Thesis for The Degree of Licentiate of Engineering

Interpretable machine learning models for predicting with missing values

Lena Stempfle

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2023

Interpretable machine learning models for predicting with missing values

Lena Stempfle

ISSN 1652-876X

Department of Computer Science and Engineering
Division of Data Science and AI
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2023.

Interpretable machine learning models for predicting with missing values

Lena Stempfle

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

Machine learning models are often used in situations where model inputs are missing either during training or at the time of prediction. If missing values are not handled appropriately, they can lead to increased bias or to models that are not applicable in practice without imputing the values of the unobserved variables. However, the imputation of missing values is often inadequate and difficult to interpret for complex imputation functions.

In this thesis, we focus on predictions in the presence of incomplete data at test time, using interpretable models that allow humans to understand the predictions. Interpretability is especially necessary when important decisions are at stake, such as in healthcare. First, we investigate, the situation where variables are missing in recurrent patterns and sample sizes are small per pattern. We propose SPSM that allows coefficient sharing between a main model and pattern submodels in order to make efficient use of data and to be independent on imputation. To enable interpretability, the model can be expressed as a short description introduced by sparsity. Then, we explore situations where missingness does not occur in patterns and suggest the sparse linear rule model MINTY that naturally trades off between interpretability and the goodness of fit while being sensitive to missing values at test time. To this end, we learn replacement variables, indicating which features in a rule can be alternatively used when the original feature was not measured, assuming some redundancy in the covariates.

Our results have shown that the proposed interpretable models can be used for prediction with missing values, without depending on imputation. We conclude that more work can be done in evaluating interpretable machine learning models in the context of missing values at test time.

Keywords

Machine learning, interpretable machine learning, missing values, healthcare

List of Publications

This thesis is based on the following appended publications:

[**Paper I**] Hákon Valur Dansson, Lena Stempfle, Hildur Egilsdóttir, Alexander Schliep, Erik Portelius, Kaj Blennow, Henrik Zetterberg and Fredrik D. Johansson for the Alzheimer's Disease Neuroimaging Initiative (ADNI), *Predicting progression and cognitive decline in amyloid-positive patients with Alzheimer's disease*
Alzheimer's Research and Therapy 13, 151 (2021).

[**Paper II**] Lena Stempfle, Ashkan Panahi, Fredrik D. Johansson, *Sharing pattern submodels for prediction with missing values*
Thirty-Seventh Conference on Artificial Intelligence (AAAI-23, To appear) (2023).

The following manuscript has been produced during my Ph.D. studies but has not been published yet.

[**Paper III**] Lena Stempfle, Fredrik D. Johansson, *Learning replacement variables in interpretable rule-based models*
Manuscript in preparation.

Acknowledgment

First of all, I would like to thank my supervisor Fredrik Johansson for his never-ending support and inspiring leadership. This thesis and our results would not exist without your guidance, knowledge, and attention to detail. Secondly, I would like to express my gratitude to my co-supervisor Devdatt Dubashi for his fruitful input and encouragement. Thank you to David Sands, my examiner, for all the interesting discussions during my follow-up meetings.

I could not have undertaken this journey without the help, and support of my colleagues: Alexander, Emilio, Filip, Daniel, Juan, Tobias, Markus, Arman, Firooz, Mehrdad, David, Hampus, Emil, Fazeleh, Tobias, Niklas, Linus, and Hanna. Thank you for creating a nice work environment. I am grateful to be able to work with and learn from the members of the HealthyAI group: Newton, Adam, Mena, and Anton. A special thanks to Lovisa and Christopher, whom I enjoy working with in the Ph.D. Council. Thank you to Kobljörn, Ashkan, Adel, Alexander, John, Dag, Rocío, and Simon for providing me with guidance, and opportunities to learn.

I have the pleasure of collaborating with the Traumabase Group in France. Thank you for the interesting discussions and for sharing your domain knowledge.

To my parents, Elisabeth and Karl, I want to thank you for your endless love and support throughout all my decisions. I am extremely grateful for my sister Johanna and my very close friends Verena, Anja, Kathi, and Theresa. Thank you for always supporting me and filling my life with joy. I could not have undertaken this journey without the support of my partner Amr. Thank you for always encouraging me and believing in me.

Our computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE) partially funded by the Swedish Research Council through grant agreement no. 2018-05973. Thank you to the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg foundation, for generously supporting my Ph.D. studies.

Lena Stempfle
Göteborg, March 2023

Contents

Part I

Introductory chapters

Chapter 1

Introduction

Making predictions on incomplete data is an ongoing research field with high relevance to practical applications. For example, medical data often suffer from incompleteness caused by tool unavailability and merging databases with heterogeneous data types [Groenwold, 2020; Madden et al., 2016]. Highly accurate predictions can significantly improve early disease detection and treatment procedures [Jayatilake and Ganegoda, 2021; Jiang et al., 2017].

Common approaches when predicting with missingness include either removing incomplete entries or imputing the missing value [Nakagawa, 2015; Rubin, 1976]. Both simple techniques such as mean imputation and more advanced ones, such as multiple imputation by chained equations (MICE), have been proposed to replace missing values [Raghunathan et al., 2001; Van Buuren, 2007]. Simpler imputation methods lead to bias in the estimate due to the variance between the imputed and the true value [Buck, 1960]. On the other hand, complicated imputation functions are not interpretable and it is difficult to trace how they recovered the missing value [White, Royston and Wood, 2011]. If values are also missing at test time, they must be imputed here as well. Moreover, if there is a distribution shift in the data between the training and the deployment, and the imputation was fit on training data it could affect the prediction [van Buuren, 2018]. Hence, imputation is powerful, but not always optimal under *test-time missingness* [Le Morvan et al., 2021].

For data-driven predictions to be effective and useful in clinical care, they must be designed to be understood by end users, i.e., medically trained personnel. The underlying challenge is that medical staff may not necessarily have a strong Machine Learning (ML) background, but should still be able to use the predictions for downstream tasks, such as making decisions on treatment or diagnosis [Ahmad, Eckert and Teredesai, 2018].

The research field that studies comprehensible ML models is called *Interpretable Machine Learning* (interpretable ML). In this thesis, "interpretable" methods and models are those that make the behavior and predictions of ML systems understandable to humans [Doshi-Velez and Kim, 2017; Molnar, 2020]. Interpretability is especially important for decisions with large implications and could help to improve transparency and strengthen trust in ML systems [Rudin,

2019]. Interpretable ML models include simpler algorithms like regression models and rule-based scoring systems as they are deemed more comprehensive than complex models like neural networks [Molnar, 2020; Tomsett et al., 2018; Ustun and Rudin, 2019].

In my research, I work on developing prediction models with high performance for problems when the input data is incomplete at the time of deployment. Moreover, I focus on creating simple and descriptive models that can be communicated and interpreted by domain experts.

This thesis compiles the results from three different papers. In Paper I (Section 3.1) we aimed to predict a clinical outcome using real-world data with high missingness. We ran experiments using traditional ML models showing that predictions heavily depend on imputation methods. We also found that missingness often occurs in patterns, which may vary between training and test time. Next, we investigated pattern missingness, and proposed *sharing pattern submodels* (`SPSM`), a linear prediction model, specialized for learning with patterns in variable missingness by sharing coefficients between patterns. The sharing is achieved by regularizing pattern specific submodels towards a main model (Section 3.2). In Paper III (Section 3.3), we investigated the case where there are no patterns in missingness. We studied interpretable rule-based models in the context of missingness and proposed to learn replacement variables during training for situations where some variables are correlated but not necessarily all observed at test time. We call the method developed in this work `MINTY`. The advantage of `MINTY` is that the methodology does not rely on imputation and it can be represented in a simple way.

Chapter 2

Background

In this chapter, I explain some of the background relevant to the contribution presented in this thesis. The focus of my research is two-fold: learning with missingness (Section 2.1), and interpretable ML (Section 2.2).

The mathematical notations for predicting with missingness at test time are presented in Section 2.1.1, followed by a background on missingness mechanisms (Section 2.1.2), which describes the relationship between the incomplete data and outcomes. A discussion on the commonly used approaches of predicting with missing values together with their limitations is mentioned in Section 2.1.3. The motivation for this thesis originates from the limitations of existing methods for predicting with missing values at test time, especially in high-stake settings such as healthcare. In such environments, it is particularly important to be able to interpret the model predictions and verify the models as needed. Therefore, I describe a class of models—interpretable ML models—and then present examples that are considered interpretable by current literature (Section 2.2.1).

2.1 Learning with missing values

In this section, I will outline the mathematical framework for prediction on incomplete input data, give an overview of the three different mechanisms of missingness and comment on missingness patterns. Then, common approaches for learning with missing values at test time and their limitations are presented. Finally, I briefly connect to the results of the first paper where we applied traditional ML to a data set with a high missingness ratio.

2.1.1 Prediction with test-time missingness

Supervised learning typically focuses on learning to predict an outcome $Y \in \mathcal{Y}$ from inputs $X \in \mathcal{X}$, where the pair (X, Y) are random variables with distribution P. Overall, the aim is to find a function $f \in F : \mathcal{X} \to \mathcal{Y}$, that minimizes the expected loss $\mathbb{E}[\ell(f(X), Y)]$, where $\ell : \mathcal{Y} \times \mathcal{X} \to R$ is the cost function [Vapnik, 1999]. Missing values are indicated as `NA` and we define an indicator matrix $\mathbf{M} \in \{0, 1\}^{n \times p}$, where, $M_{ij} = 1$ if X_{ij} is observed and $M_{ij} = 0$

otherwise. Next, let $\tilde{X} \in (\mathbb{R} \cup \{\texttt{NA}\})^{\texttt{d}}$ be the mixed observed-and-missing values of X according to $\mathbf{M}$.

Our goal is to predict Y *under missingness* $\mathbf{M}$ in X using functions $f : (\mathbb{R} \cup \{\texttt{NA}\})^{\texttt{d}} \to \mathbf{R}$. We aim to minimize the risk with respect to the expected loss

$$\min_f R(f), \text{ where } R(f) := \mathbb{E}_{\tilde{X},Y\sim P}[\ell(f(\tilde{X}) - Y)], \tag{2.1}$$

$R(f)$ is the expected loss for the future prediction. Loss functions may differ for classification (0-1 loss) or regression tasks (squared loss).

Let F be a collection of classifiers, where we want to find $f^* \in F$ such that the function is minimized by

$$f^* \in \operatorname*{argmin}_{f\in F} \mathbb{E}[\ell(f(\tilde{X}), Y)]. \tag{2.2}$$

The classifier f^* is called the Bayes classifier and it is the one with the best predictive performance for future data.

Since the distribution P is unknown in most problems, $R(f)$ is an unknown quantity as well. Instead, it is common to minimize its sampled analog

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} L(f(\tilde{x}^{(i)}), y^{(i)}). \tag{2.3}$$

$\hat{R}_n(f)$ is called the empirical risk and finding the function $\hat{f}$ that minimizes it, is known as empirical risk minimization (ERM), where

$$\hat{f} = \operatorname*{argmin}_{f\in F} \hat{R}_n(f). \tag{2.4}$$

The function f in 2.4 can be a composition of an imputation function and a prediction function, or it could be a function that makes use of the missingness mask $\mathbf{M}$.

2.1.2 Missingness mechanisms

The mechanisms generating missing values are various but are commonly classified into three main categories defined by [Rubin, 1976]: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). This categorization is based on the relationship between missing and observed values. For example, the chance that a subject will obtain measurements of an MRI image may depend on the availability of an MRI scanner in the clinic where the subject is located [Carpenter and Kenward, 2005]. We can partition the observations X into observed and missing: $X = (X^{obs}, X^{mis})$ [Sportisse, 2021].

The missingness mechanisms express the probability that a set of values are missing given the values taken by the observed and missing observations [Rubin, 1976]. It can be denoted by: $Pr(M \mid X^{obs}, X^{mis})$.

First, we define MCAR as the probability of missing an observation, independent of the observed or unobserved measurements and denote it with

$$Pr(m \mid X^{obs}, X^{mis}) = Pr(m).$$

An example that initially appears to be MCAR is taking a random sample from a population for a clinical trial, where each candidate has an equal chance of being included in the sample [van Buuren, 2018].

If given the observed data, the missingness mechanism does not depend on the unobserved data, it is possible to perform an analysis using only the observed data without information about the missingness mechanism, $Pr(M \mid X^{obs}, X^{mis})$ [Little and Rubin, 1986].

The MAR setting is described in mathematical terms by,

$$Pr(m \mid X^{obs}, X^{mis}) = Pr(m \mid X^{obs}).$$

In short, MAR conveys the idea that the missing value mechanism can be expressed exclusively in terms of observed observations. An example of MAR is when two measurements of the same variable are produced at the same time in a laboratory. If these two measurements differ by more than a specified threshold, a third measurement is taken. This third measurement is missing for all samples that did not deviate by the specified amount in the first two measurements [Carpenter and Kenward, 2005].

The third setting of missingness mechanisms is described as MCAR when neither MCAR nor MAR applies. In MCAR, even if we account for all available observational information, the reasons for the absence of variables still depend on the unseen observations themselves. A descriptive example includes when smoking status is not recorded for patients who are admitted as emergencies because these patients are more likely to have a worse surgical outcome.

We cannot verify only by observed data whether the missing observations are MCAR, NMAR or MAR. However, MCAR and MAR can be distinguished from each other [Little, 1988]. Note, in situations of MNAR it is very difficult to know the appropriate model for the missingness mechanism [Carpenter and Kenward, 2012]. We want to emphasize that the impact of the missing values on the analysis depends on the mechanism of the missing values, which is rarely known. However, in most situations, the true mechanism is probably MNAR.

Missingness Pattern Missing data can also be described by patterns of missing data, which characterizes the structure of observed and missing values in data sets [Little, 1993]. Missingness pattern should be distinguished from the missing mechanisms which explain *why* the values are missing and do not define where "the gaps" are. Pattern missingness emerges in data-generating processes where there are structural reasons for which variables are measured. Samples in data sets can be grouped by recurring patterns of observed and missing variables [Little, 1993]. Samples obtain only a fraction of the observed variables when, for example, different measurements are systematically made with different sensors or instruments in hospitals.

2.1.3 Approaches and their limitations to learning with missing values

In the following, common approaches when predicting with missing values during training and test time are presented and respective limitations are discussed. The limitations are mentioned in *italic* under each approach.

Complete case analysis The most common way of handling missing data is called *list-wise deletion*, where cases (or rows/lists) containing missing values are deleted and then a model is fit on data without any missing values [Nakagawa, 2015; Rubin, 1976].

Removing incomplete cases is not only using available data inefficiently, but it also diminishes valuable information.

Imputation methods Imputation attempts to replace the missing values, denoted by `NA` in the data set by aiming to reconstruct X from $\tilde{X}$ to create $\hat{X}$ and then use the reconstructed $\hat{X}$ to predict Y [Rubin, 1976]. This strategy is known as *impute-then-regress* estimation. The success of recovering the missing value depends on how well we can recover the information needed to make a good prediction based on the non-missing values that are still available [van Buuren, 2018].

The simplest imputation method is *zero-imputation*, where the missing values are replaced with a zero [Rubin, 1976]. Zero-imputation holds advantages for certain kinds of binary data. For instance, if a value is more likely to be present if a patient has a medical history of something (value= 1) than if they do not (value= 0) [Jamshidian and Mata, 2007]. Instead of adding a zero, we can perform *single imputation* where the missing data are filled by some means and the resulting completed data set $\hat{X}$ is used for prediction. *Mean imputation (MI)* is one such method in which the mean of the observed values is computed and then used to replace the missing values for that variable. Alternatively, [Buck, 1960] proposed imputing the missing values by predictions from regression models that are fitted using the mean and covariance matrix estimated by complete case analysis. The imputation functions can be learned by regression for continuous variables on complete observations or observations with less missingness. For categorical features, we use classification.

Zero and single imputation are rarely able to recover the true value of missing values since they do not reflect the uncertainty about the prediction of the missing values. In particular, the mean method can lead to highly biased estimates even when the data are MCAR [Jamshidian and Schott, 2007]. If the number of missing values in a variable is large and these values are replaced by the observed sample mean, the resulting variance estimate for that variable may be severely underestimated. When using deterministic functions to calculate an imputation value random, variations (i.e. an error term) around the regression slope are not considered. Imputed values are therefore often too precise and lead to an overestimation of the correlation between X and Y. Moreover, simple

imputation is based on a deterministic function that produces only one value, and that is the value taken by the missing value [Rubin, 1976]. Also note that it may be difficult to learn the missing values if all observations have some missing values in each row or if there are only a small number of complete rows ($\sim 5\%$). Especially for small data sets, the resulting reduction in statistical power is significant [Tibshirani, 1996].

A method that accounts for the variance and creates more than one sample of each missing is *multiple iterative imputation* [Rubin, 1976]. One of the most popular methods for imputation is MICE [Raghunathan et al., 2001; Van Buuren, 2007], in which a chained equation involves an iterative process that starts with a placeholder value for a missing variable and updates it in each iteration. To find the new values, a function is fit in each iteration based on imputations of other variables—this requires no full observations. To obtain multiple values, noise is added to the imputation functions to create a stochastic setting. Alternatively, the variables used for the regression can be randomly split so that slightly different functions are used each time [Rubin, 1976].

One downside of multiple imputation is if the imputation function is complex, $\hat{X}$ and the corresponding prediction might be hard to interpret [White, Royston and Wood, 2011].

Imputation methods may fail when data is not MAR—for example, missing values can not be reliably imputed. If the reason a variable is missing is connected to the outcome that we are trying to predict and that reason is not encoded in the other variables. If we impute variables based on the observed variables, the information that we need is not captured in those variables, then the imputed value might not be as good as we would like it to be to have an optimal prediction. Predicting with incomplete data can be especially beneficial, when missingness is dependent on unobserved factors that are related also to the prediction target, the fact that a variable is unmeasured can itself be predictive—so-called informative missingness [Marlin, 2008; Rubin, 1976].

Missingness indicators Alternatively, instead of replacing the missing values, the missingness mask $\mathbf{M}$ together with $\tilde{X}$ can be used to predict Y. There are models that are sensitive to the missingness mask $\mathbf{M}$. For instance, XGBoost supports missing values by default. For the tree algorithm, the branching directions for missing values are learned during training and stored as default settings for each node [Chen and Guestrin, 2016]. Missingness indicators can also be combined with any imputation method mentioned before. To generate $\hat{X}$, either a simple imputation method such as zero or mean, or even a more complicated method, like MICE is used, and the missingness mask $\mathbf{M}$ is then added to $\hat{X}$. The prediction function we try to learn is,

$$\hat{f} \in \underset{f \in F}{\operatorname{argmin}} \hat{\mathbb{E}}[\ell(f(\hat{X}, M), Y)]. \tag{2.5}$$

Although missingness indicators do not add any bias, they might hinder the interpretation of data sets with many features.

2.1.4 Connection to own research

In Paper I, we focused on applying traditional ML to the Alzheimer's Disease Neuroimaging Initiative database (ADNI) with high missingness. The missingness in the data set is most likely caused by patients dropping out of the study or different memory clinics measuring different features due to the availability of tools. To compensate for the missingness in the data, we used zero and mean imputation. We found that only a fraction of the variance could be explained by mostly highly correlated cognitive test scores. Often not all cognitive tests are measured at test time, but since they have some redundancy, they may not need to be in order to predict accurately. We come back to this finding in Paper III.

2.2 Interpretable Machine Learning

In this section, we first describe the concept and scope of interpretability in ML and then provide examples of how interpretability is implemented. Finally, we show how interpretablity was integrated into the work subject of this thesis.

There is no universally valid definition for *Interpretable Machine Learning* available. The general idea is, that interpretable ML refers to methods and models that make the behavior and predictions of ML systems understandable to humans [Doshi-Velez and Kim, 2017]. Another description by [Rudin, 2019] is:

> An interpretable ML model obeys a domain-specific set of constraints so that humans can better understand it.

The work by [Lipton, 2018] adds that interpretability is context- or domain-dependent, and thus cannot be uniformly defined. Despite this diversity of definitions and an ongoing debate about what interpretability entails, there is an agreement on what interpretability does not mean. Interpretability is NOT about understanding all bits and pieces of the model for all data points [Doshi-Velez and Kim, 2017]. It is about knowing enough for downstream tasks, especially for high-stake decisions and troubleshooting [Molnar, 2020]. Examples of such settings include criminal justice models, credit scoring, or healthcare applications [Rudin, 2019].

Situations, where interpretability is not necessarily required are, for example, ad servers, or in postal code sorting [Rudin, 2019]. In general, in any situation where no human intervention is needed, there are no consequences for unacceptable results, or when the problem is well-studied and validated in real-world applications [Doshi-Velez and Kim, 2017].

