101]. It can help in two situations: The *understanding*

phase which refers to the time during which a model is trained and evaluated, and the *explaining phase* during and after the actual deployment of the model [102]. Having such a variety of aims for XAI, one must always consider the expectations of the audience and focus on the relevant aims [84].

The relevance of XAI concerning legal guidelines was amplified in 2018 via the General Data Protection Regulation (GDPR, [103]). Article 22 concerning algorithmic decision-making in the EU states: An individual "shall have the right not to be subject to a decision based solely on automated processing, […] which produces legal effects concerning him or her or similarly significantly affects him or her". Few exceptions exist to this ban on automated decision-making that affects humans. Even then, however "suitable measures to safeguard the data subject's rights" must be taken [103]. Interpretation appendixes of this regulation explicitly mention a 'Right to an explanation' [104]. Additionally, non-discrimination obligations arise out of this article [105].

These new regulations accelerate the research for better explanation methods. Intelligible prediction systems for decision-making are required [105]. XAI methods have the potential to

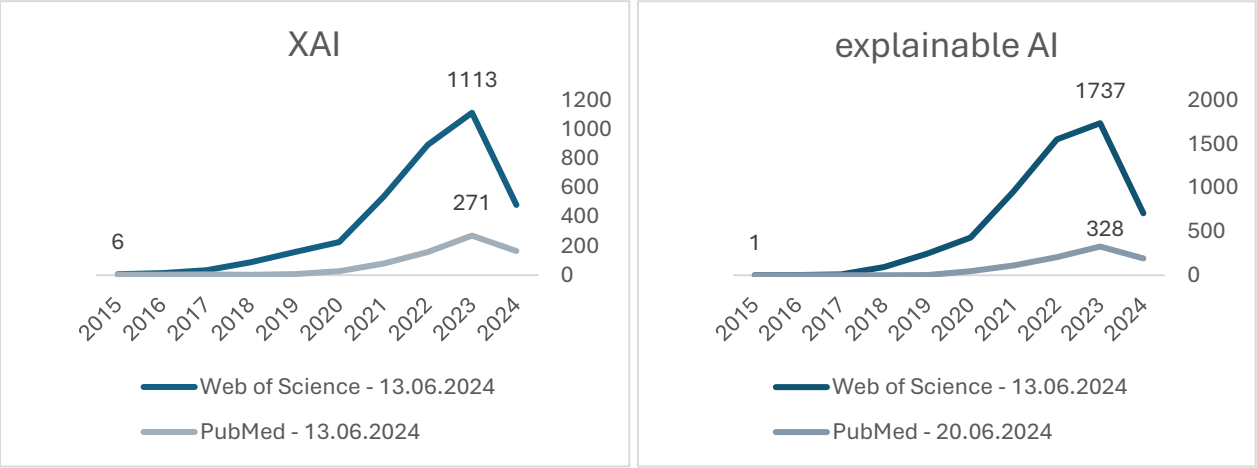

Figure 1: The counts of all papers that appear for the search queries of 'XAI' and 'explainable AI' in the last 10 years, respectively. Counts are the cumulative number of papers per year and database. Minimum and Maximum counts are displayed for each search query and database.

empower people in the automated decision process. This can create the basis for applications in compliance with these regulations [106]. A growing interest in this topic is evident in the exponential increase of research papers in this area. Figure 1 depicts the growing interest in this field over the last 10 years.

In healthcare, humans are highly affected by the potential deployment of prediction models. The goodness of health predictions can be a matter of life and death. Safety considerations must be a vital part of every effort to deploy ML tools into clinical practice [107]. This is an essential step to leverage the strength of ML models to predict diseases. Identifying these diseases before they can cause greater harm is necessary for patient recovery [108].

For successful AI systems in healthcare, one needs to be able to trust these systems. Explainability in the form of XAI can be key to achieving this [29]. Van der Veer et al. [109] performed an intriguing study where the audience could decide how accurate the prediction model was and how much they were willing to sacrifice it for enhanced explainability. Interestingly, the willingness to trade accuracy for explainability was lower in healthcare than in non-healthcare settings. This emphasizes the demand for complex ML models in combination with post-hoc explanation methods in healthcare since these methods do not impact model performance [83].

A good number of articles are available that explore the possible application of different XAI methods in healthcare settings. However, the majority of these papers only deal with the model performance and at most limited scope of explanations to retrieve insights of the model's reasoning. These articles almost solely apply a subset of the main established XAI methods SHAP, LIME, PDP, and Feature Importance (see [110], [111]), apart from neural network-specific explanations (see [112], [113], [114], [115], [116]). SHAP and LIME as stand-alone standards are problematic due to defects like occasional instability and divergence from theoretical properties [117]. Critical assessments or a wider range of explanation methods are generally not included.

There is no consensus on how XAI should be used to facilitate the implementation of ML models in clinical practice [112]. For HCPs specifically, none or only variable importance has been used so far to explain the ML prediction models (see [58], [118]), while only Langenberger, Schulte, and Groene [46] applied a limited SHAP explanation. Most existing literature has approached HCP predictions narrowly with a sole focus on predictive power. These studies highlight the promising insights these HCP prediction models offer.

There is a clear need for methods that explain and allow trusting HCP prediction models. Knowing who will account for great spending in the future can enable cost-effective, patient-tailored interventions [50]. However, a major challenge that prevents the implementation of these models is the lack of model explainability and consequentially trust [29]. XAI has the potential to close

this gap between research and clinical practice by rendering ML models more safe, transparent and trustworthy [112]. Better explanations are inevitable to leverage the superior predictive power of these ML models in the sensitive area of healthcare that will allow better care for HCPs in the future. To date, a study that delivers an HCP prediction model along with comprehensive explanations is missing. The current literature lacks research on the critical aspects of trust and the explainability of ML models for predicting HCPs.

We aim to substantially advance the healthcare field by reducing the shortcomings of current approaches regarding HCP prediction models. Our primary motivation is to facilitate the successful deployment of ML models for HCPs. The innovative character of this study is to append explanations to the model creation and validation, thus extending the procedure from other healthcare settings to HCP prediction models.

The selection of explanation methods is oriented on the comprehensive list of techniques from [100]. We approach the explanations as far as possible, following traditional logistic regression analysis procedures. We focus on coherent, rigorous model development, validation, and explanation processes. Only such a rigorous development process allows an effective deployment of models in healthcare with XAI methods (see [119], [81], [82]). Even though such novel approaches for cost saving might take years or even decades to establish themselves as a healthcare standard due to their innovative nature and unusual requirements [48], we should not be discouraged by these obstacles in trying to achieve a better and more effective healthcare system. This work helps to equip healthcare with new powerful technologies that have the potential to substantially better care for HCPs.

OBJECTIVES

This thesis aims to answer to the following research questions concerning the identification of high-cost patients:

- How do different predictive models compare?
- How do XAI methods help explaining these models?
- Are the given explanations accurate and understandable enough to support the routine use of XAI in healthcare optimisation?

The following sections will present the results obtained through a two-step approach:

- Firstly, the thesis will determine a high-performance model. A similar analysis was done by Langenberger, Schulte, and Groene [46] in a previous work on equivalently collected but older data. We perform a similar model selection on more recent data.
- Secondly, the thesis will explore the range of applicable explanations, and evaluate their results, to identify the best approach for the specific medical prediction model setting. We start by describing our data collection methods, and subsequently the process of applying and selecting a prediction model. Then, we review the explanations and assess the insights they offer to understand the model.

The research will focus on a core set of clinical and demographic characteristics identified as potential risk factors for HCPs [66], without including detailed social or behavioural information that was found predictive by the recent literature [52].

METHODS

STUDY POPULATION

The primary study units are beneficiaries of state insurance companies in the German regions Werra-Meißner and Schwalm-Eder. The intermediary OptiMedis granted access to the claims data. A bilateral contract between the state insurance companies and OptiMedis allowed the use of the data for this research.

The complete data used in this study includes the four years from 2019 to 2022. In this way, we included all available data of the insurance companies. The collected data include basic demographic information e.g. age and sex, care dependency, and participation in a Disease Management Program (DMP). DMPs aim to better chronic disease care by improving healthcare coordination across insurance companies and care providers. Chronic diseases that are included in this program are diabetes mellitus (Type 1 and 2), breast cancer, and asthma, amongst others [120]. Additionally, we recorded their healthcare utilization in the form of all inpatient and outpatient diagnoses and all prescribed medications during that year for every beneficiary. Hence, with our data structure, we continued the procedure from Langenberger, Schulte, and Groene [46].

Our goal is to predict HCPs. We define them as the patients that are in the top 5% quantile of the cumulative claims costs per year which is common in the literature [58], [48], [121]. We then predict whether the patient is in the highest 5% percentile for the cumulative costs in the subsequent year. Every individual within this percentile is referred to as an HCP; otherwise, they are considered non-HCPs. The outcome is the HCP status of a patient in the subsequent year.

Therefore, data from two subsequent years is required to predict HCPs. The first year in such a pair includes the individual healthcare records. The second year is used to retrieve the true outcome for these individuals. We used the individual healthcare records from 2019-2021 to predict the HCP status in the subsequent year, with the respective outcome indicating an HCP in the subsequent year (2020-2022). If a patient died in the first year from which predictors were retrieved, we excluded that patient. Costs for that patient would be zero for the next year. If, conversely, a patient died in the second year we included that patient.

We retrieved three data sets containing medical healthcare records for individuals:

- records from 2019 with outcomes recorded in 2020 (N=34,314)
- records from 2020 with outcomes recorded in 2021 (N=34,632)
- records from 2021 with outcomes recorded in 2022 (N=35,350)

Following the train-validate-test scheme, the data sets with records from 2019 and 2020 were combined (N= 68,946) and randomly split into 75% train (N= 51,709) and 25% validation (N= 17,237) data sets. Finally, we tested the models on an equally structured, unseen test data set. This test set contained records from 2021 with outcomes in 2022 not previously used (N= 35,350).

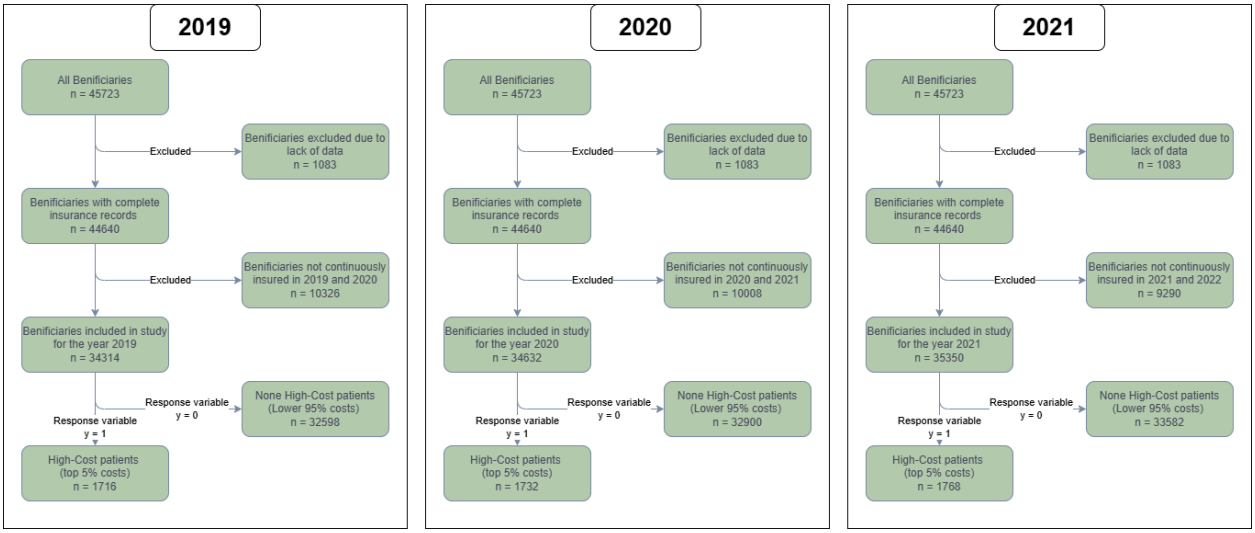

Figure 2: A flow chart indicating the number of patients being included in the data sets for each of the years 2019-2021, respectively.

STATISTICAL ANALYSIS

The study followed the steps required for effective model development to maximize chances of broader adoption in real-world applications [119]:

1. **Step 1: Problem definition and data inspection:** A critical step is to consider the prediction problem by defining the research question, evaluating known predictors, selecting patients, assessing treatment effects, measuring predictors, and including endpoint considerations.
2. **Step 2: Coding of predictors:** One must decide about the scale of continuous variables and the coding of categorical variables. User-friendly variable formats can be considered.

3. **Step 3: Model specification:** This step includes model specification the choice of which variables to include. The selection of predictors should balance statistical methods with clinical knowledge to avoid overfitting and ensure robustness.
4. **Step 4: Model estimation:** In this step, the parameters and coefficients are estimated. Adequate methods should be used for estimation.
5. **Step 5: Model performance:** Appropriate measures for model performance such as calibration and discrimination should be applied.
6. **Step 6: Model validity: T**he model is validated both internally (stability) and externally (generalizability).
7. **Step 7: Model presentation:** Prediction models should be presented in adequate formats to best meet clinical needs.

STEP 1: PROBLEM DEFINITION AND DATA INSPECTION

As an introductory step, the thesis carried out a literature review to outline the common criteria for defining HCPs. We already described predictors and endpoints in the section on the study population.

STEP 2: CODING OF PREDICTORS

Diagnoses data were reported using the German modification of the 10th revision of the *International Classification of Diseases* system (ICD-10), which is also used for billing purposes. The disease classification codes are composed hierarchically. The first three characters indicate the disease category, with increasing granularity [122]. We reported the presence or absence of each three-digit and one-digit ICD-10 code in the specific year for every individual, respectively. This includes 270 columns for both inpatient and outpatient diagnoses. Additionally, we used the *Anatomical Therapeutic Chemical* (ATC) classification system to include information about drug prescriptions. The ATC system serves as a classification tool for drug utilization [123]. Again, the three-digit level codes that indicate the second-level anatomical/pharmacological groups were used to group individual prescriptions. As in Langenberger, Schulte, and Groene [46], the 14 one-digit groups of the ATC prescription codes that indicate the main anatomical/pharmacological [124] were included. This led to 110 columns for the ATC codes. Together with the demographic and general utilization variables, the data set in this study included a total of 659 predictors.

STEP 3: MODEL SPECIFICATION

We compared popular ML classification algorithms, including random forests, gradient-boosting machines, and neural networks. Logistic regression is used to compare these models against a traditional statistical model.

Logistic regression

Traditional logistic regression was used to predict the logarithmic odds of an outcome in a binary classification task [125]. The model predicts the outcome by assigning the logarithmic ratio of the probabilities belonging to class '1' vs. class '0' [126]. All variables are assigned the label of class '1' above a model-specific threshold, otherwise, they are '0' [127]. The underlying formula assumes a linear relationship between the logarithmic odds and the independent predictors for logistic regression models [125]. Amongst various advantages offered by logistic regression is the possibility of applying score models in medical settings (see [128], [129], [130]). Logistic regression does not require any hyperparameters to be specified to fit the model.

Random Forest

The random forest classifier selects independent, similarly distributed tree-structured classifiers. Independence is achieved by randomly sampling only a subset of observations (bootstrapping) and random features that will be used to fit the individual classifiers [131]. A classification tree is grown on each of these samples and feature subsets, according to the *CART* algorithm. Based on this 'random' ensemble of trees, each of these classification trees produces a prediction for an observation [132]. The most popular class among all tree-structured classifiers is the final prediction of the random forest model [131]. This combination of statistical prediction functions trained on bootstrapped samples is called bagging [132]. As parameters, random forests require the total number of classification trees and the features used per tree-structured classifier. Choosing fewer candidate variables and greater sample sizes leads to more stable results. At the same time, it potentially worsens its predictive performance [133]. Different strategies exist for observation sampling. This includes drawing samples with and without replacement and sample sizes either the same size or smaller than the original training set. In comparison to other highly complex ML models like SVM, random forests are relatively invariant to parameter tuning [133]. Using such random tree ensembles instead of single decision trees is shown to reduce variance and thus

prevents overfitting of the model, while at the same time maintaining its high predictive power [134].

Gradient Boosting Machine

Gradient boosting is based on finding an 'optimal' element that sequentially minimizes the expected loss function. It searches the function space which includes all possible functions between the explanatory and the response variables. Additive expansion is at the heart of gradient boosting and other popular ML algorithms like neural networks and SVMs. It restricts the function estimation to a sum of simple parameterized functions:

$$F(\boldsymbol{x}; \{\beta_m, \boldsymbol{a}_m\}_1^M) = \sum_{m=1}^{M} \beta_m h(\boldsymbol{x}; \boldsymbol{a}_m)$$

Here, F is the prediction function that maps $\boldsymbol{x}$ to the outcome $\boldsymbol{y}$ with parameters β_m and $\boldsymbol{a}_m$. It is composed of simple prediction functions $h(\boldsymbol{x}; \boldsymbol{a}_m)$ with weights β_m [135]. Examples of these simple functions are tree stumps, SVMs, and wavelet functions [136]. Instead of directly solving this equation, gradient boosting optimizes the simple learners stepwise. The optimization of such a simple learner is based on the 'pseudo-residuals' (the gradient descent of the loss function) of previously assembled learners [137]. In classification, the negative binomial log-likelihood is differentiable and thus a suitable loss function [135]. We use trees as simple learners. Hyperparameter tuning involves two parameters. One parameter again is the number of simple learners included in the model. Additionally, the maximal tree depth of the simple learners can be varied.

Artificial Neural Network

Artificial neural networks (ANNs) were introduced by McCulloch and Pitts in 1943 and stem from the impulse propagation present in the neurons of our nervous system [138]. It consists of artificial neurons that are activated by the inputs they receive, leveraging an activation function. This neuron in combination with weighted inputs and an activation function is called a 'perceptron' [139]. In their most basic form without any hidden layers and sigmoid activation function, the resulting model is equivalent to predictions from a logistic regression model [140]. A layer of an ANN is a group of artificial neurons with the same inputs and varying weights. Additional 'hidden' layers between the input and output allow for non-linear functional relationships [139]. We applied such

an ANN in this study, which is referred to as a multi-layer feedforward ANN. As for logistic regression, the aim is to maximize the likelihood, leading to the *cross-entropy error* as the appropriate choice for binary classification [140]. Parameter choices include the number of hidden layers and neurons per layer as well as the activation function and the learning rate [141].

PREDICTOR SELECTION

Model specification includes the choice of predictors [119]. In the introduction, we determined various diagnostic and non-diagnostic relevant predictors for HCP identification (e.g. see [52], [50], [53], [47]). However, we rely on statistical methods to select relevant predictors. Logistic regression offers straightforward strategies to perform variable selection. It provides a direct test for the statistical significance of the model predictors [140]. We tested the full model against a model on which a stepwise backward selection of variables was performed. The statistical significance level is set to 0.05. Since backward variable selection is a discrete, stepwise process, it is highly susceptible to even changes in single observations. Variables might not be selected due to a change in a single observation. Many alternatives exist, like forward, stepwise, or purposeful selection [142]. We chose the *least absolute shrinkage and selection operator* (LASSO) approach as an alternative to compare against backward selection. LASSO mitigates such singular effects known from backward selection by continuously shrinking variables and often results in variables with a coefficient of 0 [143]. While originally developed only for linear regression, the LASSO method offers an efficient adaptation for logistic regression [144]. LASSO regression requires parametrization of the shrinkage control parameter lambda in the form of a lagrangian penalty [143]. Here we set it to 0.01 to obtain a model with a limited number of parameters. We performed LASSO variable selection and compared the resulting models via AIC, an established measure in different disciplines for non-nested regression models [145]. Additionally, a likelihood ratio test was performed. It is a common model selection procedure for nested model comparisons. It tests whether a reduced (nested) model explains the data equally well as the full model [146].

STEP 4: MODEL ESTIMATION

We followed the established cross-validation approach to select the best hyperparameters, commonly known as *tuning*. This approach is the standard method for robust accuracy estimation during training in ML in general [147], and an established procedure in healthcare during the training phase of prediction models [80]. We chose the most popular *k-fold cross-validation*

approach like Langenberger, Schulte, and Groene [46] to determine which hyperparameter set performs best. For k-fold cross-validation, the data set is split into k parts of similar size. The model is trained on all these k data subsets, except one. After training, the remaining subset is used as a test set for the trained model to estimate model accuracy or, as in our case, the *area under the receiver operating characteristic curve* (AUC). This procedure is repeated for all k subsets [148]. Finally, the overall model performance is measured by the average AUC over all subsets, a typical strategy in combination with k-fold cross-validation [133].

We specified different search spaces in the form of grids that contain hyperparameter values. For all these parameter sets, models are evaluated according to their cross-validation AUC. Consequently, the parameter set that results in the highest AUC is chosen. For the random forest model, the hyperparameter grid included the different numbers of trees (*250, 500, 1000*) and the number of randomly selected variables (*10, 20, 30*) as dimensions. For the gradient boosting machine, we varied the number of trees (*250, 500, 1000, 2000*) and the tree depth (*1, 3, 5, 10*). Finally, the search space of the ANN hyperparameter grid included 4 dimensions: the number of hidden layers (*1, 2*), the number of neurons per layer (*10, 20, 100*), the activation function (*sigmoid with/without dropout, rectifier with/without dropout, maxout with/without dropout*) and the learning rate (*0.003, 0.005, 0.007*). The H2O package in R was used for k-fold cross-validation and hyperparameter tuning.

STEP 5: MODEL PERFORMANCE

After choosing the parameters, we perform model selection using a training-validation scheme design. In contrast to different designs like *Leave-One-Out* (LOO) or *y-randomization*, this approach allows the estimation of the true performance through an external data set [149]. Our data structure is equivalent to that of the HCP analysis by Tamang et al. [51]. Three consecutive years (in our case 2019, 2020, and 2021) with labels for the subsequent year were used, respectively. However, the training-validation division differs from their approach. To ensure accurate model estimates by a greater training set, we combine samples in the years 2019 and 2020. Of this combined data set, we randomly sample 75% to create the training set. The other 25% is the validation data set as in the previous study [46].

For all four models, the AUC is reported to enable direct comparison with [46]. It is our key performance measure. The underlying receiver operating characteristic is defined by the false vs.

the true positive rate across all possible thresholds [150]. Accuracy only measures the mere agreement of the final model classifications with the true labels. Information about probability estimates for the samples is lost. In contrast, the AUC emphasizes the importance of ranking the samples according to the model predictions. It captures the model's performance over all possible thresholds and error costs [151].

We additionally include standard performance measures in ML to enable comparison across the literature. The most common measure for classification tasks is accuracy [152]. It is defined as the number of all correctly classified samples divided by the number of all samples [153]. Sensitivity or recall is the number of true positives divided by all positive samples. It measures the effectiveness of identifying positive labels. Similarly, specificity is the number of true negatives divided by all negative samples. It measures the ability of the classifier to identify negative samples [152]. In imbalanced training sets, accuracy can be misleading. The *geometric mean* of the sensitivity and specificity (g-mean) is a more appropriate measure that considers this imbalance. Sensitivity and Specificity can be seen as the accuracies of the model on only positive or negative samples. Hence, the underlying aim of the g-mean measure is to value correct predictions of positive and negative samples equally [154]. Because of these properties, we use the g-mean to determine the best cutoff threshold for each model. Such *thresholding* is a common cost-sensitive method to handle imbalanced data [155]. Possible alternative measures for cost sensitivity analysis include the *F-score* and the geometric mean of precision and recall [154], that emphasize different desirable balance properties.

As suggested by Steyerberg and Vergouwe [119], we additionally perform a *decision curves analysis* (DCA) on all models to enhance their clinical usefulness. Such a decision curve assists decision makers whether to perform a treatment or not, based on the diagnostic certainty for a condition according to the model. Decision-makers may include physicians or, as in our case, insurance company personnel. This analysis allows these decision-makers to individually choose whether to use a model and adapt their management strategies accordingly [22]. DCA assumes that a certain patient is treated only if the patient’s diagnostic certainty (prediction) is above a certain probability threshold. For all patients above this threshold, a net benefit score is computed. It takes into account the number of patients with and without the condition above the respective threshold and is computed as:

$$net\ benefit = \frac{\left(true\ positives - \frac{threshold\ probability}{1 - threshold\ probability} \times false\ positives\right)}{Total\ number\ of\ patients}$$

This net benefit balances the benefits of treating patients with the condition versus the harm inflicted by treating patients without the condition for a certain threshold. The DCA then displays the net benefit score for all probability thresholds for obtaining the treatment versus no treatment for all patients [156]. The probability threshold is the minimal ratio of true prospective HCPs among the predicted patients so that the benefit of an intervention outweighs its harm of some sort, e.g. financial costs. It can be regarded as the level of certainty we accept to target a beneficiary with some kind of intervention. For example, consider our intervention being so expensive that among three treated patients, at least one should be a prospective HCP to be financially sustainable. In that case, we can set the threshold probability to 33%. Only patients with a risk above that threshold will be treated. If instead, we target patients with an intervention where the burden of unnecessary treatment is low, we can choose a lower threshold [156]. For example, we could decide that for every treatment of a prospective HCP, we allow treating nine prospective non-HCPs. Then, our probability threshold would be at 10%.