In the context of this thesis, we differentiate between (inherently) interpretable and post-explainability [Molnar, 2020; Rudin, 2019]. Post-explainability mainly uses black box models such as neuronal networks and then creates a

second (post hoc) model to explain the first black box model. Generated explanations are often not reliable, since the explanations approximate the behavior of the system and can therefore be misleading [Rudin, 2019]. We instead use models that are inherently interpretable, meaning they provide their own explanations, which are faithful to what the model actually computes. In this thesis, for simplicity, the terms *inherently interpretable* and *interpretable* are used interchangeably. An example of interpretable models is risk scores which are widely used for clinical decision-making and are commonly generated from logistic regression models using patient data [Struck et al., 2020]. Risk scores help clinicians quickly assess the risk for a patient by adding up the point associated with key predictors which are called rules. Figure 2.1 shows such key predictors (left) and their points (right). The total score is then translated to a risk of a clinical outcome, e.g. having a seizure.

Risk Factor						Points
Frequency > 2Hz [a]						1
Sporadic Epileptiform Discharges						1
LPD/BIPD/LRDA						1
Plus Features [b]						1
Prior Seizure						1
Brief Ictal Rhythmic Discharge						2
						Total Score
Total Score:	0	1	2	3	4	5	>6
Seizure Risk:	<5%	12%	27%	50%	73%	88%	>95%

Figure 2.1: Illustration of variables used to calculate the 2HELPS2B risk score [Struck et al., 2020]. The total score is calculated by summing over the points on the right column, associated with a particular risk of experiencing seizure. The rules explain medical details such as the brief independent periodic discharge (BIPD), continuous EEG (cEEG), generalized periodic discharg (GPD), lateralized periodic discharge (LPD), lateralized rhythmic delta activity (LRDA). **Plus features** are defined as superimposed rhythmic, fast, or sharp activity for LRDA, BIPDs, LPDs, or GPDs

As the example in Figure 2.1 shows, interpretable models output humanly understandable summaries of their calculations that help us understand how predictions are produced. As a result, that leads to better transparency, and humans may increase their trust in ML systems [Rudin, 2019]. A remaining challenge of interpretability is to design models that are simple enough to be understood by users while maintaining high predictive power.

2.2.1 Examples of interpretablity

Next, I present what is widely understood to be interpretable by current literature. In general, simpler model classes like regression, sparse/small decision trees, rule models, and scoring systems are seen to be more interpretable than complex models like neural networks [Adadi and Berrada, 2018; Molnar, 2020].

Linear Models A linear regression model, e.g., $y = a_0 + a_1x_1 + ... + a_dx_d + \epsilon$ aims to predict the target y as a weighted sum of its j features. a_js define the coefficients or learned feature weights and a_0 denotes the intercept. The ϵ is the error resulting from the difference between the prediction and the actual outcome. The model is linear if the association between features x and target y is modeled linearly [Hastie, Tibshirani and Friedman, 2009]. The linearity of the learned relationship makes the interpretation easy. The interpretation of weight in the linear regression model depends on the type of the corresponding feature. For a numerical feature, an increase of feature x_j by one unit increases the prediction for output y by a_j units when all other feature values remain fixed. Changing a categorical feature x_j from the reference category to the other category increases the prediction for y by a_j when all other features remain fixed [Hastie, Tibshirani and Friedman, 2009]. However, not all relationships between features are linear, interactions between characteristics or a nonlinear relationship between features and the target value can be compensated by adding interaction terms or regression splines [Molnar, 2020].

Sparsity In linear models, sparsity is integrated by enforcing feature selection and regularization of the selected feature weights which results in only a subset of the input characteristics [Tibshirani, 1996]. Sparsity is introduced by utilizing a penalty on a loss function, such as *LASSO* [Tibshirani, 1996]. *LASSO* stands for "least absolute shrinkage and selection operator" and regularizing a linear models results in the following cost function,

$$\frac{1}{n}\sum_{i=1}^{n}(a^\top x^{(i)} - y^{(i)})^2 + \alpha\|a\|_1, \tag{2.6}$$

$a > 0$ is a vector of coefficients. The last term is called the ℓ_1-norm or *LASSO* penalty [Tibshirani, 1996] and leads to the penalization of large weights and α is a hyperparameter used to set the intensity of this penalty term. Mathematically, we can describe sparsity as being the zeros in the coefficient vector by $\|a\|_1 = \sum_{j=1}^{d}|a_j|$ [Tibshirani, 1996]. The *LASSO* penalty forces the less important coefficients down to zero, and removes these variables from the model, hence the sparsity and only the best features are selected. By reducing the number of parameters to be analyzed, sparse models may be easier to understand, resulting in higher descriptive accuracy.

Decision trees Decision trees (DTs) are useful when relationships between features and outcomes are non-linear or when features interact with each other [Tibshirani, 1996]. Trees exist for classification and regression. Tree-structured classifiers split the data several times according to certain threshold values in the covariates. Through splitting, different subsets of the data set are created, indicating which instance belongs to one subset. The final subsets are named leaf nodes and the average outcome of the training data in a node is used to predict the outcome in each leaf node. A tree can be interpreted by following from the root node, through the edges and the subsets, to the leaf node with the predicted outcome. All edges are connected by "AND" [Molnar,

2020]. In addition, feature importance in DTs is computed and interpreted as a share of the overall model importance. Note that, DTs are human interpretable, as long as they are short. Gradient boosting machines [Chen and Guestrin, 2016], and random forests are typically inscrutable to humans due to their complexity and, often big size.

Decision Rules A decision rule is a simple IF-THEN statement consisting of a condition and a prediction [Molnar, 2020]. Rules are learned from data by using e.g. RuleFit [Friedman and Popescu, 2008], and Bayesian modeling [Chen and Guestrin, 2016]. Note that any tree can be transformed into a rule set, while the opposite is not possible [Margot and Luta, 2021]. An advantage of the IF-THEN structure is that, if the conditions are expressed in understandable terms, they are semantically similar to natural language and there are not too many rules, they are easy for humans to interpret [Margot and Luta, 2021].

Scoring systems/Risk scores Traditionally, scoring systems have been designed using manual feature elimination on logistic regression models, with rounded coefficients. Today, we can learn scoring systems, for instance, as sparse nonlinear models with integer coefficients for risk assessment (i.e., risk scores) from data [Ustun and Rudin, 2019]. Risk scores represent the majority of scoring systems that are currently used in medicine, such as sleep apnea screening [Ustun et al., 2016], or Alzheimer's diagnosis [Souillard-Mandar et al., 2016] and criminal justice (recidivism prediction) [Rudin, Wang and Coker, 2020].

Generalized Linear Models Linear models can be extended to Generalized Linear Models (GLMs) when the outcome, given the features, has a non-Gaussian distribution, or the features interact and the relationship between the features and the outcome is nonlinear [Nelder and Wedderburn, 1972].

In GLM, the weighted sum of the features (aX^T) is linked to the mean value of the assumed distribution using the link function g, depending on the type of outcome. We denote E_Y defined by the probability distribution from the exponential family [McCullagh, 2019] such that,

$$g(E_Y(y|x)) = a_0 + a_1x_1 ... + a_dx_d. \tag{2.7}$$

The assumed distribution together with the link function, determines how the estimated feature weights are interpreted. The interpretability decreases when the feature dimensions are too large, which may be beyond the comprehension ability of humans [Wei et al., 2019].

2.2.2 Connection to own research

To create interpretable models, I made use of the methods mentioned in the previous section in my research. In Paper II, we incorporated sparsity into `SPSM` and applied it to tabular data to create short descriptions of pattern specialization that help construct a simple and expressive model. We enforced

sparsity to limit the number of differences between submodels. If the number of nonzero coefficients is sufficiently small, a practitioner can interpret the variables corresponding to those coefficients to be meaningfully related to the outcome, and can also interpret the magnitude and direction of the coefficients.

In Paper III, we proposed an interpretable rule-based model `MINTY` which is used when covariates are redundant and can be used as substitutes for each other when one or more of them are missing at test time. Literal disjunctions are used, and *LASSO* regularization leads to a sparse solution that naturally allows for a trade-off between interpretability and predictive power.

Chapter 3

Summary of Included Papers

3.1 Paper I: Predicting progression and cognitive decline in amyloid-positive patients with Alzheimer's disease

In Paper I, we studied the problem of predicting disease progression and cognitive decline of potential Alzheimer's disease (AD) patients with established amyloid-β (Aβ) pathology using ADNI.

In AD, Aβ peptides aggregate in the brain and form amyloid concentrations in the cerebrospinal fluid, which is an important pathological characteristic of the disease. Although, amyloid concentrations in CSF may also be observed in cognitively unimpaired elderly participants. In this work, we thus aim to explain the variance in disease progression in patients with Aβ pathology.

From ADNI, we selected a cohort of n=2293 participants, of whom n=749 were Aβ positive to study heterogeneity in disease progression for patients with $A\beta$ pathology. Clinical variables including demographics, genetic markers, and neuropsychological data were used in the analysis. We trained statistical and supervised ML models to predict cognitive decline within two and four years from baseline, and a model for predicting worsened diagnosis status after two years. To compensate for cohorts that were too small and unbalanced, we designed our experiments to use weighted groups. The use of non-$A\beta$-positive subjects in the derivation of progression prediction models decreases variance by increasing the sample size of cohorts with a small number of individuals.

When predicting the change in cognitive test scores in Aβ-positive subjects during the 2-year follow-up period, we achieved an R^2 value of 0.388, whereas the best model for predicting negative changes in diagnosis produced a weighted F_1 value of 0.791. Conforming to expectations, Aβ-positive subjects declined faster on average than those without Aβ pathology, but the specific CSF Aβ

was not predictive of progression rate. The best model achieved an R^2 score of 0.325, when predicting cognitive score change four years after baseline and it was found that fitting models to the extended cohort improved performance. We identified that in order to achieve the most accurate predictions, the models combine clinial variables measured at baseline. In this regard, the results of the cognitive tests at baseline proved to be the strongest predictors, explaining most of the variance in all models.

We also realized that the data suffered from a significant amount of missingness in the covariates but also the outcome variable. The missingness in the outcome variable is partly explained by subjects leaving the study before follow-up. The reason for subjects who ended their participation in the study is not known but may be connected to disease development [Larson et al., 2004]. This phenomenon can bias the trend of the Aβ positive subjects decreasing their MMSE score (Figure 3.1). As a result, if more people with lower cognitive function were included, the slope of the graph would be slightly steeper, resulting in an even lower average MMSE score. To compensate for the missingness we imputed missing values in the data set during pre-processing using zero and mean imputation.

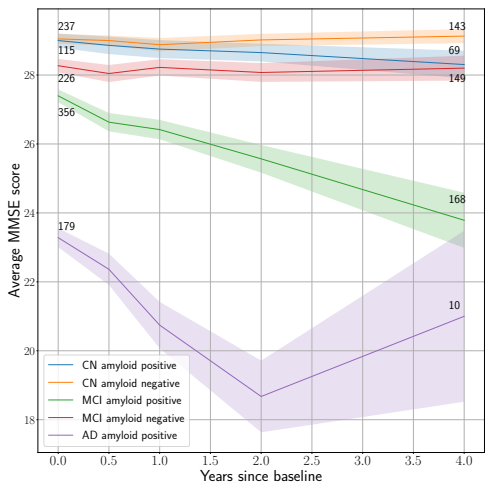

Figure 3.1: Graph showing the MMSE score development for CN, MCI and AD subjects split by Aβ-status. The shaded areas represent 95% confidence intervals for the mean values. The number of subjects decreases over time, hence the growing uncertainty bands.

In summary, we first found that most predictive covariates are highly correlated, mainly due to the fact that cognitive tests capture similar information. Second, we concluded that a high missingness rate affects the prediction when using imputation. This led to further investigations in methods that do not require imputation and utilities the redundancy in predictive features, which build the motivation for Paper III.

3.2 Paper II: Sharing pattern submodels for prediction with missing values

In the presence of missingness patterns at the time of deployment and small sample sizes per pattern, we proposed an approach called *Sharing Pattern Submodels* (`SPSM`). `SPSM` produces accurate predictions based on incomplete input data and has a short description that allows for better interpretability by domain experts.

In this work, we focused on settings where inputs are partially missing both during training and at the time of prediction [Little and Rubin, 1986]. If missing values are not handled appropriately, they can result in increased bias or in models that are inapplicable in deployment without imputing the values of unobserved variables [Le Morvan et al., 2020; Liu, Zachariah and Stoica, 2020]. To avoid the limitations of imputation, it can be beneficial to let models make predictions based on both the partially observed covariates and on missingness indicators [Groenwold, 2020; Jones, 1996].

In Figure 3.2, we show an example of observing patients from three different clinics, each systematically taking slightly different measurements [Stempfle and Johansson, 2022].

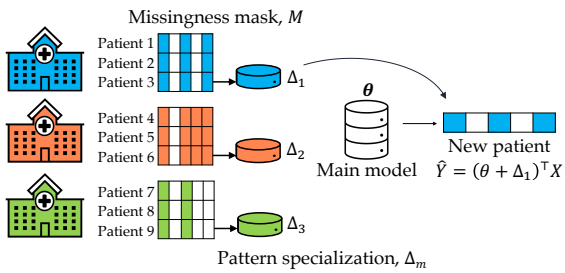

Figure 3.2: Coefficient sharing between a main model θ and pattern submodels for three clinics with different patterns in missing values. Without specialization, Δ_m, an average prediction shared by clinics with different patterns may not lead to an optimal solution for any of them. Conversely, fitting separate models for each clinic does not use all of the available data efficiently and leads to high variance.

For this setting, *Pattern submodels* have been proposed, fitting a separate model to samples from each pattern [Marshall et al., 2002; Mercaldo and Blume, 2020]. This solution does not rely on imputation and can improve interpretability compared to black-box methods. Although, it might led to high variance, especially when the number of unique patterns is large and the number of samples for a individual pattern is small. Notably, pattern submodels do not account for the fact that the prediction task is shared between each pattern. However, in situations such as in Figure 3.2, using one shared model for all clinics may also be suboptimal if clinics take different measurements, or treat patients differently (high bias).

In this work, we proposed SPSM, in which submodels for different missingness patterns share coefficients across patterns, while allowing limited specialization. Sharing is accomplished by regularizing submodels towards a main model and solving the resulting coupled optimization problem in 3.2. Moreover, sharing coefficients encourages efficient use of information across submodels leading to a beneficial tradeoff between predictive power and variance in the case where similar submodels are desired.

Fitting SPSM Let $\theta \in \mathbb{R}^d$ represent the *main model* coefficients used in prediction under all missingness patterns and defined as subset of coefficients corresponding to variables observed under m. To emphasize, $\theta_{\neg m}$ depends only on m in selecting a subset of θ—the coefficients are shared across patterns. Similarly, define $\Delta_{\neg m} \in \mathbb{R}^{d_m}$ to be *pattern-specific specialization* of these coefficients to m. In contrast to $\theta_{\neg m}$, the values of $\Delta_{\neg m}$ are unique to each pattern m. Note, a model f_m depends only on the observed components of X. In regression tasks, we learn **sharing** pattern submodels on the form

$$f_m(x) \coloneqq (\theta_{\neg m} + \Delta_{\neg m})^\top x_{\neg m}, \;\; \forall \;\; m \in \mathcal{M} \tag{3.1}$$

by solving the problem,

$$\begin{aligned} \underset{\theta, \{\Delta_{\neg m}\}}{\text{minimize}} \;\; & \frac{1}{n} \sum_{i=1}^{n} \left((\theta_{\neg m^{(i)}} + \Delta_{\neg m^{(i)}})^\top x^{(i)}_{\neg m^{(i)}} - y^{(i)} \right)^2 \\ & + \frac{\gamma}{n} \|\theta\| + \sum_{m \in \mathcal{M}} \frac{\lambda_m}{n_m} \|\Delta_{\neg m}\|_1 \;, \end{aligned} \tag{3.2}$$

n_m is the number of samples of pattern m and $\lambda_m \geq 0$ and $\gamma \geq 0$ are regularization parameters. The learning problem is taken from our paper in [Stempfle and Johansson, 2022]. While we show a regression task in 3.2, SPSM can also be learned for classification tasks by replacing the square loss with the logistic loss. The penalty for $\|\theta\|$, can either be the ℓ_1 or ℓ_2 norm to tradeoff bias and variance in the main model. A high value for λ_m regularizes the specialization of model coefficients to missingness pattern m in a way that high λ_m encourages smaller $\|\Delta_m\|_1$ which leads to greater coefficient sharing. Our Δ is regularized by a ℓ_1-norm since we aim for a sparse solution with a small number of non-zero specialization coefficients.

We evaluated the SPSM model[1] on simulated and real-world data. Experimental results show that SPSM performs comparably or slightly better than baselines across all data sets without relying on imputation (Figure 3.3). The results demonstrate that the proposed method never performs worse than non-sharing pattern submodels as these do not make efficient use of the available data. Our theoretical analysis shows that, in the linear-Gaussian setting, our method also recovers the sparsity of the true process. In the large-sample limit, this may not be beneficial for variance reduction, but sparsity contributes to interpretability.

[1] Code to reproduce experiments is available at https://github.com/Healthy-AI/spsm.

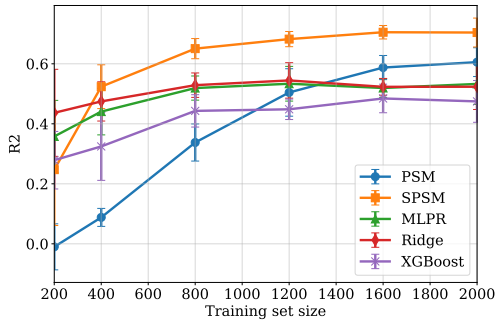

Figure 3.3: Performance on simulated data Setting A (higher is better). Error bars show standard deviation over 5 random data splits. The full data set has $n = 2000$ samples.

Table 3.1: Example of Δ_4 for regression using `SPSM` using ADNI. `SPSM` takes $\gamma = 10$ and $\lambda = 13$ as parameters for a single seed. There are 10 missingness patterns in total, while 4 of them have non-zero coefficients for Δ and pattern-specific intercept. Coefficients are for standardized variables.

Missing features in pattern 4: ABETA, TAU, and PTAU at baseline (bl)

Feature	Δ_4	θ	$\theta + \Delta_4$
Age	-0.140	0.121	-0.019
FDG-PET	-0.090	-0.039	-0.129
Whole Brain (bl)	0.000	-0.045	-0.044
Fusiform	0.016	0.021	0.037
ICV	0.001	0.093	0.094
Intercept	-0.10	0.18	

Why is `SPSM` Interpretable? We improve interpretablity in `SPSM` by allowing domain experts to compare pattern specializations and reason about how similar submodels are, and how they are affected by missing values (Table 3.1). We claim that a set of submodels is easier to interpret if specializations contain fewer non-zero coefficients, $\Delta_{\neg m}$ is sparse. We achieved that by adding regularization in `SPSM` which leads to a sparse model that only comprises a subset of the input features affecting predictions, and reducing the effective complexity of the model [Cowan, 2010; Miller, 1956].

Note, `SPSM` is limited to learning linear models, but it is not limited to learning from linear systems. For future work, we aim to identify other models in interpretable ML that could benefit from this type of sharing.

3.3 Paper III: Learning replacement variables in interpretable rule-based models

In Paper III, we further explored the model class of interpretable ML models when predicting with test-time missingness. For this purpose, we made use of rule-based models, which are among the most interpretable models, due to their use of natural language and simple representation. However, with incomplete input data during testing time, the predictions of standard rule models are undefined or ambiguous.

In this work, we learned accurate rule-based models with missing values at both training and testing times, based on the notion of replacement variables. We defined replacement variables in settings where there is redundancy in the covariates, meaning two features have similar associations with the outcome, but they do not both need to be observed to predict accurately. Rather, the redundant variables A and B could be used as replacements to each other when one of the two variables is missing: "If A is not available, use B", or "If B is not available, use A".

Below we show a case of rules that demonstrates how replacement variables can be utilized in the context of linear rule models with binary variables. We focused on situations, where if at least one variable in each rule is observed and active, the prediction is the same independent of whether other covariates are missing from the rule.

$$\begin{aligned}\textbf{prediction} = &\ \text{Coefficient}_1(\text{Variable}_1 \text{ OR } \text{Variable}_2) \\ &+ \text{Coefficient}_2(\text{Variable}_3 \text{ OR } \text{Variable}_4) \\ &+ \text{Coefficient}_3(\text{Variable}_5)\end{aligned}$$

Imputation or missingness indicators are not required, when using replacement variables when making predictions.