The net benefit itself is hard to interpret directly. However, a positive net benefit tells us that the ratio of prospective HCPs compared to prospective non-HCPs is higher than the probability threshold. Hence, one would profit from applying the model. Decision-makers can individually determine the trade-off between harm and benefit of the treatment by choosing the threshold. For each threshold, one should choose the model with the highest net benefit for an optimal outcome. As in [58], we use decision curves to compare our HCP prediction models against each other. Different models can be displayed and thus compared in a single plot. Steyerberg and Vergouwe [119] suggest the additional display of two trivial management strategies that do not depend on risk assignments, namely treating all and no patients.

Another measure to evaluate models in this medical context is the *cost capture* from Tamang et al. [51]. We report this cost capture for all our models to support comparison across HCP analyses in the literature. Cost capture describes the ratio of HCP costs the model can predict and is defined as:

$$Cost\ Capture = 100 \times \frac{\begin{array}{c}\text{Total costs of the } k^{th} \text{ percentile} \\ \text{with the highest predicted probabilities for being an HCP}\end{array}}{\begin{array}{c}\text{The 'true' total costs of the } k^{th} \text{ percentile} \\ \text{with the highest costs (HCP)}\end{array}}$$

[51]. In our case, we are interested in the highest 5% percentile of total costs. This measure allows a direct approximation of the model's cost-saving potential and effectiveness.

All measures are reported with their 95% confidence interval. To assess the confidence of the accuracy, sensitivity, specificity, and g-mean, we use the Wald standard interval for asymptotically binomially distributed properties as in [157]. Even though this interval has been criticized due to its erratic behaviour for rather small sample sizes or values close to 0 or 1 [158], due to the vast number of samples and values not close to 0 or 1, the Wald interval is expected to yield accurate confidence estimates in our case. The cost capture uses non-parametric bootstrapping as introduced by [159] to retrieve estimates for the confidence interval. Such empirical approximations are known to result in good confidence interval estimates, often superior to standard intervals [160]. For the computation of the ROC-AUC, a computationally efficient alternative to bootstrapping is used as in LeDell et al. [161].

STEP 6: MODEL VALIDITY

We ensure the internal validity of our model by the previous model specification and evaluation design. It includes cross-validation and a split into train and validation data sets. In addition, this paper constitutes an external model validation in the form of an independent test set. As in Tamang et al. [51], our test set consists of the last year, with data from 2021 and outcomes for 2022 for the final model validation. Since all medical records from 2021 are in the test set and all previous records in the full training set, our model is validated via temporal variation. In addition, even stronger external validation applies to the model concept previously presented in Langenberger, Schulte, and Groene [46]. We used the same model creation process and data structure but performed our analysis in different regions (geographical variation) and years (temporal variation) compared to this study. Such an external validation ensures even stronger generalizability than internal validation by emphasizing transportability [119]. All final models are implemented and evaluated within the R ML platform H2O [162].

STEP 7: MODEL PRESENTATION

One of the most important aspects of the model development cycle is the question of how to present the knowledge of the model to address the actual clinical needs [119]. Model presentation is indispensable for successful model deployment in clinical practice. A helpful perspective here is the notion of model transparency [82]. Logistic regression models are always transparent, while levels of transparency range can vary, depending on the interpretability and number of predictors: *simulatable models* are comprehendible and can be simulated by a human; *Algorithmically transparent models* on the other hand contain too many predictors to be analysed by a human without additional tools [82]. In logistic regression, model presentation includes the display of statistically significant predictors and their coefficients and embedding them in appropriate formulas for seamless interpretation. Common alternatives to such formulas are score charts or web-based calculators [119].

In contrast, the complex models used in this study (random forest, gradient boosting machine, neural network) are considered all non-transparent [82]. Simple formulas that condense the learned knowledge of these complex black-box models into a few lines of formula on paper are not obvious. Their presentation designs in healthcare must consider interpretability and explainability [80]. It is believed that carefully designed model presentations in the form of explanations can facilitate clinicians using complex ML models [106]. The field of XAI can be key in creating explanations that close the gap between ML research and clinical practice [112]. In the following, different approaches and methodologies in the field of XAI for model explanations are presented.

XAI DIMENSIONS

Different categorizations and especially different naming conventions for the same XAI concepts are pervasive. Two main categories can be distinguished in most of the recent literature. We refer to them as model interpretability and post-hoc explainability (see [82], [81]), respectively. Some authors refer to model interpretability as *model explainability* [81], *transparent models* [82], [100], or *interpretable models* [85]. This field of XAI is concerned with building accurate ML models that are intrinsically transparent, including methods like logistic regression, decision trees, and k-nearest-neighbour models [100]. A decisive characteristic of interpretable models is the presence of some kind of constraint like monotonicity or linearity for regression models [84]. Similar terminology diversions exist for post-hoc explainability which is also described instead as *post-hoc*

interpretability [83] or of *post-modelling explainability* [85]. Post-hoc methods do not build interpretable models but encompass all means to explain complex models after they were trained [99]. In this study, we are only concerned with post-hoc explainability due to the complex nature of the best-performing model we aim to explain.

Several categorizations are proposed for post-hoc explainability methods. Almost always they are primarily categorized as either *model-agnostic* or *model-specific* (see [81], [82], [85], [98], [100], [102], [106], [163]). Model-agnostic methods refer to XAI methods that do not depend on the inner workings of a model to generate its explanations. Model-specific XAI methods, on the other hand, make use of the properties of certain ML models for their explanations [100]. Hence, they are only applicable to certain model types. This naming convention is consistent throughout the literature.

Another popular categorization is the one of *global* versus *local explanation* methods (see [81], [98], [102], [106], [163]). While this distinction is also known as *dataset-level* versus *prediction-level* interpretation [83], others list local explanations just as one type of explanation method amongst many others like *visual* or *textual explanations* (see [82], [100]). Global explanations focus on explaining the whole model behaviour. On the other hand, local explanations aim to unveil the reasons for the model's prediction of a single instance [106].

A third category dimension regards the design in which XAI methods convey their explanations. This is referred to as the *type of explanation* (see [82], [85], [97], [164]). Most existing taxonomies list explanation types only for model-agnostic explanation methods [82], but not all of them [164]. Examples of explanation types include *explanation by example*, *contrastive explanations*, *textual explanation*, *visualizations*, and *simplification* or *surrogate model explanations* (see [85], [97], [164]). Categorizations and naming conventions differ fairly, however. For example, explanations by example [82] are referred to as *data point* explanations [164] or *what-else* explanations [97].

The most diverse and broadest field of interest in XAI concerns Neural Networks and deep learning. (see [116]). However, we focus on model-agnostic as well as tree-based explanation approaches.

XAI METHODS

We adopt similar explanation methodologies as suggested by [100]. Their procedure includes common XAI methods for which established implementations exist. Our structure for explaining

the complex model takes logistic regression models as a standard. We explore explanation methods that offer similar insights as can be drawn from logistic regression analysis, proceeding from more general global explanations to local explanations for a single sample. This way we explore a unified explanation approach for ML models that resembles the established procedure for simple models.

In logistic regression, as a first step significant predictors are determined, and model reduction techniques are used to fit models with a reduced number of predictors [119]. In addition, it is obvious to observe the impact magnitude a predictor has on the model predictions by simply looking at its coefficient. At least for situations with comparable units as in our case with diagnosis and prescription indicators (either 0 or 1), coefficients can be compared directly. For complex models, however, the importance of predictors is not trivial. Different global methods exist to measure variable importance [165]. Such variable importance indicates how much a variable changed the model outcome on average. SHAP summary plots are an increasingly popular global method with the rise of the XAI that extends sole variable importance plots [99]. This method offers more granular insights into the type of variable influences.

Variable importance

Different variable or feature importance measures exist. We focus here on a classical feature importance method for tree ensembles. A famous permutation-based method was introduced in the original random forest paper. It uses out-of-the-bag error estimates and randomly permutates values of a specific variable to compare them with the error estimates of the original values [131]. Recently, a model-agnostic version of this measure was introduced [166]. We use a different method, however, called Gini variable importance [165]. The H2O package implements this variable importance measure. It uses the internal tree structure of the random forest model by comparing the node purity of the tree predictions before and after using the specific variable as a splitting criterion. Its importance is determined by the total reduction of the squared error between the node and the child in a tree for all nodes that use that certain variable as a split criterion. This is done over all trees in the model. The node purity can be interpreted as a measure of the discriminative power of our random forest, and the importance of a variable is then the relative contribution of the variable to the accuracy of the model [165].

SHAP summary

SHapley Additive exPlanations (SHAP) is a game-theoretic approach that uniquely determines 'fair' feature contributions of the model prediction. It is based on Shapley values. Three desirable properties uniquely define Shapley values: Local accuracy, missingness, and consistency [167]. The unique solution to these Shapley values always exists [168]. For the set of all features F and a feature $i \in F$, its Shapley value $\emptyset_i$ is defined as:

$$\emptyset_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!} [f_{S\cup\{i\}}(\boldsymbol{x}_{S\cup\{i\}}) - f_S(\boldsymbol{x}_S)]$$

Here, $f_S(\boldsymbol{x}_S)$ is the model trained on the values from $\boldsymbol{x}_S$ on all features in S [167]. The values for the remaining features are drawn repeatedly from the deterministic conditional distribution. The resulting model prediction for the set S is then the average model prediction on all drawn samples. The innovative character of SHAP is to transfer the properties of Shapley values to an additive feature contribution method to explain any ML model [81]. SHAP values are the Shapley values of the model's conditional expectation function. Each feature is assigned its 'fair' (game-theoretical) contribution for a single observation. Generally, approximations are required to retrieve the SHAP values [167]. Despite its recent introduction, SHAP is by far the most popular method for post-hoc explanations [114]. SHAP values can be used for global (SHAP summary) and local (SHAP local) explanations. While SHAP local only uses the SHAP values for a single sample, SHAP summary combines the SHAP values of all samples in a single plot. The variables can then be ranked by their average SHAP absolute value, resulting in a feature importance plot. Unlike the previous feature importance method, in such a SHAP summary plot one can inspect the impact qualities and magnitudes of these features by the x-axis position of the samples, which are displayed as dots [81].

Next, we aim to investigate the relationship between a feature and the model prediction, the *feature effect* of the model. For simple regression models, relationships between predictors and the model response are easily interpretable. In linear regression, the coefficient of a predictor is just the response increase if this predictor is increased by one unit. For logistic regression, a linear relationship is constructed via the logit function. This function allows direct interpretation of the coefficient. If the predictor is increased by one unit, the natural logarithm of the odds of the outcome (ratio of the probability of positive outcome divided by negative outcome) increases by the amount of the coefficient [169]. While feature effects for logistic regression are hence defined

by its design and are linear apart from some predefined transformations, this is not the case for more complex models. For these models, relationships between variables and model prediction can be highly non-linear. Even after knowing which variables are important for the complex model predictions, it remains unclear how exactly these variables influence the predictions. Global explanation methods like *Partial Dependence Plots* (PDP) [135] and *Accumulated Local Effects* (ALE) [170] unveil the variable effects on the model predictions. Single and pairs of variables can be inspected this way. We focus on the most important variables following the previous variable importance plots. In addition, we include the age variable in our explanations as a demographic variable.

PDP

Friedman was the first to describe *Partial Dependence Plots* (PDP) to accomplish insightful visualizations [135]. The main idea is to display how the model's prediction depends on one or a few features (hence the name *partial dependence*) [99]. We call this feature set $S \subset \{1, \dots, p\}$, where p is the total number of features. If the dependence between features in S and its complement feature set $\{1, \dots, p\} \backslash S := C$ is not strong, the conditional probability for $\boldsymbol{x}_C$ can be approximated well by its marginal distribution [135]. Given specific values x_S for the features in S, the model's prediction function f then is marginalized over C to obtain the expected value of f for $\boldsymbol{x}_S$ [171]. The results $f_{S,PD}$ is a low-dimensional representation of the average model prediction for the specific values in x_S [102]. This way, the behaviour of the model can be described for a manageable number of features. The underlying theoretical foundation for such plots is given by the formula:

$$f_{S,PD}(\boldsymbol{x}_S) = E_C[f(\boldsymbol{x}_S, \boldsymbol{X}_C)] = \int f(\boldsymbol{x}_S, \boldsymbol{x}_C) p(\boldsymbol{x}_C) d\boldsymbol{x}_C$$

[135]. For actual computation, the marginal and joint probabilities are generally unknown, but can be approximated empirically by using the empirical distribution:

$$\hat{f}_{S,PD}(\boldsymbol{x}_S) = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_S, \boldsymbol{x}_{i,C})$$

PDPs reveal general relations between the response and the selected features [135]. However, they are simplistic and generally ignore feature interactions [99]. Hence, independence between the feature sets is required for such trends to be reliable. If this assumption is violated, heterogenous

effects might exist but will not be visible in the PDP [102]. A related limitation caused by dependent features regards the extrapolation of the feature values, which can lead to averaging over unrealistic data [170]. However, this can also be seen as a strength of this explanation method. A major problem for many ML models is the sometimes erratic and unplausible predictions on data that are substantially different from the training data [86]. PDPs let the user inspect the model dependence on the complete data set. Hence it shows how features affect the model even when the data might change in the future.

ALE

A major shortcoming of PDP plots is their incapability to account for correlated features [172]. An alternative that handles this shortcoming is *Accumulated Local Effects* (ALE) plots [102]. It uses the conditional instead of the marginal probability density function. This way the extrapolation of data based on the conditional probability is restricted to samples inside the training space and hence closer to actual observations than in the PDP. In addition, instead of averaging directly over the model predictions, ALE averages and accumulates the changes in prediction, which are also called *local effects*. Such extrapolation and focus on local effects instead of sole model predictions prevents the inference of correlated features [99]. The ALE effect on $\boldsymbol{x}_S$ is then defined as:

$$f_{S,ALE}(\boldsymbol{x}_S) = \int_{\boldsymbol{x}_{min}}^{\boldsymbol{x}_S} E_C\left[f'(\boldsymbol{X}_S, \boldsymbol{X}_C)|\boldsymbol{X}_S = \boldsymbol{z}_S\right] d\boldsymbol{z}_S - constant$$

[170]. The constant is chosen so that the plot will be centered vertically. Local effects are accounted for by the first-order derivative of the prediction function f'. The accumulation of the local effects then shows how the feature affects the model response across all its values [81]. The y-axis shows these accumulated local effects in comparison to the expected (average) accumulated local effects of the feature. Apley and Zhu [170], the inventors of the ALE explanations method argue that especially more flexible, non-linear models are likely to profit from this method that constrains the data to be more realistic. In opposition to simpler models, their predictions of samples outside of the training space are generally much less reliable. This assures reliable feature effect estimation even when correlation is present. In addition to these benefits, ALE plots require generally less computation power than PDP plots [102].

Finally, we aim to investigate the decisive features and their influences on the model prediction of a single sample via *feature attributions*. While global explanation methods aim to reveal as much of the model's behaviour as possible, local methods are characterized by their focus on individual-level explanations. For simple logistic regression models, the predictions for a single sample can easily be investigated by inserting all predictor values along with their coefficients into the logit formula and comparing the result to the decision threshold of the model. The impact of every

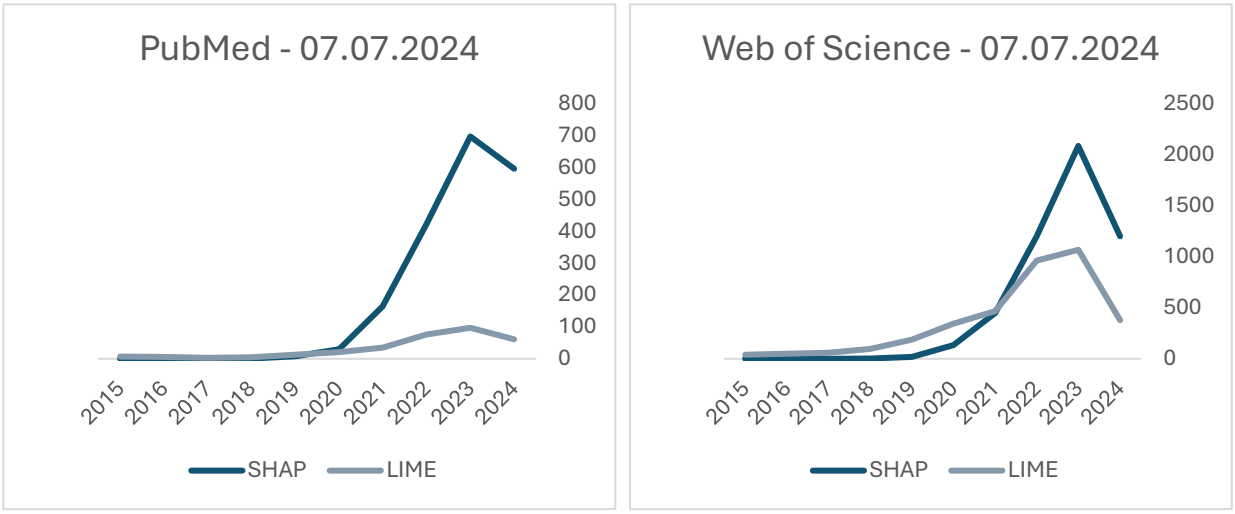

Figure 3: Popularity of the common local XAI methods SHAP and LIME in the literature in comparison over the time.

predictor is clear. In contrast, for complex models, it is generally not clear how single feature values of a sample contributed to the model's prediction. Various local explanation methods are dedicated to investigating these feature attributions for black-box models. We focus on the most important local explanation methods, including local SHAP [167] as well as Local-Interpretable Model-agnostic Explanations (LIME) [173], which are both feature-level explanation methods [106]. SHAP is a *feature contribution* method that quantifies the relevance each feature had on the model prediction for a single sample [82]. Additionally, LIME generates a simplified local approximation of the model to explain the model predictions, which makes it a *simplification* method [85]. Both methods indicate which feature values determined the prediction for a specific sample in an additive manner, making them *additive feature attribution methods* [99]. This means that adding the SHAP values for all variables of a single sample yields the model's prediction. Such a link between input data and prediction is paramount to achieving interpretable model representations in healthcare [80]. These two methods accounted for 45% and 7% of all XAI applications in healthcare in the last years, respectively [114], and have received mounting attention in the last

years (see Figure 3). We exemplarily explain a sample with an actual positive outcome and a medium predicted probability (43%) of having a positive outcome.

SHAP local

For the SHAP method exists probably the largest variety of plots and methods, including SHAP local. Its underlying SHAP values are always calculated for single samples and are the same as for the SHAP summary method. SHAP local allows the calculation of feature contributions of a single sample by ordering and visualizing the SHAP values of all its variables. This way, one can inspect which features influenced the model prediction compared to the baseline feature effects [102]. Such feature effect shows the change in prediction when one conditions on that feature [167]. This additive feature attribution plot emphasizes the local character of the SHAP explanation method. All tested libraries except the *alibi* python package offer the application of using SHAP to explain a single sample.

LIME

The *Local Interpretable Model-agnostic Explanation* (LIME) was developed in 2016 by Ribeiro, Singh, and Guestrin [173]. They place *local fidelity* and *interpretability* at the center of attention. LIME is based on the idea of approximating the original complex model locally by an interpretable model (e.g. linear models or decision trees). It additionally incorporates a measure of the interpretable model's complexity to ensure interpretability. The result is a model that optimally weighs both the goodness of approximation (local fidelity) and model complexity (interpretability). Better approximations come at the cost of a higher complexity of the simplified model [173]. The most common is the LASSO regression approximation with a low number of non-zero regression coefficients. The approximation is based on the generation of new, weighted artificial samples in the vicinity of the sample we aim to explain, making it a perturbation-based method [102]. All numerical variables are dichotomized before the LASSO regression is applied. The feature contributions are therefore not feature contribution averages over different variable sets like for SHAP, but are the product of the LASSO coefficient and its dichotomized variable value, -1 or 1. Consequently, the basic difference between LIME and SHAP is that the SHAP constraints ensure some preferable theoretical properties based on its extensive iterations [100], while the LIME algorithm creates a local simplified model that is more intuitive and computationally less expensive.

In addition, we implemented the XAI methods *Anchors* [174] and *Counterfactual Explanations* (see e.g. [175], [176]). However, due to too long computation times, explanations from these methods could not be included in this study.

EVALUATION OF EXPLANATIONS

So far, there is no consensus in the scientific community on appropriate measures that assess the goodness of an explanation. The predictive accuracy of a model can be easily reported. In contrast, the quantification of the descriptive accuracy (i.e. measure of the degree to which XAI methods unveil the relationships the model learned) remains an open challenge [83]. No established framework for evaluating explanations exists so far [177]. In a conceptual paper, Carvalho et al. [164] identified three approaches to evaluate the descriptive accuracy of an explanation: *Functionally grounded*, *human-grounded*, and *application-grounded*.

Functionally grounded evaluations objectify explainability and thus allow a formal quality assessment of the explanation. An example that resembles the procedures for performance evaluation is the fidelity score. This score measures the goodness of approximation of the surrogate model in comparison to the original (complex) model [81] and hence allows us to estimate the accuracy of the generated explanations. Such a score is only applicable to models that approximate the whole model at least partly and its computation requires access to the approximations. The packages used in this study for local approximations of the original model do not measure the fidelity score, and neither do they give access to the required data to allow its external implementation.

Another functionally grounded evaluation is offered by the *Individual Conditional Expectation* (ICE) method. Even though ICEs are generally considered an XAI explanation method themselves, they are also a powerful tool to evaluate PDP plots by displaying the different variants of the conditional relationship [171]. The diversity of feature effects can be obscured by the averaging of PDP plots.

ICE

Individual Conditional Expectation (ICE) displays analogously to PDP marginal model behaviour for selected features. The main difference is the focus on individual samples. The averaged marginal expectation is disaggregated to show the variation of the response in dependence on a

single feature [171]. This way, ICEs can be used to gain a more granular and detailed view of feature impacts on the model prediction [98]. One is limited to displaying only one feature at a time [81], which prevents the examination of the relationship between two or three features simultaneously as is possible for PDP. As for PDP, also the assumption of independence must hold to obtain meaningful results [102]. Both ICD and PDP do not rely on approximations but can be computationally expensive due to iterations over both samples and feature values [102]. Belle and Papantonis [100] suggest a combination of these ICE and PDP plots as they offer complementary insights with similar underlying characteristics.

Including yet another functionally grounded evaluation, we assess the stability of the local explanations in this thesis. Explanation methods should follow the axiom of *identity* to conform to human intuition. According to that axiom, the same sample should yield the same explanation [164]. We compare outputs for local explanations for the same sample to assess their consistency. These comparisons are limited by the different model implementations underlying our explanations.

Human-grounded approaches are the second method to assess the goodness of explanations in the XAI literature. One practiced method within this approach is to leverage questionnaires to evaluate explanations. Jung et al. [177] reviewed several such studies. The median number of participants was 6.5, a rather small study size. Such an approach was not feasible within this thesis.

The third evaluation approach (application-grounded) is considered the most appropriate type of evaluation. It requires the application of the explanations in real-world applications [164]. Such an evaluation was beyond the scope of a thesis. However, part of application-grounded evaluations is the assessment of form, prerequisites, and obstacles of these explanations in the real world. The format in which AI tools are presented determines their usefulness. Therefore, human-computer interactions must be considered [29]. For this purpose, we reported the relevant implementation characteristics of our explanations concerning their applicability in real-world settings along with their execution times.

PROGRAMMING PACKAGES

This thesis aims to facilitate the employment of our ML model in a real-world application. Because of this, we explore existing libraries that combine several XAI methods and provide a unified interface that paves the way for user-friendly model applications. To implement these methods, no

software solution is available that offers all XAI of the mentioned XAI methods. Therefore, different approaches are leveraged. Firstly, the native explanation methods of the *H2O* package are used [162]. This is most intuitive since our evaluated models are implemented with this package. Additionally, only a very limited number of alternative packages that implement XAI methods in R are available, with *DALEX* being the only established one. It was introduced by Baniecki et al. [178] and offers a unified programming interface to implement various XAI methods. It natively supports R models e.g. from the *randomForest* package in R [179]. In theory, also H2O models can be explained with functions of this model. However, the execution of the H2O models took too long (>24h) on the machine used in this study. Therefore, a native random forest model in R was implemented as a substitute for the original H2O random forest model. This way, we can demonstrate the explanation capabilities of the DALEX package, even though a different model is explained.