We proposed a method called `MINTY` which learns replacement variables in the form of disjunctions when one or more is missing. This results in a sparse linear rule-based model that naturally allows a trade-off between interpretability and predictive performance, while being sensitive to missing values at test time.

In preliminary experiments, we compared `MINTY` with baselines in predictive performance and interpretability. In future work, we will further develop the methodology by exploring the models' limits. This includes the model's behavior when no substitute is found for rare but highly predictive covariates, or the model is very uncertain in its prediction. One alternative way to determine replacement variables is to incorporate concept associations by domain knowledge into the constraints [Lage and Doshi-Velez, 2020]. We will also investigate to what extent the model performance can be improved by some kind of imputation without affecting the degree of interpretability. Moreover, we will conduct more experiments with simulated and real-world data sets to generalize our results more.

Chapter 4

Concluding Remarks and Future Directions

In this thesis, we studied methods that enable prediction with missing values at test time. In the context of healthcare, values can be missing at test time for several reasons such as incomplete data entry, and tool unavailability [Groenwold, 2020; Madden et al., 2016]. To provide clinical relevance and support medical staff in their decision-making, the proposed methods need to be interpretable [Doshi-Velez and Kim, 2017; Rudin, 2019].

In Paper I (Section 3.1), we used traditional machine learning methods to predict the disease progression of AD in amyloid-β subjects using real-world data from ADNI. We identified that the data suffers from high missingness and predictions might be affected by imputation of missing values. We also found that a fraction of the variance can be explained by only cognitive test scores which are among the most predictive features but not always all of them are measured at test time. The cognitive test scores contain similar information and are therefore somewhat redundant to each other.

In Paper II (Section 3.2), we focused on missingness occurring in patterns, where we learned prediction models from data with pattern missingness with different sets of incomplete predictors at test time. We targeted the case where small, interpretable models are desired, but sample sizes per missingness pattern are small. Therefore, in Paper II we proposed `SPSM` as a method that utilizes coefficient sharing over patterns. `SPSM` does not rely on imputation and provides a small description of shared coefficients introduced by regularizing the sharing parameter λ to achieve a sparse solution. The model representation allows domain experts to reason about how similar submodels are, and how they are affected by missing values.

Next, we looked at situations where missingness does not occur in patterns, but we still aim to provide an interpretable model. Based on the findings in Paper I, where we learned that some main predictors are strongly correlated but not always available for all subjects at test time, we hypothesize that there is a way to exploit this property. In Paper III (Section 3.3), we proposed the method `MINTY` which learns replacement variables in disjunctive combinations

within rules.

There are some limitations in this work that call for future work. We limit `SPSM` to learn linear models, although in theory it is not limited to learning from linear systems. In future work, we plan to investigate other classes of models developed with interpretability that take advantage from sharing information.

So far, in Paper III we have shown how `MINTY` can be learned, but to generalize the results more experiments including several other baselines and data sets are needed. In addition, the methodology `MINTY` itself can be further developed which I discuss in more detail in the next section.

Second, although we strive for interpretability in the models we have developed and have conducted initial assessments, user-studies can help to evaluate our models in order to make more general statements about their usefulness in an applied context.

4.1 Future Directions

The methodology in `MINTY` can be extended by integrating various ways of determining replacement variables. For instance, a replacement can be found if variables are based on similar measurements or come from a pre-defined concept. An example of a concept in our situation is cognitive risk scores, including different questionnaires and test procedure to evaluate similar cognitive abilities. To this end, one idea is to involve humans in an interactive approach to define the concepts and then use the concepts to make final predictions [Koh et al., 2020; Lage and Doshi-Velez, 2020]. A potential model improvement in `MINTY` tackles the situations where there is uncertainty in the prediction. An existing solution is to learn predictors that choose to defer the decision to a downstream expert [Mozannar and Sontag, 2020]. More investigation is needed on how *learning to defer* is useful in the presence of missing values at test time.

Another research direction is the evaluation of interpretable ML models when predicting with missing values. The evaluation of interpretable ML models can be extended beyond traditional prediction metrics, such as accuracy or area under the curve (AUC) to quantitatively and qualitatively decide whether the model helps stakeholders achieve their downstream tasks [Futoma et al., 2020; Lipton, 2018]. Evaluation frameworks including application-grounded, human-grounded, and functionally grounded perspectives have been proposed [Doshi-Velez and Kim, 2017]. One approach to include end users in the model development and evaluation process is called *human-in-the-loop* [Monarch, 2021]. In an upcoming project, we will perform a human-in-the-loop evaluation on `MINTY` within an international collaboration with domain experts from the TraumaBase Group[1]. Traumabase Group is a database originated in France which collects data from Trauma patients. We hypothesize that, especially for the replacement variables, clinicians have a mental model that they use in clinical practice. We plan to investigate how users interpret different classes of machine learning models with missing values at test time and identify their model representation preferences.

[1] https://www.traumabase.eu/en_US

Bibliography

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052 (cit. on p. 11)

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 559–560. https://doi.org/10.1145/3233547.3233667 (cit. on p. 3)

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society: Series B (Methodological)*, *22*(2), 302–306 (cit. on pp. 3, 8).

Carpenter, J., & Kenward, M. (2012). *Multiple imputation and its application*. John Wiley & Sons. (Cit. on p. 7).

Carpenter, J., & Kenward, M. (2005). Missing value jargon. https://www.lshtm.ac.uk/media/38311. (Cit. on pp. 6, 7)

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (cit. on pp. 9, 13).

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science*, *19*(1), 51–57 (cit. on p. 19).

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning* (cit. on pp. 3, 10, 21, 22).

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The annals of applied statistics*, 916–954 (cit. on p. 13).

Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2020). The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, *2*(9), e489–e492 (cit. on p. 22).

Groenwold, R. (2020). Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic and prognostic research*, 8. https://doi.org/10.1186/s41512-020-00077-0 (cit. on pp. 3, 17, 21)

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer. (Cit. on p. 12).

Jamshidian, M., & Mata, M. (2007). Advances in analysis of mean and covariance structure when data are incomplete. In *Handbook of latent variable and related models* (pp. 21–44). Elsevier. (Cit. on p. 8).

Jamshidian, M., & Schott, J. R. (2007). Testing equality of covariance matrices when data are incomplete. *Computational statistics & data analysis*, *51*(9), 4227–4239 (cit. on p. 8).

Jayatilake, S. M. D. A. C., & Ganegoda, G. U. (2021). Involvement of machine learning tools in healthcare decision making. *Journal of healthcare engineering*, *2021* (cit. on p. 3).

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and vascular neurology*, *2*(4) (cit. on p. 3).

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, *91*(433), 222–230 (cit. on p. 17).

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. https://doi.org/10.48550/ARXIV.2007.04612 (cit. on p. 22)

Lage, I., & Doshi-Velez, F. (2020). Learning interpretable concept-based models with human feedback. *arXiv preprint arXiv:2012.02898* (cit. on pp. 20, 22).

Larson, E. B., Shadlen, M.-F., Wang, L., McCormick, W. C., Bowen, J. D., Teri, L., & Kukull, W. A. (2004). Survival after initial diagnosis of alzheimer disease. *Annals of internal medicine*, *140*(7), 501–509 (cit. on p. 16).

Le Morvan, M., Josse, J., Moreau, T., Scornet, E., & Varoquaux, G. (2020). NeuMiss networks: differentiable programming for supervised learning with missing values. *arXiv:2007.01627* (cit. on p. 17).

Le Morvan, M., Josse, J., Scornet, E., & Varoquaux, G. (2021). What's a good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, *34* (cit. on p. 3).

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57 (cit. on pp. 10, 22).

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, *83*(404), 1198–1202 (cit. on p. 7).

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*(421), 125–134 (cit. on p. 7).

Little, R. J., & Rubin, D. B. (1986). *Statistical analysis with missing data*. John Wiley & Sons, Inc. (Cit. on pp. 7, 17).

Liu, X., Zachariah, D., & Stoica, P. (2020). Robust prediction when features are missing. *IEEE Signal Processing Letters*, *27*, 720–724 (cit. on p. 17).

Madden, J. M., Lakoma, M. D., Rusinak, D., Lu, C. Y., & Soumerai, S. B. (2016). Missing clinical and behavioral health data in a large electronic health record (ehr) system. *Journal of the American Medical Informatics Association*, *23*(6), 1143–1149 (cit. on pp. 3, 21).

Margot, V., & Luta, G. (2021). A new method to compare the interpretability of rule-based algorithms. *AI*, *2*(4), 621–635 (cit. on p. 13).

Marlin, B. M. (2008). *Missing data problems in machine learning* (Doctoral dissertation). University of Toronto. (Cit. on p. 9).

Marshall, G., Warner, B., MaWhinney, S., & Hammermeister, K. (2002). Prospective prediction in the presence of missing data. *Statistics in medicine*, *21*(4), 561–570 (cit. on p. 17).

McCullagh, P. (2019). *Generalized linear models*. Routledge. (Cit. on p. 13).

Mercaldo, S. F., & Blume, J. D. (2020). Missing data and prediction: the pattern submodel. *Biostatistics*, *21*(2), 236–252 (cit. on p. 17).

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81 (cit. on p. 19).

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com. (Cit. on pp. 3, 4, 10–13).

Monarch, R. M. (2021). *Human-in-the-loop machine learning: Active learning and annotation for human-centered ai*. Simon; Schuster. (Cit. on p. 22).

Mozannar, H., & Sontag, D. A. (2020). Consistent estimators for learning to defer to an expert. *International Conference on Machine Learning* (cit. on p. 22).

Nakagawa, S. (2015). Missing data: Mechanisms, methods and messages. *Ecological statistics: Contemporary theory and application*, 81–105 (cit. on pp. 3, 8).

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384 (cit. on p. 13).

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., Solenberger, P., et al. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, *27*(1), 85–96 (cit. on pp. 3, 9).

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592 (cit. on pp. 3, 6, 8, 9).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206–215 (cit. on pp. 3, 10, 11, 21).

Rudin, C., Wang, C., & Coker, B. (2020). The Age of Secrecy and Unfairness in Recidivism Prediction [https://hdsr.mitpress.mit.edu/pub/7z10o269]. *Harvard Data Science Review*, *2*(1) (cit. on p. 13).

Souillard-Mandar, W., Davis, R., Rudin, C., Au, R., Libon, D. J., Swenson, R., Price, C. C., Lamar, M., & Penney, D. L. (2016). Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test. *Machine learning*, *102*, 393–441 (cit. on p. 13).

Sportisse, A. (2021). *Handling heterogeneous and mnar missing data in statistical learning frameworks: Imputation based on low-rank models, online linear regression with sgd, and model-based clustering* (Doctoral dissertation). Sorbonne université. (Cit. on p. 6).

Stempfle, L., & Johansson, F. (2022). Sharing pattern submodels for prediction with missing values. *arXiv preprint arXiv:2206.11161* (cit. on pp. 17, 18).

Struck, A. F., Tabaeizadeh, M., Schmitt, S. E., Ruiz, A. R., Swisher, C. B., Subramaniam, T., Hernandez, C., Kaleem, S., Haider, H. A., Cissé, A. F., et al. (2020). Assessment of the validity of the 2helps2b score for inpatient seizure risk prediction. *JAMA neurology*, *77*(4), 500–507 (cit. on p. 11).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. Retrieved February 2, 2023, from http://www.jstor.org/stable/2346178 (cit. on pp. 9, 12)

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. https://doi.org/10.48550/ARXIV.1806.07552. (Cit. on p. 4)

Ustun, B., & Rudin, C. (2019). Learning optimized risk scores. *J. Mach. Learn. Res.*, *20*(150), 1–75 (cit. on pp. 4, 13).

Ustun, B., Westover, M. B., Rudin, C., & Bianchi, M. T. (2016). Clinical prediction models for sleep apnea: The importance of medical history over symptoms. *Journal of Clinical Sleep Medicine*, *12*(2), 161–168 (cit. on p. 13).

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, *16*(3), 219–242 (cit. on pp. 3, 9).

van Buuren, S. (2018). *Flexible imputation of missing data, second edition (2nd ed.)* hapman; Hall. https://doi.org/9780429492259. (Cit. on pp. 3, 7, 8)

Vapnik, V. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999. https://doi.org/10.1109/72.788640 (cit. on p. 5)

Wei, D., Dash, S., Gao, T., & Gunluk, O. (2019). Generalized linear rule models. *International Conference on Machine Learning*, 6687–6696 (cit. on p. 13).

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine*, *30*(4), 377–399 (cit. on pp. 3, 9).

Part II

Appended Papers

Paper I

Predicting progression and cognitive decline in amyloid-positive patients with Alzheimer's disease

Hákon Valur Dansson, Lena Stempfle, Hildur Egilsdóttir, Alexander Schliep, Erik Portelius, Kaj Blennow, Henrik Zetterberg and Fredrik D. Johansson for the Alzheimer's Disease Neuroimaging Initiative (ADNI)

Alzheimer's Research and Therapy 13, 151 (2021)

https://doi.org/10.1186/s13195-021-00886-5

RESEARCH

Predicting progression & cognitive decline in amyloid-positive patients with Alzheimer's disease

Hákon Valur Dansson[1], Lena Stempfle[1*], Hildur Egilsdóttir[1], Alexander Schliep[1], Erik Portelius[2,3], Kaj Blennow[2,3], Henrik Zetterberg[2,3,4,5], Fredrik D. Johansson[1]
and for the Alzheimer's Disease Neuroimaging Initiative (ADNI)[6]

Abstract

Background: In Alzheimer's disease, amyloid-β (Aβ) peptides aggregate in the brain forming CSF amyloid levels, which are a key pathological hallmark of the disease. However, CSF amyloid levels may also be present in cognitively unimpaired elderly individuals. Therefore, it is of great value to explain the variance in disease progression among patients with Aβ pathology.

Methods: A cohort of n=2293 participants, of whom n=749 were Aβ positive, was selected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database to study heterogeneity in disease progression for individuals with Aβ pathology. The analysis used baseline clinical variables including demographics, genetic markers and neuropsychological data to predict how the cognitive ability and AD diagnosis of subjects progressed using statistical models and machine learning. Due to the relatively low prevalence of Aβ pathology, models fit only to Aβ-positive subjects were compared to models fit to an extended cohort including subjects without established Aβ pathology, adjusting for covariate differences between the cohorts.

Results: Aβ pathology status was determined based on the $A\beta_{42}/A\beta_{40}$ ratio. The best predictive model of change in cognitive test scores for Aβ-positive subjects at the two-year follow-up achieved an R^2 score of 0.388 while the best model predicting adverse changes in diagnosis achieved a weighted F_1 score of 0.791. Aβ-positive subjects declined faster on average than those without Aβ pathology, but the specific level of CSF Aβ was not predictive of progression rate. When predicting cognitive score change four years after baseline, the best model achieved an R^2 score of 0.325 and it was found that fitting models to the extended cohort improved performance. Moreover, using all clinical variables outperformed the best model based only on a suite of cognitive test scores which achieved an R^2 score of 0.228.

Conclusion: Our analysis shows that CSF levels of Aβ are not strong predictors of the rate of cognitive decline in Aβ-positive subjects when adjusting for other variables. Baseline assessments of cognitive function accounts for the majority of variance explained in the prediction of two-year decline but is insufficient for achieving optimal results in longer-term predictions. Predicting changes both in cognitive test scores and in diagnosis provides multiple perspectives of the progression of potential AD subjects.

Keywords: Alzheimer's disease; Amyloid-beta; Progression; Prediction; Machine learning

Background

About 50 million people worldwide suffer from some form of dementia and 60–80% of all cases have Alzheimer's disease (AD) [1]. Patients who already suffer from mild cognitive impairment (MCI) are at higher risk of developing AD [2, 3]. Studies have shown that the conversion rate from MCI to AD is between 10% to 15% per year with 80% of these MCI patients progressing to AD after approximately six years of follow-up [4, 5]. Identifying those who are at greatest risk of progression to AD is a central problem.

A key pathological hallmark, required for an AD diagnosis, is the accumulation of Aβ peptides into plaques, located extracellularly, and in intracellular tangles, consisting of phosphorylated tau (p-tau) protein [6, 7]. The precipitation of Aβ in the brain appears decades before the patient shows symptoms during the so-called preclinical stage of AD [8, 9, 10]. Lower levels

*Correspondence: stempfle@chalmers.se
[1]Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Gothenburg, SE
Full list of author information is available at the end of the article

of the aggregation-prone peptide Aβ_{42} (or Aβ_{42}/Aβ_{40} ratio) together with increased levels of p-tau and total-tau (t-tau) are a core cerebrospinal fluid (CSF) signature of AD [6]. However, despite strong evidence for association between these biomarkers and AD, individuals with significantly lowered Aβ ratio do not necessarily exhibit any cognitive impairment [11, 12]. Therefore, Aβ pathology alone is not sufficient as a predictor of disease progression [13, 14].

Although AD predictors and pathological hallmarks have been researched for many years, today there is still no drug available that cures AD or drastically changes its course. New drug candidates that have potential disease-modifying effects [15] are currently in development and recently, the FDA approved aducanumab for the treatment of patients with AD under the Accelerated Approval process. The FDA concluded that the benefits of Aduhelm for patients with Alzheimer's disease outweigh the risks of the therapy.

If a successful treatment is developed, it is of utmost importance that a prognostic tool is available to identify the patients most likely to decline towards AD, to implement preventive treatments and interventions. This leaves the challenge of predicting how patients with Aβ pathology will progress, explaining the variation in cognitive function of such subjects. As a result, a recent focus area in applid statistical and computational research is predicting a change in diagnosis for patients progressing from cognitively normal (CN) to MCI and from MCI to AD [16, 17, 5, 18, 19].

Most predictive models of neurodegenerative diseases are based on recent advances in machine learning models by obtaining data sets with measurements of cognition and neuropathology from large cohorts [20, 21, 16, 22]. In this context, classification methods such as random forest[13, 23, 24, 21], and logistic regression (LR) [25, 26, 27, 21] have been used to predict whether individuals will decline or remain stable in their diagnosis.

Classification approaches are dependent on the availability of clinical labels and do not focus on capturing patient-specific disease trajectories. To overcome this limitation, disease progression has also been studied with respect to continuous measures of the disease severity [28, 29]. Previous works employed an elastic net linear regression model [30, 31] to predict changes in cognitive test scores to capture the patient's cognitive ability over time. The most common targets when predicting cognitive decline are the Mini Mental Status Test (MMSE) [32] and the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) [33] scores [34, 35, 36].

In prediction modelling, the question arises as to which of the considered input variables are particularly predictive. In addition to predictors of AD diagnosis, relationships between CSF biomarkers (CSF p-tau/Aβ42 ratio and several other biomarkers) and prediction of cognitive decline have been explored [37, 38, 26, 39]. However, even though Aβ-positivity has been identified as a strong predictor of disease status, little is known about what determines the disease progression of Aβ-positive subjects [27, 40].

This study aims to predict the future severity of dementia for subjects with established presence of low Aβ levels in CSF. We propose and demonstrate several predictive models of disease progression for three different cohorts , studying two primary aspects of progression: cognitive decline and change in diagnosis. For the former, we predict the change in the MMSE cognitive test score both two and four years after baseline (the first visit of each patient). For the latter, we use a classification approach to predict whether subjects will have a worse diagnosis two years after baseline. Both tasks are addressed using linear and non-linear prediction models, the parameters of which were selected using ML methodology.

A predictive approach could be used to assist healthcare professionals in evaluating and prioritizing patients for treatment. Given that our model builds on only a small set of biomarkers and demographic data, available for most patients, the methodology is widely applicable.

Methods

Subjects and ADNI

The data used in this study were obtained from the publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`http://adni.loni.usc.edu`). ADNI collects clinical data, neuroimaging data, genetic data, biological markers, and clinical and neuropsychological assessments from participants at different sites in the USA and Canada to study MCI and AD. Since its inception in 2003, several releases have been made; the cohorts used in this work were assembled from ADNI 1,2,3 and GO.

The compiled data set used in this project includes 2293 subjects that were further filtered by eligibility criteria, such as availability of diagnostic labels and on Aβ ratios. Among the 2293 subjects, there were 749 Aβ-positive positive subjects. The exclusion flowchart (see Figure 2) describes how many subjects are assigned to perform a prediction task for the all-subjects and Aβ-positive cohorts. For baseline statistics of the processed Aβ cohort, see Table 1. Tables 6 and 7 in the supplementary material show the characteristics of All Subjects and Aβ-positive subjects for the three prediction tasks.

Determination of amyloid-positive status

The presence of Aβ plaques can be detected at a preclinical stage years before the patient shows any symptoms [9, 10]. While Aβ plaques (and tau-levels) may not be the root cause of disease development [41], their abnormal deposits in the brain uniquely define AD [6, 7]. However, even among subjects with Aβ pathology, there is significant variability in symptoms, such as cognitive function. For this reason, our work is focused on predicting progression for subjects with lowered CSF levels of Aβ indicating plaque formations in the brain. Subjects were evaluated for Aβ pathology based on their Aβ_{42}/Aβ_{40} ratio (hereinafter simply Aβ ratio) as measured in CSF at baseline. The full cohort was split into three groups: those who had a baseline Aβ ratio lower than 0.13 (Aβ-positive), those who had a higher ratio (Aβ-negative), and those with unknown status. The threshold used in this work is slightly higher than in some other works. For example, a threshold of 0.0975 proposed in [42] for the diagnosis of AD. However, as diagnosing AD was not our primary concern, we let the distribution of ratios themselves decide the threshold, see Figure 1, rather than tying it to a particular prediction target.

Progression outcomes

We studied the progression of Aβ-positive subjects with respect to two principal outcomes: change in cognitive function relative to baseline and change in clinical dementia diagnosis.

Cognitive function was assessed using the widely adopted MMSE scale [32]. The MMSE score is commonly used as a target variable in clinical trials analysing the treatment effects of drugs aimed at enhancing cognition for AD patients and in ML for predicting change in patient's cognitive ability [43, 44]. The MMSE comprises a series of 20 individual tests covering 11 domains for a total of 30 items. The test covers the person's orientation to time and place, recall ability, short-term memory, and arithmetic ability.The MMSE score takes values on a scale from 0 to 30 where a lower score represents worse cognitive function [45]. The specific targets of prediction were the changes in MMSE score measured at follow-up visits two years after baseline and four years, relative to baseline.

Changes in dementia diagnoses were determined by comparing the disease status (CN/MCI/AD) recorded in ADNI at follow-up visits to the status at baseline. For the corresponding prediction task, a binary variable was created, indicating whether or not a subject's diagnosis had worsened in two years. Due to the low number of available subjects after four years, changes were evaluated only two years after baseline. The models were used to predict whether Aβ-positive subjects would transfer from the CN group at baseline to either MCI or AD, or convert from MCI at baseline to AD at a follow-up visit after two years.

Potential predictors

The covariates available at baseline (enrolment in ADNI) contain analyzed biofluid samples from CSF, plasma and serum including different biochemical-markers such as proteins, hormones and lipids. Additionally, features extracted from brain imaging biomarkers, such as positron emission tomography (PET) scan and MRI were included. Demographic data such as age and gender were also considered.

The CSF samples include measurements of both Aβ_{42} and Aβ_{40}, which are Aβ peptides ending at positions 42 and 40 respectively. Their ratio in CSF measurements has been proposed to better reflect brain amyloid production than their individual measures [46, 47]. Therefore, the ratio Aβ ratio in CSF is calculated and added as a new feature for all subjects with both measurements available.

Predictive models were built on two different sets of features. The first set of features (all features) was preselected following [48] and expanded to include key features from the ADNI TadPole competition [49] in addition to a few features that were available for over 90% of the ADNI cohort. This resulted in a set of 37 features including biomarkers tau, ptau and Aβ_{42} in CSF, the PET measures of AV45 and FDG, seven different size measurements of brain regions and 15 different cognitive tests. Moreover, the FDG-PET data has been measured by a research group of UC Berkeley. The MPRAGEs (Magnetization Prepared Rapid Acquisition Gradient Echo) for each subject is segmented and parcellated with Freesurfer (version 5.3.0) to define a variety of regions of interest in each subject's native space. The second feature set (cognitive tests only) consists only of the 15 cognitive tests also present in the all feature set. A full list and descriptions of the features are given in Table 1 and Table 2 in the supplementary material. When building models for predicting the change in MMSE score, the MMSE measures at baseline were not included in the predictions since the target output itself was calculated from the change in its baseline value.

Statistical analyses

We used machine learning (ML) methods to train predictive models of cognitive decline within two (task A1) and four years (task A2) from baseline, as well as a model for predicting worsened diagnosis status (task B) after two years. The full procedure, described further below, involved cohort sample splitting and weighting, model selection and fitting, and evaluation.

Derivation & evaluation cohorts

Due to the small number of Aβ-positive subjects available for each task (500/230/398 for tasks A1/A2/B, respectively), see the exclusion flowchart in Figure 2), we compared training predictive models from only Aβ-positive subjects to two ways of training using All Subjects, irrespective of Aβ status. All models were evaluated only on Aβ-positive subjects, as they are the primary target of this work.

The first derivation setting (Aβ Only) used only Aβ-positive subjects for model derivation. This ensures that model parameters are unbiased with respect to the Aβ-positive cohort but may suffer from high variance due to a small sample size. The second setting (All Subjects) combined Aβ-positive and Aβ-negative subjects and those without Aβ measurements into one derivation set. Consequently, the derivation sample size has been increased substantially, at the cost of introducing bias into the sample, while the evaluation cohort remains the same.

In the third setting (All Subjects, Weighted), we applied sample weighting to the All Subjects cohort to mimic a larger sample of Aβ-positive subjects. Each subject i was assigned a weight $w_i > 0$ based on the probability that their individual Aβ-ratio r_i would be observed for an average hypothetical Aβ-positive subject, as estimated using a two-component Gaussian mixture model (GMM) [50] fit to observed ratios.

We let the latent state $C \in \{0, 1\}$ of a GMM, fit to the Aβ-ratios of the All Subjects cohort, represent Aβ-positivity. The weight w_i was computed as

$$w_i = \hat{p}(R = r_i \mid C = 1)/\hat{p}(R = r_i) \ .$$

This is the ratio of the estimated probability to observe the Aβ-ratio r_i for an average Aβ-positive subject and the overall probability of observing that ratio. This procedure is described further in the supplementary material. The weighting scheme assigns a higher weight to subjects with Aβ-ratio more like that of Aβ-positive subjects and lower to those with higher or unobserved ratios. The weight was clamped between 0.2 and 1.0 so that subjects with unmeasured or very high ratios were given small but non-negligible influence and so that decidedly $A\beta$-positive subjects would be given the weight 1.0. Each prediction model was then fit to the weighted full sample but evaluated only on held-out (unweighted) Aβ-positive subjects.

Prediction models & learning objectives

First, we predicted the change in MMSE score relative to baseline at the two-year follow-up (task A1) and four-year follow-up (task A2) visits using two separate regressions. Second, prediction of change in diagnosis after two years (task B) was treated as a binary classification problem (worse diagnosis/not worse diagnosis). For each task, we considered both linear and non-linear estimators.

The first model type used for the MMSE prediction was ordinary least squares linear regression. Similarly, for the classification task, a logistic regression model was used. The second model type used both for regression and classification was tree-based gradient boosting [51]. Gradient boosting is an ensemble method where many weak learners, in our case decision and regression trees, are combined in an iterative fashion to create a strong one. The trees are fit to the negative gradient of the loss function (mean squared error and logistic loss): iteratively, the remaining residual error from the current tree model is the target of the next model. The trained trees are then combined together to form the final model. Our estimates were made using the scikit-learn [52] library.

Model selection & evaluation

In this work, we are primarily interested in evaluating how well machine learning models perform for previously unseen subjects. To this end, sample splitting was used to produce an unbiased estimate of the out-of-sample performance of our models. We used k-fold cross-validation to divide the Aβ-positive subjects into training and test sets. Selection of hyperparameters for the gradient boosting models then used a nested k-fold cross-validation scheme, i.e. cross validation was further performed only on the training samples to select hyperparameters from a grid of possibilities to give a good trade-off between bias and variance.

Cross-validation was used to divide the sample into k outer folds of approximately the same size, $k - 1$ of which were used for model derivation and 1 for validation. The out-of-sample performance was measured by the average across each combination of k derivation and validation folds. In this work, 5-fold cross-validation ($k = 5$) was used, training the model on 80% of the data and testing it on the other 20%. This was repeated so that each subset is used exactly once as a validation set and therefore giving a better indication of how well the model performs on unseen data. The overall performance can then be estimated by averaging over the k folds [53].

Hyperparameter search was performed within each of the k folds; each derivation set was further split again into k inner folds, $k - 1$ of which were used to select a set of model hyperparameters and 1 fold used to validate these. Once the best set was identified, according to the average of the inner held-out folds, the model was retrained on the entire outer derivation fold and tested on the held-out data.

To get a robust and consistent evaluation this procedure was repeated 10 times for different five-fold cross-validation splits and the average test score given as the final performance i.e., 50 held out test score measures from models with (possibly) different hyperparameters are behind the average score and standard deviation reported. As such, it is indicative of the average quality we can expect from a model trained on a new similarly-sized sample and evaluated on a held-out similarly sized sample.

The classification models were evaluated using the weighted F_1 score while the regression models used the coefficient of determination—the R^2 score—as a criterion. The F_1 score contained the weighted average of precision and recall. Consequently, this score took both false positives and false negatives into account. The F_1 was chosen since it is usually more useful than accuracy, especially if the data show an uneven class distribution [54]. The R^2 measures how well the independent variables are capable of explaining the variance of the dependent variable and is defined by $R^2 = 1 - S_{res}/S_{tot}$ where $S_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)$ is the residual sum of squares and $S_{tot} = \sum_{i=1}^{n}(y_i - \overline{y})$ is the total sum of squares. An R^2 value of 0 indicates that performance is as good as predicting the mean of the variable; higher values are better. This definition of R^2 takes values in $[-\infty, 1]$ where negative values represent predictions worse than the mean [55].

Results

We first report the results of the data preprocessing steps, present cohort statistics and describe the imputation approach of variables in the ADNI data set. Second, we present the average rates of cognitive decline over time for the CN, MCI and AD groups, including both Aβ-positive and Aβ-negative subjects. We then inspect the results for models predicting change in MMSE relative to baseline (A1, A2) and change in diagnosis (B). Finally, we study the relationship between predictions in tasks A1 and B with respect to the 2-year follow-ups.

Preprocessing

Preprocessing of the data started with zero-mean normalization of continuous variables and one-hot encoding (dichotomization) of categorical variables to reduce variation in the variables' scales. A simple imputation scheme was used to address missingness in the covariate set. For continuous features, missing values were imputed using mean-imputation while categorical, one-hot encoded features were zero-imputed. These preprocessing steps are performed to maximize the size of the available data and have all features on a similar scale. Since the available cohort for each task was fairly small, and our focus was on held-out prediction risk, which can be estimated in an unbiased way irrespective of the imputation method, model-based imputation was not used. Subjects with missing outcomes were excluded from each corresponding prediction task.

As our main focus is to study the progression of subjects with Aβ pathology, we identified an Aβ-positive cohort by examining the recorded ratio of Aβ_{40} and Aβ_{42} at baseline. To avoid introducing bias in the analysis, the ratio was not imputed. For 1279 subjects measurements of both Aβ_{40} and Aβ_{42} were available which resulted in an Aβ-positive cohort of $n = 749$ subjects (see Figure 2). It should be noted that, over time, some participants left the study. Consequently, different numbers of subjects were available at follow-ups two and four years after baseline. The number of Aβ-positive subjects with an MMSE test score available two years after baseline was 500 and 230 after four years. A total of 398 subjects remained for the diagnosis change prediction task.

The subgroup of Aβ-positive subjects had a mean age of 73.7 years (std. of 7.2) over all the diagnosis groups. The gender distribution over all groups was 55.4% male and 44.6% female. The MMSE score was available for All Subjects at baseline: the CN group had a mean value of 29.0 (1.2), the MCI subgroup a mean of 27.4 (1.9) and AD subjects 23.3 (2.0). Another important feature was the tau variable, where measurements were available nearly for all (99%) of the Aβ-positive subjects. Additionally, the main genetic risk factor for AD, the APOE4 gene, of which a person can have zero, one or two copies, was included for almost all of the Aβ-positive cohort (only 39 were missing) [56].

FDG, measured by positron emission tomography and shown to be a strong marker for AD [47], was absent in 20.9% of the cohort. The statistics of key features used in the three prediction tasks are presented in Table 1 for the subgroup of Aβ-positive subjects and in Table 6 in the supplementary file for the cohort of All Subjects.

Average rates of cognitive decline

For each visit at $t = 1/2, 1, 2$ and 4 years after baseline, the average MMSE score was calculated for observations of different groups divided based on baseline diagnosis (CN, MCI, AD) and Aβ-cohorts ($A\beta$-positives and $A\beta$-negatives). The results are shown in Figure 3. While we observe a noticeable difference in the rate of cognitive decline between the $A\beta$-positive and negative groups for the MCI subjects, the two CN groups differ only slightly in their trajectories. For the group of AD-positive participants the mean MMSE score increases again after two years. However, it is likely that

this change is due to the dropout of a significant number of study participants around this time, resulting in a cohort with different characteristics than at baseline.

The average MMSE score for the Aβ-positive MCI group was 23.79 four years after the baseline visit, while it was initially 27.40—a decrease on average by 3.61. In contrast, the average score of the MCI Aβ-negative group started at 28.27 and averaging 28.20 score points after four years, showing an average decrease of only 0.07. The analysis shows for the CN Aβ-positive and negative groups a decrease in the average score of 0.70 and an increase of 0.08 respectively.

As expected, Aβ-positivity was strongly correlated with faster progression. Although there were remarkable differences in the average deterioration of the MMSE score between the Aβ groups, it should be noted that there was a significant number of missing observations for each group and time point after the baseline visit, due to subjects not undergoing a certain inspection or dropping out of the study. For reference, there were fewer Aβ-positive subjects involved in the study in total ($n = 230$) after four years than at the beginning of the study ($n = 749$) (Figure 2). The number of participants in the CN Aβ-positive (and negative) groups decreased from 115 and (237) at the beginning of the study to 69 and (143) after four years, respectively, while the number of subjects in the MCI Aβ-positive (and negative) groups started at 356 and (226), and declined to 168 and (149) after four years. The AD group has a massive drop from 179 at baseline to only 10 subjects after four years.

Task A: Predicting change in MMSE score

In Table 2, we report the performance of the linear regression and the gradient boosting models that predict the change in MMSE scores after two and four years, respectively, as measured using the average cross-validated R^2 score and standard deviation. The standard deviation was computed across the held-out validation sets corresponding to different cross-validation folds. We compare models fit using only cognitive test scores measured at baseline as predictors, to models fit using a preselected feature set described previously.

The best two-year MMSE prediction model achieved an R^2 of 0.388 (std. 0.073) using all features and a linear regression model utilizing All Subjects but weighted during training. This model scored marginally higher than restricting the training data to only Aβ subjects with a R^2 of 0.372 (std. 0.081). The gradient boosting models performed worse across the three cohort selections compared with their linear regression counterparts. These results do not indicate any immediate benefit from using nonlinear estimators to model cognitive score change in this sample. The best prediction for the two-year follow-up using only cognitive tests resulted in an R^2 of 0.350 (std. 0.079) which is only slightly lower than the best model using all features.

The best cross-validated R^2 score for predicting change in MMSE after four years was 0.325 (std. 0.134), using all features and a linear regression model using the equally weighted cohort in the training. Using only cognitive tests for this task gives a lower score indicating that other biomarkers offer more than in the two-year case. Using only the Aβ subjects for this task results in quite poor predictions with high variability compared to utilizing the weighted sample cohort or the weighted equally cohort while training, indicating that more data can significantly improve the training of these models. Similarly to the two-year setting, the gradient boosting models showed lower performance than the linear models.

Across both tasks A1 and A2, linear models using the larger feature selection and utilizing more subjects than just the cohort containing only Aβ positive subjects performed considerably better in predicting the change of the MMSE score.

Figure 4 shows a calibration plot for held-out data corresponding to a single fold from the cross-validation from a linear regression model predicting MMSE change after two years. Calibration was good for smaller declines but worse for faster-declining subjects, for which the predictions underestimated the change. This trend was consistent across the two follow-up lengths; there are a few subjects whose change in the MMSE score is significantly larger than others and therefore are more difficult to predict. These outliers may potentially also have decreased the quality of predictions of other data points.

In Table 4 in the supplementary material, we list the importance measures of features across the two-year prediction models using all features. For predicting change in the MMSE score, the most important features were baseline cognitive scores, with ADAS13, TRABSCORE, and ADAS11 being the most predictive. The linear models additionally selected the mPAACCtrailsB, LDELTOTAL and ADASQ4 while other cognitive tests such as FAQ and RAVLT_immediate were chosen by the gradient boosting models as part of the most predictive features. This is expected since subjects with early disease status (e.g., with high baseline MMSE score) tend to change less rapidly than already progressing subjects [57]. For this reason, we included also the results of estimators predicting change in MMSE based only on baseline cognitive scores in Table 2. However, we see that across all models and tasks, the performance improved slightly by using additional predictors.

Several features were only identified as important by one or two models across the cohorts. For instance, the volume measurement of WholeBrain was selected by two gradient boosting models including All Subjects equally weighted and the Aβ only cohort. Moreover, the FDG feature, obtained by PET and known to be a strong marker for AD [47] is selected in the cohort including only Aβ positives among the five most important features.

The estimated levels of Aβ measured through Aβ_{42} in CSF and AV45 PET scans showed low predictive power in the context of other features across all cohorts and models. For example, the Aβ_{42} measurements were only included with a coefficient of 0.30 in the linear regression model using All Subjects equally weighted and -0.01 when training with only Aβ subjects and the AV45 is rated even less predictive.

For the four-year predictions, the features that are rated most important in the linear regression models are a dementia diagnosis, TAU and PTAU proteins in CSF followed by the mPACCtrailsB and ADASQ cognitive tests. The gradient boosting models however deem FDG along with the cognitive scores ADAS13, FAQ and mPACCtrailsB to be of most importance for making predictions. Comparing to the two-year predictions it is interesting to see the increased value in using biomarkers other than cognitive tests. The four-year predictions also indicate low predictability by Aβ related features when predicting the rate of decline in Aβ-positive individuals.

Task B: Predicting diagnosis change

In Table 3, we report the results of predicting a worsened diagnosis at the two-year follow-up visit. Gradient boosting using all features and an equally weighted cohort during training resulted in the best performance, achieving a cross-validated weighted F_1 score of 0.791 with a standard deviation of 0.042. However, the gradient boosting model with weighted subjects in the cohort reaches only a slightly lower weighted F_1 score of 0.782 with a standard deviation of 0.040. The logistic regression models consistently perform worse than the gradient boosting ones on the diagnosis prediction for the two-years follow-up.

When using only the cognitive tests, the best performing model also uses gradient boosting and a cohort including All Subjects weighted equally achieving a weighted F_1 score of 0.787 with a standard deviation of 0.043. This is very close to the previous result using all features. Similarly, the other models using only cognitive tests performed marginally worse than their counterparts using all features. Similarly, the models using only cognitive tests performed marginally worse than their counterparts using all features. In summary, additional features lead to only a slight improvement in the performance for both logistic regression and gradient boosting.

The most important features for the diagnosis models over all three training cohorts are LDELTOTAL and mPACCtrailsB. This result demonstrates that the two most important features in progression prediction belong to the group of cognitive assessment. The logistic regression models also selected as important: TAU, PTau, two $APOE4$ genes and $DX\_NUM\_1.0$ which represents the MCI diagnosis at baseline. However, the gradient boosting models identified several other cognitive test scores as important features, for example, FAQ, TRABSCOR, and ADAS13. Similarly, to the prediction of the change of MMSE score, one can conclude that the Aβ_{42} obtained by CSF as well as the AV45 retrieved by PET are not among the most important features for any of the diagnosis change models.

We can conclude that the logistic regression models and the gradient boosting models rely on similar features. There are bigger differences between important features in logistic regression models than those using gradient boosting.

Relating predicted cognitive decline & diagnosis change

In Figure 5, we plot the predictions made by models for tasks A1 and B for the same set of baseline-MCI subjects. Overall, we see a strong correlation between predicted cognitive decline (negative change in MMSE) and predicted change from MCI to AD status. The variance in predicted MMSE change is larger for AD-transitioning subjects than for MCI-stable subjects.

Discussion

Formation of amyloid-beta plaques in the brain is a hallmark of Alzheimer's disease. Only recently, the first drug which may mitigate or slow down the formation of these plaques was approved by the FDA[58, 59]. To best target future interventions of this kind, it is of great interest to identify individuals who are most likely to suffer rapid cognitive decline. Since presence of Aβ plaques is required for an AD diagnosis and can be detected early in CSF and plasma, successful prediction of who among Aβ-positive subjects are likely to deteriorate first could have significant clinical implications.