Far more packages in the field of XAI are available in Python, which also has much greater community support. Because of this, we decided to extend our exploration of explanation methods to methods implemented in Python. Namely, we used the resourceful, well-documented, and actively supported *interpretML* [180] and the *alibi* [181] Python packages. This way, we ensure wide compatibility of our classification model with existing XAI libraries and frameworks. Again, we had to implement a new random forest model to be able to make use of their explanation methods.

Therefore, it is important to remember that these XAI methods (except the H2O native explanations) explain a model different from the previously evaluated random forest model. We present explanation capabilities and explore different approaches to model transparency. However, we do not attempt to fully explain the previously chosen and evaluated model.

RESULTS

DESCRIPTIVE STATISTICS

Characteristic		2019	2020	2021	p-value	
N=		34,314	34,632	35,350		
Demographics						
Sex						
	m	16,577	16,716	17,076	0.993	
	f	17,737	17,916	18,274	0.993	
Age groups						
	0-30	11,042	11,002	11,197	0.056	.
	30-65	17,947	18,088	18,419	0.056	.
	65-80	4,206	4,293	4,439	0.056	.
	>80	1,119	1,249	1,295	0.056	.
Average age		41.686	41.967	42.071	0.057	.
Utilization						
HC-Patient						
	no	32,598	32,900	33,582	1.000	
	yes	1,716	1,732	1,768	1.000	
Average need of care duration (years)		0.110	0.132	0.151	0.000	***
Average DMP duration (years)		0.700	0.763	0.817	0.000	***
Average total costs in the current year (Euro)		1,832.696	1,860.798	1,918.211	0.142	
Average total costs in the following year (Euro)		1,965.691	2,059.432	2,143.968	0.003	**
Average inpatient number of diagnoses		0.639	0.555	0.562	0.000	***
Average outpatient number of diagnoses		30.493	31.479	33.508	0.000	***
Average number of prescriptions		7.823	7.672	7.687	0.170	

Table 1: Frequencies for the three years included in this study by demographic and main utilization variables are displayed. Significance levels for the variables are reported to test for statistically significant differences between the years. Welch's two sample t-test and ANOVA were used to compute the p-value for numerical variables. For categorical variables, we used Pearson's Chi-squared test.

Firstly, the data sets for the different years are examined in Table 1. Consistent with our definition, 5% of all the patients were HCPs each year. A significant (slight) increase was visible in most of the reported continuous medical indicators between 2019 and 2021. While a plausible explanation for the statistically inconclusive increase in total costs could be inflation over the years, this does not hold for the other indicators. This includes both the average need of care and DMP duration as well as diagnoses and prescription numbers. The only exceptions are the number of prescriptions

and hospital diagnoses (inpatient number of diagnoses) per year, which did not show a clear trend. The sex ratio was constant over the years with slightly more female (51.7%) than male (48.3%) individuals. Hence, the demographic variation including age and sex was not statistically significant. The patient groups for each year are hence comparable.

Characteristics	**Prospective non-HCP**	**Prospective HCP**	**p-value**	
N=	99,080	5,216		
Demographics				
Sex				
m	47,956	2,413	0.003	**
f	51,124	2,803	0.003	**
Age groups				
0-30	32,768	473	0.000	***
30-65	51,803	2,651	0.000	***
65-80	11,471	1,467	0.000	***
>80	3,038	625	0.000	***
Average age	41.049	58.254	0.000	***
Utilization				
Current HCP				
no	96,026	3,054	0.000	***
yes	3,054	2,162	0.000	***
Average need of care duration (years)	0.109	0.550	0.000	***
Average DMP duration (years)	0.675	2.387	0.000	***
Average total costs in the current year (Euro)	1,362.190	11,536.300	0.000	***
Average total costs in the following year (Euro)	1,019.828	21,763.360	0.000	***
Average inpatient number of diagnoses	0.474	2.691	0.000	***
Average outpatient number of diagnoses	29.596	74.517	0.000	***
Average number of prescriptions	6.863	24.121	0.000	***

Table 2: Contingency table for demographic and utilization variables by the actual outcome (HCP in the prospective year). Chi-squared and two-sample (Welch) t-tests are performed to determine statistically significant differences for categorical and numerical variables, respectively. For categorical variables, Pearson's Chi-squared test was used.

Secondly, we examined the differences between prospective HCPs and non-HCPs in Table 2. In contrast to patients from different years, demographic differences between HCP groups were statistically meaningful. Both groups were profoundly different. Our data supports conclusions from the literature that HCPs tend to be of higher age (58.3 years for HCPs in comparison to 41 years for non-HCPs). In addition, the percentage of men in the prospective HCP group was significantly lower than in the non-HCP group.

All utilization indicators showed significantly higher values for patients with high costs for the following year. The differences ranged from an increase of more than twice the amount of outpatient diagnoses for HCPs to an almost 9-fold increase in total costs in the current year (from 1,362€ to 11,536€) and a more than 20-fold (from 1,019€ to 21,763€) increase in the following year. Regarding the costs of actual HCPs, with average total costs of 21,763€, they accounted for about 53% of all healthcare expenditures in our study population. The majority of HCPs did not continue to have high costs in the following year. From the beneficiaries who were HCPs in the current year, only 2162 were in the top 5% cost percentile in the subsequent year, while 3054 did not become HCPs again. However, interpretation must be conducted carefully, since also people who died that year were included.

Figure 4 suggests that prospective HCPs accounted for much higher accumulative costs already in the current year before becoming an HCP. Even though a logarithmic scale was used, the differences between prospective HCPs and non-HCPs are evident.

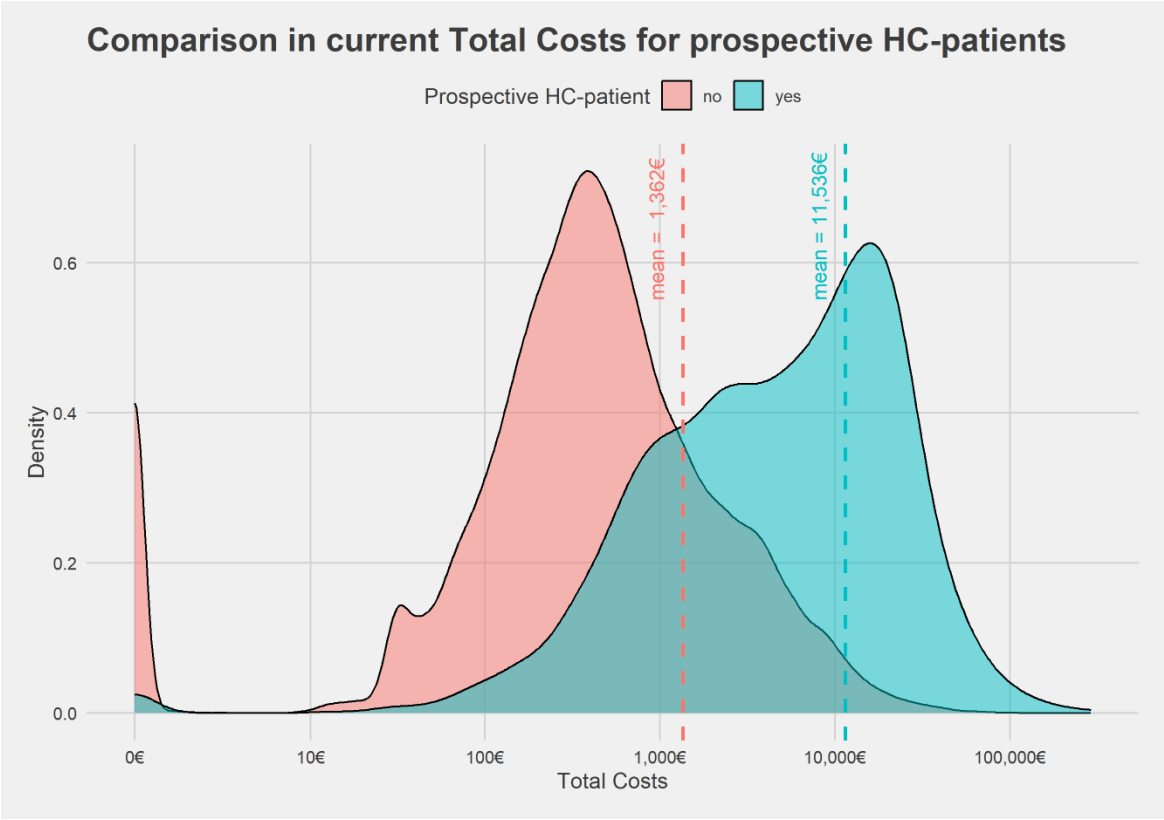

Figure 4: The distributions of Total Costs for prospective HCPs and non-HCPs in the current year. The colour indicates the HCP status for the subsequent year. A logarithmic scale is used. The averages for each distribution are displayed. An almost 9-fold magnitude difference in costs is present between HCPs and non-HCPs.

MODEL RESULTS

MODEL SPECIFICATION

We fitted three logistic regression models. Firstly, after the exclusion of collinear variables, a full model with 617 variables was trained, resulting in AIC_{full}=17007.52. Secondly, LASSO regularization was used as a strategy for variable selection. The penalization parameter lambda was set to 0.01, yielding an interpretable model with a limited number of 11 predictors. The number of predictors in this section always refers to the predictors excluding the intercept. The AIC of this model on this training set was reported as AIC_{lasso}=15298.79, which was lower than the one for the full model. The LASSO regression model contained a subset of the full model. Therefore, the likelihood ratio test for nested models was applicable here. For the negative 2*log-likelihood on the training set the reduced LASSO model resulted in ll_{lasso}=15136.71. For the full model, we obtained ll_{full}=13783.58 and hence a likelihood-ratio $LR=ll_{lasso}-ll_{full}$=1353.13. The p-value of this chi-squared distributed ratio with 606 degrees of freedom was 0<0.05. This suggests that the full model outperformed the reduced LASSO regression model significantly. Consequently, comparisons of the goodness of fit for the models via the likelihood-ratio test and AIC were inconclusive and did not allow for a clear model preference between these two.

Models	Number of predictors	Log-Likelihood	-2 * Log-Likelihood	AIC	p-value
Full model	617	-6891.79	13783.58	17007.52	-
Backward selection	90	-7105.00	14209.99	14658.24	0.9995
LASSO regularization	11	-7568.36	15136.71	15298.79	0.0000

Table 3: Logistic regression models with different predictor sets. All measures stem from models trained on the training set. Two variable reduction techniques were applied. The reported measures offer a ground for model comparison and selection. The p-value refers to the chi-squared distribution of the likelihood-ratio test with the full model.

Finally, we performed a different variable selection strategy for logistic regression, namely the stepwise backward with a significance level of 0.05. Starting from the full model with 617 predictors after dropping constant and collinear columns, this led to 90 predictors. All these remaining variables were hence significant. As before, we computed the AIC of this model on the training set as $AIC_{backward}$=14658.24, which was lower than the AIC_{full}=17007.52 of the full and AIC_{lasso}=15298.79 of the LASSO regression model. The backward selection model contained a subset of the predictors of the full model. Therefore, we additionally applied the likelihood-ratio test that compares the goodness of fit for nested models here. We obtained the -2*log-likelihood

values of $ll_{backward}$= 14209.99 for the backward selection model. Again, for the full model, we obtained ll_{full}= 13783.58, resulting in the likelihood-ratio LR=$ll_{backward}$-ll_{full}=426.41. The p-value of this chi-squared distributed ratio with 527 degrees of freedom was 0.999 > 0.05. Therefore, the likelihood-ratio test similarly suggests that the reduced model did not perform significantly worse than the full model. Consequently, the backward selection model should be preferred over the full model.

In opposition to the previous comparisons, the likelihood-ratio test was not applicable to compare the backward selection and LASSO regression model. The variable *Prescription_C* was one of the 11 predictors which were included in the LASSO regression model, but not in the model with a backward selection strategy. Inversely, 80 predictors were included in the predictors of the model with backward selection, but not in the LASSO regression model. Therefore, the models were not nested and thus prevented the use of the likelihood-ratio test. This forced us to rely solely on the AIC, on which the backward selection method performed best. Hence, the reduced model with 90 predictors selected via the backward reduction strategy was the best among all models and comparisons and was thus chosen.

All predictor sets can be found in the appendix.

MODEL ESTIMATION

Following our hyperparameter tuning design, we performed 5-fold cross-validation for hyperparameter tuning on the training data set (N= 51,709). All model parameters except the ones for logistic regression were tuned in a grid-search manner and sorted according to their AIC. For the random forest model, a total number of 1000 trees and a maximum number of 30 splits showed the best performance and resulted in an AUC of 0.871. The gradient boosting machine performed similarly well with 250 trees and a maximum tree depth of 3 splits, resulting in an AUC=0.873. For the artificial neural network, tan-with-drop-out as the activation function, one hidden layer with 100 neurons, and a learning rate of 0.03 were selected in the hyperparameter tuning process, leading to an AUC of 0.856. Logistic regression with backward variable selection and 90 predictors showed an AUC of 0.87.

MODEL PERFORMANCE

According to the train-validate scheme, we retrained the tuned models and evaluated them on a previously unseen validation data set for model comparison and selection. The train data set consisted of 51,709 individuals, while the validation data set contained 17,237 individuals.

TRAINING-VALIDATION EVALUATION

	ROC-AUC			Accuracy			Sensitivity			Specificity			G-mean			Cost Capture		
Model	*Value*	*lower*	*upper*	*Value*	*lower*	*upper*	*Value*	*lower*	*upper*	*Value*	*lower*	*upper*	*Value*	*lower*	*upper*	*Value*	*lower*	*upper*
Logistic regression	0.867	0.854	0.880	0.789	0.782	0.795	0.775	0.769	0.781	0.789	0.783	0.795	0.782	0.776	0.788	0.587	0.545	0.629
Random forest	0.885	0.874	0.897	0.830	0.824	0.835	0.771	0.764	0.777	0.833	0.827	0.838	0.801	0.795	0.807	0.626	0.586	0.666
Gradient boosting machine	0.883	0.872	0.895	0.801	0.795	0.807	0.805	0.799	0.811	0.800	0.794	0.806	0.803	0.797	0.809	0.633	0.589	0.669
Artificial neural network	0.798	0.782	0.813	0.731	0.724	0.738	0.723	0.716	0.730	0.731	0.725	0.738	0.727	0.721	0.734	0.421	0.376	0.483

Table 4: Performance measures of all models trained on 75% of the data from 2019 & 2020 combined and evaluated on the left-over 25%. This is the basis for model selection.

Evaluation results are displayed in Table 4. On the key performance measure in this study, the AUC, the random forest model was best with AUC=0.885, but not significantly different from the gradient boosting machine (AUC=0.883). Logistic regression performed slightly worse with AUC=0.867. Evaluation of the artificial neural network revealed inferior performance to all other models with an AUC of only 0.782. Regarding model performances for accuracy and specificity, the model results rank similarly: The random forest model was on top with 0.830 and 0.833, respectively. This was about 3% better than the second-best model, the gradient boosting machine.

In contrast, the highest sensitivity, g-mean, and cost capture were achieved by the gradient boosting machine on the validation data set (0.805, 0.803, and 0.633, respectively). While the random forest followed closely in second place for the g-mean (0.2% less) and the cost capture with 0.626 (0.6% less), logistic regression showed a higher sensitivity (0.775) than the random forest model (0.771). However, the difference was not statistically significant. The artificial neural network did not show competitive performance for any performance measure.

To facilitate the deployment of our ML model, we computed the decision curves for all our models in Figure 5. The greatest benefit was achieved with a threshold at and just above 0. All models achieved a net benefit of 0.05. For higher thresholds, the acceptable ratio between true and false positives changes towards penalizing false positives more strongly. The net benefit decreased monotonically. Random forest consistently outperformed all other models over all thresholds below 22% and should be chosen in that range to achieve maximal net benefit. For higher thresholds up to 35%, the gradient boosting machine was as good or even outperformed the random

forest slightly. The users only need to determine which threshold or minimum net benefit they require. Above a threshold of 35%, none of the models showed a positive net benefit.

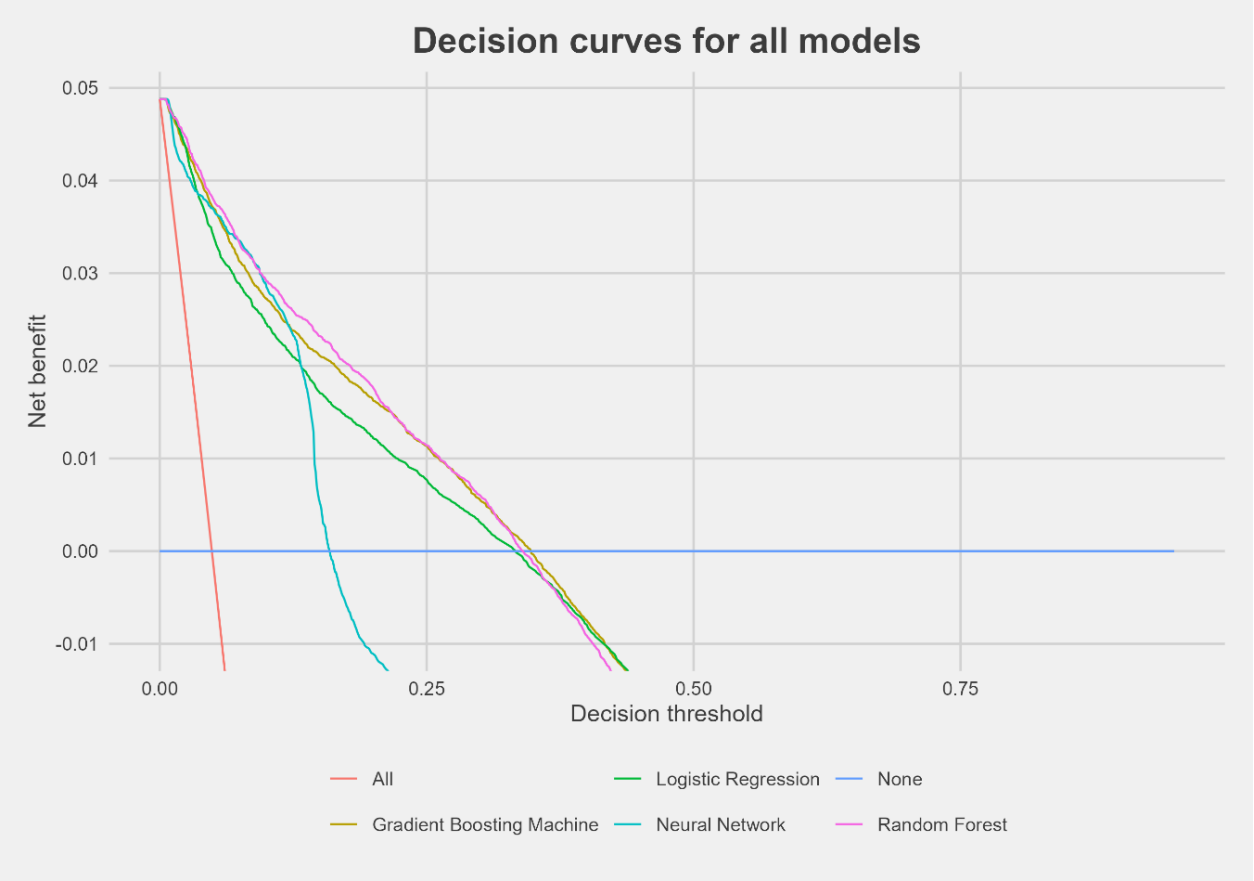

Figure 5: Decision curves for the different models in this study. The x-axis shows different probability thresholds. All patients with a predicted risk above that threshold receive treatment. The y-axis shows the net benefit according to the formula in the methods section. In addition to the four statistical models, treating all and treating none are included as additional intervention management strategies.

MODEL VALIDITY

The train-test validation design as described in the methods section was applied to validate the selected model. For training, we used the combined data set with all samples (N=68,946) from the

years 2019 and 2020. We evaluated all models on the test set consisting of previously unseen data from the year 2021 (N=35,350). The results are shown in Table 5.

TRAIN-TEST EVALUATIONS																		
	ROC-AUC			Accuracy			Sensitivity			Specificity			G-mean			Cost Capture		
Model	*Value*	*lower*	*upper*	*Value*	*lower*	*upper*	*Value*	*lower*	*upper*	*Value*	*lower*	*upper*	*Value*	*lower*	*upper*	*Value*	*lower*	*upper*
Logistic regression	0.873	0.864	0.882	0.822	0.818	0.826	0.769	0.765	0.774	0.824	0.820	0.828	0.796	0.792	0.801	0.588	0.557	0.621
Random forest	0.884	0.876	0.893	0.829	0.825	0.833	0.779	0.775	0.783	0.831	0.827	0.835	0.805	0.801	0.809	0.610	0.579	0.642
Gradient boosting machine	0.885	0.876	0.893	0.832	0.828	0.836	0.780	0.776	0.784	0.834	0.831	0.838	0.807	0.803	0.811	0.612	0.580	0.644
Artificial neural network	0.814	0.804	0.824	0.748	0.744	0.753	0.776	0.772	0.780	0.747	0.742	0.751	0.761	0.757	0.766	0.312	0.280	0.344

Table 5: Performance measures of all models trained on data from 2019 & 2020 and evaluated on data from 2021. These results are part of the model validation.

According to the AUC, the gradient boosting machine was the best-performing model in this study on the test data set with the highest AUC of 0.885. Its 95% confidence interval was [0.876;0.893]. However, in the second place, the random forest model lied well within this interval with an AUC of 0.884 and a statistically insignificant difference in performance of 0.1%. Logistic regression performed significantly worse with an AUC=0.873 and more than 1% lower than the random forest model. Similar patterns could be observed for all other traditional performance measures: The accuracy of the random forest model was 0.829 and thus 0.3% lower than that of the gradient boosting machine, while the sensitivity was 0.779 and thus 0.1% lower than the gradient boosting machine. Logistic regression performed worse in both measures by approximately 1%. The specificities of all these three models were about 5% higher than their sensitivities, making it more likely to receive a correct classification as a prospective non-HCP than as an HCP. Also here, the gradient boosting machine and random forest models performed best with a specificity of 0.834 and 0.831, respectively. Finally, for the g-mean again these two models were best, with 0.807 and 0.805. In both cases, logistic regression was again about 1% worse than the two best models. In contrast, the neural network again did not show competitive performance. This confirmed the previous model performance evaluation.

Interestingly, the gradient boosting machine performed slightly better on the test data set than the random forest model, even though these differences were not statistically significant. In contrast, on the validate data set the random forest model was better in some measures, especially the ones that included correct classification of prospective non-HCPs (higher specificity and accuracy). This was probably because the gradient boosting machine profited more from more training data, which led to an improvement in the sensitivity-specificity balance. With a further massive increase in

data, also the neural network is expected to improve its performance significantly. Between the evaluation on the validate and test data set, performance in all traditional performance measures for the ANN increased by about 1.5%-5%.

To further support comparison throughout works in this area across literature, we reported the cost capture. Similar to the traditional measures, the gradient boosting machine showed the best performance, with a proportion of 61.17% detected costs. The random forest and logistic regression models were second and third place, respectively. Their performance differences were not statistically significant. The decision curve on the test set showed a very similar behaviour as on the validation set and can be found in the appendix.

EXPLANATION METHODS

We implemented native R and Python random forests to apply various explanation methods. Their performance was comparable with the H2O model that we evaluated in detail before. The native R random forest achieved an AUC of 0.881 and the random forest model in Python had an AUC of 0.882. These were slightly worse than the H2O random forest but well within its confidence interval.

Variable importance

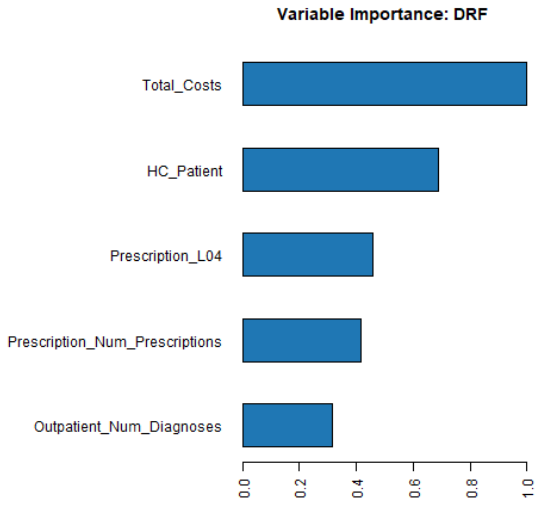

Figure 6: Model-specific variable importance plot from the H2O package, according to the Gini method. Relative importances are displayed.

Arbitrarily many variables can be displayed in a variable importance plot. To assure clarity, we only included 5 features in the final plot in Figure 6. The cumulative costs in the current year was the most important variable, according to this variable importance definition, followed by the current HCP status. The third variable indicated whether an immunosuppressant was prescribed, which has the ATC code L04 [182]. The variables that indicate the number of prescriptions and ambulatory diagnoses were in the fourth and fifth positions, respectively.