Machine learning approaches, including classification [23, 24] and regression [26, 28] methods, have been used to predict progression of patients from CN to MCI and from MCI to AD. The results show that subjects who already have cognitively declined are most likely to deteriorate more rapidly. However, although such studies have shown that Aβ levels among others are strong predictors of the transition from MCI to an AD

diagnosis [13, 20, 27], prediction of progression specifically for patients with established amyloid pathology is so far unexplored.

In this work, we studied prediction of cognitive decline in an Aβ-positive cohort using machine learning methods. We applied multivariate statistical analyses to explain the variation in changes in cognitive scores and diagnoses, between subjects in the ADNI dataset, as a function of commonly available clinical variables. We found that the predictability of changes in cognitive test scores was low, leaving a large portion of variance unexplained. Our results complement previous works which show good discrimination of progressing and non-progressing subjects [16, 21] in cohorts comprising both Aβ-positive and Aβ-negative subjects. In particular, we show that discriminating between subjects who are potential candidates for drugs designed to reverse or slow down Aβ plaque formation presents a harder prediction task.

Predictors of progression in Amyloid-positive subjects

Confirming previous results, we found that the ratio of Aβ_{42} and Aβ_{40} CSF level is a good first-line predictor of decline in the MMSE score [43]. However, when limiting the cohort to only the Aβ-positive subjects, the predictive power of the levels of Aβ_{42} and Aβ_{40} was substantially reduced. In other words, the Aβ biomarkers served predominantly to produce a binary grouping of subjects.

The most important features for predicting disease progression in all considered tasks were baseline cognitive test scores. Although related work has not focused specifically on the Aβ-positive cohort, these results are consistent with previous results in selecting cognitive tests such as the MMSE and ADAS13 tests as important predictive features [44, 29]. Our analysis demonstrated that cognitive test results indicate well how the individual will progress and that those who were already cognitively impaired would likely deteriorate more. Since most of the cognitive test scores are highly correlated, several cognitive scores could perhaps be combined and summarized in a joint variable rather than using all of them separately. Apart from cognitive scores, some of the CSF biomarkers, brain scans and other biomarkers showed lower average importance as predictors for progression when including All Subjects. This can partially be explained due to the higher missingness of these features when viewing All Subjects.

Increasing training cohort

Increasing the number of subjects by adding those that were not in the Aβ-positive cohort to the training set consistently increased the predictions for that group. Therefore, it seems the Aβ-negative subjects have fairly similar characteristics that determine their cognitive decline. A weighting procedure allowing us to include more subjects in the training gave a better performance than using only the subjects we were interested in predicting. The increased performance from the addition of out-of-cohort samples also indicates that more data would increase the quality of the prediction tasks even further. In the case of predicting MMSE change after four years using a small cohort of only Aβ-positive subjects gave a drastically worse performance.

MMSE as target variable

The MMSE score has been used frequently in dementia research for grading the cognitive state of patients[60, 61]. For this reason, the change in MMSE score was used in this work as a target variable and thus as a proxy for a person's cognitive change. The test benefits from high practicability as the typical administration time is only 8 minutes for cognitively unimpaired individuals and increases to 15 minutes for individuals with dementia. Internal consistency appears to be moderate and test-retest reliability good [62].

The MMSE is neither the most accurate nor the most efficient instrument for assessing cognitive impairment, nor is it designed specifically for AD. Despite its frequent use, the MMSE lacks sensitivity in patients with high levels of premorbid education and suspected early impairment [63]. Especially for studies that screen cognitively normal populations for evidence of cognitive impairment, the Montreal Cognitive Assessment (MOCA) may be better able to detect age-related cognitive decline in adults since it eliminates the ceiling effects of MMSE [64]. The ADAS13 cognitive test which we used in the primitive studies, could also function as a target variable. The ADAS13 test is also commonly used in clinical trials to thoroughly identify incremental improvements or deteriorations in cognitive performance. Although the ADAS is genuinely accurate in distinguishing individuals with normal cognition from those with impaired cognition, some research studies indicate that the ADAS test may not be difficult enough to consistently detect only mild cognitive impairment [33, 65, 66]. Alternatively, for future work, the outcome variable could be a combination of several cognitive tests, which outweighs the individual characteristics of the cognitive test.

Clinical implications

Prediction of cognitive decline among Aβ-positive subjects could have clinical implications in a scenario where a disease-modifying drug becomes available on the market. In this case, our approach could be used to

assess how an Aβ-positive person, either unimpaired or already in cognitive decline, might develop in the near future. With a further developed predictive approach, physicians could be supported in the prioritization and evaluation of patients for treatment. In particular, models with interpretability aspects may encourage clinicians to use machine learning-based decision-making methods in a clinical context. Further, our approach benefits from relying only on a small number of biomarkers and demographic data that are widely available for many patients and therefore provides high practical relevance. In order to be able to generalize results even better, more accessible patient data will be needed in the future. For an efficient, timely and practical approach to predicting disease development in Alzheimer's patients, the approach of precision medicine could be important. With the goal of improving the health of well-defined patient populations, precision medicine will affect all stakeholders in the healthcare system at multiple levels, from the individual perspective to the societal perspective[67].

Limitations

Our study should be viewed in light of the following limitations. First, there was significant missingness in the target outcome variables, MMSE and diagnosis status, for all prediction tasks. Since these are the targets of prediction, they were not imputed and only subjects with the available output variables were included. Consequently, the cohorts for tasks A1, A2 and B were all different and potentially biased subsets of the initial cohort. For example, the cohort sizes for the regression tasks differ based on whether the MMSE test score variable was available after two years (A1, $n = 500$) or after four years (A2, $n = 230$).

The missingness of outcome variables at follow-up time is partly explained by subjects leaving the study before follow-up. The reason for subjects to end their participation in the study is not known but may be related to disease progression [68]. This phenomenon can bias the trend of the Aβ positive subjects decreasing their MMSE score (Figure 3). However, the dropout rate of people was around 40% in both CN groups and the MCI Aβ-negative one while there was more dropout in the MCI Aβ-positive group where it was 55% and a staggering 94% for the AD group. Consequently, if more people with lower cognitive function would have been included, the average MMSE score would be lower and therefore, the slope of the graph would be slightly steeper and result in an even lower average MMSE score.

As a consequence of the prohibitively small and imbalanced cohorts, we performed a grouped analysis. The use of non-$A\beta$-positive subjects in deriving progression prediction models reduces variance by increasing the sample size of cohorts that had small numbers of subjects. However, this risks bias in terms of the best model for $A\beta$-positive subjects. Note that $A\beta$-positive-negative subjects were used in the derivation of predictive models, but not in evaluation.

Conclusions

We studied the problem of predicting disease progression and cognitive decline of potential AD patients with established Aβ pathology in the ADNI database. The best performing model achieved a performance of $R^2 = 0.388$ predicting the change in MMSE scores two years after baseline using a linear regression model based on a cohort with weighted samples in the training cohort using all features at baseline. Similarly, a gradient boosting model with all subjects weighted equally predicted the change in diagnosis with high accuracy ($F_1 = 0.791$) when using all features. For the most accurate predictions, our models combine variables measured at the baseline such as cognitive tests, CSF biomarkers, proteins and genetic markers. Among these, baseline cognitive tests scores were found to be the strongest predictors, accounting for most of the variance explained by all features, across models. Finally, we identified that even though the $A\beta_{42}/A\beta_{40}$ ratio is a good predictor for AD in the preclinical phase, the respective levels of Aβ are less useful in predicting progression among only Aβ-positive subjects.

Acknowledgements
The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE) partially funded by the Swedish Research Council through grant agreement no. 2018-05973. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Funding
KB is supported by the Swedish Research Council (#2017-00915), the Alzheimer Drug Discovery Foundation (ADDF), USA (#RDAPB-201809-2016615), the Swedish Alzheimer Foundation (#AF-742881), Hjärnfonden, Sweden (#FO2017-0243), the Swedish state under the agreement between the Swedish government and the County Councils, the ALF-agreement (#ALFGBG-715986), and European Union Joint Program for Neurodegenerative Disorders (#JPND2019-466-236).
HZ is a Wallenberg Scholar supported by grants from the Swedish Research Council (#2018-02532), the European Research Council (#681712), Swedish State Support for Clinical Research (#ALFGBG-720931), the Alzheimer Drug Discovery Foundation (ADDF), USA (#201809-2016862), the AD Strategic Fund and the Alzheimer's Association (#ADSF-21-831376-C, #ADSF-21-831381-C and #ADSF-21-831377-C), the Olav Thon Foundation, the Erling-Persson Family Foundation, Hjärnfonden, Sweden (#FO2019-0228), the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860197 (MIRIADE), and the UK Dementia Research Institute at UCL.
LS and FJ are supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. HD is supported by Chalmers AI Research Centre co-seed project (#CHAIR-CO-AIMDAD-2020-012-1).
Data collection and sharing was funded by ADNI (NIH #U01 AG024904) and DOD ADNI (#W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Abbreviations
AD: Alzheimer's disease dementia
ADAS-Cog: Alzheimer's Disease Assessment Scale-Cognitive Subscale
ADNI: Alzheimer's Disease Neuroimaging Initiative
Aβ: Amyloid-β
Aβ ratio: Aβ_{42}/Aβ_{40} ratio
Aβ-positive: Aβ ratio lower than 0.13
Aβ-negative: Aβ ratio higher than 0.13
A1: Predicting cognitive decline after two years
A2: Predicting cognitive decline after four years
B: predicting worsened diagnosis status
CN: Cognitively normal
CSF: Cerebrospinal fluid
EMA: European Medicines Agency
FDA: Food and Drug Administration
GB: Gradient boosting
GMM: Gaussian mixture model
LR: Logistc regression
MCI: Mild cognitive impairment
ML: Machine learning
MMSE: Mini Mental Status Test
MRI: Magnetic resonance imaging
PET: Positron emission tomography
std: Standard deviation

Availability of data and materials
I can confirm I have included a statement regarding data and material availability in the declaration section of my manuscript.

Ethics approval and consent to participate
All procedures were approved by the Institutional Review Boards of all participating institutions. Written informed consent was obtained from every research participant according to the Declaration of Helsinki and the Belmont Report.

Competing interests
KB has served as a consultant, at advisory boards, or at data monitoring committees for Abcam, Axon, Biogen, JOMDD/Shimadzu. Julius Clinical, Lilly, MagQu, Novartis, Roche Diagnostics, and Siemens Healthineers, and is a co-founder of Brain Biomarker Solutions in Gothenburg AB (BBS), which is a part of the GU Ventures Incubator Program.
The other authors declare that they have no competing interests.

Consent for publication
Not applicable

Authors' contributions
All authors have read and approved the final manuscript.

Author details
[1]Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Gothenburg, SE. [2]Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, The Sahlgrenska Academy at the University of Gothenburg, , Mölndal, Sweden. [3]Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Mölndal, Sweden. [4]Department of Neurodegenerative Disease, UCL Institute of Neurology, , London, UK. [5]UK Dementia Research Institute, UCL, , London, UK. [6]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: London http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

References
1. Association, A.: Alzheimer's disease facts and figures. Alzheimers Dement 2020 **16(3)**, 391 (2020)
2. Lee, G., Nho, K., Kang, B., Sohn, K.-A., Kim, D.: Predicting alzheimer's disease progression using multi-modal deep learning approach. Scientific reports **9**(1), 1–12 (2019)
3. Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., Iwatsubo, T., Jack Jr, C.R., Kaye, J., Montine, T.J., *et al.*: Toward defining the preclinical stages of alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. Alzheimer's & dementia **7**(3), 280–292 (2011)
4. Tábuas-Pereira, M., Baldeiras, I., Duro, D., Santiago, B., Ribeiro, M.H., Leitão, M.J., Oliveira, C., Santana, I.: Prognosis of early-onset vs. late-onset mild cognitive impairment: Comparison of conversion rates and its predictors. Geriatrics **1**(2), 11 (2016)
5. Mitchell, A., Shiri-Feshki, M.: Temporal trends in the long term risk of progression of mild cognitive impairment: a pooled analysis. Journal of Neurology, Neurosurgery & Psychiatry **79**(12), 1386–1391 (2008)
6. Jack, C., Bennett, D., Blennow, K., Carrillo, M., Dunn, B., Haeberlein, S., Holtzman, D., Jagust, W., Jessen, F., Karlawish, J., et al.: NIA-AA research framework: toward a biological definition of Alzheimer's disease. Alzheimer's Dement 14: 535–562 (2018)

7. Soto, C.: Unfolding the role of protein misfolding in neurodegenerative diseases. Nature Reviews Neuroscience **4**(1), 49–60 (2003)
8. Solomon, A., Mangialasche, F., Richard, E., Andrieu, S., Bennett, D.A., Breteler, M., Fratiglioni, L., Hooshmand, B., Khachaturian, A.S., Schneider, L.S., *et al.*: Advances in the prevention of alzheimer's disease and dementia. Journal of internal medicine **275**(3), 229–250 (2014)
9. Bondi, M.W., Jak, A.J., Delano-Wood, L., Jacobson, M.W., Delis, D.C., Salmon, D.P.: Neuropsychological contributions to the early identification of alzheimer's disease. Neuropsychology review **18**(1), 73–90 (2008)
10. Murphy, M.P., LeVine III, H.: Alzheimer's disease and the amyloid-β peptide. Journal of Alzheimer's disease **19**(1), 311–323 (2010)
11. Crystal, H., Dickson, D., Fuld, P., Masur, D., Scott, R., Mehler, M., Masdeu, J., Kawas, C., Aronson, M., Wolfson, L.: Clinico-pathologic studies in dementia: nondemented subjects with pathologically confirmed alzheimer's disease. Neurology **38**(11), 1682–1682 (1988)
12. Braak, H.: Neuropathological staging of alzheimer-related changes correlates with psychometrically assessed intellectual status. In: Alzheimer's Disease: Advances in Clinical and Basic Research. Third International Conference on Alzheimer's Disease and Related Disorders, 1993 (1993). John Wiley & Sons
13. Hammond, T.C., Xing, X., Wang, C., Ma, D., Nho, K., Crane, P.K., Elahi, F., Ziegler, D.A., Liang, G., Cheng, Q., *et al.*: β-amyloid and tau drive early alzheimer's disease decline while glucose hypometabolism drives late decline. Communications biology **3**(1), 1–13 (2020)
14. Henriques, A.D., Benedet, A.L., Camargos, E.F., Rosa-Neto, P., Nóbrega, O.T.: Fluid and imaging biomarkers for alzheimer's disease: Where we stand and where to head to. Experimental gerontology **107**, 169–177 (2018)
15. Machado, A., Ferreira, D., Grothe, M.J., Eyjolfsdottir, H., Almqvist, P.M., Cavallin, L., Lind, G., Linderoth, B., Seiger, Å., Teipel, S., *et al.*: The cholinergic system in subtypes of alzheimer's disease: an in vivo longitudinal mri study. Alzheimer's Research & Therapy **12**, 1–11 (2020)
16. Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., Initiative, A.D.N., *et al.*: Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects. Neuroimage **104**, 398–412 (2015)
17. McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack Jr, C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., *et al.*: The diagnosis of dementia due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. Alzheimer's & dementia **7**(3), 263–269 (2011)
18. Petersen, R.C.: Early diagnosis of alzheimer's disease: is mci too late? Current Alzheimer Research **6**(4), 324–330 (2009)
19. Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., Initiative, A.D.N., *et al.*: Brainage in mild cognitive impaired patients: predicting the conversion to alzheimer's disease. PloS one **8**(6), 67346 (2013)
20. Tanveer, M., Richhariya, B., Khan, R., Rashid, A., Khanna, P., Prasad, M., Lin, C.: Machine learning techniques for the diagnosis of alzheimer's disease: A review. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **16**(1s), 1–35 (2020)
21. Beltrán, J.F., Wahba, B.M., Hose, N., Shasha, D., Kline, R.P., Initiative, A.D.N.: binexpensive, non-invasive biomarkers predict alzheimer transition using machine learning analysis of the alzheimer's disease neuroimaging (adni) database. PloS one **15**(7), 0235663 (2020)
22. Giorgio, J., Landau, S.M., Jagust, W.J., Tino, P., Kourtzi, Z., Initiative, A.D.N., *et al.*: Modelling prognostic trajectories of cognitive decline due to alzheimer's disease. NeuroImage: Clinical **26**, 102199 (2020)
23. Satone, V., Kaur, R., Faghri, F., Nalls, M.A., Singleton, A.B., Campbell, R.H.: Learning the progression and clinical subtypes of alzheimer's disease from longitudinal clinical data. arXiv preprint arXiv:1812.00546 (2018)
24. Bucholc, M., Ding, X., Wang, H., Glass, D.H., Wang, H., Prasad, G., Maguire, L.P., Bjourson, A.J., McClean, P.L., Todd, S., *et al.*: A practical computerized decision support system for predicting the severity of alzheimer's disease of an individual. Expert Systems with Applications **130**, 157–171 (2019)
25. Shaffer, J.L., Petrella, J.R., Sheldon, F.C., Choudhury, K.R., Calhoun, V.D., Coleman, R.E., Doraiswamy, P.M., Initiative, A.D.N.: Predicting cognitive decline in subjects at risk for alzheimer disease by using combined cerebrospinal fluid, mr imaging, and pet biomarkers. Radiology **266**(2), 583–591 (2013)
26. Tanaka, T., Lavery, R., Varma, V., Fantoni, G., Colpo, M., Thambisetty, M., Candia, J., Resnick, S.M., Bennett, D.A., Biancotto, A., *et al.*: Plasma proteomic signatures predict dementia and cognitive impairment. Alzheimer's & Dementia: Translational Research & Clinical Interventions **6**(1), 12018 (2020)
27. Pascoal, T.A., Therriault, J., Mathotaarachchi, S., Kang, M.S., Shin, M., Benedet, A.L., Chamoun, M., Tissot, C., Lussier, F., Mohaddes, S., *et al.*: Topographical distribution of aβ predicts progression to dementia in aβ positive mild cognitive impairmentb. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **12**(1), 12037 (2020)
28. Casanova, R., Barnard, R.T., Gaussoin, S.A., Saldana, S., Hayden, K.M., Manson, J.E., Wallace, R.B., Rapp, S.R., Resnick, S.M., Espeland, M.A., *et al.*: Using high-dimensional machine learning methods to estimate an anatomical risk factor for alzheimer's disease across imaging databases. NeuroImage **183**, 401–411 (2018)
29. Geifman, N., Kennedy, R.E., Schneider, L.S., Buchan, I., Brinton, R.D.: Data-driven identification of endophenotypes of alzheimer's disease progression: implications for clinical trials and therapeutic interventions. Alzheimer's research & therapy **10**(1), 1–7 (2018)
30. Moradi, E., Hallikainen, I., Hänninen, T., Tohka, J., Initiative, A.D.N., *et al.*: Rey's auditory verbal learning test scores can be predicted from whole brain mri in alzheimer's disease. NeuroImage: Clinical **13**, 415–427 (2017)
31. Thabtah, F., Spencer, R., Ye, Y.: The correlation of everyday cognition test scores and the progression of alzheimer's disease: a data analytics study. Health Information Science and Systems **8**(1), 1–11 (2020)
32. Galea, M., Woodward, M.: Mini-mental state examination (mmse). Australian Journal of Physiotherapy **51**(3), 198 (2005)
33. Mohs, R.C., Cohen, L.: Alzheimer's disease assessment scale (adas). Psychopharmacol Bull **24**(4), 627–628 (1988)
34. Hua Wang, Nie, F., Huang, H., Risacher, S., Ding, C., Saykin, A.J., Shen, L.: Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In: 2011 International Conference on Computer Vision, pp. 557–562 (2011). doi:10.1109/ICCV.2011.6126288
35. Zhang, D., Shen, D., Initiative, A.D.N., *et al.*: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. NeuroImage **59**(2), 895–907 (2012)
36. Zhang, D., Shen, D., Initiative, A.D.N., *et al.*: Predicting future clinical changes of mci patients using longitudinal and multimodal biomarkers. PloS one **7**(3), 33182 (2012)
37. Guo, T., Korman, D., La Joie, R., Shaw, L.M., Trojanowski, J.Q., Jagust, W.J., Landau, S.M.: Normalization of csf ptau measurement by aβ 40 improves its performance as a biomarker of alzheimer's disease. Alzheimer's research & therapy **12**(1), 1–15 (2020)
38. Bouallègue, F.B., Mariano-Goulart, D., Payoux, P., (ADNI, A.D.N.I., *et al.*: Comparison of csf markers and semi-quantitative amyloid pet in alzheimer's disease diagnosis and in cognitive impairment prognosis using the adni-2 database. Alzheimer's research & therapy **9**(1), 32 (2017)
39. Hampel, H., Toschi, N., Baldacci, F., Zetterberg, H., Blennow, K., Kilimann, I., Teipel, S.J., Cavedo, E., Melo dos Santos, A., Epelbaum, S., *et al.*: Alzheimer's disease biomarker-guided diagnostic workflow using the added value of six combined cerebrospinal fluid candidates: Aβ1–42, total-tau, phosphorylated-tau, nfl, neurogranin, and ykl-40. Alzheimer's & Dementia **14**(4), 492–501 (2018)
40. Schenker-Ahmed, N.M., Bulsara, N., Yang, L., Huang, L., Iranmehr, A., Wu, J., Graff, A.M., Dadakova, T., Chung, H.-K., Tkach, D., et al.: Addition of genetics to quantitative mri facilitates earlier prediction of dementia: A non-invasive alternative to amyloid measures. bioRxiv, 731661 (2019)
41. Spires-Jones, T.L., Attems, J., Thal, D.R.: Interactions of pathological