SHAP summary

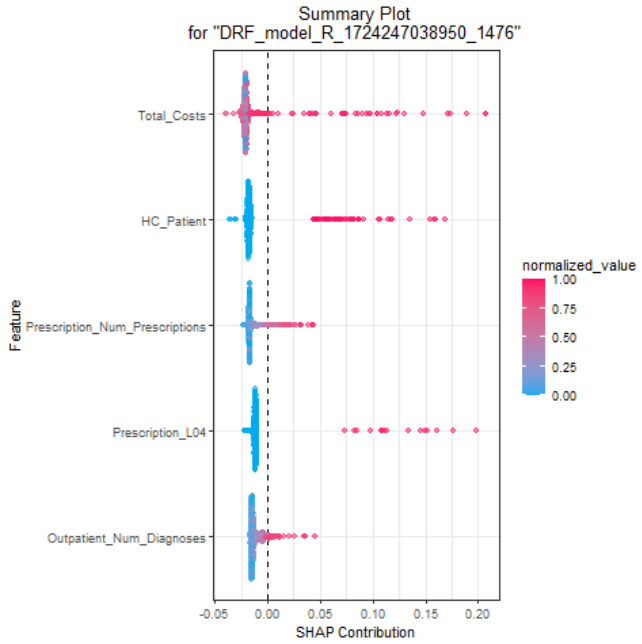

Figure 7: SHAP summary plot with the five most important variables according to the SHAP values, implemented with H2O. Each dot represents a sample. The colour indicates the normalized value of the variable. The x-axis shows the amount by which the respective variable increased or decreased the predicted risk in form of the SHAP value.

The most common XAI method SHAP offers an extended variable importance plot. For each variable in the plot, the position on the x-axis indicates the SHAP value. The colour represents the variable value of the sample on a normalized scale. The variables were ordered by the average amount of their SHAP values over all samples. This variable importance definition led to a slightly different ranking of the same variables. While the current total costs and HCP status still were the most important variables, the number of prescriptions and prescription L04 changed positions here. The presence of the prescription with an ATC code of L04 that is indicated by red dots seemed to cause a clear increase in predicted risk. In addition, the HCP status again displayed by the red dots seemed to increase predictions, but for most samples only about 4% to 8%, while L04 rather uniformly increased risk by values between 7% to 20%. Almost all costs that significantly increased risk are at the high end of the cost scale. Higher numbers of prescriptions, as well as outpatient diagnoses, led to a rather gradual and light increase. Even very high values only moderately increased the risk by less than 5%.

Next, we focus on the *feature effects*, the relationship between the model response, and the most important variables by the previous variable importance plots, the total costs and the current HCP status, as well as the demographic variable age.

PDP

The PDP reveals the dependence of the model prediction on certain values of a specified feature. It allows the examination of the feature effects of one or two features in a joined plot. PDPs are

implemented in all the XAI packages included in this study. Functionalities concerning joined plots and additional distribution information differ, however. In opposition to a linear regression model that would always display a part of the logistic curve, the influence of feature values for the random forest model is much more heterogeneous.

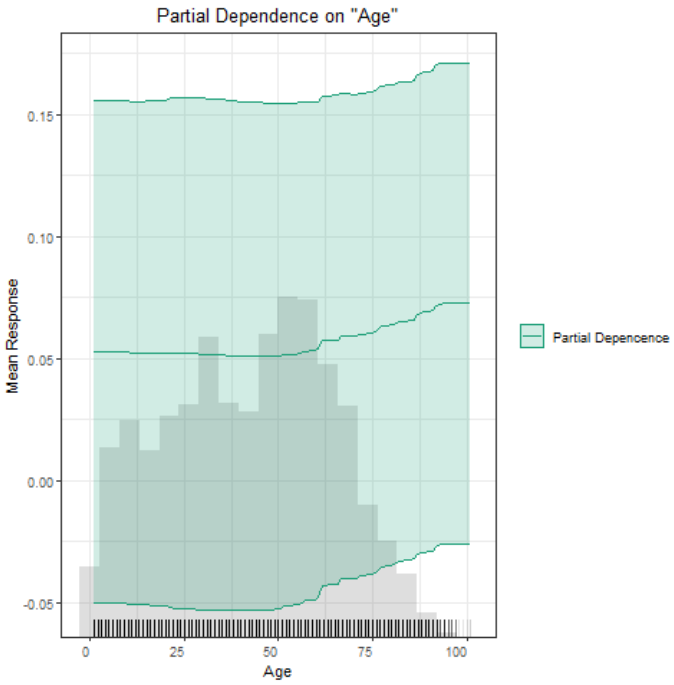

Figure 8: Partial Dependence Plot of the variable age from the native H2O explanation functions. The green band displays a confidence band with a preset confidence level. In addition to the PDP, the distribution of the age variable is displayed via both a histogram and a tick bar.

Figure 8 displays the PDP for the *age* variable. The coloured area around the line depicts a confidence interval. However, the documentation of the function is not clear about its computation and does not allow the specification of a confidence level. The bar and histogram showed the age distribution. The mode was the age group between 50 and 54. Before that age, there seemed to be a general stagnation or even decline in the predicted risk of becoming an HCP in the next year. After the age of 50, the predicted risk level steeply increased for individuals above that threshold. However, the increase was not smooth but was characterized by small perturbations.

The PDP from the alibi package for the Python model is illustrated in Figure 9 and gives a similar picture, but with age increasing the risk up to 9% and not only 7.5% like in the PDP from the H2O model. Alibi allowed the customization and a shared y-axis among its plots. Minimal distributional information was displayed in the tick bars. Certain age values seemed to increase the average predicted risk, while slightly lower or higher ages showed lower risk. In addition to the age, we showed the two most important variables according to the H2O variable importance plot: total costs and current HCP. Even though this involved a different random forest implementation, the pathway of the predicted risk in comparison to the age resembled the previous PDP of the original H2O model. However, this explanation method revealed magnitude differences in the marginal influence

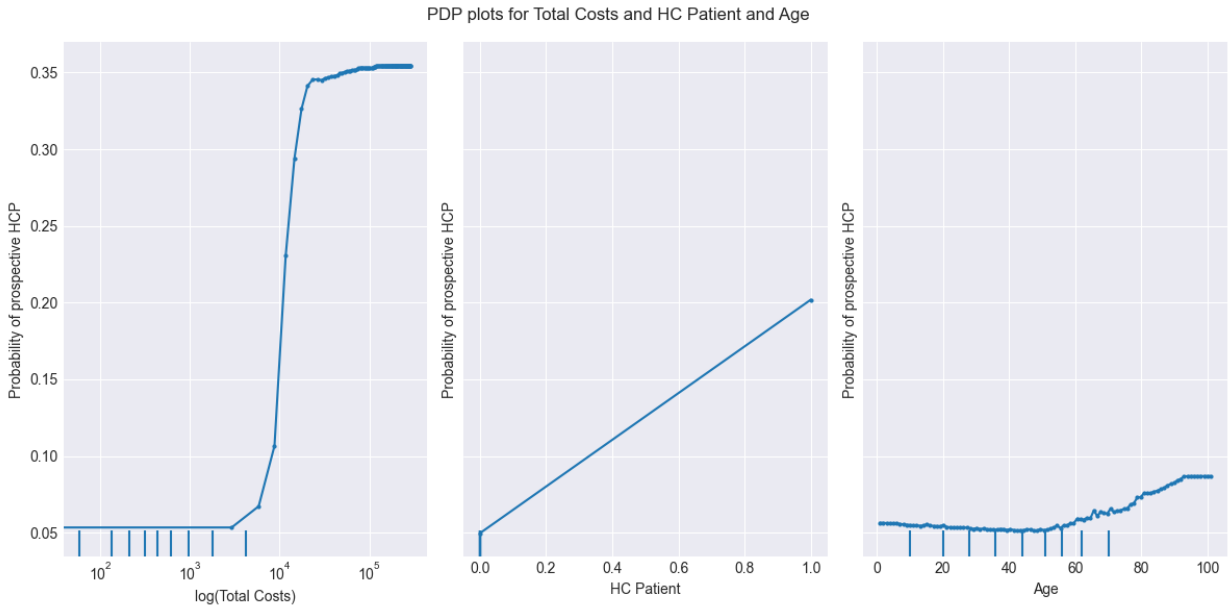

Figure 9: The profile of the average risk distribution of the variables total costs (log), current HCP and age according to PDP explanation method. Implementation stemmed from the alibi package.

of the different features. While even a person with an age of 100 was predicted to become an HCP in the next year with only about 9%, being an HCP in the current year led to a predicted risk of more than 20% regardless of age. For total costs, the increase was even more evident. The predicted risk monotonically increased the higher the cumulative costs for that year until reaching a plateau-like state at about 25,000€ at around 35% predicted risk. After that, the risk increased only mildly. These plots for the continuous variables of total costs and age revealed a non-linear relationship with the predicted risk. The PDP explanation from the DALEX package for the variable age was included in the appendix and showed a very similar picture. Additionally, an exemplary interpretML explanation for the variable age can be found in the appendix. To demonstrate the capabilities of PDP plots, a joined plot for the variables age and total costs from the alibi package is included in the appendix as well.

ALE

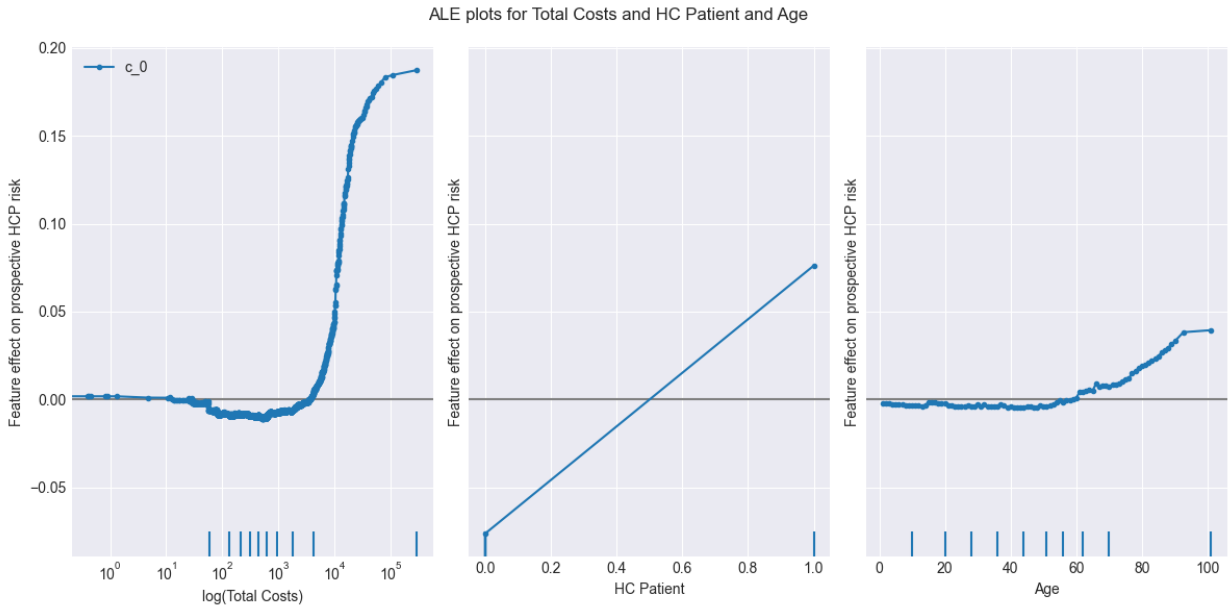

Figure 10: The profile of the local feature effects on the predicted risk of the variables total costs (log), current HCP, and age. Implementation according to the ALE explanation method came from the alibi package.

The displayed relationships in the ALE plots in Figure 10 resemble the closely related PDP explanations. Both current HCP status and age showed very similar trajectories. Risk levels for these two variables had similar ranges as in the PDP plot with minimum and maximum risk about 0.15 and 0.04 apart, respectively. In contrast, the total costs show a qualitatively different picture. For very small costs until 100€ in the current year, an increase in costs appeared to decrease the local effect slightly. Hence, until that threshold, more expenditure seemed to decrease the model's risk prediction. After that threshold, it resembled the PDP plot, but with a lower plateau and maximum. The range of risk levels was about 0.2 and thus much smaller than the 0.3 range from the PDP plot. While PDPs show the feature effect over the whole range of its values, ALE plots are only faithful locally. E.g., Figure 10 tells us that the predicted risk of a 40-year-old person would not or barely change with an age of 39 or 41 and all other variables unchanged. However, ALE does allow us to make assumptions about the predicted risk for the same person at the age of 80. Some variables might be strongly correlated with age, making such a sample unrealistic.

The DALEX package offers the possibility to display both PDP and ALE plots in the same plot. This enables users to directly compare the two plots and check for differences. Such a plot can be found in the appendix.

In this thesis, the final focus is on the feature attribution methods SHAP local and the simplification method LIME. The results for different explanation methods on the same sample are displayed.

SHAP local

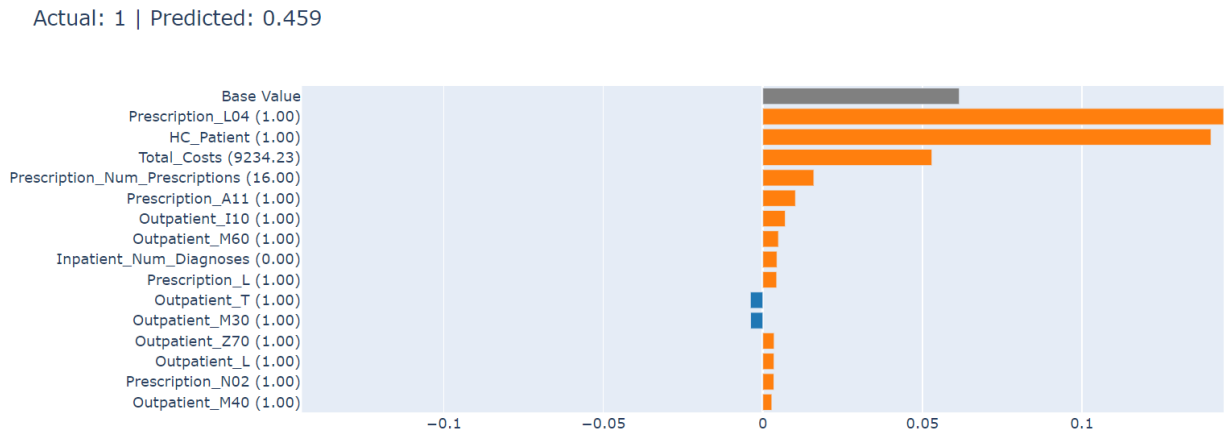

Figure 11: SHAP local explanations for the sample are displayed. The implementation came from the interpretML package.

Figure 11 shows the SHAP local explanations of the interpretML package. The SHAP contributions on the x-axis indicate how much the feature value increased or decreased the model prediction, according to the SHAP method. The base value was the same for all samples and with 0.06 a bit higher than the actual risk, which was exactly 0.05 by definition. The presence of a prescription with an ATC of L04 increased the predicted risk of being a prospective HCP the most with about 0.15. It is important to remember that this was not the increase of risk of a present in comparison to an absent L04 prescription. Rather, it was the comparison of the average background risk of having an L04 prescription compared to the presence of the L04 prescription in this case. The second most important variable for this sample was the current HCP status, which contributed about 0.14 to the predicted risk. It was followed by the total costs, which increased the predicted risk by about 0.06. After that, the contributions were rather marginal, all lower than 0.02. The most important variables were on top, which were similar to the most important variables from the global

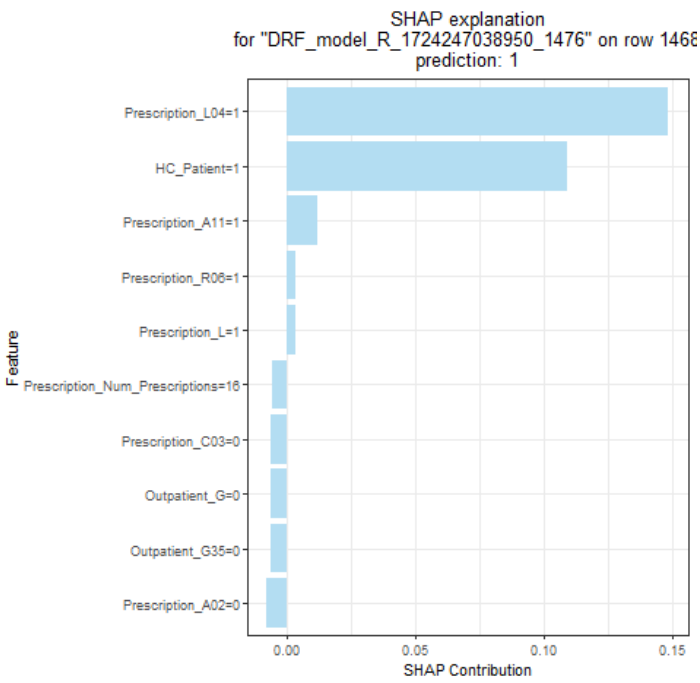

Figure 12: SHAP local explanation for the sample are displayed. Implementation came from the H2O package.

variable importance methods. Notable were the outpatient diagnoses with ICD-10 codes T and M30. The presence of these diagnoses decreased the predicted risk.

The same sample and explanation method were implemented in the H2O package and shown in Figure 12. In the plot title, only the final prediction was displayed and not the predicted risk. What can be noticed first is the absence of the intercept in the H2O plot. As in the interpretML plot, the prescription with ATC code L04 (immunosuppressants) and the current HCP status were the most decisive features for the sample with risk increases of about 0.15 and 0.11, respectively. However, the total costs were completely missing in this plot, indicating substantial differences in the different random forest models or XAI implementations. All other features only had a minor impact on the model's prediction. The majority of these variables were different from the ones in Figure 11.

Additionally, the DALEX package in R implements SHAP for the visualization of feature contributions for a single sample. The credibility of these feature contributions could be assessed by boxplots on the feature bars that indicated the range of the feature effect values across different variable orderings. It showed the same five most important variables as the SHAP local explanation from interpretML in Figure 11. See the appendix for details.

LIME

Similar to the local SHAP explanation, the LIME methods in the *DALEX* and *interpretML* packages depict single variables' contribution to the prediction outcome, ordered by their influence magnitude. The LIME plot of the interpretML package in Figure 13 shows the variable values along with their contributions, while the LIME explanation from the *DALEX* package in Figure 14

shows the discretized variables. In the interpretML package, it is possible to show both the predicted risk and the actual prediction (0 or 1), but not to choose the number of variables to display.

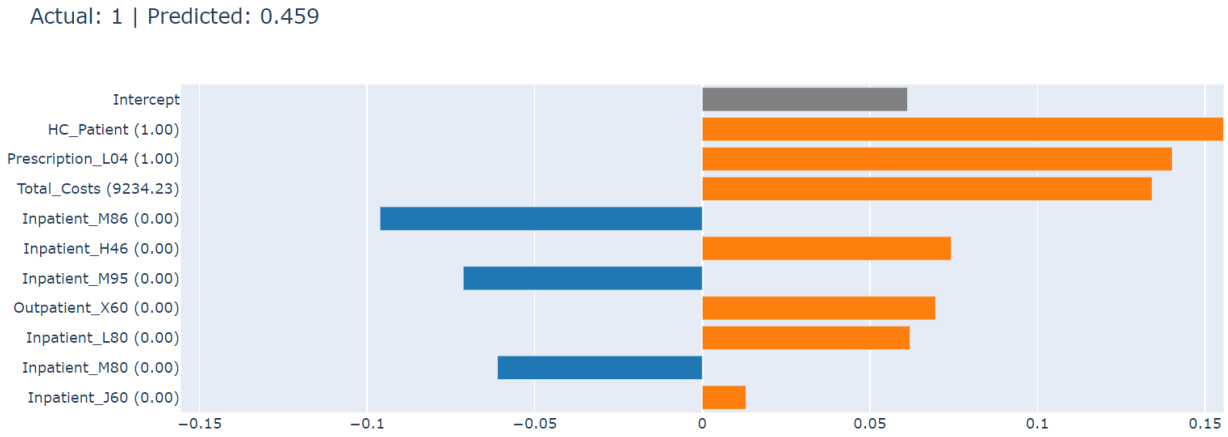

Figure 13: LIME explanation model for the sample. The interpretML Python package was used for implementation. A lasso regression model on dichotomized data obtained the feature contributions.

After all continuous variables were dichotomized, the intercept of the locally fitted LASSO regression model for the LIME explanations of the interpretML package in Figure 13 was about 0.06 for the whole data set. The current HPC status increased the risk with its presence the most by about 0.16. In addition, prescription L04 and total costs were the second and third most decisive variables, with positive contributions of about 0.14. This was in accordance with the global variable importance explanations and the interpretML SHAP explanation, with the same three most important variables on top. The first variable value that decreased the predicted risk locally was the absence of the hospital diagnosis M86. Interestingly, having no outpatient diagnosis with an ICD code of X60 and hospital diagnoses with L80 and J60 increased the predicted risk, according to the LIME plot. This was surprising and might depict shortcomings in the model. It is not clear why a risk should increase in the absence of a diagnosis. However, the effects for the most important four variables seemed reasonable since the qualitative influence of their risk was what we would have expected from these variables.

The LIME plot for the native R random forest model in the DALEX package is illustrated in Figure 14. The most important variables differed from the LIME plot in the interpretML package which explained a random forest model in Python. No intercept was included in the plot. In comparison to the LIME in the interpretML package, Prescription L04 was missing in the DALEX plot. Total

costs were by far the most important variable and significantly increased the predicted risk by about 26%. The number of prescriptions and outpatient diagnoses were ranked second and third but were both not included in the LIME plot of the interpretML package. They increased the risk by 8% and 4%, respectively. Only the age between 24 and 44 decreased the risk of becoming an HCP slightly, according to the native R random forest model. Except qualitatively for the total costs and HCP variables, the LIME explanations by the interpretML and DALEX package did not agree on the important variables and their attributions.

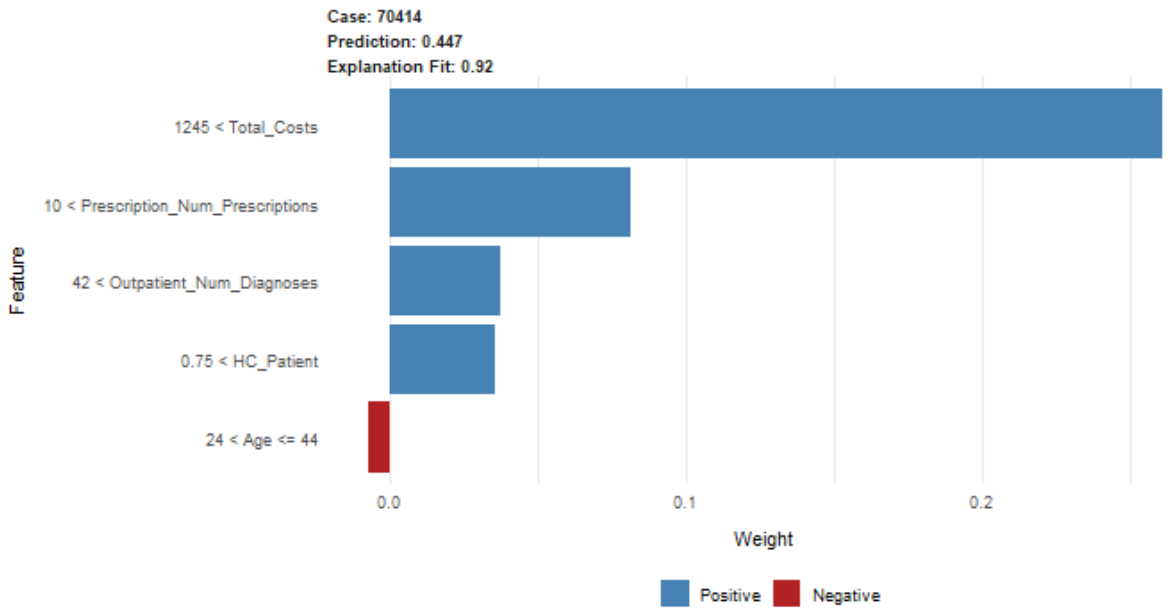

Figure 14: LIME explanation model for the sample. The DALEX R package was used for implementation. Feature contributions were obtained by a LASSO regression model on dichotomized data.

EVALUATION OF EXPLANATIONS

A combined plot as in Figure 15 lets us examine the ICE plot from the alibi package and compare it with the PDP directly. The PDP is by definition the average line of all individual lines in the ICE plots. Individual lines that each depict a different sample can be on different levels. These levels are the result of different ground-level risks in the underlying samples due to their remaining variables. While some overlapping lines infer the possibility of following single lines exactly, the relationship between total costs and the predicted risk for all lines seemed to be fairly similar to its average, the PDP line. If all individual lines follow the same trend as the PDP line, this indicates

that the relationship between variable and response as displayed by the PDP is like this or similar for all samples. Hence, this indicated that the previously identified feature effect is reliable.

In opposition to total costs, the age and to a lesser extent the current HCP variable showed more heterogeneity in their individual relationships with the predicted risk level. The risk seemed to increase for all samples with a positive current HCP status. However, while some lines were almost horizontal, others showed a steeper increase than the average PDP line. Horizontal lines were observed both at high- and low-risk levels. Therefore, it seemed that the influence of the current HCP status depended highly on the other features, but it was always non-negative. The individual lines in dependence on the age variable did not all follow the trend of the average PDP line either. While almost all lines with relatively low-risk levels (<20%) seemed to follow the trend of the PDP line but on different base risk levels, a few lines showed great perturbations above the age of 50

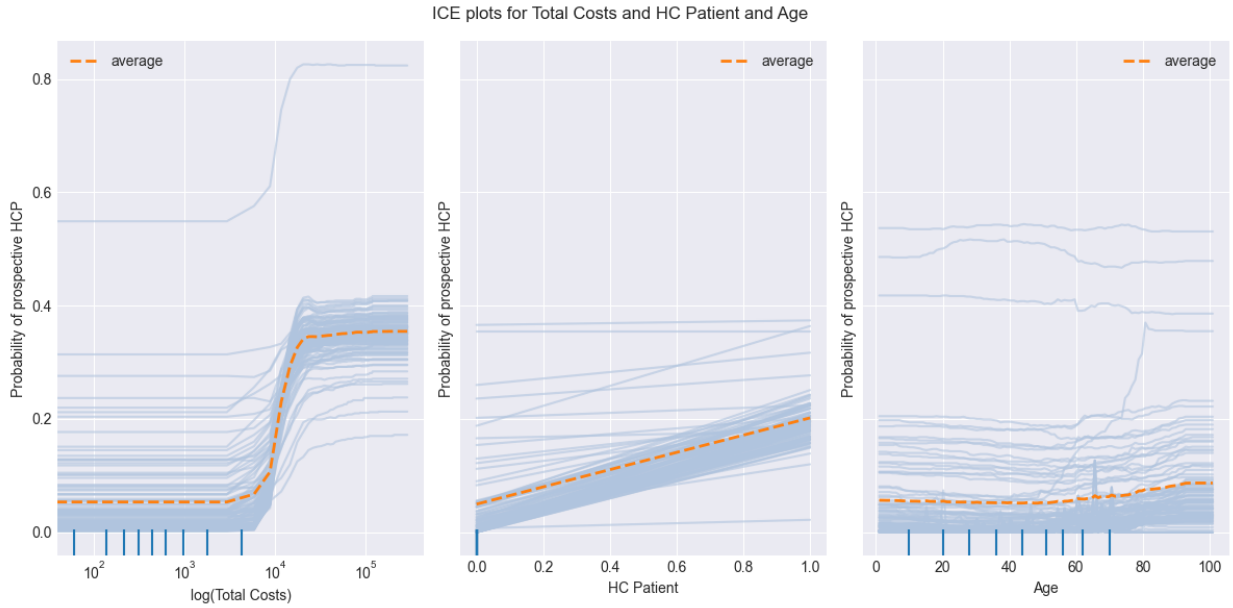

Figure 15: ICE plots for the variables total costs (log), current HCP and age. Implementation according to ICE explanation method from the alibi package.

years. In addition, for the individual lines with a high base risk, the age variable showed a different effect on the risk, for some samples a higher age even decreased the risk. During the execution of the ICE explanation method in alibi, evoking the plot function on the same explanation object resulted in different ICE plots, with some plots depicting more heterogeneous curves than others. ICE explanations from the H2O package can be found in the appendix.

Regarding the stability of explanations, the global methods in this study agreed on most of the variable rankings and feature effects. Additionally, the feature contributions from the SHAP local explanations were comparable among the native R and Python random forest model implementations and DALEX and interpretML packages. For the sample we investigated, both explanations agreed on their five most important and most of the other variables. Additionally, the H2O plot for the same method showed different feature contributions, agreeing with the explanations from the other packages on only three out of the five most important variables.

The most important feature according to all SHAP local explanations was the prescription with an ATC code of L04. This feature was not present in the LIME explanation from the DALEX package. Similarly, SHAP local and LIME explanations differed for the Python model in the interpretML package. While the three most important variables were the same, the explanations did not agree on the contributions of the other variables.

Looking at the characteristics of the programming packages, the H2O package came with several advantages. It allowed seamless and efficient incorporation of the explanation methods into the previously trained H2O random forest model. The computation times were the fastest for all methods it implemented except the PDP plot. The DALEX package was theoretically also able to explain H2O models. However, the computation times for this model for all explanation methods were too long (>24 hours). A disadvantage of the H2O model was its limited adaptability and documentation. The model threshold was chosen automatically according to the maximal F1 score and is used for the explanation methods. In addition, the PDP plot showed an unclear confidence interval, and the intercept was omitted in the SHAP local plot.

The DALEX package is the only established solution in R and offers a wide range of methods. However, many of these methods are not established in the literature. The computational costs varied strongly depending on the explanation method. The alibi package offered a great selection of the most established XAI methods in Python, while not inventing or including more experimental methods. However, Anchors and Counterfactual Examples could not be included due to the long execution times in this package. The interpretML package offers XAI methods with

very limited functionalities for black-box models. In opposition to all the other packages, explanations could be combined and presented in an interactive dashboard.

We recorded the execution times for the different explanation methods and packages in Table 6. Not all explanation methods were implemented for all packages. Only the H2O package generated a SHAP summary plot, with an execution time of about 13 minutes, making it the longest minimal time for any explanation method. Together with PDP and ICE plots, these explanation methods had median execution times of more than five minutes. However, the DALEX package computed the PDP plots in less than a minute. The variable importance, SHAP local, ALE, and LIME explanations all had median execution times of less than 5 minutes and minimum execution times of only a few seconds. Exact computation times also differed between variables and chosen samples for the global and local methods, respectively, and hence must be treated with care. Nonetheless, they can indicate what magnitude of execution time to expect.

	EXECUTION TIMES						
XAI Method	**VARIABLE IMPORTANCE**	**SHAP SUMMARY**	**SHAP LOCAL**	**PDP**	**ALE**	**LIME**	**ICE**
Package							
alibi	n.i.	n.i.	n.i.	1540.03	1.50	X	39.67
interpretML	X	X	4.90	2652.09	X	0.35	X
DALEX	8.47	X	2530.44	0.11	0.11	0.19	n.i.
H2O	0.01	13.06	0.24	5.34	X	X	6.34
Minimum	0.01	13.06	0.24	0.11	0.11	0.19	6.34
Median	4.24	13.06	4.90	772.68	0.80	0.27	23.00

Table 6: Computation times across different methods and packages for the creation of explanations for the different XAI methods. Execution times were recorded in minutes. Some explanation methods were not included in the package (X), and others were included in the package, but we did not implement them due to time constraints (n.i.).

DISCUSSION

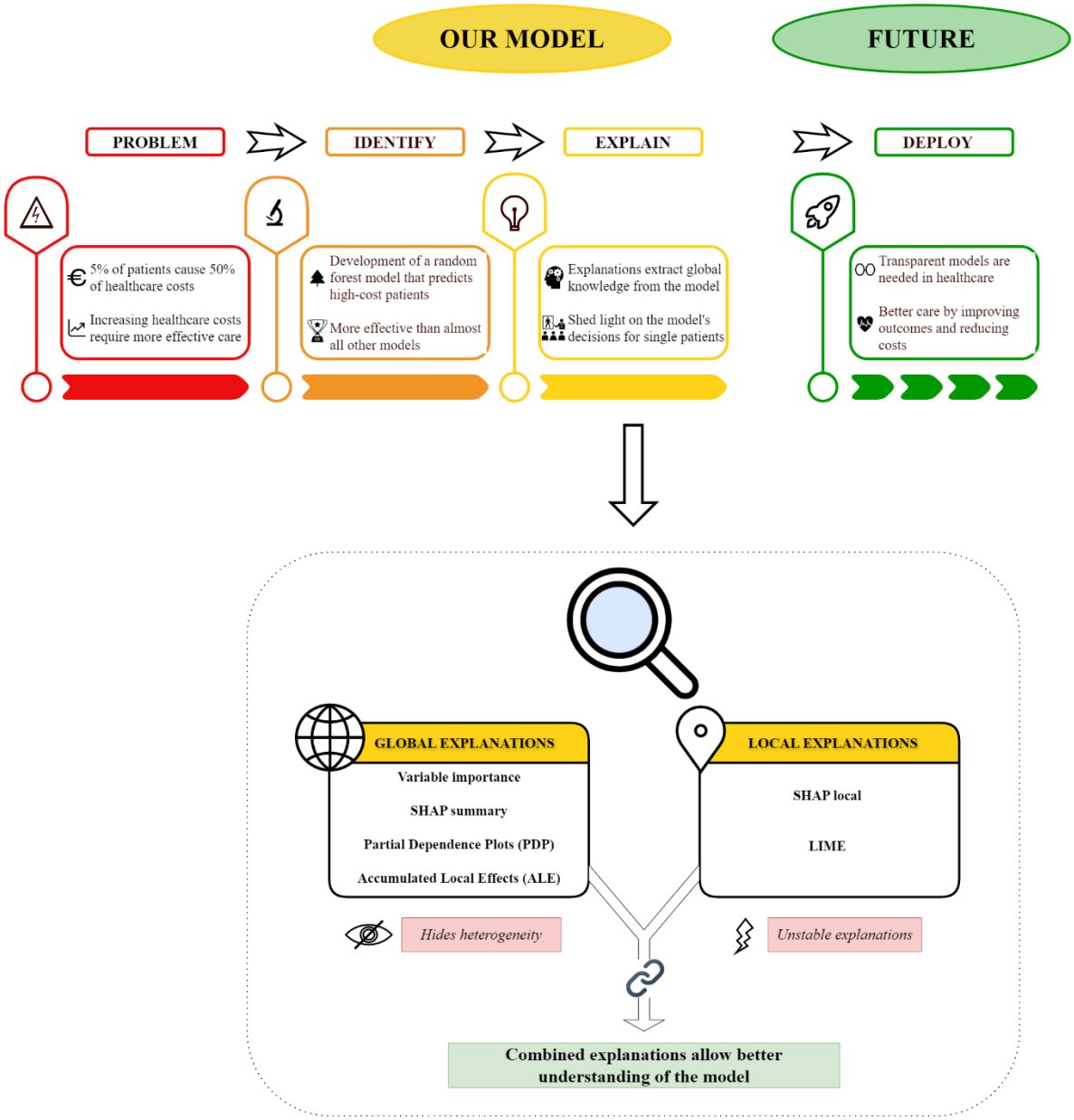

Figure 16: Chart of progress and findings in this work. The four model stages included determining the problem, identifying HCPs by the random forest model, and explaining the model's decisions. Deployment as a last step should be part of future efforts. Diverse global and local methods were applied to render the model more transparent.

In this thesis, we compared the usage and performance of HCP predictions of models using explainable AI. This thesis aimed to facilitate their routine use in healthcare. A major critique of current AI systems is the impact that they will have on physicians and patients in the form of a shift of responsibility [29]. Especially in the medical area, black box models that are too complicated

for humans to understand and difficult to troubleshoot are particularly problematic [87]. It seems like users will be forced to choose between explainable AI or blindly trust black-box models that work most efficiently [87]. The main challenges to be resolved are the lack of trust, accountability, and fairness that come along with impaired understandability [29].

We set our focus on better identifying patients that will become high-cost in the future. Interventions that are tailored to HCPs can be targeted more efficiently with the use of new technologies in statistics and ML. To improve healthcare for HCPs cost-effectively, ML models and methods need to overcome the lack of interpretation that seems to be associated with their use. We focused on facilitating the deployment of medical models, following the seven steps suggested by Steyerberg and Vergouwe [119], with an emphasis on model explanations as part of the model presentation. This emphasis led us to consider using model-agnostic XAI methods, to keep the review as generally applicable to different settings using a range of supervised learning models.

This thesis aimed to validate and extend the previous findings obtained by Langenberger, Schulte, and Groene [46]. The authors found that random forests were the best-performing class of models. The present study trained an equivalent model on more recent data, exploring different XAI methods to explain model predictions. Finally, we evaluated the explanations according to their reliability, robustness, and applicability.

The following sections will provide detailed responses to the research questions formulated at the outset.

RESEARCH QUESTION 1: HOW DO DIFFERENT PREDICTIVE MODELS COMPARE?

The question was tackled by comparing three logistic regression models obtained using two different variable reduction techniques and subsequently benchmarking a random forest, gradient boosting machine, and ANN model against the best-performing logistic regression model.

The full and backward selection and LASSO regression models relied on 617, 90, and 11 variables, respectively. A logistic regression model with 11 variables could be still considered sufficiently transparent (simulatable), but models with 90 and 617 variables require decomposition to be interpreted by humans [82]. We identified the backward selection model with 90 variables as the best model, both according to the AIC and likelihood ratio test. Remarkably, the vast majority of

variables did not seem to achieve a higher accuracy. Less than 15% of the variables contain all relevant information for accurate logistic regression models. However, this interpretable model was outperformed by more complex black-box models used in this study.

In accordance with findings obtained by Langenberger, Schulte, and Groene [46], the random forest model performed best on our train-validate scheme used for model selection. In contrast, in our validation on the test data set, we found that the random forest (AUC=0.884) did not achieve the highest AUC. In fact, its AUC was not significantly different from the best-performing model, the gradient boosting machine (AUC=0.885). In general, the random forest model used in this study performed as well as the model in Langenberger, Schulte, and Groene [46] where an AUC=0.883 was recorded. Consequently, the result does not change significantly using different years, insurance funds, and model implementations. Such remarkable consistency throughout temporal and geographical variation emphasizes its validity. Our findings are consistent with previous HCP prediction models, reaching similar heights of accuracy. The AUC of such prediction models ranged from 0.78 [183], 0.79 [50], 0.84 [51], and [58] to 0.89 [73]. In a systematic review by de Ruijter et al. [55], only one study reported a model with a higher AUC than our random forest model, out of 60 studies in the review. Despite different data and HCP definitions (5% or 10% of the costs), almost all HCP prediction models seem to lead to similar performances. ML models consistently outperform traditional logistic regression models.

A crucial difference to the study in Langenber et al. [46] was the performance of the simple logistic regression model. In our case, it achieved an AUC of 0.873 on the test data set, which is more than 3% higher than their logistic regression model. This makes logistic regression a competitive model with a worse but comparable performance than the complex random forest and gradient-boosting machine models. This finding is consistent with Osawa et al. [58] who found that LASSO regression performed slightly worse than ML models with an AUC of 0.82. Their findings indicate that our model could even perform better with the inclusion of additional clinical data.

The tree ensemble models had not only the highest AUC but were also more accurate, sensitive, and specific in their predictions than logistic regression. Sensitivity in different HCP prediction models in the literature ranged from 0.71 [58] and [183] to 0.75 [50]. Our model was the most sensitive model with a sensitivity of 0.78, reproducing the same sensitivity as in Langenberger, Schulte, and Groene [46].

Specificity for HCP models ranges from, 0.70 [50], 0.73 [183], to 0.85 [58]. We achieved an even higher specificity than Langenberger, Schulte, and Groene [46] of 0.83. The combination of both highly sensitive and specific predictions of our model is remarkable. The only model that achieved a higher performance in one of these measures is the specificity of Osawa [58], which is more than outweighed by the below-average sensitivity of its model.

Only Langenberger, Schulte, and Groene [46] reported the accuracy (0.82). However, due to the constant ratio of 5% in most cases between positive and negative outcomes, accuracy could be derived from sensitivity and specificity values for some models, leading to the following estimated accuracy: 0.73 [183] and 0.84 [58]. However, accuracy can be misleading especially with highly unbalanced classes like in our case [83] and should only be considered in combination with additional performance measures for HCP models. A model predicting all patients as non-HCPs would achieve an outstanding accuracy of 0.95%. In comparison, we achieved again a top-performance model with an accuracy of 0.83.

To further facilitate the deployment of our ML, we extended the performance evaluation framework. Two additional measures were reported, namely decision curves and cost capture. Decision curves give more flexibility and autonomy to the model users (like physicians, insurance company personnel, or the patient) so that they can decide which model to use or whether they want to rely on a model at all. On both the validation and test data set and for almost all probability thresholds below 35%, the tree ensemble models offered the best net benefit. This leaves the user with the choice of whether to use the random forest or gradient boosting machine or not use a model at all. However, above a threshold of 35%, all models exhibited a negative net benefit. A negative net benefit indicates that we do not reach the ratio of actual prospective HCPs against all predicted HCPs of the specified probability threshold. Therefore, our random forest model is an effective tool to target intervention if the acceptable minimal ratio of true prospective HCPs among all predicted prospective HCPs is lower than 35%. However, if e.g. at least every second predicted prospective HCP should be a true prospective HCP, the probability threshold would be 50%. For this threshold, all our models showed a negative net benefit. Hence, none of the models is applicable in that situation.

Another indicator that gives the model users more information about the medical usefulness of the model is the cost capture reported for each of the models in this thesis. With a proportion of 61.17%

detected costs, our best model was consistent and even slightly superior to the best model found in Tamang et al. [51] with 60%. The random forest model was superior to their model as well with a cost capture of 61.02%. This percentage gives model users e.g. insurance companies an idea of the potential cost savings that could be achieved with this model.

Our results and especially the decision curve highlight the superior power of our complex ML models, even in comparison to a complex traditional logistic regression model. As observed in previous studies like by Goals et al. [17], the complex ML models offer more powerful and precise predictions than traditional models. Especially our superior accuracy and sensitivity allow targeting HCPs more efficiently, overcoming the previous obstacles to identifying them reliably as remarked by Berkman et al. [52]. With *value* as the measure for successful healthcare [71], the model could enable more efficient interventions targeting HCPs and potentially lead to better outcomes cost-effectively. Bremer et al. [78] already showed that such an ultimate increase in *value* is possible with the deployment of AI, and our high proportion of captured costs emphasizes the cost-saving potential of our model.

RESEARCH QUESTION 2: HOW DO XAI METHODS HELP EXPLAINING THESE MODELS?

We aimed to facilitate the deployment of our model into the real world, a task on which almost all developed medical models fail [31]. Understanding the model is essential to comply with the basic ethical guidelines of fairness and accountability in AI [35]. The logistic regression model does not require such explanations, since it is directly interpretable apart from its high number of variables. Our explanations for the complex ML models, in particular the random forest in this study, are an important step in this direction. We explored how to inform the user about the model predictions in a similar manner as logistic regression models can be interpreted.

Our starting point for explaining the random forest model was the notion of variable importance. When interpreting a logistic regression model, the user inspects the significance levels of different variables. Complex ML models do generally not possess such an inherent definition of significance calculation. The XAI method variable importance tells us which variables were relevant for the predictions of our random forest model. While variable importance is popular among tree ensembles with many model-specific measures [165], the SHAP summary explanations can be applied to all types of models.

Both explanation methods list total costs and current HCP status as the most important variables. The same variables were found most important in Langenberger, Schulte, and Groene [46], while in Osawa et al. [58] total costs were by far the most important variable. More than 40% of the HCPs in this study were also HCPs in the subsequent year. Therefore, it is not surprising that total costs and current HCP were most important. The only diagnosis or prescription indicator in the five most important variables was a prescription with ATC code L04, namely Immunosuppressants. Both the number of prescriptions and outpatient diagnoses influenced the model predictions strongly.

A major shortcoming of variable importance is the lack of a measure for the impact of each variable. In logistic regression, the sign and value of the variable coefficient provide a usable measure for the impact of each significant variable on model prediction. To determine such effects in most complex models, additional model explanation methods are needed.

SHAP summary explanations already offer some limited additional information on the variable impact. As expected, the red dots on the positive part of the x-axis for total costs indicate that only high costs increase the risk. The two binary variables current HCP and ATC prescription L04 in the SHAP summary plot lead to a clear increase in risk if present. However, this summary plot offers only vague connections between variable values and the increase in risk. Many red dots (high-cost values) for the total costs variable have a negative SHAP contribution and hence decrease the predicted risk. This is where the feature effect methods PDP and ALE can help us to better investigate the impact variables have on the model predictions.

We investigated the influence of the two most important variables total costs and current HCP status. In addition, we included age in this analysis as a demographic variable. Its influence was inconclusive in several studies about HCPs [52], [50], but was one of the most predictive for the prospective HCP status in Langenberger, Schulte, and Groene [46]. We included both PDP and ALE explanations. Both methods exhibited very similar trends for these variables. The relationships between the model predictions and the two continuous variables total costs and age were non-linear or sigmoid. According to these explanations, the lowest risk for becoming an HCP is at the age of 50-54 years. After that age, the risk increases continuously until it reaches a plateau for ages above 90 years. Interestingly, the predicted risk decreases slightly with growing age before the age of 50. Such behaviour could not be captured with logistic regression. This underlines the complexity of the influence of age on the HCP status. It offers a possible explanation as to why

several studies contradict each other, some associating lower ages and others higher ages with an increased risk of becoming an HCP [50].

It is important to remember that these explanations do not show variable-outcome relationships that are necessarily true or present in the data. Instead, the focus is only on revealing the learned knowledge of the model. The ALE method ensures through its sample generation that the displayed relationships are based only on data in the vicinity of the sample. PDPs, on the other hand, use the marginal distribution of variables, which may lead to unrealistic data generation and therefore include irrelevant data in its explanations that influence the relationship.

As an example, consider a person without any present prescriptions and diagnoses. When looking at the PDP for total costs, these costs are varied from 0€ to the maximum value of just under 3×10^5€ and are included in the PDP plot. However, since the person did not utilize any healthcare services, the costs are always 0€ for such a person. The model could make arbitrary predictions for such a person with high costs since no such sample could occur in the real world. On the other hand, the inclusion of such unpredictable or unrealistic data might also help reveal and solve a major concern in ML: The erratic and unplausible behaviour of ML models on unforeseen data [87]. PDPs can offer a robust view of the learned relationships of the model and show a reliable feature effect even when data is changing in the future. In the case of many correlated features, ALE explanations should be preferred. Its promising theoretical properties make it robust regarding correlated features and show only the effect that a feature has on the predictions and not the combined effect of correlated features. The variable total costs clearly illustrate this advantage.

Looking at the total costs, the predicted risk between these two methods changes drastically. This clear difference between the risk increase between PDP and ALE plots suggests a strong correlation between total costs and at least one other feature. We know that the variables current HCP and total costs are strongly correlated by definition. The current HCP status restricts the total costs to be either above or below a certain cost threshold. ALE plots take this into account, while PDPs ignore that relation. Therefore, ALE is helpful when one is interested in the additional increase that is only caused by total costs with known current HCP status. Total costs itself can change the risk by about 20% on average according to ALE, and not 30% as indicated by PDP. On the other hand, in contrast to PDPs, ALE plots do not offer a clear interpretation of the y-axis, making direct conclusions about the risk assessment harder to interpret.

A clear advantage of our complex model is its flexibility to include heterogeneous feature effects like for total costs, which cannot be modelled by logistic regression. The random forest caps the increase of model predictions for values higher than a certain threshold (around 25,000€) and at around 35%. In logistic regression, every feature with a non-zero coefficient could in theory outweigh all other features held constant. Therefore, the feature effect in logistic regression for both PDP and ALE is always a sigmoid curve with a range of all values between the asymptotes 0 and 1. As confirmed both by the PDP and ALE plot, neither of the learned relationships by the more flexible random forest model for the two continuous variables total costs and age reveals such a sigmoid shape. In general, PDP and ALE plots can be interpreted similarly to coefficients in logistic regression.

Global explanation methods like the variable importance and PDPs were among the first XAI methods that have been around for more than two decades [131], [135]. They aim to extract knowledge from the whole model [163] and allow us to learn about the general relationships between features and responses that we expect to observe by depicting the average increase in risk [106]. However, in an actual deployment of this model, users are probably more interested in why a single patient is predicted to become an HCP. For such single samples, the predicted risk can increase more or less for a certain feature, depending on the other variable values. Explaining the prediction of a single sample can be achieved via local explanations. Such methods only operate on the input space close to the sample to reveal how the model operates in this area. Global relationships can be very different from local ones [100]. These local explanations are often easier to understand due to their very constrained scope [163]. The SHAP and LIME explanations in this study both show feature contributions for single samples. While SHAP is considered a feature contribution method [82], LIME is a simplification method [85].

For the sample investigated here, the SHAP values align mostly with the global relationships. No unexpected model behaviour occurs for these models. The variables that were globally most important were also similarly important for this sample for two out of three SHAP explanations. Their influences were in accordance with the average feature effects identified by the PDP and ALE explanations. Other prescription and diagnosis indicators like a prescription with the ATC code A11 or the outpatient ICD diagnosis I10 only influenced the sample prediction slightly. This

explanation confirmed the globally identified trends. However, this congruity must not hold for all samples.