proteins in neurodegenerative diseases. Acta neuropathologica **134**(2), 187–205 (2017)
42. West, T., Kirmess, K.M., Meyer, M.R., Holubasch, M.S., Knapik, S.S., Hu, Y., Contois, J.H., Jackson, E.N., Harpstrite, S.E., Bateman, R.J., *et al.*: A blood-based diagnostic test incorporating plasma aβ42/40 ratio, apoe proteotype, and age accurately identifies brain amyloid status: findings from a multi cohort validity analysis. Molecular neurodegeneration **16**(1), 1–12 (2021)
43. Balceiras, I., Santana, I., Leitão, M.J., Gens, H., Pascoal, R., Tábuas-Pereira, M., Beato-Coelho, J., Duro, D., Almeida, M.R., Oliveira, C.R.: Addition of the aβ42/40 ratio to the cerebrospinal fluid biomarker profile increases the preadictive value for underlying alzheimer's disease dementia in mild cognitinve impaierfrment. Alzheimer's research & therapy **10**(1), 1–15 (2018)
44. Li, K., Chan, W., Doody, R.S., Quinn, J., Luo, S.: Prediction of conversion to alzheimer's disease with longitudinal measures and time-to-event data. Journal of Alzheimer's Disease **58**(2), 361–371 (2017)
45. Dick, J., Guiloff, R., Stewart, A., Blackstock, J., Bielawska, C., Paul, E., Marsden, C.: Mini-mental state examination in neurological patients. Journal of Neurology, Neurosurgery & Psychiatry **47**(5), 496–499 (1984)
46. Lewczuk, P., Esselmann, H., Otto, M., Maler, J.M., Henkel, A.W., Henkel, M.K., Eikenberg, O., Antz, C., Krause, W.-R., Reulbach, U., *et al.*: Neurochemical diagnosis of Alzheimer's dementia by CSF Aβ42, Aβ42/Aβ40 ratio and total tau. Neurobiology of aging **25**(3), 273–281 (2004)
47. Lewczuk, P., Riederer, P., O'Bryant, S.E., Verbeek, M.M., Dubois, B., Visser, P.J., Jellinger, K.A., Engelborghs, S., Ramirez, A., Parnetti, L., Jr, C.R.J., Teunissen, C.E., Hampel, H., Lleó, A., Jessen, F., Glodzik, L., de Leon, M.J., Fagan, A.M., Molinuevo, J.L., Jansen, W.J., Winblad, B., Shaw, L.M., Andreasson, U., Otto, M., Mollenhauer, B., Wiltfang, J., Turner, M.R., Zerr, I., Handels, R., Thompson, A.G., Johansson, G., Ermann, N., Trojanowski, J.Q., Karaca, I., Wagner, H., Oeckl, P., van Waalwijk van Doorn, L., Bjerke, M., Kapogiannis, D., Kuiperij, H.B., Farotti, L., Li, Y., Gordon, B.A., Epelbaum, S., Vos, S.J.B., Klijn, C.J.M., Nostrand, W.E.V., Minguillon, C., Schmitz, M., Gallo, C., Mato, A.L., Thibaut, F., Lista, S., Alcolea, D., Zetterberg, H., Blennow, K., Kornhuber, J., on Behalf of the Members of the WFSBP Task Force Working on this Topic: Peter Riederer, F.T. Carla Gallo Dimitrios Kapogiannis Andrea Lopez Mato: Cerebrospinal fluid and blood biomarkers for neurodegenerative dementias: An update of the consensus of the task force on biological markers in psychiatry of the world federation of societies of biological psychiatry. The World Journal of Biological Psychiatry **19**(4), 244–328 (2018). doi:10.1080/15622975.2017.1375556. PMID: 29076399. https://doi.org/10.1080/15622975.2017.1375556
48. Nguyen, M., He, T., An, L., Alexander, D.C., Feng, J., Yeo, B.T.T.: Predicting alzheimer's disease progression using deep recurrent neural networks. NeuroImage **222**, 117203 (2020). doi:10.1016/j.neuroimage.2020.117203
49. Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W., Barkhof, F., Fox, N.C., Eshaghi, A., Toni, T., et al.: The alzheimer's disease prediction of longitudinal evolution (tadpole) challenge: Results after 1 year follow-up. arXiv preprint arXiv:2002.03419 (2020)
50. Reynolds, D.A.: Gaussian mixture models. Encyclopedia of biometrics **741**, 659–663 (2009)
51. Friedman, J.H.: Stochastic gradient boosting. Computational statistics & data analysis **38**(4), 367–378 (2002)
52. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
53. Raschka, S.: Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808 (2018)
54. Lipton, Z.C., Elkan, C., Naryanaswamy, B.: Optimal thresholding of classifiers to maximize f1 measure. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 225–239 (2014). Springer
55. Dancer, D., Tremayne, A.: R-squared and prediction in regression with ordered quantitative response. Journal of Applied Statistics **32**(5), 483–493 (2005)
56. Liu, C.-C., Kanekiyo, T., Xu, H., Bu, G.: Apolipoprotein e and alzheimer disease: risk, mechanisms and therapy. Nature Reviews Neurology **9**(2), 106–118 (2013)
57. Caldwell C. C. Yao, J..B.R.D.: Targeting the prodromal stage of alzheimer's disease: bioenergetic and mitochondrial opportunities. Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics **12**(1), 66–80 (2015)
58. Dr. Patrizia Cavazzoni, F.C.f.D.E. Director, Research: FDA's Decision to Approve New Treatment for Alzheimer's Disease (Accessed: 06/07/2021). https://www.fda.gov/drugs/news-events-human-drugs/fdas-decision-approve-new-treatment-alzheimers-disease
59. Parkins, K.: Alzheimer's trials: Biogen and Lilly's amyloid-targeting drugs race for FDA approval (Accessed: 18/04/2021). https://www.clinicaltrialsarena.com/analysis/alzheimers-biogen-eli-lilly-amyloid-targeting-therapy-fda-approval/
60. Folstein, M.F., Folstein, S.E., McHugh, P.R.: "mini-mental state": A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research **12**(3), 189–198 (1975). doi:10.1016/0022-3956(75)90026-6
61. Mcguinness, B., Todd, S., Passmore, P., Bullock, R.: Blood pressure lowering in patients without prior cerebrovascular disease for prevention of cognitive impairment and dementia. Cochrane database of systematic reviews (Online) **7**, 004034 (2006). doi:10.1002/14651858.CD004034.pub2
62. Mitchell, A.J.: The mini-mental state examination (mmse): update on its diagnostic accuracy and clinical utility for cognitive disorders. In: Cognitive Screening Instruments, pp. 37–48. Springer, ??? (2017)
63. Martin R., O.D.: Taxing your memory. London, England (373(9680)) (2009-2010). doi:10.1016/S0140-6736(09)60349-4
64. Gluhm, S., Goldstein, J., Loc, K., Colt, A., Van Liew, C., Corey-Bloom, J.: Cognitive performance on the mini-mental state examination and the montreal cognitive assessment across the healthy adult lifespan. Cognitive and behavioral neurology : official journal of the Society for Behavioral and Cognitive Neurology **26**, 1–5 (2013). doi:10.1097/WNN.0b013e31828b7d26
65. Kueper, J.K., Speechley, M., Montero-Odasso, M.: The alzheimer's disease assessment scale–cognitive subscale (adas-cog): modifications and responsiveness in pre-dementia populations. a narrative review. Journal of Alzheimer's Disease **63**(2), 423–444 (2018)
66. Podhorna, J., Krahnke, T., Shear, M., Harrison, J.E.: Alzheimer's disease assessment scale–cognitive subscale variants in mild cognitive impairment and mild alzheimer's disease: change over time and the effect of enrichment strategies. Alzheimer's research & therapy **8**(1), 1–13 (2016)
67. Faulkner, E., Holtorf, A.-P., Walton, S., Liu, C.Y., Lin, H., Biltaj, E., Brixner, D., Barr, C., Oberg, J., Shandhu, G., *et al.*: Being precise about precision medicine: what should value frameworks incorporate to address precision medicine? a report of the personalized precision medicine special interest group. Value in Health **23**(5), 529–539 (2020)
68. Larson, E.B., Shadlen, M.-F., Wang, L., McCormick, W.C., Bowen, J.D., Teri, L., Kukull, W.A.: Survival after initial diagnosis of alzheimer disease. Annals of internal medicine **140**(7), 501–509 (2004)

Figures

Figure 1 Histogram showing the Aβ ratio of subjects at baseline. The different coloured groups represent different diagnoses: dementia, MCI and CN. The upper left histogram shows all diagnoses together and overlaps have blended colours like green and dark red.

Figure 2 Exclusion flowchart showing the $A\beta$-positive (left) and All Subjects (right) cohort. The graphs present the cohorts used for the prediction in the change in the MMSE score (A1 and A2) for the a change in the diagnosis after two years (B).

Figure 3 Graph showing the MMSE score development for CN, MCI and AD subjects split by Aβ-status. The shaded areas represent 95% confidence intervals for the mean values. The number of subjects decreases over time, hence the growing uncertainty bands.

Figure 4 A calibration plot (true vs predicted values) for a linear regression model that predicts the change in MMSE score two years from baseline.

Figure 5 Predicted change in MMSE score and the predicted probability of a change in diagnosis after two years in the baseline-MCI group. Points are color-labeled based on their observed change in diagnosis.

Tables

Table 1 Baseline demographic and clinical characteristics of the ADNI cohort for Aβ-positive subjects and All Subjects.
Column 'Overall $A\beta$-positive' and 'Missingness' have been swapped.

		Overall $A\beta$-positive	Missing	CN	MCI	AD	All subjects	All missing
n		749		132	411	206	2293	
AGE, mean (SD)		73.7 (7.2)	1	75.2 (6.0)	73.2 (7.0)	73.6 (8.1)	73.2 (7.2)	9
Gender, n (%)	m	415 (55.4)	0	55 (41.7)	247 (60.1)	113 (54.9)	1217 (53.2)	5
	f	334 (44.6)		77 (58.3)	164 (39.9)	93 (45.1)	1071 (46.8)	
MMSE, mean (SD)		26.5 (2.8)	0	29.0 (1.2)	27.4 (1.9)	23.3 (2.0)	27.4 (2.66)	5
ADAS13, mean (SD)		20.1 (9.6)	5	10.2 (4.6)	18.4 (6.7)	30.3 (7.7)	17.0 (9.25)	29
TAU, mean (SD)		337.1 (139.2)	7	281.5 (102.5)	334.3 (141.9)	378.6 (141.5)	285.6 (132.9)	1010
ABETA42, mean (SD)		754.0 (320.0)	0	885.3 (396.5)	761.5 (311.6)	654.8 (240.7)	1090.7 (607.5)	1014
FDG, mean (SD)		1.19 (0.15)	157	1.29 (0.11)	1.22 (0.14)	1.06 (0.13)	1.23 (0.15)	806
APOE4, n (%)	0	245 (34.5)	39	63 (50.4)	133 (34.5)	49 (24.6)	1162 (54.1)	147
	1	345 (48.6)		56 (44.8)	187 (48.4)	102 (51.3)	780 (36.4)	
	2	120 (16.9)		6 (4.8)	66 (17.1)	48 (24.1)	204 (9.5)	
Hippocampus, mean (SD)		6517.6 (1091.6)	168	7300.8 (788.8)	6585.0 (1029.8)	5883.8 (1008.4)	6794.0 (1185.7)	806
AV45, mean (SD)		1.37 (0.20)	312	1.29 (0.21)	1.36 (0.20)	1.44 (0.18)	1.21 (0.23)	1205
ABETARatio, mean (SD)		0.087 (0.029)	0	0.094 (0.020)	0.086 (0.023)	0.084 (0.023)	0.129 (0.057)	1014

Table 2 Performance of the linear and gradient boosting regressions, predicting change in MMSE two and four years after baseline for three different cohort selections. We compare models trained on features a) the all features set from baseline and b) from baseline cognitive scores only.

R^2 (SD)	2-year follow-up	4-year follow-up
all features		
Linear regression, $A\beta$ Only	0.372 (0.081)	0.205 (0.227)
Linear regression, All Subjects	0.354 (0.083)	0.325 (0.134)
Linear regression, All Subjects, Weighted	0.388 (0.073)	0.304 (0.152)
Gradient boosting, $A\beta$ Only	0.287 (0.124)	0.156 (0.244)
Gradient boosting, All Subjects	0.356 (0.108)	0.252 (0.191)
Gradient boosting, All Subjects, Weighted	0.338 (0.950)	0.263 (0.192)
Cognitive tests only		
Linear regression, $A\beta$ Only	0.343 (0.087)	0.178 (0.203)
Linear regression, All Subjects	0.333 (0.081)	0.228 (0.143)
Linear regression, All Subjects, Weighted	0.350 (0.079)	0.225 (0.160)
Gradient boosting, $A\beta$ Only	0.272 (0.133)	-0.050 (0.358)
Gradient boosting, All Subjects	0.323 (0.118)	0.149 (0.224)
Gradient boosting, All Subjects, Weighted	0.293 (0.114)	0.118 (0.227)

Table 3 Performance of the classification models in predicting change in diagnosis two years after baseline for three different cohort selections.

weighted F_1 (SD)	follow-up after 2-year	
	all features	cognitive tests only
LR, $A\beta$ Only	0.763 (0.050)	0.761 (0.046)
LR, All Subjects	0.781 (0.044)	0.762 (0.050)
LR, All Subjects, Weighted	0.776 (0.047)	0.770 (0.046)
GB, $A\beta$ Only	0.770 (0.043)	0.784 (0.041)
GB, All Subjects	0.788 (0.045)	0.793 (0.039)
GB, All Subjects, Weighted	0.786 (0.046)	0.787 (0.037)

Supplementary material

0.1 Supplementary material 1

Supplementary material The supplementary material includes a method describing weighting of cohorts, along with lists of the cognitive tests and other features. Hyperparameter values, and two tables showing the feature importance for two and four years after baseline are presented.

Paper II

Sharing pattern submodels for prediction with missing values

Lena Stempfle, Ashkan Panahi, Fredrik D. Johansson

Thirty-Seventh Conference on Artificial Intelligence (AAAI-23, To appear)
(2023)

Sharing Pattern Submodels for Prediction with Missing Values

Lena Stempfle, Ashkan Panahi, Fredrik D. Johansson

Chalmers University of Technology
Department of Computer Science and Engineering, Gothenburg, Sweden
stempfle@chalmers.se, ashkan.panahi@chalmers.se, fredrik.johansson@chalmers.se

Abstract

Missing values are unavoidable in many applications of machine learning and present challenges both during training and at test time. When variables are missing in recurring patterns, fitting separate pattern submodels have been proposed as a solution. However, fitting models independently does not make efficient use of all available data. Conversely, fitting a single shared model to the full data set relies on imputation which often leads to biased results when missingness depends on unobserved factors. We propose an alternative approach, called sharing pattern submodels, which i) makes predictions that are robust to missing values at test time, ii) maintains or improves the predictive power of pattern submodels, and iii) has a short description, enabling improved interpretability. Parameter sharing is enforced through sparsity-inducing regularization which we prove leads to consistent estimation. Finally, we give conditions for when a sharing model is optimal, even when both missingness and the target outcome depend on unobserved variables. Classification and regression experiments on synthetic and real-world data sets demonstrate that our models achieve a favorable tradeoff between pattern specialization and information sharing.

1 Introduction

Machine learning models are often used in settings where model inputs are partially missing either during training or at the time of prediction (Rubin 1976). If not handled appropriately, missing values can lead to increased bias or to models that are inapplicable in deployment without imputing the values of unobserved variables (Liu, Zachariah, and Stoica 2020; Le Morvan et al. 2020). When missingness is dependent on unobserved factors that are related also to the prediction target, the fact that a variable is unmeasured can itself be predictive—so-called *informative missingness* (Rubin 1976; Marlin 2008). Often, imputation of missing values is insufficient, and it can be beneficial to let models make predictions based on both the partially observed data and on indicators for which variables are missing (Jones 1996; Groenwold et al. 2012). As mentioned in Morvan et al. (2020), even the linear model—the simplest of all regression models—has not yet been thoroughly investigated with missing values and still reveals unexpected challenges.

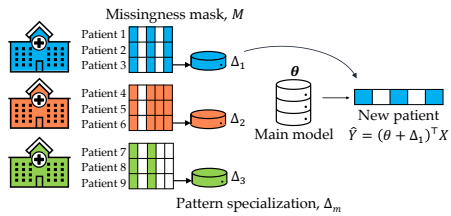

Figure 1: Coefficient sharing between a main model θ and pattern submodels for three clinics with different patterns in missing values. Without specialization, Δ_m, an average prediction shared by clinics with different patterns may not lead to an optimal solution for any of them. Conversely, fitting separate models for each clinic does not use all of the available data efficiently and leads to high variance.

Pattern missingness emerges in data generating processes (DGPs) where there are structural reasons for which variables are measured—samples are grouped by recurring patterns of measured and missing variables (Little 1993). In Figure 1, we illustrate an example of this when observing patients from three different clinics, each systematically collecting slightly different measurements. Assume for simplicity that the pattern of missing values is unique to each clinic. In this way, a pattern-specific model is also site-specific.

Pattern submodels have been proposed for this setting, fitting a separate model to samples from each pattern (Mercaldo and Blume 2020; Marshall et al. 2002). This solution does not rely on imputation and can improve interpretability over black-box methods (Rudin 2019), but can suffer from high variance, especially when the number of distinct patterns is large and the number of samples for a given pattern is small. Moreover, if the fitted models differ significantly between patterns, it may be hard to compare or sanity-check their predictions. Notably, pattern submodels disregard the fact that the prediction task is shared between each pattern. However, in the context of Figure 1, using a shared model for all clinics may also be suboptimal if clinics take different measurements, or treat patients differently (high bias).

We propose the *sharing pattern submodel* (`SPSM`) in which submodels for different missingness patterns share

coefficients while allowing limited specialization. This encourages efficient use of information across submodels leading to a beneficial tradeoff between predictive power and variance in the case where similar submodels are desired and sample sizes per pattern are small. Additionally, models with few and small differences between patterns are easier for domain experts to interpret.

We describe `SPSM` in Section 3, and we prove that in linear-Gaussian systems, a model which shares coefficients between patterns may be optimal—even when the prediction target depends on missing variables and on the missingness pattern (Section 4). Finally, we find in an experimental evaluation on real-world and synthetic data that `SPSM` compares favorably to baseline classifiers and regression methods, paying particular attention to how `SPSM` boosts sample efficiency and model sparsity (Section 5).

2 Prediction with Test-Time Missingness

Let $X = [X_1, ..., X_d]^\top$ be a vector of d random variables taking values in $\mathcal{X} \subseteq \mathbb{R}^d$, and $M = [M_1, ..., M_d]^\top$ be a random missingness mask in $\mathcal{M} \subseteq \{0, 1\}^d$ where $M_j = 1$ indicates that variable X_j is missing. Next, let $\tilde{X} \in (\mathbb{R} \cup \{\texttt{NA}\})^d$ be the mixed observed-and-missing values of X according to M and define $X_{\neg M} = [X_j : M_j = 0]^\top \in \mathbb{R}^{d - \|M\|_1}$ to be the vector of *observed* covariates under M. The outcome of interest, $Y \in \mathbb{R}$, may depend on all of X, observed or missing, as well as on M. Let $k = |\mathcal{M}|$ denote the number of possible missingness patterns.[1] Further, assume that variables X, M, Y are distributed according to a *fixed, unknown* joint distribution p. The assumed (causal) dependencies of the variables used, coincide most closely with *selection missingness* (Little 1993) (Figure 4 in the appendix).

Our goal is to predict Y *under missingness* M in X using functions $f : (\mathbb{R} \cup \{\texttt{NA}\})^d \to \mathbb{R}$. We aim to minimize risk with respect to the squared loss on p,

$$\min_f R(f), \text{ where } R(f) \coloneqq \mathbb{E}_{\tilde{X}, Y \sim p}[(f(\tilde{X}) - Y)^2] \,. \quad (1)$$

Under the assumption that Y has centered, additive noise,

$$Y = g(X, M) + \epsilon \;\text{ where }\; \mathbb{E}[\epsilon] = 0, \quad (2)$$

the Bayes-optimal predictor of Y is $f^* = \mathbb{E}[Y \mid X_{\neg M}, M]$. In general, observed values $X_{\neg M}$ are insufficient for predicting Y; f^* may depend directly on the mask M, *even if Y does not depend directly on M* (Le Morvan et al. 2021).

A common strategy to learn f is to first impute the missing values in $\tilde{X}$ and then fit a model on the observed-or-imputed covariates $X^I \in \mathbb{R}$—so-called *impute-then-regress* estimation. Even though imputation is powerful, it is not always optimal under test-time missingness (Le Morvan et al. 2021) and often assumes that data is missing at random (MAR) (Carpenter and Kenward 2012; Seaman et al. 2013).

2.1 Pattern Submodels

In cases where the number of distinct missingness patterns k is small, it is possible to learn separate predictors f_m for each pattern. This idea has been called *pattern submodels* (PSM) (Mercaldo and Blume 2020; Marshall et al. 2002), a set of models which aim to minimize the empirical risk under each missingness pattern. Let $D = \{(\tilde{x}^{(1)}, m^{(1)}, y^{(1)}), ..., (\tilde{x}^{(n)}, m^{(n)}, y^{(n)})\}$ be a data set of n samples, with partially observed features $\tilde{x}^{(i)}$, corresponding to missingness patterns $m^{(i)}$, drawn independently and identically distributed from p. PSM may be learned by minimizing the regularized empirical risk,

$$\min_{\{f_m\} \in \mathcal{F}^k} \; \frac{1}{n} \sum_{i=1}^{n} L(f_{m^{(i)}}(\tilde{x}^{(i)}), y^{(i)}) + \sum_{m \in \mathcal{M}} \mathcal{R}(f_m) \quad (3)$$

over a suitable class of models $\mathcal{F}$ and regularization $\mathcal{R}$. Mercaldo and Blume (2020) considered linear and logistic regression models, $f_m = \sigma(\theta_m^\top x)$ with σ either the identity or logistic function and loss L chosen to match. The objective in (3) is separable in m and can be solved independently for each pattern. However, this often leads to high variance in the small-sample regime since each pattern accounts for only a subset of the available samples. Without structural assumptions, the number of patterns k grows exponentially with d (see discussion in Section 6).

PSM allows for prediction under test-time missingness which adapts to the pattern m without relying on imputation or assumptions on missingness mechanisms like MAR. However, the prediction target (and the Bayes-optimal model f^*) may have only a small dependence on the pattern m; *the optimal submodels for all m may share significant structure*. Next, we propose estimators that exploit such structures to reduce variance and increase interpretability.

3 Sharing Pattern Submodels

We propose *sharing pattern submodels* (`SPSM`), linear prediction models, specialized for patterns in variable missingness, which share information during learning. Sharing is accomplished by regularizing submodels towards a main model and solving the resulting coupled optimization problem. While linear models are limited in expressive power, they are often found to be useful approximations of nonlinear functions due to their superior interpretability.

Fitting SPSM Let $\theta \in \mathbb{R}^d$ represent *main model* coefficients used in prediction under all missingness patterns, and define $\theta_{\neg m} = [\theta_j : m_j = 0]^\top \in \mathbb{R}^{d_m}$ to be the subset of coefficients corresponding to variables observed under m. To emphasize, $\theta_{\neg m}$ depends only on m in selecting a subset of θ—the coefficients are shared across patterns. Similarly, define $\Delta_{\neg m} \in \mathbb{R}^{d_m}$ to be *pattern-specific specialization* of these coefficients to m. In contrast to $\theta_{\neg m}$, the values of $\Delta_{\neg m}$ are unique to each pattern m. Note, a model f_m depends only on the observed components of X. In regression tasks, we learn **sharing** pattern submodels on the form

$$f_m(x) \coloneqq (\theta_{\neg m} + \Delta_{\neg m})^\top x_{\neg m}, \;\; \text{for all } \; m \in \mathcal{M} \quad (4)$$

[1] In practical scenarios, we expect k to be much smaller than the worst-case number, 2^d.

by solving the following problem with $\lambda_m \geq 0$ and $\gamma \geq 0$,

$$\underset{\theta,\{\Delta_{\neg m}\}}{\text{minimize}} \ \frac{1}{n}\sum_{i=1}^{n}\left((\theta_{\neg m^{(i)}} + \Delta_{\neg m^{(i)}})^\top x^{(i)}_{\neg m^{(i)}} - y^{(i)}\right)^2 + \frac{\gamma}{n}\|\theta\| + \sum_{m\in\mathcal{M}}\frac{\lambda_m}{n_m}\|\Delta_{\neg m}\|_1 \ . \quad (5)$$

where n_m is the number of samples of pattern m. $\lambda_m > 0$ and $\gamma > 0$ are regularization parameters. Intercepts (pattern-specific and shared) are left out for brevity. The optimization problem is convex, and we find optimal values for θ and Δ_m using L-BFGS-B (Byrd et al. 1995) in experiments. In classification tasks, the square loss is replaced by the logistic loss. In either case, we call the solution to (5) `SPSM`.

For the penalty $\|\theta\|$, we use either the ℓ_1 or ℓ_2 norm to tradeoff bias and variance in the main model. A high value for λ_m regularizes the specialization of model coefficients to missingness pattern m such that high λ_m encourages smaller $\|\Delta_m\|_1$ and greater coefficient sharing. In experiments, we let λ_m take the same value λ for all patterns. ℓ_1-regularization is used for Δ as we aim for a sparse solution where the majority of specialization coefficients are zero.

Consistency For fixed λ, γ, sums of the minimizers of (5), $\theta^*_{\neg m} + \Delta^*_{\neg m}$, converge to the best linear approximations of the Bayes-optimal predictors f^*_m for each pattern m in the large-sample limit. We state this formally and sketch a proof in Appendix A.2 using standard arguments. This result is agnostic to parameter sharing; Δ^* may not be sparse. In Section 4, we prove that, in the linear-Gaussian setting, our method also recovers the sparsity of the true process. In the large-sample limit, this may not be beneficial for variance reduction, but sparsity contributes to interpretability.

Why is SPSM Interpretable? Comparing pattern specializations allows domain experts to reason about how similar submodels are, and how they are affected by missing values. We argue that a set of submodels is more interpretable if specializations contain fewer non-zero coefficients, $\Delta_{\neg m}$ is sparse. Sparsity is a generally useful measure of interpretablity (Rudin 2019), since it results in only a subset of the input features affecting predictions, reducing the effective complexity of the model (Miller 1956; Cowan 2010).

4 Optimality of Sharing Models

In this section, we give conditions under which an optimal pattern submodel has sparse specializations (shares parameters between patterns) and when `SPSM` converges to such a model in the large-sample limit. We analyze DGPs where the outcome Y depends linearly on *all* components of X (*models* have access only the observed subset of these) and on the pattern M, but not on interactions between X and M,

$$Y = \theta^\top X + \alpha_M + \epsilon \ , \text{ with } \epsilon \sim \mathcal{N}(0, \sigma_Y^2). \quad (6)$$

Here, α_M is a pattern-specific intercept. Without α_M, this is a setting often targeted by imputation methods, since the outcome is a parametric function of the full X. However, we know that X will be partially missing also at test time, and M is allowed to have arbitrary dependence on X. In this case, imputation need not be necessary or sufficient.

Next, we study this setting with Gaussian X, where we can precisely characterize optimal models and their sparsity.

4.1 Sparsity in Linear-Gaussian DGPs

Recall that $X_{\neg m}$ and $\theta_{\neg m}$ denote covariates and coefficients restricted to *observed* variables under pattern m, and define X_m and θ_m analogously for missing variables. For outcomes which obey (6), the Bayes-optimal model under m is

$$\mathbb{E}[Y \mid X_{\neg m}, M = m] = {\theta_{\neg m}}^\top X_{\neg m} + \xi_m \quad (7)$$

where $\xi_m = {\theta_m}^\top \mathbb{E}_{X_m}[X_m \mid X_{\neg m}] + \alpha_m$ is the bias of the naïve prediction made using the coefficients $\theta_{\neg m}$ of the true system but restricted to observed variables. Ignoring ξ_m coincides with performing prediction following zero-imputation and is biased in general. ξ_m thus captures the specialization required for pattern submodels to be unbiased. For closer analysis, we study the following setting.

Condition 1 (Linear-Gaussian DGP)**.** *Covariates $X = [X_1, ..., X_d]^\top$ are Gaussian, $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ. The outcome Y is linear-Gaussian as in* (6) *with parameters $(\theta, \{\alpha_m\}, \sigma_Y)$. M is arbitrary.*

In line with Condition 1, let $\Sigma_{\neg m,m}$ be the submatrix of Σ restricted to the rows corresponding to *observed* variables under m and columns corresponding to variables *missing* under m. Define $\Sigma_{\neg m,\neg m}$ and $\Sigma_{m,\neg m}$ analogously. Throughout, we assume that Σ is invertible so that the distribution is non-degenerate. In practice, the non-degenerate case can be handled through ridge regularization.

Proposition 1. *Suppose covariates X and outcome Y obey Condition 1 (are linear-Gaussian). Then, the Bayes-optimal predictor for an arbitrary missingness mask $m \in \mathcal{M}$, is*

$$f^*_m = \mathbb{E}[Y \mid X_{\neg m}, m] = (\theta_{\neg m} + \Delta_{\neg m})^\top X_{\neg m} + C_m$$

where $C_m \in \mathbb{R}$ is constant with respect to $X_{\neg m}$ and

$$\Delta_{\neg m} = (\Sigma^{-1}_{\neg m,\neg m})\Sigma_{\neg m,m}\theta_m \ .$$

Proposition 1 states that, for a linear-Gaussian system, the Bayes-optimal model under missingness pattern m has the same form as `SPSM` with pattern-specific intercept, combining coefficients of a main model θ and specializations $\Delta_{\neg m}$. The result is proven in Appendix A.3.

In nonlinear DGPs, the optimal correction term $\Delta_{\neg m}$ may not be constant with respect to $X_{\neg m}$. The NeuMiss model by Le Morvan et al. (2020) learns such corrections as functions of the input and missingness mask using deep neural networks. However, this method lacks the interpretability of sparse linear models sought here. Even in this more general case, `SPSM` may achieve a good bias-variance tradeoff. Indeed, we find on real-world data, which may not be linear, that `SPSM` is often preferable to strong nonlinear baselines.

When is Sparsity Optimal? Like other sparsity-inducing regularized estimators, such as `LASSO` (Tibshirani 1996), `SPSM` reduces variance by shrinking some model parameters to zero. Under appropriate conditions, when the training

set grows large, we expect the learned sparsity to correspond to properties inherent to the DGP. For LASSO, this means recovering zeros in the coefficient vector of the outcome. For SPSM, objective (5) is used to learn submodels on the form $(\theta_{\neg m}+\Delta_{\neg m})^\top X_{\neg m}$ where θ is shared between patterns and $\Delta_{\neg m}$ is sparse. It is natural to ask: When can we expect the "true" or an "optimal" $\Delta_{\neg m}$ to be sparse and, if it is, when can we recover this sparsity with SPSM? Surprisingly, as we will see, the optimal specialization $\Delta_{\neg m}$ may be sparse even if Y depends on *all* covariates in X.

Assume that Condition 1 (Linear-Gaussian DGP) holds with system parameters $(\mu, \Sigma, \theta, \{\alpha_m\}, \sigma_Y)$. We can characterize sparsity in the Bayes-optimal model $(\theta, \{\Delta_{\neg m}\})$, see Proposition 1, by the interactivity of covariates. We say that variables X_j and $X_{j'}$ are non-interactive if they are statistically independent given all other covariates. As is well-known, for Gaussian X, X_j and $X_{j'}$ are non-interactive if $S_{j,j'} = 0$, where $S = \Sigma^{-1}$ is the precision matrix.

Proposition 2 (Sparsity in optimal model)**.** *Suppose that a covariate $j \in [d]$ is observed under pattern m, i.e., $m_j = 0$, and assume that X_j is non-interactive with every covariate $X_{j'}$ that is missing under m. Then $(\Delta_{\neg m})_j = 0$.*

Proposition 2 states that the sparsity in Δ is partially determined by the covariance pattern of observed and unobserved covariates. For example, specialization is *not needed* for a variable j under pattern m if it is uncorrelated with all missing variables under m. Conversely, specialization, i.e., $(\Delta_{\neg m})_j \neq 0$, *is needed* for features j that are predictive ($\theta_j \neq 0$) and redundant (replicated well by unobserved features which are also predictive). This is because in the main model, redundant variables may share the predictive burden, but when they are partitioned by missingness, they have to carry it alone. This shows that prediction with a single model and zero-imputation is sub-optimal in general.

Consistency of SPSM In the large-sample limit, under Condition 1, we can prove that SPSM recovers maximally sparse optimal model parameters. If the true system parameters are also sparse, SPSM learns these.

Theorem 1. *Suppose that Condition 1 holds with parameters $(\theta, \{\Delta_{\neg m}\})$ as in Proposition 1, such that, for each covariate j, the number of patterns m for which $m_j = 0$ and $(\Delta_{\neg m})_j = 0$ is strictly larger than the number of patterns m' for which $m'_j = 0$ and $(\Delta_{\neg m'})_j \neq 0$. Then, with $\gamma = 0$ and fixed $\lambda > 0$, the true parameters $(\theta, \{\Delta_{\neg m}\})$ are the unique solution to* (5) *in the large-sample limit, $n \to \infty$.*

Proof sketch. We provide a full proof in Appendix A.5. The main steps involve showing that the SPSM objective (5) is asymptotically dominated by the risk term, and the *sums* of its minimizers $(\theta^*_{\neg m} + \Delta^*_{\neg m})$ coincide with optimal regression coefficients $(\hat{\theta}_{\neg m})$ fit independently for each missingness pattern m. For any $\lambda > 0$, regularization steers the solution towards one which is maximally sparse in $\Delta^*_{\neg m}$. □

4.2 Relationship to Other Methods

For particular extreme values of the regularization parameters γ, λ_m, SPSM coincides with other methods (Table 1).

	$\gamma < \infty$	$\gamma \to \infty$
$\lambda_m \to \infty$	Zero imputation	Constant
$0 < \lambda_m < \infty$	Sharing model	Pattern submodel
$\lambda_m = 0$	No sharing	Pattern submodel

Table 1: Extreme cases and equivalences of SPSM, provided that no pattern-specific intercept is used.

First, the full-sharing model ($\lambda_m \to \infty, \gamma < \infty$) coincides with fitting a single model to all samples after zero-imputation. To see this, set $\Delta_{\neg m} = 0$ for all m and note

$$\theta_{\neg m^{(i)}}{}^\top x^{(i)}_{\neg m^{(i)}} = \theta^\top I_0(\tilde{x}^{(i)})$$

where $I_0(\tilde{x})$ replaces missing values in $\tilde{x}$ with 0. In this setting, submodel coefficients cannot adapt to m. In the implementation, we allow the fitting of pattern-specific intercepts which are not regularized by λ_m. Second, ($\lambda_m < \infty, \gamma \to \infty$) corresponds with the standard PSM without parameter sharing (Mercaldo and Blume 2020) or the ExtendedLR method of (Morvan et al. 2020). The precise nature of this equivalence depends on the choice of regularization.[2] In this setting, each submodel $\hat{f}_m$ is fit completely independently of every other. Finally, an SPSM model with optimal parameters $(\theta, \{\Delta_{\neg m}\})$, in the linear-Gaussian case, implicitly makes a perfect single linear imputation,

$$\mathbb{E}[X_m \mid X_{\neg m}] = X_{\neg m}\Sigma^{-1}_{\neg m,\neg m}\Sigma_{\neg m,m},$$

and applies the main model's parameters θ_m to the imputed values. If many samples are available, it may be feasible to learn the imputation directly. However, if the variables in $X_{\neg m}$ and X_m are never observed together, imputation is no longer possible. In contrast, SPSM could still learn an optimal submodel for each pattern, given enough samples.

5 Experiments

We evaluate the proposed SPSM model[3] on simulated and on real-world data, aiming to answer two main questions: How does the accuracy of SPSM compare to baseline models, including impute-then-regress, for small and larger samples; How does sparsity in pattern specializations Δ affect performance and interpretation?

Experimental Setup In the SPSM algorithm, before one-hot-encoding of categorical features, all missingness patterns in the training set are identified. At test time, patterns that did not occur during training, variables are removed until the closest training pattern is recovered. Both linear and logistic variants of SPSM were trained using the L-BFGS-B solver provided as part of the SciPy Python package (Virtanen et al. 2020). Our implementation supports both ℓ_1 and ℓ_2-regularization of the main model parameters θ and ℓ_1-regularization of pattern-specific deviations Δ. This includes both the no-sharing pattern submodel ($\lambda_m < \infty, \gamma \to \infty$) and full-sharing model ($\lambda_m \to$

[2] Mercaldo and Blume (2020) adopted a two-stage estimation procedure, the relaxed LASSO (Meinshausen 2007).

[3] Code to reproduce experiments and the appendix is available at https://github.com/Healthy-AI/spsm.

$\infty, \gamma < \infty$) as special cases. In the experiments, γ can take values within $[0, 0.1, 1, 5, 10, 100]$, and we used a shared $\lambda_m = \lambda \in [1, 5, 10, 100, 1000, 1e^8]$ for all patterns. Intercepts were added for both the main model and for each pattern without regularization. We do not require patterns to have a minimum sample size but support this functionality (appendix Table 8). For missingness patterns at test time that did not occur in the training data, variables were removed until the closest training pattern was recovered.

We compare linear and logistic regression models to the following baseline methods: Imputation + Ridge / logistic regression (`Ridge`/`LR`), Imputation + Multilayer perceptron (`MLP`) with a single hidden layer, and XGBoost (`XGB`), where missing values are supported by default (Chen et al. 2019). Last, we compare the Pattern Submodel (`PSM`) (Mercaldo and Blume 2020). Note, our implementation of `PSM` is based on a special case of our `SPSM` implementation where regularization is applied over all patterns and not in each pattern separately. Hyperparameters are based on the validation set. For imputation, we use zero (I_0), mean (I_μ) or iterative imputation (I_{it}) from SciKit-Learn (Pedregosa et al. 2011b; Van Buuren 2018). `XGB`'s handling of missing values is denoted I_n. Details about method implementations, hyperparameters and evaluation metrics are given in Appendix B.2.

5.1 Simulated Data

We use simulated data to illustrate the behavior of sharing pattern submodels and baselines in relation to Proposition 1, focusing on bias and variance. We sample d input features X from a multivariate Gaussian $\mathcal{N}(0, \Sigma)$ with covariance matrix Σ specified by a cluster structure; the features are partitioned into k clusters of equal size. The covariance is defined as $\Sigma_{ii} = 1$, $\Sigma_{i \neq j} = 0$ if i, j are in different clusters, and $\Sigma_{i \neq j} = c$ if i, j are in the same cluster, where c is chosen as large as possible so that Σ remains positive semidefinite.

Each cluster $c \in \{1, ..., k\}$ is represented in the outcome function $Y = \theta^\top X + \epsilon$ by a single feature $i(c)$, such that $\beta_{i(c)} \sim \mathcal{N}(0, 1)$ and $\theta_j = 0$ for other features. We let $\epsilon \sim \mathcal{N}(0, 1)$, independently for each sample. We consider three missingness settings: In Setting A, each variable in cluster c is missing if $X_{i(c)} > -0.5$. In Setting B, each variable in cluster c—except one chosen uniformly at random—is missing if $X_{i(c)} > -0.5$. Both settings satisfy the conditions of Proposition 1 but are designed to violate MAR by letting the outcome variable depend directly on missing values which may not be recovered from observed ones. In Setting C, we follow missing-completely-at-random (MCAR), where variables are missing independently with probability 0.2. We generate samples with $d = 20$ and $k = 5$.

In Figure 2, we show the test set coefficient of determination (R^2) for Setting A. Note, that the methods which use imputation (imputation method selected based on validation error at each data set size) perform well initially but plateau quickly, indicating relatively high bias. `SPSM` and `PSM` both achieve a higher R^2 for the full sample. `SPSM` performs better than `PSM` for small samples indicating lower variance. The `SPSM` model includes 42 non-zero pattern-specific coefficients when the training set size is 0.2 and 68 with the fraction is 0.8. Results for Setting B and C are presented in Appendix C.1. Even in the MCAR setting C, `PSM` performs considerably worse than alternatives due to excessive variance from fitting independent pattern-specific models.