LIME explanations showed a slightly different picture. Even though the globally most important variables were on top, the plot from the interpretML package revealed a relatively strong and unexpected influence of utilization indicators e.g. the hospital diagnosis with ICD code H46. The absence of this diagnosis significantly increased the predicted risk. Such counterintuitive feature influences highlight that the investigator should cautiously use local explanations. Such discrepancies between different local explanations are caused either by deficiencies in the model or inadequate local representation through the explanation method. In addition, the LIME explanations from the DALEX package showed different variable influences for the same sample.

RESEARCH QUESTION 3: ARE THE GIVEN EXPLANATIONS ACCURATE AND UNDERSTANDABLE ENOUGH TO SUPPORT THE ROUTINE USE OF XAI IN HEALTHCARE OPTIMISATION?

This thesis answered the third question by focusing on the notion of reliability [82]. Clinical decision-makers increasingly need tools that allow them to assess the reliability of explanations for patient predictions [106]. ICE explanations offer a functionally grounded evaluation for the PDP plots. ICE plots reveal heterogeneity in the model predictions whereas PDPs obscure this by averaging the responses [171].

In this thesis, we observed a high variation for age and current HCP status variables and a low variation in the responses for the total costs variable. This suggests caution when using the PDP relationship between variable and response for age and current HCP. With the Python random forest model, being an HCP always increases the probability of remaining in that group in the next year. The magnitude of this increase differs, depending on the demographic and other health utilization variables. For total costs, however, the PDP relationship can generally be trusted and generalized. ICE returns substantially different curves in the alibi package when the plot function is evoked repeatedly on the same generated explanation object. Such unpredictable behaviour adds uncertainty to the plot. The stochastic nature of this method is considered the main obstacle to obtaining high-quality explanations [164].

For the H2O model, the ICE plot shows a line for each decile. Since the H2O implementation only shows 10 lines, this plot only gives a limited overview of the variability of the feature effect. For example, in opposition to the alibi ICE plot, in the ICE plot of the H2O package, the age variable did not show substantial deviations from the PDP trend line, possibly obscuring a more heterogeneous relationship with the model prediction. However, H2O plots were stable, with repeatedly the same lines obtained for the same model.

ICE plots are generally not considered an evaluation method of PDPs, but an explanation method on their own. In our study, we showed that ICE plots extend the display of PDPs sensibly when these two explanations are combined as suggested by Belle and Papantonis [100]. They allow the users and decision-makers to assess the reliability of the relationships of the PDP plots. As for PDPs, ICE plots leverage their strengths best for continuous response variables, but might rely on unrealistic and hence invalid data due to their use of marginal distributions [102].

To be able to trust the explanations, they need to be robust against small perturbances or changes [83]. Especially local explanations tend to have stability issues [100]. The combination and comparison of local explanation methods that show feature contributions decreases the risk of relying on an inadequate local explanation. If the methods show similar feature contributions, we can infer that these explanations are robust and likely to represent the real local behaviour of the model, which is necessary to achieve trustworthy XAI explanations.

The SHAP local explanations are semi-robust for the local explanations in our study and agree on most variable influences. Such agreement indicates that the predictions are stable even across different models and implementations and supports the reliability of the model. However, we just investigated a single example. Conducting further analysis is necessary to better investigate the stability of these models and explanations.

LIME does not guarantee the theoretical properties as SHAP does, including local accuracy, missingness, and consistency [167]. It is susceptible to manipulation as it highly depends on random sampling. Different sampling schemes are needed to mitigate the unreliability of current LIME explanations [184]. In our study, LIME explanations differed substantially from the SHAP local feature contributions in DALEX for the same model. Such unreliability might especially impact the usability of these explanations for model users who are not ML or AI experts, for whom trust and reliability are often very important [97].

A conclusion in accordance with relevant literature is that global explanations are more stable across different models than local explanations [100]. We received similar variable importance and even feature effect explanations from the global explanations for different random forest models in R and Python. However, local explanations differ, like the LIME plot in the interpretML and DALEX packages. While the model user might be more interested in local explanations for specific patients, one should inspect the global explanations first as they are more reliable.

Application grounded evaluations are an important part of facilitating our model deployment. In this work, such evaluations included the assessment of form, prerequisites, and obstacles of the explanations in the real world. Several authors state that good XAI explanations need to be interactive to close the gap between theory and practice in some fields, like clinical support decision systems for physicians [106]. Interactivity is crucial for effective explanations when their success is determined by their usability to the end user [82]. A useful tool to achieve such interactivity is the dashboard functionality in the interpretML package. All explanations can be displayed in an interactive dashboard, where the user can easily select which sample and variable should be explained. Due to the preprocessing, no additional computation time is required to show its explanations. These properties allow explanations generated by interpretML to be transported outside of the development environment. The other packages we explored in this study only produce static explanations in the form of plots and png-files.

For the actual deployment of these models, machine-aided decision-making should be as clear as possible to mitigate trust lost [81]. Different explanations require different levels of user specifications. For the global explanation methods PDP and ALE, the user must choose which variables should be explained. This is not a trivial task, given the 659 predictors included in this study. On the other hand, variable importance and the presented local explanations automatically choose which factors are important, minimizing the need for variable pre-selection by the user.

Another factor influencing the usability and feasibility of these explanations is their algorithmic complexity, determining the computational efficiency [164]. ALE and LIME explanations are the fastest and can be computed in seconds. This makes them suitable for the creation of instant explanations and ad-hoc requests. For some packages, also variable importance, SHAP local and PDP explanations can be computed instantly. However, computation times differ strongly and for other packages, it might take too long to create these explanations instantly. One possible reason

could be the different extent of the created explanations. The interpretML package creates a PDP plot for all variables. Once computed, the user can thus select the variables of interest without further computations. SHAP summary and ICE plots take several minutes across all implementations. Execution times of this magnitude allow request-answer explanation schemes, but not immediate, interactive explanations.

Most packages are available only in Python. Hence Python models can leverage most of these powerful XAI tools and additional XAI methods can be integrated more seamlessly.

LIMITATIONS

This work has provided an overview of an HCP prediction models in combination with XAI methods that present several limitations.

Firstly, we identified the main drivers for the model predictions but did not directly classify HCPs patients into different subgroups. Such division is crucial to effectively target these patients with tailored interventions [48]. For this, an unsupervised learning approach could be used [50].

Secondly, robustness has been mentioned as an important characteristic of a trustworthy ML model [83]. However, to evaluate and increase robustness, model external validation was performed (both geographical and temporal). More robustness evaluation could have included, e.g. sensitivity analysis (predict top 1% or 10% cost quantile patients instead of top 5%) like in [58]. Moreover, along with decision curve analysis, [119] suggest the inclusion of calibration and discrimination that measure the agreement between predictions and observed endpoints. This ensures that the predicted probabilities are comparable to real risks and should be included to guarantee the comparability of the model's net benefits.

Finally, underprivileged populations might not access healthcare as frequently and thus might be excluded from this analysis. Rosella et al. [54] observed an increased risk for white and non-immigrant individuals, which might be due to limited service availability for other ethnic groups. Contrarily, in another study, low income and little education greatly increased the risk of becoming an HCP [63]. Our data selection strategy does not allow the assessment of such differences between population groups. More demographic data like family income and living situation should be included to account for potential biases in the model. Additionally, one must be careful to generalize our results to other countries or even all of Germany. Our study population is only

recruited from individuals in two rural regions in Germany that are not representative of the German population.

CONCLUSIONS

Our findings confirm previous results that complex machine learning models, like tree ensemble models, identify prospective high-cost patients more effectively than traditional logistic regression models. The random forest model developed in this study is consistent with a similar model on older, equivalent data built and stands out as one of the best prediction models reported in the literature. Furthermore, integrating cost capture and decision curve analysis provides valuable insights into the usefulness and potential of the model for real-world applications.

The black-box nature of machine learning models poses a significant challenge to their successful application in healthcare, particularly in the context of high-cost patient models, where limited efforts have been made in this direction. To address this issue, we leveraged global and local explainable AI (XAI) methods to explain the inner workings of our random forest model, aligning them with procedures from traditional logistic regression analysis. Using variable importance measures alongside partial dependence plots (PDP) and accumulated local effects (ALE), we identified the most influential variables consistent with related literature and their non-linear effects. In addition, by applying SHAP and LIME explanation methods, we obtained granular insights into the model's reasoning for a single sample, revealing some unexpected variable influences, including certain diseases and prescriptions.

To foster trust in our model, we evaluated its explanations using the scarce statistically sound methods available. Individual conditional expectations (ICE) plots revealed heterogeneity in the feature effects obscured by PDP and ALE explanations. Stability comparisons indicate that global methods are more robust than local ones, suggesting that these should be inspected first even for patient-specific insights. Adequate tools are essential for real-world applications and evaluating the characteristics of our utilized XAI programming packages H2O, DALEX, alibi, and interpretML highlights significant differences in execution times, with the latter presenting a user-friendly interactive dashboard suited for deployment in non-expert settings.

This thesis offered an opportunity to those interested in using novel XAI models to understand and familiarize themselves with the opportunities and risks presented by these models. Future applied work should focus on how to investigate the current utilization of healthcare services to highlight actionable measures. These approaches will improve outcomes and encourage cost-saving

incentives in high-cost patients that are highly needed to relieve increasing financial pressure on health systems [61].

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my professor and supervisor of this thesis, Fabrizio Carinci, for his invaluable feedback and guidance during my research. I am especially thankful for apl.-Professor Oliver Gröne who supported me during the beginning of this endeavour in shaping the idea for this topic and sharing his expertise with me. I also could not have undertaken this journey without the members and coworkers at OptiMedis AG who warmly welcomed me during my two-month internship and generously provided the foundational data for this thesis.

I am also grateful for my family and friends who did not hesitate to provide me with profound feedback and motivational words when I needed them. Special thanks also go to my fellow students with whom I shared a lot of time at the library for the regular coffee breaks and inspiration between study sessions.

REFERENCES

[1] C. Janiesch, P. Zschech, and K. Heinrich, 'Machine learning and deep learning', *Electron. Mark.*, vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/s12525-021-00475-2.

[2] M. Baek *et al.*, 'Accurate prediction of protein structures and interactions using a 3-track neural network', *Science*, vol. 373, no. 6557, pp. 871–876, Aug. 2021, doi: 10.1126/science.abj8754.

[3] L. M. F. Bertoline, A. N. Lima, J. E. Krieger, and S. K. Teixeira, 'Before and after AlphaFold2: An overview of protein structure prediction', *Front. Bioinforma.*, vol. 3, Feb. 2023, doi: 10.3389/fbinf.2023.1120370.

[4] Z. Yao *et al.*, 'Machine learning for a sustainable energy future', *Nat. Rev. Mater.*, vol. 8, no. 3, pp. 202–215, Mar. 2023, doi: 10.1038/s41578-022-00490-5.

[5] T. R. Jena, S. S. Barik, and S. K. Nayak, 'Electricity Consumption & Prediction Using Machine Learning Models', *Acta Tech. Corviniensis - Bull. Eng.*, vol. 14, no. 1, p. 61,67-68, Mar. 2021.

[6] D. Silver *et al.*, 'Mastering the game of Go with deep neural networks and tree search', *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016, doi: 10.1038/nature16961.

[7] N. Farnaaz and M. A. Jabbar, 'Random Forest Modeling for Network Intrusion Detection System', *Procedia Comput. Sci.*, vol. 89, pp. 213–217, Jan. 2016, doi: 10.1016/j.procs.2016.06.047.

[8] J. Ma *et al.*, '"Bingo"—a large language model- and graph neural network-based workflow for the prediction of essential genes from protein data', *Brief. Bioinform.*, vol. 25, no. 1, p. bbad472, Dec. 2023, doi: 10.1093/bib/bbad472.

[9] H. Wang, X. Yang, S. Ma, K. Zhu, and S. Guo, 'An Optimized Radiomics Model Based on Automated Breast Volume Scan Images to Identify Breast Lesions', *J. Ultrasound Med.*, vol. 41, no. 7, pp. 1643–1655, 2022, doi: 10.1002/jum.15845.

[10] Y. Belsti *et al.*, 'Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model', *Int. J. Med. Inf.*, vol. 179, p. 105228, Nov. 2023, doi: 10.1016/j.ijmedinf.2023.105228.

[11] B. Langenberger, D. Schrednitzki, A. M. Halder, R. Busse, and C. M. Pross, 'Predicting whether patients will achieve minimal clinically important differences following hip or knee arthroplasty', *Bone Jt. Res.*, vol. 12, no. 9, pp. 512–521, Sep. 2023, doi: 10.1302/2046-3758.129.BJR-2023-0070.R2.

[12] S. K. Ramu and A. Byale, 'S1417 Comparison of Logistic Regression Model With Machine Learning Models to Predict Acute Liver Injury in Patients Hospitalized With COVID-19', *Off. J. Am. Coll. Gastroenterol. ACG*, vol. 118, no. 10S, p. S1081, Oct. 2023, doi: 10.14309/01.ajg.0000955308.75334.21.

[13] G. Forkuor, O. K. L. Hounkpatin, G. Welp, and M. Thiel, 'High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models', *PLoS One*, vol. 12, no. 1, Jan. 2017, doi: 10.1371/journal.pone.0170478.

[14] M. A. Khalil, M. R. Fatmi, and M. Orvin, 'Developing and microsimulating demographic dynamics for an integrated urban model: a comparison between logistic regression and machine learning techniques', *Transportation*, Feb. 2024, doi: 10.1007/s11116-024-10468-7.

[15] A. Zahlan, R. P. Ranjan, and D. Hayes, 'Artificial intelligence innovation in healthcare: Literature review, exploratory analysis, and future research', *Technol. Soc.*, vol. 74, p. 102321, Aug. 2023, doi: 10.1016/j.techsoc.2023.102321.

[16] X. Du-Harpur, F. M. Watt, N. M. Luscombe, and M. D. Lynch, 'What is AI? Applications of artificial intelligence to dermatology', *Br. J. Dermatol.*, vol. 183, no. 3, pp. 423–430, Sep. 2020, doi: 10.1111/bjd.18880.

[17] S. B. Golas *et al.*, 'A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data', *BMC Med. Inform. Decis. Mak.*, vol. 18, no. 1, p. 44, Jun. 2018, doi: 10.1186/s12911-018-0620-z.

[18] H. A. Haenssle *et al.*, 'Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists', *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, Aug. 2018, doi: 10.1093/annonc/mdy166.

[19] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, 'Applications of Support Vector Machine (SVM) Learning in Cancer Genomics', *Cancer Genomics Proteomics*, vol. 15, no. 1, pp. 41–51, Jan. 2018.

[20] C. Iwendi *et al.*, 'COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm', *Front. Public Health*, vol. 8, Jul. 2020, doi: 10.3389/fpubh.2020.00357.

[21] F. Schwendicke *et al.*, 'Cost-effectiveness of Artificial Intelligence for Proximal Caries Detection', *J. Dent. Res.*, vol. 100, no. 4, pp. 369–376, Apr. 2021, doi: 10.1177/0022034520972335.

[22] K. N. Kunze, E. M. Polce, A. J. Sadauskas, and B. R. Levine, 'Development of Machine Learning Algorithms to Predict Patient Dissatisfaction After Primary Total Knee Arthroplasty', *J. Arthroplasty*, vol. 35, no. 11, pp. 3117–3122, Nov. 2020, doi: 10.1016/j.arth.2020.05.061.

[23] O. Bayar Kapici, Y. Kapici, A. Tekın, and M. Şırık, 'A novel diagnosis method for schizophrenia based on globus pallidus data', *Psychiatry Res. Neuroimaging*, vol. 336, p. 111732, Dec. 2023, doi: 10.1016/j.pscychresns.2023.111732.

[24] S. Yue *et al.*, 'Machine learning for the prediction of acute kidney injury in patients with sepsis', *J. Transl. Med.*, vol. 20, no. 1, p. 215, May 2022, doi: 10.1186/s12967-022-03364-0.

[25] C. Preiksaitis *et al.*, 'The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review', *JMIR Med. Inform.*, vol. 12, no. 1, p. e53787, May 2024, doi: 10.2196/53787.

[26] R. Li, A. Kumar, and J. H. Chen, 'How Chatbots and Large Language Model Artificial Intelligence Systems Will Reshape Modern Medicine: Fountain of Creativity or Pandora's Box?', *JAMA Intern. Med.*, vol. 183, no. 6, pp. 596–597, Jun. 2023, doi: 10.1001/jamainternmed.2023.1835.

[27] S. A. Alowais *et al.*, 'Revolutionizing healthcare: the role of artificial intelligence in clinical practice', *BMC Med. Educ.*, vol. 23, no. 1, p. 689, Sep. 2023, doi: 10.1186/s12909-023-04698-z.
[28] A. Alanazi, 'Using machine learning for healthcare challenges and opportunities', *Inform. Med. Unlocked*, vol. 30, p. 100924, Jan. 2022, doi: 10.1016/j.imu.2022.100924.
[29] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, 'AI in health and medicine', *Nat. Med.*, vol. 28, no. 1, pp. 31–38, Jan. 2022, doi: 10.1038/s41591-021-01614-0.
[30] M. M. Voets, J. Veltman, C. H. Slump, S. Siesling, and H. Koffijberg, 'Systematic Review of Health Economic Evaluations Focused on Artificial Intelligence in Healthcare: The Tortoise and the Cheetah', *Value Health*, vol. 25, no. 3, pp. 340–349, Mar. 2022, doi: 10.1016/j.jval.2021.11.1362.
[31] I. El Naqa *et al.*, 'Translation of AI into oncology clinical practice', *Oncogene*, vol. 42, no. 42, pp. 3089–3097, Oct. 2023, doi: 10.1038/s41388-023-02826-z.
[32] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, 'A Dataset for Breast Cancer Histopathological Image Classification', *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455–1462, Jul. 2016, doi: 10.1109/TBME.2015.2496264.
[33] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, 'Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis', *Proc. Natl. Acad. Sci.*, vol. 117, no. 23, pp. 12592–12594, Jun. 2020, doi: 10.1073/pnas.1919012117.
[34] K. Goddard, A. Roudsari, and J. C. Wyatt, 'Automation bias: a systematic review of frequency, effect mediators, and mitigators', *J. Am. Med. Inform. Assoc.*, vol. 19, no. 1, pp. 121–127, Jan. 2012, doi: 10.1136/amiajnl-2011-000089.
[35] T. Hagendorff, 'The Ethics of AI Ethics: An Evaluation of Guidelines', *Minds Mach.*, vol. 30, no. 1, pp. 99–120, Mar. 2020, doi: 10.1007/s11023-020-09517-8.
[36] J. Waring, C. Lindvall, and R. Umeton, 'Automated machine learning: Review of the state-of-the-art and opportunities for healthcare', *Artif. Intell. Med.*, vol. 104, p. 101822, Apr. 2020, doi: 10.1016/j.artmed.2020.101822.
[37] OECD, *Health at a Glance 2023: OECD Indicators*. in Health at a Glance. OECD, 2023. doi: 10.1787/7a7afb35-en.
[38] A. Dreher, × Mirjam Theune, C. Kersting, F. Geiser, and B. Weltermann, 'Prevalence of burnout among German general practitioners: Comparison of physicians working in solo and group practices', *PLoS One*, vol. 14, no. 2, Feb. 2019, doi: 10.1371/journal.pone.0211223.
[39] B. Pantenburg, K. Kitze, M. Luppa, H.-H. König, and S. G. Riedel-Heller, 'Physician emigration from Germany: insights from a survey in Saxony, Germany', *BMC Health Serv. Res.*, vol. 18, p. 341, May 2018, doi: 10.1186/s12913-018-3142-6.
[40] J. Lermann *et al.*, 'The work and training situation for young physicians undergoing specialty training in gynecology and obstetrics in Germany: an assessment of the status quo', *Arch. Gynecol. Obstet.*, vol. 302, no. 3, pp. 635–647, Sep. 2020, doi: 10.1007/s00404-020-05616-0.
[41] H. Sturm *et al.*, 'Do perceived working conditions and patient safety culture correlate with objective workload and patient outcomes: A cross-sectional explorative study

from a German university hospital', *PLoS One*, vol. 14, no. 1, Jan. 2019, doi: 10.1371/journal.pone.0209487.

[42] A. Powell, S. Savin, and N. Savva, 'Physician Workload and Hospital Reimbursement: Overworked Physicians Generate Less Revenue per Patient', *Manuf. Serv. Oper. Manag.*, vol. 14, no. 4, pp. 512–528, Oct. 2012, doi: 10.1287/msom.1120.0384.

[43] OECD, *Fiscal Sustainability of Health Systems: How to Finance More Resilient Health Systems When Money Is Tight?* OECD, 2024. doi: 10.1787/880f3195-en.

[44] G. Iacobucci, 'Budget cuts have worsened quality of mental healthcare, think tank warns', *BMJ Br. Med. J. Online*, vol. 351, Nov. 2015, doi: 10.1136/bmj.h6121.

[45] O. Iban, R. Gemma, and X. Angels, 'Effects of Public Healthcare Budget Cuts on Life Satisfaction in Spain', *Soc. Indic. Res.*, vol. 156, no. 1, pp. 311–337, Jul. 2021, doi: 10.1007/s11205-021-02624-8.

[46] B. Langenberger, T. Schulte, and O. Groene, 'The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data', *PloS One*, vol. 18, no. 1, p. e0279540, 2023, doi: 10.1371/journal.pone.0279540.

[47] P. M. Reardon *et al.*, 'Characteristics, Outcomes, and Cost Patterns of High-Cost Patients in the Intensive Care Unit', *Crit. Care Res. Pract.*, vol. 2018, no. 1, p. 5452683, Jan. 2018, doi: 10.1155/2018/5452683.

[48] D. Blumenthal, C. Bruce, T. Fulmer, J. Lumpkin, and J. Selberg, 'Caring for High-Need, High-Cost Patients — An Urgent Priority', *N. Engl. J. Med.*, vol. 375, no. 10, pp. 909–911, Sep. 2016, doi: 10.1056/NEJMp1608511.

[49] J. M. McWilliams and A. L. Schwartz, 'Focusing on High-cost Patients: the Key to Addressing High Costs?', *N. Engl. J. Med.*, vol. 376, no. 9, pp. 807–809, Mar. 2017, doi: 10.1056/NEJMp1612779.

[50] S. H. X. Ng *et al.*, 'Characterising and predicting persistent high-cost utilisers in healthcare: a retrospective cohort study in Singapore', *BMJ Open*, vol. 10, no. 1, p. e031622, Jan. 2020, doi: 10.1136/bmjopen-2019-031622.

[51] S. Tamang *et al.*, 'Predicting patient "cost blooms" in Denmark: a longitudinal population-based study', *BMJ Open*, vol. 7, no. 1, p. e011580, Jan. 2017, doi: 10.1136/bmjopen-2016-011580.

[52] N. D. Berkman, E. Chang, J. Seibert, and R. Ali, 'Characteristics of High-Need, High-Cost Patients', *Ann. Intern. Med.*, Nov. 2022, doi: 10.7326/M21-4562.

[53] T. L. Johnson *et al.*, 'For Many Patients Who Use Large Amounts Of Health Care Services, The Need Is Intense Yet Temporary', *Health Aff. (Millwood)*, Aug. 2017, doi: 10.1377/hlthaff.2014.1186.

[54] L. C. Rosella, K. Kornas, J. Sarkar, and R. Fransoo, 'External Validation of a Population-Based Prediction Model for High Healthcare Resource Use in Adults', *Healthcare*, vol. 8, no. 4, Art. no. 4, Dec. 2020, doi: 10.3390/healthcare8040537.

[55] U. W. de Ruijter *et al.*, 'Prediction Models for Future High-Need High-Cost Healthcare Use: a Systematic Review', *J. Gen. Intern. Med.*, vol. 37, no. 7, pp. 1763–1770, May 2022, doi: 10.1007/s11606-021-07333-z.

[56] I. Papanicolas *et al.*, 'Differences in health outcomes for high-need high-cost patients across high-income countries', *Health Serv. Res.*, vol. 56, no. Suppl 3, pp. 1347–1357, Dec. 2021, doi: 10.1111/1475-6773.13735.

[57] C. de Oliveira, J. Cheng, S. Vigod, J. Rehm, and P. Kurdyak, 'Patients With High Mental Health Costs Incur Over 30 Percent More Costs Than Other High-Cost Patients', *Health Aff. (Millwood)*, Aug. 2017, doi: 10.1377/hlthaff.2015.0278.

[58] I. Osawa, T. Goto, Y. Yamamoto, and Y. Tsugawa, 'Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data', *Npj Digit. Med.*, vol. 3, no. 1, pp. 1–9, Nov. 2020, doi: 10.1038/s41746-020-00354-8.

[59] S. Rais, A. Nazerian, S. Ardal, Y. Chechulin, N. Bains, and K. Malikov, 'High-Cost Users of Ontario's Healthcare Services', *Healthc. Policy*, vol. 9, no. 1, pp. 44–51, Aug. 2013.

[60] N. S. Lee, N. Whitman, N. Vakharia, G. B. T. PhD, and M. B. Rothberg, 'High-Cost Patients: Hot-Spotters Don't Explain the Half of It', *J. Gen. Intern. Med.*, vol. 32, no. 1, pp. 28–34, Jan. 2017, doi: 10.1007/s11606-016-3790-3.

[61] K. Geurts, M. Bruijnzeels, and E. Schokkaert, 'Do we care about high-cost patients? Estimating the savings on health spending by integrated care', *Eur. J. Health Econ.*, vol. 23, no. 8, pp. 1297–1308, Nov. 2022, doi: 10.1007/s10198-022-01431-3.

[62] J. J. G. Wammes, M. Tanke, W. Jonkers, G. P. Westert, P. Van Der Wees, and P. P. Jeurissen, 'Characteristics and healthcare utilisation patterns of high-cost beneficiaries in the Netherlands: a cross-sectional claims database study', *BMJ Open*, vol. 7, no. 11, p. e017775, Nov. 2017, doi: 10.1136/bmjopen-2017-017775.

[63] T. Fitzpatrick *et al.*, 'Looking Beyond Income and Education: Socioeconomic Status Gradients Among Future High-Cost Users of Health Care', *Am. J. Prev. Med.*, vol. 49, no. 2, pp. 161–171, Aug. 2015, doi: 10.1016/j.amepre.2015.02.018.

[64] J. Yan *et al.*, 'Applying Machine Learning Algorithms to Segment High-Cost Patient Populations', *J. Gen. Intern. Med.*, vol. 34, no. 2, pp. 211–217, Feb. 2019, doi: 10.1007/s11606-018-4760-8.

[65] O. Kieu Nguyen, N. Tang, J. M. Hillman, and R. Gonzales, 'What's cost got to do with it? Association between hospital costs and frequency of admissions among "high users" of hospital care', *J. Hosp. Med.*, vol. 8, no. 12, pp. 665–671, 2013, doi: 10.1002/jhm.2096.

[66] S. V. Nuti, P. Doupe, B. Villanueva, J. Scarpa, E. Bruzelius, and A. Baum, 'Characterizing Subgroups of High-Need, High-Cost Patients Based on Their Clinical Conditions: a Machine Learning-Based Analysis of Medicaid Claims Data', *J. Gen. Intern. Med.*, vol. 34, no. 8, pp. 1406–1408, Aug. 2019, doi: 10.1007/s11606-019-04941-8.

[67] OECD, *Rethinking Health System Performance Assessment: A Renewed Framework*. Paris: Organisation for Economic Co-operation and Development, 2024. Accessed: Jul. 04, 2024. [Online]. Available: https://www.oecd-ilibrary.org/social-issues-migration-health/rethinking-health-system-performance-assessment_107182c8-en

[68] D. Rajan, I. Papanicolas, M. Karanikolos, K. Koch, K. Rohrer-Herold, and J. Figueras, 'Health system performance assessment', *Primer Policy-Mak. WHO Cph.*, 2022.

[69] F. Carinci *et al.*, 'Towards actionable international comparisons of health system performance: expert revision of the OECD framework and quality indicators', *Int. J. Qual. Health Care*, vol. 27, no. 2, pp. 137–146, Apr. 2015, doi: 10.1093/intqhc/mzv004.

[70] C. Crowley, J. Perloff, A. Stuck, and R. Mechanic, 'Challenges in predicting future high-cost patients for care management interventions', *BMC Health Serv. Res.*, vol. 23, p. 992, Sep. 2023, doi: 10.1186/s12913-023-09957-9.
[71] M. E. Porter, 'What Is Value in Health Care?', *N. Engl. J. Med.*, vol. 363, no. 26, pp. 2477–81, Dec. 2010, doi: 10.1056/NEJMp1011024.
[72] D. Blumenthal and M. K. Abrams, 'Tailoring Complex Care Management for High-Need, High-Cost Patients', *JAMA*, vol. 316, no. 16, pp. 1657–1658, Oct. 2016, doi: 10.1001/jama.2016.12388.
[73] L. S. F. de Carvalho *et al.*, 'Machine Learning Improves the Identification of Individuals With Higher Morbidity and Avoidable Health Costs After Acute Coronary Syndromes', *Value Health*, vol. 23, no. 12, pp. 1570–1579, Dec. 2020, doi: 10.1016/j.jval.2020.08.2091.
[74] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, 'A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models', *J. Clin. Epidemiol.*, vol. 110, pp. 12–22, Jun. 2019, doi: 10.1016/j.jclinepi.2019.02.004.
[75] J. Wolff, J. Pauling, A. Keck, and J. Baumbach, 'The Economic Impact of Artificial Intelligence in Health Care: Systematic Review', *J. Med. Internet Res.*, vol. 22, no. 2, p. e16866, Feb. 2020, doi: 10.2196/16866.
[76] O. Ericson, J. Hjelmgren, F. Sjövall, J. Söderberg, and I. Persson, 'The Potential Cost and Cost-Effectiveness Impact of Using a Machine Learning Algorithm for Early Detection of Sepsis in Intensive Care Units in Sweden', *J. Health Econ. Outcomes Res.*, vol. 9, no. 1, pp. 101–110, 2022, doi: 10.36469/jheor.2022.33951.
[77] J. de Vos *et al.*, 'The Potential Cost-Effectiveness of a Machine Learning Tool That Can Prevent Untimely Intensive Care Unit Discharge', *Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res.*, vol. 25, no. 3, pp. 359–367, Mar. 2022, doi: 10.1016/j.jval.2021.06.018.
[78] V. Bremer *et al.*, 'Predicting Therapy Success and Costs for Personalized Treatment Recommendations Using Baseline Characteristics: Data-Driven Analysis', *J. Med. Internet Res.*, vol. 20, no. 8, p. e10275, Aug. 2018, doi: 10.2196/10275.
[79] A. L. Nelson, J. T. Cohen, D. Greenberg, and D. M. Kent, 'Much Cheaper, Almost as Good: Decrementally Cost-Effective Medical Innovation', *Ann. Intern. Med.*, Nov. 2009, Accessed: Jun. 28, 2024. [Online]. Available: https://www.acpjournals.org/doi/10.7326/0003-4819-151-9-200911030-00011
[80] A. A. H. de Hond *et al.*, 'Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review', *Npj Digit. Med.*, vol. 5, no. 1, pp. 1–13, Jan. 2022, doi: 10.1038/s41746-021-00549-7.
[81] S. Ali *et al.*, 'Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence', *Inf. Fusion*, vol. 99, p. 101805, Nov. 2023, doi: 10.1016/j.inffus.2023.101805.
[82] A. Barredo Arrieta *et al.*, 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[83] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, 'Definitions, methods, and applications in interpretable machine learning', *Proc. Natl. Acad. Sci.*, vol. 116, no. 44, pp. 22071–22080, Oct. 2019, doi: 10.1073/pnas.1900654116.

[84] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, 'XAI-Explainable artificial intelligence', *Sci. Robot.*, vol. 4, no. 37, p. eaay7120, Dec. 2019, doi: 10.1126/scirobotics.aay7120.

[85] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, 'Explainable artificial intelligence: a comprehensive review', *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 3503–3568, Jun. 2022, doi: 10.1007/s10462-021-10088-y.

[86] C. Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[87] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, 'Interpretable machine learning: Fundamental principles and 10 grand challenges', *Stat. Surv.*, vol. 16, no. none, 2022, doi: 10.1214/21-SS133.

[88] M. Stefik *et al.*, 'Explaining autonomous drones: An XAI journey', *Appl. AI Lett.*, vol. 2, no. 4, p. e54, 2021, doi: 10.1002/ail2.54.

[89] J. Dong, S. Chen, M. Miralinaghi, T. Chen, P. Li, and S. Labi, 'Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems', *Transp. Res. Part C Emerg. Technol.*, vol. 156, p. 104358, Nov. 2023, doi: 10.1016/j.trc.2023.104358.

[90] H.-S. Kim and I. Joe, 'An XAI method for convolutional neural networks in self-driving cars', *PLOS ONE*, vol. 17, no. 8, p. e0267282, Aug. 2022, doi: 10.1371/journal.pone.0267282.

[91] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, 'Unmasking Clever Hans predictors and assessing what machines really learn', *Nat. Commun.*, vol. 10, no. 1, p. 1096, Mar. 2019, doi: 10.1038/s41467-019-08987-4.

[92] A. Dikshit and B. Pradhan, 'Interpretable and explainable AI (XAI) model for spatial drought prediction', *Sci. Total Environ.*, vol. 801, p. 149797, Dec. 2021, doi: 10.1016/j.scitotenv.2021.149797.

[93] A. Deeks, 'The Judicial Demand for Explainable Artificial Intelligence', *Columbia Law Rev.*, vol. 119, no. 7, pp. 1829–1850, 2019.

[94] B. Predić, M. Ćirić, and L. Stoimenov, 'Business Purchase Prediction Based on XAI and LSTM Neural Networks', *Electronics*, vol. 12, no. 21, 2023, doi: 10.3390/electronics12214510.

[95] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, 'Explainable artificial intelligence (XAI) in deep learning-based medical image analysis', *Med. Image Anal.*, vol. 79, p. 102470, Jul. 2022, doi: 10.1016/j.media.2022.102470.

[96] S. A. Martin, F. J. Townend, F. Barkhof, and J. H. Cole, 'Interpretable machine learning for dementia: A systematic review', *Alzheimers Dement.*, vol. 19, no. 5, pp. 2135–2149, May 2023, doi: 10.1002/alz.12948.

[97] S. Mohseni, N. Zarei, and E. D. Ragan, 'A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems', *ACM Trans Interact Intell Syst*, vol. 11, no. 3–4, p. 24:1-24:45, Sep. 2021, doi: 10.1145/3387166.

[98] A. Rai, 'Explainable AI: from black box to glass box', *J. Acad. Mark. Sci.*, vol. 48, no. 1, pp. 137–141, Jan. 2020, doi: 10.1007/s11747-019-00710-5.
[99] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, 'Explainable AI: A Review of Machine Learning Interpretability Methods', *Entropy*, vol. 23, no. 1, Art. no. 1, Jan. 2021, doi: 10.3390/e23010018.
[100] V. Belle and I. Papantonis, 'Principles and Practice of Explainable Machine Learning', *Front. Big Data*, vol. 4, Jul. 2021, doi: 10.3389/fdata.2021.688969.
[101] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, 'Principles of Explanatory Debugging to Personalize Interactive Machine Learning', in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, in IUI '15. New York, NY, USA: Association for Computing Machinery, März 2015, pp. 126–137. doi: 10.1145/2678025.2701399.
[102] R. Dwivedi *et al.*, 'Explainable AI (XAI): Core Ideas, Techniques, and Solutions', *ACM Comput Surv*, vol. 55, no. 9, p. 194:1-194:33, Jan. 2023, doi: 10.1145/3561048.
[103] *Regulation (EU) 2016/679*, vol. 119. 2016. Accessed: Jun. 16, 2024. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj/eng
[104] M. E. Kaminski, 'The Right to Explanation, Explained', *Berkeley Technol. Law J.*, vol. 34, no. 1, pp. 189–218, 2019.
[105] B. Goodman and S. Flaxman, 'European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"', *AI Mag.*, vol. 38, no. 3, pp. 50–57, Fall 2017.
[106] A. M. Antoniadi *et al.*, 'Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review', *Appl. Sci.*, vol. 11, no. 11, Art. no. 11, Jan. 2021, doi: 10.3390/app11115088.
[107] J. Wiens *et al.*, 'Do no harm: a roadmap for responsible machine learning for health care', *Nat. Med.*, vol. 25, no. 9, pp. 1337–1340, Sep. 2019, doi: 10.1038/s41591-019-0548-6.
[108] E. Tjoa and C. Guan, 'A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI', *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.
[109] S. N. van der Veer *et al.*, 'Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries', *J. Am. Med. Inform. Assoc. JAMIA*, vol. 28, no. 10, pp. 2128–2138, Aug. 2021, doi: 10.1093/jamia/ocab127.
[110] J. Peng *et al.*, 'An Explainable Artificial Intelligence Framework for the Deterioration Risk Prediction of Hepatitis Patients', *J. Med. Syst.*, vol. 45, no. 5, p. 61, Apr. 2021, doi: 10.1007/s10916-021-01736-5.
[111] I. Neves *et al.*, 'Interpretable heartbeat classification using local model-agnostic explanations on ECGs', *Comput. Biol. Med.*, vol. 133, p. 104393, Jun. 2021, doi: 10.1016/j.compbiomed.2021.104393.
[112] B. M. de Vries, G. J. C. Zwezerijnen, G. L. Burchell, F. H. P. van Velden, C. W. Menke-van der Houven van Oordt, and R. Boellaard, 'Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review', *Front. Med.*, vol. 10, May 2023, doi: 10.3389/fmed.2023.1180773.

[113] J. Tritscher, A. Krause, and A. Hotho, 'Feature relevance XAI in anomaly detection: Reviewing approaches and challenges', *Front. Artif. Intell.*, vol. 6, Feb. 2023, doi: 10.3389/frai.2023.1099521.
[114] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, 'Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)', *Comput. Methods Programs Biomed.*, vol. 226, p. 107161, Nov. 2022, doi: 10.1016/j.cmpb.2022.107161.
[115] J. Born *et al.*, 'Accelerating Detection of Lung Pathologies with Explainable Ultrasound Image Analysis', *Appl. Sci.*, vol. 11, no. 2, Art. no. 2, Jan. 2021, doi: 10.3390/app11020672.
[116] K. Borys *et al.*, 'Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches', *Eur. J. Radiol.*, vol. 162, p. 110787, May 2023, doi: 10.1016/j.ejrad.2023.110787.
[117] E. Amparore, A. Perotti, and P. Bajardi, 'To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods', *PeerJ Comput. Sci.*, vol. 7, p. e479, 2021, doi: 10.7717/peerj-cs.479.
[118] C. Yang, C. Delcher, E. Shenkman, and S. Ranka, 'Machine learning approaches for predicting high cost high need patient expenditures in health care', *Biomed. Eng. OnLine*, vol. 17, no. 1, p. 131, Nov. 2018, doi: 10.1186/s12938-018-0568-3.
[119] E. W. Steyerberg and Y. Vergouwe, 'Towards better clinical prediction models: seven steps for development and an ABCD for validation', *Eur. Heart J.*, vol. 35, no. 29, pp. 1925–1931, Aug. 2014, doi: 10.1093/eurheartj/ehu207.
[120] 'DMP - Bundesamt für Soziale Sicherung'. Accessed: Aug. 12, 2024. [Online]. Available: https://www.bundesamtsozialesicherung.de/de/themen/disease-management-programme/dmp-grundlegende-informationen/
[121] L. C. Rosella *et al.*, 'Predicting High Health Care Resource Utilization in a Single-payer Public Health Care System: Development and Validation of the High Resource User Population Risk Tool', *Med. Care*, vol. 56, no. 10, p. e61, Oct. 2018, doi: 10.1097/MLR.0000000000000837.
[122] M. J. Baker and J. L. Bonkowsky, 'An introduction to ICD code development for pediatric neurology', *Ann. Child Neurol. Soc.*, vol. 1, no. 3, pp. 180–185, Sep. 2023, doi: 10.1002/cns3.20028.
[123] Norwegian Institute of Public Health, WHO Collaborating Centre for Drug Statistics Methodology, 'ATCDDD - Purpose of the ATC/DDD system'. Accessed: Sep. 19, 2024. [Online]. Available: https://atcddd.fhi.no/atc_ddd_methodology/purpose_of_the_atc_ddd_system/
[124] Norwegian Institute of Public Health, WHO Collaborating Centre for Drug Statistics Methodology, 'ATCDDD - Structure and principles'. Accessed: Sep. 19, 2024. [Online]. Available: https://atcddd.fhi.no/atc/structure_and_principles/
[125] M. P. LaValley, 'Logistic Regression', *Circulation*, vol. 117, no. 18, pp. 2395–2399, May 2008, doi: 10.1161/CIRCULATIONAHA.106.682658.
[126] S. Sperandei, 'Understanding logistic regression analysis', *Biochem. Medica*, vol. 24, no. 1, pp. 12–18, Feb. 2014, doi: 10.11613/BM.2014.003.

[127] M. J. Kist and R. T. Silvestrini, 'Incorporating Confidence Intervals on the Decision Threshold in Logistic Regression', *Qual. Reliab. Eng. Int.*, vol. 32, no. 5, pp. 1769–1784, 2016, doi: 10.1002/qre.1912.
[128] J. Lindström and J. Tuomilehto, 'The diabetes risk score: a practical tool to predict type 2 diabetes risk', *Diabetes Care*, vol. 26, no. 3, pp. 725–731, Mar. 2003, doi: 10.2337/diacare.26.3.725.
[129] Y. Fu, L. Yang, J. Du, R. Khan, and D. Liu, 'Establishment of HIV-negative neurosyphilis risk score model based on logistic regression', *Eur. J. Med. Res.*, vol. 28, no. 1, p. 200, Jun. 2023, doi: 10.1186/s40001-023-01177-5.
[130] M. Polavarapu, H. Klonoff-Cohen, D. Joshi, P. Kumar, R. An, and K. Rosenblatt, 'Development of a Risk Score to Predict Sudden Infant Death Syndrome', *Int. J. Environ. Res. Public. Health*, vol. 19, no. 16, 2022, doi: 10.3390/ijerph191610270.
[131] L. Breiman, 'Random Forests', *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[132] A. Buja and W. Stuetzle, 'Observations on Bagging', *Stat. Sin.*, vol. 16, no. 2, pp. 323–351, 2006.
[133] P. Probst, M. N. Wright, and A.-L. Boulesteix, 'Hyperparameters and tuning strategies for random forest', *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 3, p. e1301, 2019, doi: 10.1002/widm.1301.
[134] G. Biau, 'Analysis of a random forests model', *J. Mach. Learn. Res.*, vol. 13, pp. 1063–1095, 2012.
[135] J. H. Friedman, 'Greedy Function Approximation: A Gradient Boosting Machine', *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.
[136] H. Lu and R. Mazumder, 'Randomized Gradient Boosting Machine', *SIAM J. Optim.*, vol. 30, no. 4, pp. 2780–2808, Jan. 2020, doi: 10.1137/18M1223277.
[137] B. Candice, C. Anna, and M.-M. Gonzalo, 'A comparative analysis of gradient boosting algorithms', *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.
[138] W. S. McCulloch and W. Pitts, 'A logical calculus of the ideas immanent in nervous activity', *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, Dec. 1943, doi: 10.1007/BF02478259.
[139] I. A. Basheer and M. Hajmeer, 'Artificial neural networks: fundamentals, computing, design, and application', *J. Microbiol. Methods*, vol. 43, no. 1, pp. 3–31, Dec. 2000, doi: 10.1016/S0167-7012(00)00201-3.
[140] S. Dreiseitl and L. Ohno-Machado, 'Logistic regression and artificial neural network classification models: a methodology review', *J. Biomed. Inform.*, vol. 35, no. 5, pp. 352–359, Oct. 2002, doi: 10.1016/S1532-0464(03)00034-0.
[141] M. G. M. Abdolrasol *et al.*, 'Artificial Neural Networks Based Optimization Techniques: A Review', *Electronics*, vol. 10, no. 21, Art. no. 21, Jan. 2021, doi: 10.3390/electronics10212689.
[142] Z. Bursac, C. H. Gauss, D. K. Williams, and D. W. Hosmer, 'Purposeful selection of variables in logistic regression', *Source Code Biol. Med.*, vol. 3, no. 1, p. 17, Dec. 2008, doi: 10.1186/1751-0473-3-17.

[143] R. Tibshirani, 'Regression Shrinkage and Selection via the Lasso', *J. R. Stat. Soc. Ser. B Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.
[144] S. K. Shevade and S. S. Keerthi, 'A simple and efficient algorithm for gene selection using sparse logistic regression', *Bioinformatics*, vol. 19, no. 17, pp. 2246–2253, Nov. 2003, doi: 10.1093/bioinformatics/btg308.
[145] W. A. Link and R. J. Barker, 'Model Weights and the Foundations of Multimodel Inference', *Ecology*, vol. 87, no. 10, pp. 2626–2635, 2006.
[146] F. Lewis, A. Butler, and L. Gilbert, 'A unified approach to model selection using the likelihood ratio test', *Methods Ecol. Evol.*, vol. 2, no. 2, pp. 155–162, 2011, doi: 10.1111/j.2041-210X.2010.00063.x.
[147] J. Wainer and G. Cawley, 'Nested cross-validation when selecting classifiers is overzealous for most practical applications', *Expert Syst. Appl.*, vol. 182, p. 115222, Nov. 2021, doi: 10.1016/j.eswa.2021.115222.
[148] T.-T. Wong and P.-Y. Yeh, 'Reliable Accuracy Estimates from k-Fold Cross Validation', *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1586–1594, Aug. 2020, doi: 10.1109/TKDE.2019.2912815.
[149] A. Golbraikh and A. Tropsha, 'Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection', *J. Comput. - Aided Mol. Des.*, vol. 16, no. 5–6, pp. 357–69, May 2002, doi: 10.1023/A:1020869118689.
[150] T. Saito and M. Rehmsmeier, 'The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets', *PLoS ONE*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.
[151] C. X. Ling, J. Huang, and H. Zhang, 'AUC: A Better Measure than Accuracy in Comparing Learning Algorithms', in *Advances in Artificial Intelligence*, Y. Xiang and B. Chaib-draa, Eds., Berlin, Heidelberg: Springer, 2003, pp. 329–341. doi: 10.1007/3-540-44886-1_25.
[152] M. Sokolova and G. Lapalme, 'A systematic analysis of performance measures for classification tasks', *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.
[153] Y. Jiao and P. Du, 'Performance measures in evaluating machine learning based bioinformatics predictors for classifications', *Quant. Biol.*, vol. 4, no. 4, pp. 320–330, Dec. 2016, doi: 10.1007/s40484-016-0081-2.
[154] M. Kubat, S. Matwin, and others, 'Addressing the curse of imbalanced training sets: one-sided selection', in *Icml*, Citeseer, 1997, p. 179.
[155] C. Esposito, G. A. Landrum, N. Schneider, N. Stiefl, and S. Riniker, 'GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning', *J. Chem. Inf. Model.*, vol. 61, no. 6, pp. 2623–2640, Jun. 2021, doi: 10.1021/acs.jcim.1c00160.
[156] M. Fitzgerald, B. R. Saville, and R. J. Lewis, 'Decision Curve Analysis', *JAMA*, vol. 313, no. 4, pp. 409–410, Jan. 2015, doi: 10.1001/jama.2015.37.
[157] P. G. Andersson, 'Approximate Confidence Intervals for a Binomial p—Once Again', *Stat. Sci.*, vol. 37, no. 4, pp. 598–606, Nov. 2022, doi: 10.1214/21-STS837.
[158] L. D. Brown, T. T. Cai, and A. DasGupta, 'Interval Estimation for a Binomial Proportion', *Stat. Sci.*, vol. 16, no. 2, pp. 101–117, 2001.
[159] B. Efron, 'Bootstrap Methods: Another Look at the Jackknife', *Ann. Stat.*, vol. 7, no. 1, pp. 1–26, 1979.

[160] B. Efron, 'Second Thoughts on the Bootstrap', *Stat. Sci.*, vol. 18, no. 2, pp. 135–140, 2003.
[161] E. LeDell, M. Petersen, and M. van der Laan, 'Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates', *Electron. J. Stat.*, vol. 9, no. 1, pp. 1583–1607, Jan. 2015, doi: 10.1214/15-EJS1035.
[162] T. Fryda, 'h2o package - RDocumentation'. Accessed: Sep. 20, 2024. [Online]. Available: https://www.rdocumentation.org/packages/h2o/versions/3.44.0.3
[163] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, 'Interpretability of machine learning-based prediction models in healthcare', *WIREs Data Min. Knowl. Discov.*, vol. 10, no. 5, p. e1379, 2020, doi: 10.1002/widm.1379.
[164] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, 'Machine Learning Interpretability: A Survey on Methods and Metrics', *Electronics*, vol. 8, no. 8, Art. no. 8, Aug. 2019, doi: 10.3390/electronics8080832.
[165] P. Wei, Z. Lu, and J. Song, 'Variable importance analysis: A comprehensive review', *Reliab. Eng. Syst. Saf.*, vol. 142, pp. 399–432, Oct. 2015, doi: 10.1016/j.ress.2015.05.018.
[166] A. Fisher, C. Rudin, and F. Dominici, 'All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously', *J. Mach. Learn. Res.*, vol. 20, no. 177, pp. 1–81, 2019.
[167] S. M. Lundberg and S.-I. Lee, 'A Unified Approach to Interpreting Model Predictions', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Jul. 01, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
[168] L. S. Shapley, '17. A Value for n-Person Games', in *Contributions to the Theory of Games (AM-28), Volume II*, H. W. Kuhn and A. W. Tucker, Eds., Princeton University Press, 1953, pp. 307–318. doi: 10.1515/9781400881970-018.
[169] J. P. Chao-Ying, L. L. Kuk, and G. M. Ingersoll, 'An introduction to logistic regression analysis and reporting', *J. Educ. Res.*, vol. 96, no. 1, p. 3, Oct. 2002.
[170] D. W. Apley and J. Zhu, 'Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models', *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 82, no. 4, pp. 1059–1086, Sep. 2020, doi: 10.1111/rssb.12377.
[171] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, 'Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation', *J. Comput. Graph. Stat.*, vol. 24, no. 1, pp. 44–65, Jan. 2015, doi: 10.1080/10618600.2014.907095.
[172] H. Shi, N. Yang, X. Yang, and H. Tang, 'Clarifying Relationship between PM2.5 Concentrations and Spatiotemporal Predictors Using Multi-Way Partial Dependence Plots', *Remote Sens.*, vol. 15, no. 2, Art. no. 2, Jan. 2023, doi: 10.3390/rs15020358.
[173] M. T. Ribeiro, S. Singh, and C. Guestrin, '"Why Should I Trust You?": Explaining the Predictions of Any Classifier', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[174] M. T. Ribeiro, S. Singh, and C. Guestrin, 'Anchors: High-Precision Model-Agnostic Explanations', *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, Art. no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11491.
[175] S. Wachter, B. Mittelstadt, and C. Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR', *Harv. J. Law Technol. Harv. JOLT*, vol. 31, p. 841, 2018 2017.
[176] M. T. Keane and B. Smyth, 'Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI)', in *Case-Based Reasoning Research and Development*, I. Watson and R. Weber, Eds., Cham: Springer International Publishing, 2020, pp. 163–178. doi: 10.1007/978-3-030-58342-2_11.
[177] J. Jung, H. Lee, H. Jung, and H. Kim, 'Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review', *Heliyon*, vol. 9, no. 5, p. e16110, May 2023, doi: 10.1016/j.heliyon.2023.e16110.
[178] H. Baniecki, W. Kretowicz, P. Piątyszek, J. Wiśniewski, and P. Biecek, 'dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python', *J. Mach. Learn. Res.*, vol. 22, no. 214, pp. 1–7, 2021.
[179] A. Liaw, 'randomForest package - RDocumentation'. Accessed: Sep. 20, 2024. [Online]. Available: https://www.rdocumentation.org/packages/randomForest/versions/4.7-1.1
[180] H. Nori, S. Jenkins, P. Koch, and R. Caruana, 'InterpretML: A Unified Framework for Machine Learning Interpretability', Sep. 19, 2019, *arXiv*: arXiv:1909.09223. doi: 10.48550/arXiv.1909.09223.
[181] J. Klaise, A. Van Looveren, G. Vacanti, and A. Coca, 'Alibi Explain: Algorithms for Explaining Machine Learning Models', *J. Mach. Learn. Res.*, vol. 22, no. 181, pp. 1–7, Jun. 2021.
[182] Norwegian Institute of Public Health, WHO Collaborating Centre for Drug Statistics Methodology, 'ATCDDD - ATC/DDD Index'. Accessed: Sep. 03, 2024. [Online]. Available: https://atcddd.fhi.no/atc_ddd_index/?code=L04&showdescription=no
[183] D. W. Frost, S. Vembu, J. Wang, K. Tu, Q. Morris, and H. B. Abrams, 'Using the Electronic Medical Record to Identify Patients at High Risk for Frequent Emergency Department Visits and High System Costs', *Am. J. Med.*, vol. 130, no. 5, p. 601.e17-601.e22, May 2017, doi: 10.1016/j.amjmed.2016.12.008.
[184] D. Vreš and M. Robnik-Šikonja, 'Preventing deception with explanation methods using focused sampling', *Data Min. Knowl. Discov.*, vol. 38, no. 5, pp. 3262–3307, Sep. 2024, doi: 10.1007/s10618-022-00900-w.

APPENDIX

CODE

The code for the training and analysis of all our models alongside their respective explanations is publicly available on GitHub under the following link:

https://github.com/SebastianVeuskens/High-Cost-patient-analysis

VARIABLE SELECTION

The logistic regression models were built on different predictor sets. All these three sets can be found in the file attached under the following link in GitHub:

https://github.com/SebastianVeuskens/High-Cost-patient-analysis/tree/master/results/Appendix

MODEL PERFORMANCE

The DCA for all models on the test set can be found in Figure 17. The resulting curves are qualitatively similar to the ones in Figure 5 on the validate data set. The ANN shows a different curve, but as before performs worse than all other models and seems to be rather unstable.

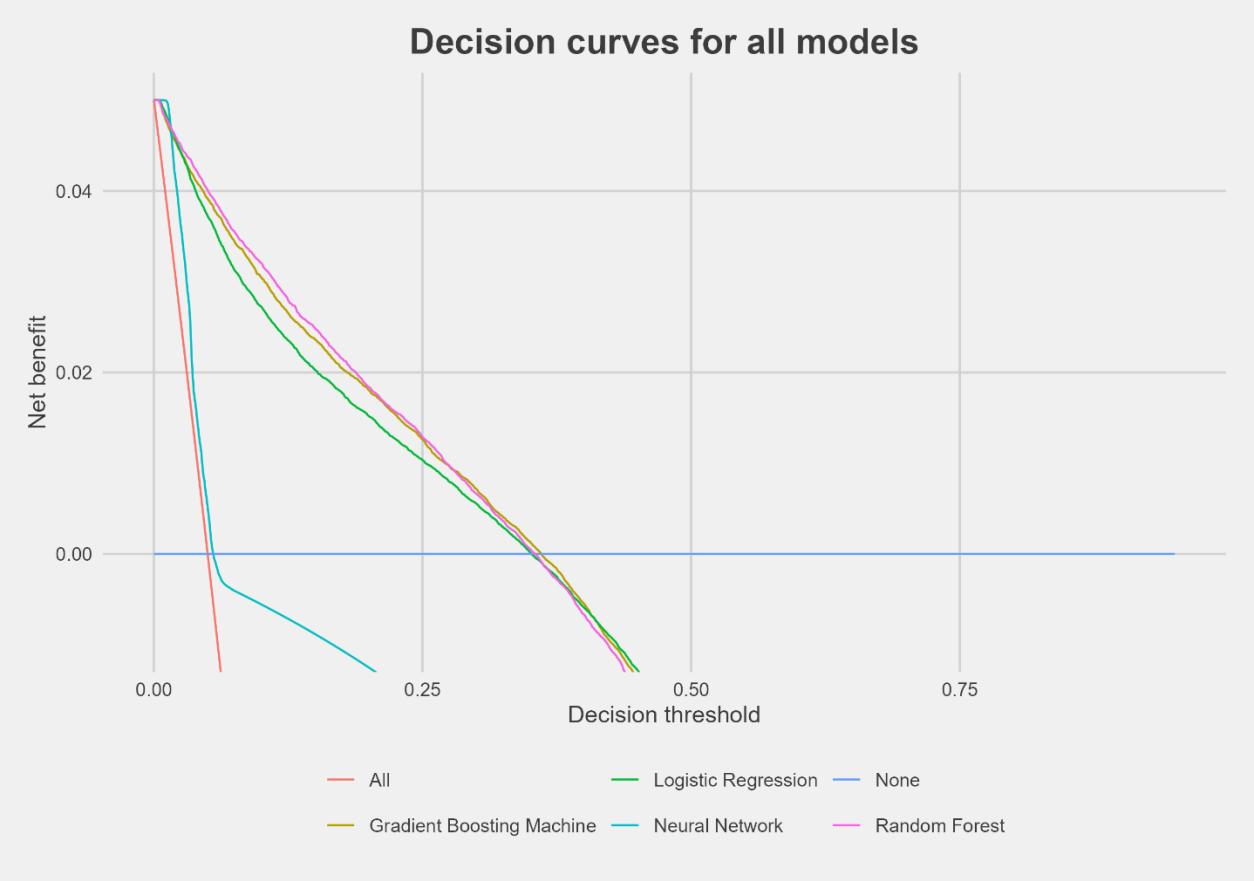

Figure 17: Decision curve for all models on the test data set. Results are similar to the decision curves on the validation data set. The random forest model performs best until a threshold of around 20%, after which it performs identically to the gradient boosting machine. Again, above a threshold of about 35%, none of the models has a positive net benefit. The neural network does perform very poorly, even for small probability thresholds.

EXPLANATION METHODS

The PDP explanations from the interpretML for the variable age is displayed in Figure 18. The documentation is not clear about the computation of the different lines that probably indicate different samples or risk levels. In addition to the PDP lines, the distribution information is depicted in a histogram, similarly to the H2O explanation.

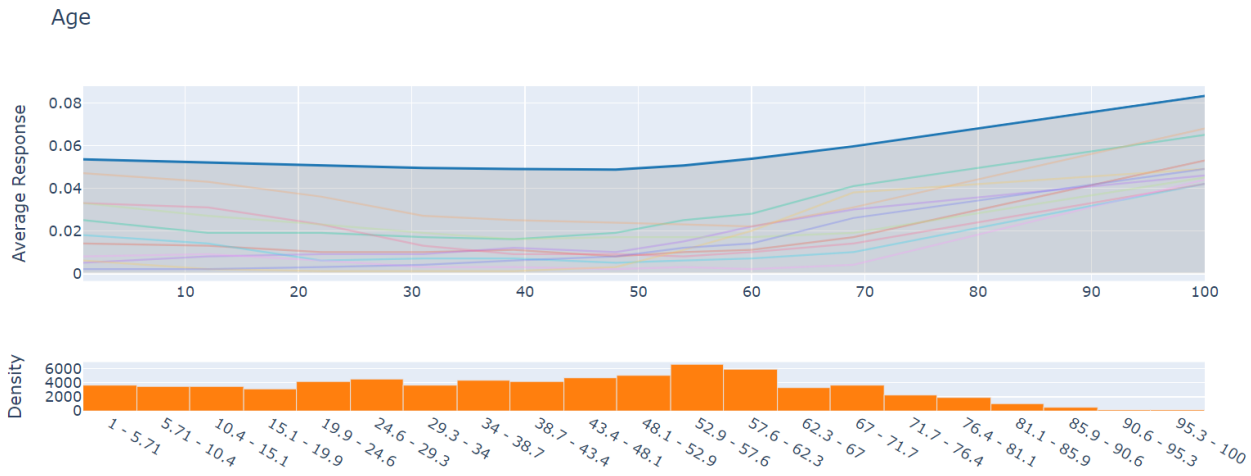

Figure 18: Exemplary, the PDP from the interpretML package is displayed. Several lines were shown, but the documentation was not clear about their meaning. In general, the lines showed a similar increase in risk as the PDP from the other models.

In Figure 19, the joint PDP of two variables age and total costs is displayed. Total costs was the more decisive variable with the most variation in the direction of its value, which was in accordance with the previous PDP plots for the single variables. The only visible interaction is that for an individual with an age above 60 years, the predicted risk reaches 9% with lower total costs than for individuals below that age.

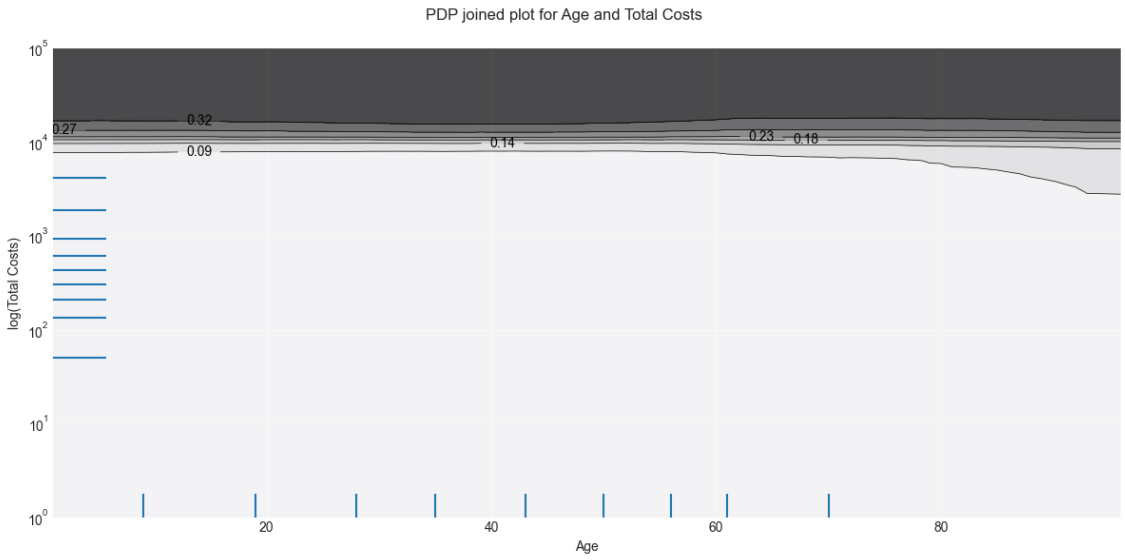

Figure 19: An exemplary PDP plot for two variables from the alibi package. Combined feature effects of the two variables Age and current Total costs can be inspected this way.

In Figure 20 the combination of both PDP and ALE plots yields a possibility to directly compare feature contributions and check for discrepancies. While the y-axis values for the PDP plot are well-defined by the definition of PDPs, this does not hold for ALE plots. ALE plots depend on an integral with a gradually expanding border. However, the starting point is not uniquely defined and is chosen rather arbitrarily in the DALEX package. That is why the PDP and ALE curves follow very similar trends but on different levels. Hence, one should inspect the difference between curve dynamics and not the absolute difference between the curves.

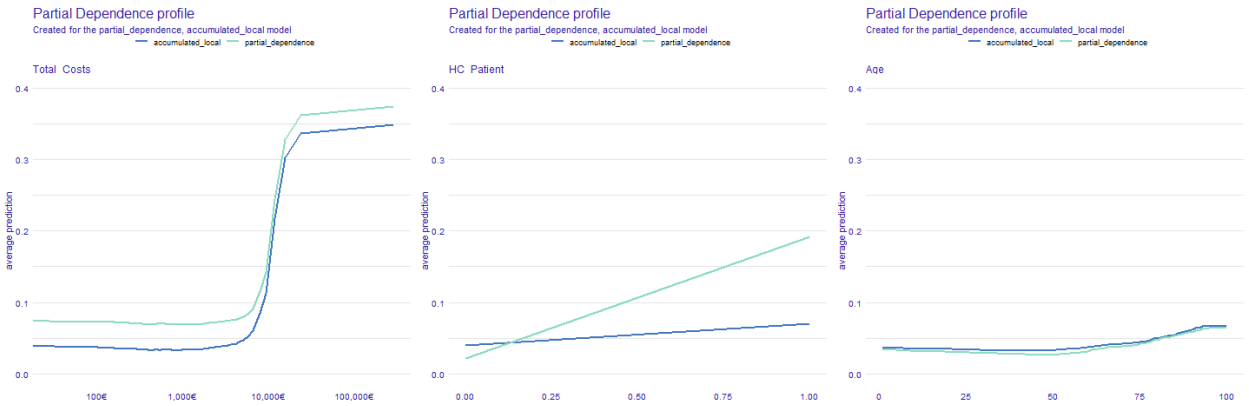

Figure 20: Combined PDP and ALE plot from the DALEX package. On the x-axis are the current total costs, HCP status, and age, and on the y-axis the corresponding risk that is assigned by the model. Plots cannot be adjusted and combined as in the alibi package. Results are similar to the alibi and H2O packages, with the HCP status in the ALE explanation showing a slower increase.

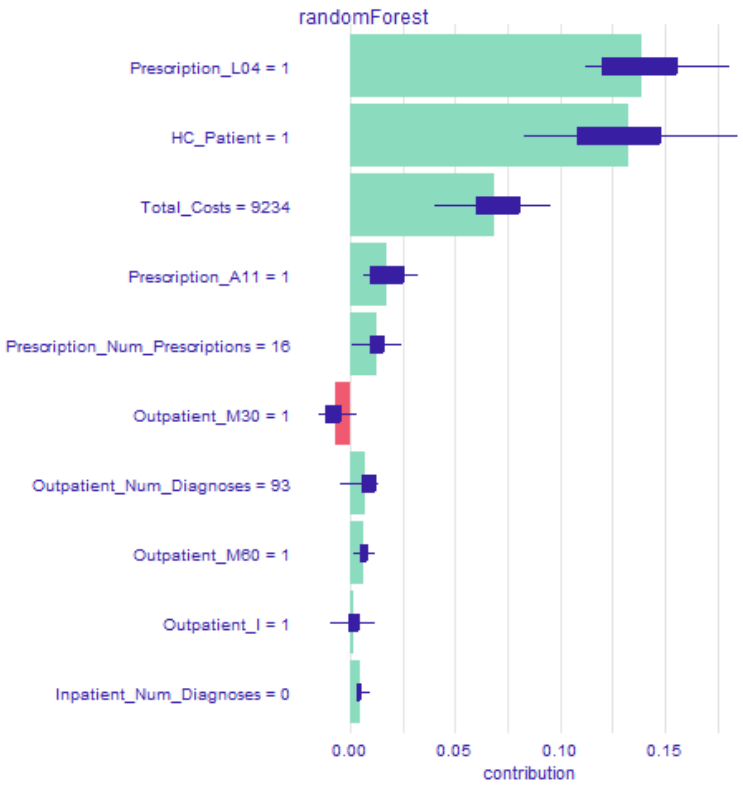

Figure 21: SHAP local explanation from the DALEX package in R. Contributions resemble the feature attributions from the equivalent explanations from the interpretML package.

Figure 21 contains the SHAP explanation from the DALEX package for the native R model. Explanations are very similar to the ones from the interpreML package in Figure 11. Again, the three most important that increase the risk by around 6-13% agree with the globally most important variables as identified in Figure 6.

The variable importance explanation in the DALEX package unfortunately does not offer a convenient possibility to limit the number of variables to be displayed. All variable names are displayed in a single plot, making it impossible to inspect the importances due to heavy overplotting with the presence of more than 600 variables. Such a dysfunctional explanation is displayed in Figure 22.

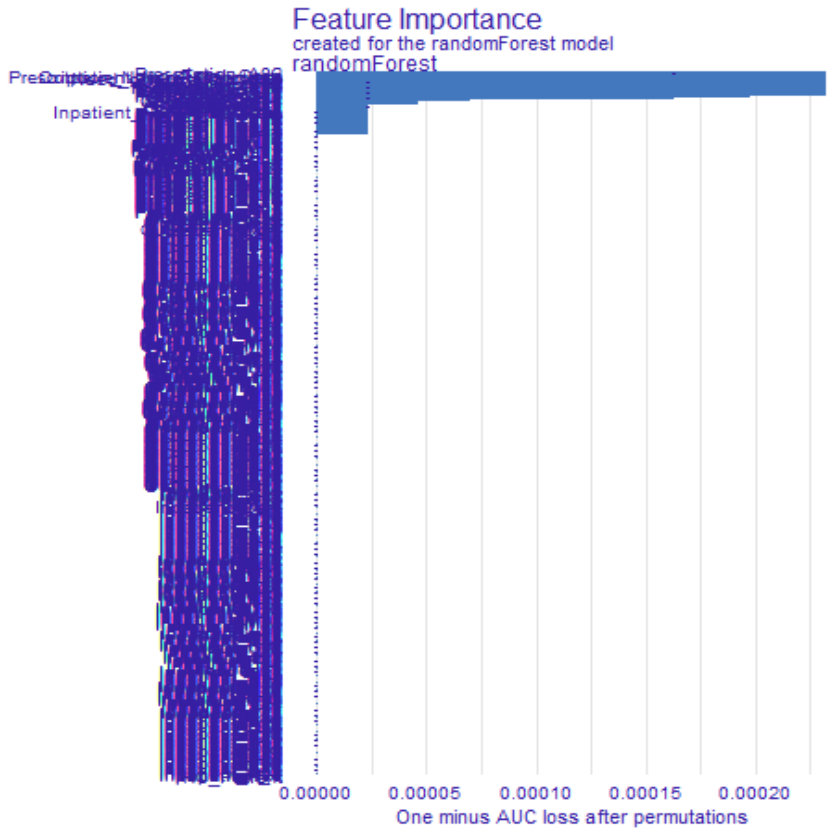

Figure 22: Variable importance explanation according to the DALEX package. The package did not offer the possibility to limit number of variables. More than 600 variables were hence included.

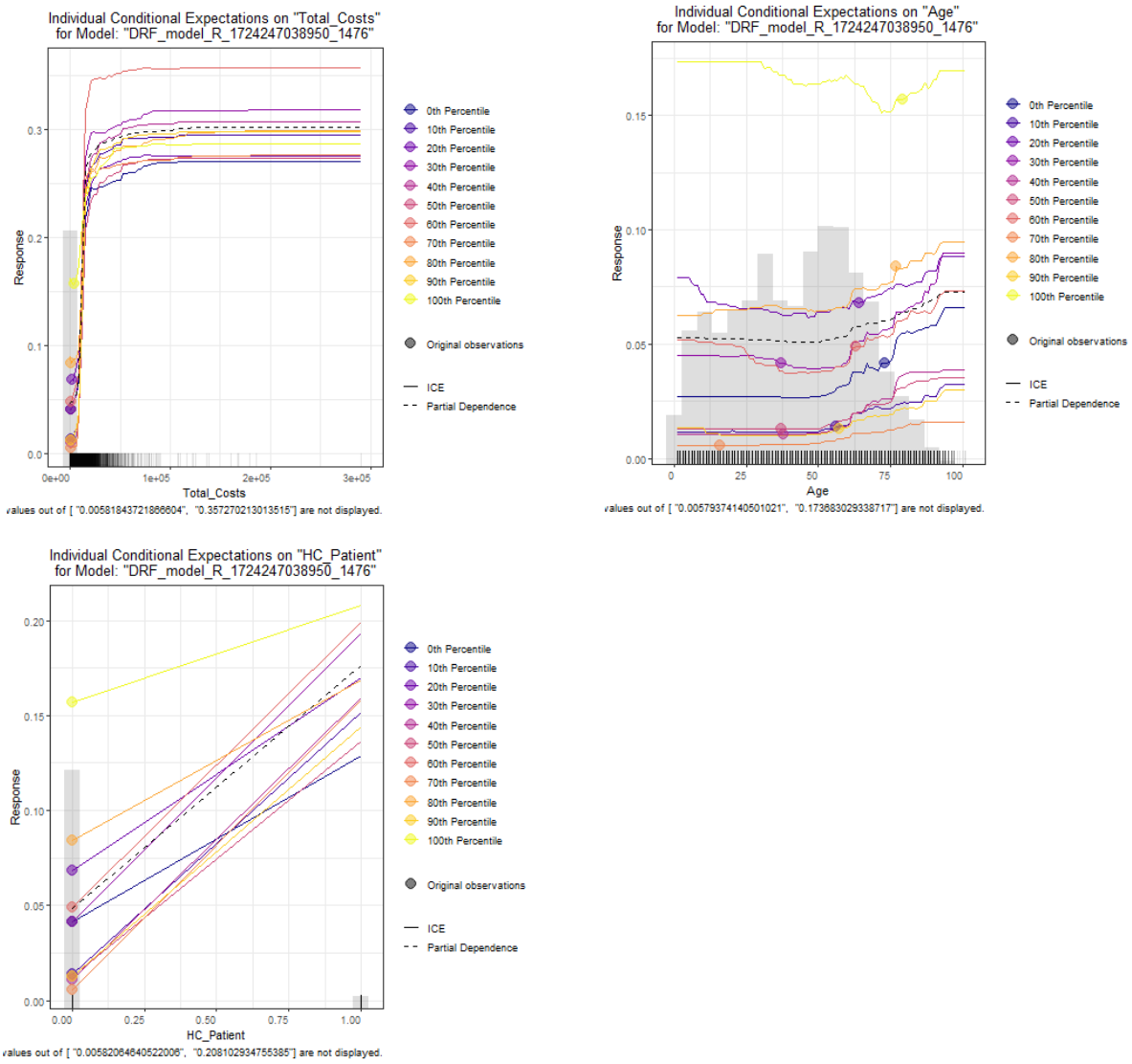

Figure 23: The ICE plots from the H20 package. As for the interpretML package, several lines are displayed, depicting every decile. The general trends were similar to the ICE plots from the alibi package. However, no strong perturbations were present for the age and current HCP variable. All variable seems to be rather stable and follow the trends from the PDP explanations.

Figure 23 the ICE explanations from the H2O package are displayed. As for the interpretML explanations, different lines are displayed. The line between ICE and PDP explanations for these two plots is blurry. No clear perturbations were visible. While this indicates stability for all three features, due to the limited amount of displayed curves in this plot, the explanations from the alibi plot were more likely to reveal instabilities and should be preferred.