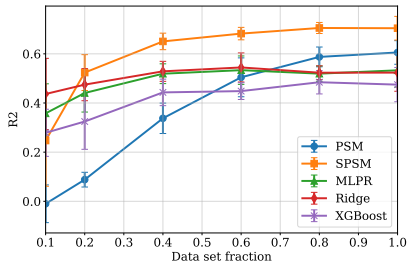

Figure 2: Performance on simulated data Setting A (higher is better). Error bars show standard deviation over 5 random data splits. The full data set has $n = 2000$ samples.

5.2 Real-World Tasks

We describe two health care data sets used for classification and regression. More information on the non-health related HOUSING (De Cock 2011) data is shown in Appendix C.3.

ADNI The data is obtained from the publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) database.[4] ADNI collects clinical data, neuroimaging and genetic data. In the classification task, we predict if a patient's diagnosis will change 2 years after baseline diagnosis. The regression task aims to predict the outcome of the ADAS13 (Alzheimer's Disease Assessment Scale) (Mofrad et al. 2021) cognitive test at a 2-year follow-up based on available data at baseline.

SUPPORT We use data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) (Knaus et al. 1995), which aims to model survival over a 180-day period in seriously ill hospitalized adults using the Physiology Score (SPS). Following Mercaldo and Blume (2020), in the regression task we predict the SPS while for the classification task, we predict if a patient's SPS is above the median; the label rate is 50/50 by definition. We mimic their MNAR setting by adding 25 units to the SPS values of subjects missing the covariate "partial pressure of oxygen in the arterial blood".

5.3 Results

We report the results on health care data in Table 2. For regression tasks, we provide the number of non-zero coefficients used by the linear models. In addition, we study prediction performance as a function of data set size in Figure 3 and in the appendix Figure 7. The statistical uncertainty of the average error is measured with its square root, which is a standard deviation and expressed by 95% confidence intervals over the test set. Results of HOUSING data are presented in Appendix C.3.

[4] http://adni.loni.usc.edu

Regression	R^2	**# Coefficients**
ADNI		
`Ridge`, I_μ	0.66 (0.59, 0.73)	37 + 0
`XGB`, I_μ	0.41 (0.31, 0.50)	—
`MLP`, I_0	0.62 (0.55, 0.69)	—
`PSM`	0.51 (0.43, 0.60)	0 + 430
`SPSM`	0.66 (0.59, 0.73)	37 + 21
SUPPORT		
`Ridge`, I_0	0.38 (0.35, 0.42)	11 + 0
`XGB`, I_n	0.30 (0.27, 0.34)	—
`MLP`, I_μ	0.56 (0.53, 0.59)	—
`PSM`	0.52 (0.49, 0.56)	0 + 188
`SPSM`	0.53 (0.50, 0.56)	11 + 91
Classification	**AUC**	**Accuracy**
ADNI		
`LR`, I_0	0.85 (0.80, 0.90)	0.85 (0.74, 0.94)
`XGB`, I_n	0.80 (0.74, 0.86)	0.84 (0.73, 0.94)
`MLP`, I_0	0.86 (0.78, 0.89)	0.84 (0.73, 0.94)
`PSM`	0.81 (0.75, 0.87)	0.84 (0.74, 0.95)
`SPSM`	0.86 (0.81, 0.90)	0.85 (0.75, 0.96)
SUPPORT		
`LR`, I_0	0.83 (0.81, 0.85)	0.77 (0.74, 0.79)
`XGB`, I_0	0.85 (0.83, 0.87)	0.78 (0.75, 0.81)
`MLP`, I_0	0.86 (0.85, 0.88)	0.79 (0.76, 0.81)
`PSM`	0.84 (0.83, 0.86)	0.78 (0.75, 0.81)
`SPSM`	0.85 (0.83, 0.86)	0.78 (0.75, 0.80)

Table 2: Results for ADNI and SUPPORT tasks along with the respective imputation method (see setup). We also report the number of non-zero coefficients in shared (k) and pattern-specific models (l) as $k + l$.

For ADNI regression, `SPSM` and `Ridge` are the best performing models with R^2 of 0.66 showing the same confidence in the prediction. Validation performance resulted in selecting $\gamma = 10.0, \lambda = 50$ for `SPSM`. With an R^2 score of 0.51, `PSM` seems not able to benefit from pattern-specificity in ADNI. In contrast, `SPSM` makes use of coefficient sharing which results in a significantly smaller number of coefficients compared to `PSM`. For SUPPORT regression, `PSM` achieves almost the same result as `SPSM` (R^2 of 0.52–0.53) with partly overlapping confidence intervals for the predictions. Although, the number of coefficients used in `SPSM` is smaller than in `PSM` due to the coefficient sharing between submodels. The best regularization parameter values for `SPSM` were $\gamma = 0.1, \lambda = 5.0$ which is lower than for ADNI, consistent with the larger data set size. The best performing model is `MLP` (R^2 of 0.56) for SUPPORT regression. However, the black-box nature of `MLP` is not conducive to reasoning about the influence of the missingness pattern. Mean and zero imputation have the best validation performance for `Ridge`, `XGB` and `MLP`. In summary, `SPSM` is consistently among the best-performing models in both data sets, with fairly tight confidence intervals. In ADNI classification, `SPSM`, `MLP` and `LR` achieve the highest prediction accuracy (0.84–0.85) and Area Under the ROC Curve (AUC) (0.85–0.86). All methods perform similarly well on ADNI.

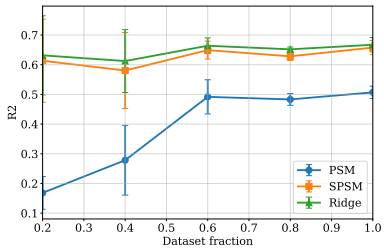

Figure 3: Performance on ADNI for the regression task. Error bars indicate standard deviation over 5 random subsamples of the data. Equal performance for `SPSM` and `Ridge` and subpar performance for `PSM` indicates that for ADNI regression, pattern specialization is mostly irrelevant.

`SPSM` selected $\gamma = 0$ and $\lambda = 1.0$ which indicates moderate coefficient sharing. For SUPPORT data, all models perform almost at the same level. `XGB` and `MLP` perform slightly better than `SPSM` (γ = 0.1, $\lambda = 10.0$) and `PSM`. Across ADNI and SUPPORT `LR`, `XGB` and `MLP` predominantly use zero imputation. In all tasks, `SPSM` performs comparably or favorably to all other methods. The tight confidence intervals for classification in both data sets indicate high certainty in the result averages.

Non-Healthcare Data and Coefficient Specialization In contrast to the previous data sets, where sharing coefficients is beneficial, we see for the HOUSING data, a large advantage from nonlinear estimation: the tree-based approach `XGB` (Table 9). It shows an R^2 of 0.76 and outperforms the other baseline methods for the regression task confirming the non-linearity of that data set. We also do not see the same positive effect in specializing (`PSM`, `SPSM` not better than `Ridge` with imputation). None of the missing value indicators show a significant feature importance level in `XGB` which might indicate that pattern specialization is not necessary. For results on the HOUSING data, see Appendix C.3.

Performance with Varying Training Set Size Figure 3 shows the test R^2 for linear models trained on different fractions of ADNI data. Each set was subsampled into fractions $0.2, 0.4, 0.6, 0.8, 1.0$ of the full data set. Especially for small fractions, `SPSM` benefits from coefficient sharing and lower variance data compared to `PSM`. `Ridge` with mean imputation performs comparably. A similar figure for the SUPPORT is presented in appendix Figure 7. `SPSM` and `PSM` perform equally well across the fractions, whereas `Ridge` shows high error compared to both pattern submodels.

Pattern Specialization in `SPSM` We inspect pattern specializations Δ for `SPSM` in the ADNI regression task with respect to interpretablity. In Table 3, we present the main model θ and pattern-specific coefficients Δ_4 for pattern 4. Table 7 in the appendix shows all patterns m with $\Delta_{\neg m} \neq 0$. For pattern 4, measurements of the amyloid-β (ABETA) peptide and the proteins TAU and PTAU are missing in the

Missing features in pattern 4:
ABETA, TAU and PTAU at baseline (bl)

Feature	Δ_4	θ	$\theta + \Delta_4$
Age	-0.140	0.121	-0.019
FDG-PET	-0.090	-0.039	-0.129
Whole Brain (bl)	0.000	-0.045	-0.044
Fusiform	0.016	0.021	0.037
ICV	0.001	0.093	0.094
Intercept	-0.10	0.18	

Table 3: Example of Δ_4 for regression using SPSM using ADNI. SPSM takes $\gamma = 10$ and $\lambda = 13$ as parameters for a single seed. There are 10 missingness pattern in total, while 4 of them have non-zero coefficients for Δ and pattern-specific intercept. Coefficients are for standardized variables.

baseline diagnostics (ADNI 2012). The absence of these three features affects pattern specialization: For an imaging test FDG-PET (fluorodeoxyglucose), the magnitude of its coefficient is increased, placing heavier weight on the feature in prediction. Similarly, the coefficients for Fusiform (brain volume), and ICV (intracranial volume) increase in magnitude and predictive significance when ABETA, TAU, and PTAU are absent. In contrast, for the feature AGE, the resulting coefficient of -0.019 (compared to 0.121 in the main model) means that the predictive influence of this feature decreases under pattern 4. As Table 3 shows, SPSM applied to tabular data allows for short descriptions of pattern specialization, which helps construct a simple and meaningful model. We enforce sparsity in Δ to limit the number of differences between submodels, and present all features j with specialized coefficients $\Delta_{\neg m}(j) \neq 0$, five in the example case. In this way, the set of submodels is more interpretable and the user, e.g., a medical staff member can be supported in decision-making. For a more detailed analysis on interpretability properties of SPSM, see Appendix C.4.

Tradeoff between Interpretability and Accuracy The interpretability-accuracy tradeoff is especially crucial for practical use of SPSM. The empirical results do not show any significant evidence that our proposed sparsity regularization hurts prediction accuracy (Table 2, Figure 3). Nevertheless, in a practical scenario, domain experts may choose a simpler model at a slight cost in performance. Then, we can measure the tradeoff by varying values of hyperparameters to find an adequate balance (Figure 8). The parameter selection is based on the validation set and aligns with the test set results. We see some parameter sensitivity in SUPPORT that supports sharing, but only in a moderate way.

6 Related Work

Pattern-mixture missingness refers to distributions well-described by an independent missingness component and a covariate model dependent on this pattern (Rubin 1976; Little 1993). In this work, *pattern missingness* refers to emergent patterns which may or may not depend on observed covariates (Marshall et al. 2002). Mercaldo and Blume (2020); Morvan et al. (2020) and Bertsimas, Delarue, and Pauphilet (2021) define pattern submodels for flexible handling of test time missingness. The ExtendedLR method of Morvan et al. (2020) represents a related method to pattern submodels. However, they neither study coefficient sharing between models nor provide a theoretical analysis of when optimal submodels have partly identical coefficients (sharing, sparsity in specialization). Marshall et al. (2002) describes the one-step sweep method using estimated coefficients and an augmented covariance matrix obtained from fully observed and incomplete data at test time. In very recent and so far unpublished work, Bertsimas, Delarue, and Pauphilet (2021) present two methods for predicting with test time missingness. First, *Affinely adaptive regression* specializes a shared model by applying a coefficient correction given by a linear function of the missingness pattern. When the number of variables d is smaller than the number of patterns (which could grow as 2^d), and the outcome is not smooth in changes to missingness mask, this may introduce significant bias. The resulting bias-variance tradeoff differs from our method, and unlike our work, is not justified by theoretical analysis. Second, *Finitely adaptive regression* starts by placing each pattern in the same model, recursively partitioning them into subsets.

Several deep learning methods which are applicable under test time missingness with or without explicitly attempting to impute missing values have been proposed (Bengio and Gingras 1995; Che et al. 2018; Le Morvan et al. 2020; Morvan et al. 2020; Nazabal et al. 2020). The NeuMiss network, discussed briefly in Section 4.1, proposes a new type of non-linearity: the multiplication by the missingness indicator (Le Morvan et al. 2020). NeuMiss approximates the specialization term $\Delta_{\neg m}^\top X_{\neg m}$ (along with per-pattern biases) using a deep neural network where both covariates and missingness mask are given as input, sharing parameters across patterns. NeuMiss and Affinely adaptive regression (see above) are similar since their pattern specializations are functions of the inputs and the masks, both in contrast to SPSM. Moreover, neither method attempts to learn sparse specialization terms (e.g., no ℓ_1 regularization of Δ).

7 Conclusion

We have presented sharing pattern submodels (SPSM) for prediction with missing values at test time. We enforce parameter sharing through sparsity in pattern coefficient specializations via regularization and analyze SPSM's consistency properties. We have described settings where information sharing is optimal even when the prediction target depends on missing values and the missingness pattern itself. Experimental results using synthetic and real-world data confirm that SPSM performs comparably or slightly better than baselines across all data sets without relying on imputation. Notably, the proposed method never performs worse than non-sharing pattern submodels as these do not use the available data efficiently. While SPSM is limited to learning linear models, it is not limited to learning from linear systems. An interesting direction is to identify other classes of models developed with interpretability that could benefit from this type of sharing.

Acknowledgements

We want to thank Devdatt Dubhashi and Marine Le Morvan for their support and fruitful discussions.

This work was partly supported by WASP (Wallenberg AI, Autonomous Systems and Software Program) funded by the Knut and Alice Wallenberg foundation.

The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE) partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

References

ADNI, A., Alzheimer's Disease Neuroimaging Initiative. 2012.

Bengio, Y.; and Gingras, F. 1995. Recurrent neural networks for missing or asynchronous data. *Advances in neural information processing systems*, 8.

Bertsimas, D.; Delarue, A.; and Pauphilet, J. 2021. Prediction with Missing Data. *ArXiv*, abs/2104.03158.

Byrd, R. H.; Lu, P.; Nocedal, J.; and Zhu, C. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5): 1190–1208.

Carpenter, J.; and Kenward, M. 2012. *Multiple imputation and its application*. John Wiley & Sons.

Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; and Liu, Y. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1): 1–12.

Chen, T.; He, T.; Benesty, M.; and Khotilovich, V. 2019. Package 'xgboost'. *R version*, 90.

Cowan, N. 2010. The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science*, 19(1): 51–57.

Dancer, D.; and Tremayne, A. 2005. R-squared and prediction in regression with ordered quantitative response. *Journal of Applied Statistics*, 32(5): 483–493.

De Cock, D. 2011. Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3).

Fagerland, M. W.; Lydersen, S.; and Laake, P. 2015. Recommended confidence intervals for two independent binomial proportions. *Statistical methods in medical research*, 24(2): 224–254.

Groenwold, R. H.; White, I. R.; Donders, A. R.; Carpenter, J. R.; Altman, D. G.; and Moons, K. G. 2012. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, 184(11): 1265–1269.

Hanley, J.; and McNeil, B. 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3): 839–843.

Henelius, A.; Puolamäki, K.; and Ukkonen, A. 2017. Interpreting classifiers through attribute interactions in datasets. *arXiv preprint arXiv:1707.07576*.

Jones, M. P. 1996. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, 91(433): 222–230.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knaus, W. A.; Harrell, F. E.; Lynn, J.; Goldman, L.; Phillips, R. S.; Connors, A. F.; Dawson, N. V.; Fulkerson, W. J.; Califf, R. M.; Desbiens, N.; et al. 1995. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3): 191–203.

Le Morvan, M.; Josse, J.; Moreau, T.; Scornet, E.; and Varoquaux, G. 2020. NeuMiss networks: differentiable programming for supervised learning with missing values. *arXiv:2007.01627*.

Le Morvan, M.; Josse, J.; Scornet, E.; and Varoquaux, G. 2021. What'sa good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, 34.

Lipton, Z. C.; Kale, D. C.; Wetzel, R.; et al. 2016. Modeling missing data in clinical time series with rnns. *Machine Learning for Healthcare*, 56.

Little, R. J. 1993. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421): 125–134.

Liu, X.; Zachariah, D.; and Stoica, P. 2020. Robust Prediction When Features are Missing. *IEEE Signal Processing Letters*, 27: 720–724.

Marlin, B. M. 2008. *Missing Data Problems in Machine Learning*. Ph.D. thesis, University of Toronto.

Marshall, G.; Warner, B.; MaWhinney, S.; and Hammermeister, K. 2002. Prospective prediction in the presence of missing data. *Statistics in medicine*, 21(4): 561–570.

Meinshausen, N. 2007. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1): 374–393.

Mercaldo, S. F.; and Blume, J. D. 2020. Missing data and prediction: the pattern submodel. *Biostatistics*, 21(2): 236–252.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2): 81.

Mofrad, S. A.; Lundervold, A. J.; Vik, A.; and Lundervold, A. S. 2021. Cognitive and MRI trajectories for prediction of Alzheimer's disease. *Scientific Reports*, 11(1): 1–10.

Morvan, M. L.; Prost, N.; Josse, J.; Scornet, E.; and Varoquaux, G. 2020. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In Chiappa, S.; and Calandra, R., eds., *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 3165–3174. PMLR.

Nazabal, A.; Olmos, P. M.; Ghahramani, Z.; and Valera, I. 2020. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107: 107501.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; ; and Duchesnay, E. 2011a. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011b. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Rubin, D. B. 1976. Inference and missing data. *Biometrika*, 63(3): 581–592.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215.

Seaman, S.; Galati, J.; Jackson, D.; and Carlin, J. 2013. What is meant by "missing at random"? *Statistical Science*, 28(2): 257–268.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288.

Van Buuren, S. 2018. *Flexible Imputation of Missing Data (2nd ed.)*. Chapman and Hall/CRC, Boca Raton, FL.

Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, 17(3): 261–272.

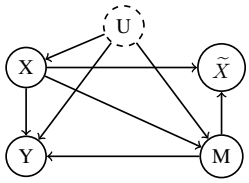

Figure 4: Directed graph showing assumed probabilistic dependencies. $\tilde{X}$ is a deterministic function of X, M. Unobserved variables U may influence both covariates X, missingness M and the outcome Y, ruling out 'missing at random' (MAR).

A Technical appendix

A.1 Variable dependencies

The assumed (causal) dependencies of the variables X, M, Y are represented in a directed graph in Figure 4.

A.2 Consistency in the general case

Proposition 3. *For each pattern m, the minimizers $(\theta^*, \Delta^*_{\neg m})$ of* (5) *are consistent estimators of the best linear approximation to $\mathbb{E}[Y \mid X_{\neg m}, M = m]$,*

$$\lim_{n\to\infty} (\theta^*_{\neg m} + \Delta^*_{\neg m}) = \min_{\eta} \mathbb{E}[(\eta^\top X_{\neg m} - Y)^2 \mid M = m] \,.$$

When the true outcome is linear, $Y = \eta^\top_{\neg m} X_{\neg m} + \epsilon$ with Gaussian errors ϵ, $\lim_{n\to\infty}(\theta^*_{\neg m} + \Delta^*_{\neg m}) = \eta_{\neg m}$.

Proof sketch. Minimizers Δ^* and θ^* will have bounded norm due to the quadratic form of the objectives. This, in the limit $n \to \infty$, regularization terms vanish due to normalization with n and the minimizers $(\theta^*, \{\Delta^*_{\neg m}\})$ are invariant to additive transformations; with $c \in \mathbb{R}^{d_m}$, $\theta'_{\neg m} = \theta^*_{\neg m} + c$ and $\Delta'_{\neg m} = \Delta^*_{\neg m} - c$ also minimize the objective. Choosing $c = -\theta^*_{\neg m}$, we get $\theta'_{\neg m} = 0$ and the objective becomes separable in m. As a result, the objective can be written as k standard least squares problems, one for each pattern. As is well known, for additive sub-Gaussian noise, the minimizers of these problems are consistent for the best linear approximation to the corresponding conditional mean. □

A.3 Proof of Proposition 1

Proposition (Proposition 1 Restated)**.** *Suppose covariates X and outcome Y obey Condition 1 (are linear-Gaussian). Then, the Bayes-optimal predictor for an arbitrary missingness mask $m \in \mathcal{M}$, is*

$$f^*_m = \mathbb{E}[Y \mid X_{\neg m}, m] = (\theta_{\neg m} + \Delta_{\neg m})^\top X_{\neg m} + C_m$$

where $C_m \in \mathbb{R}$ is constant with respect to $X_{\neg m}$ and

$$\Delta_{\neg m} = (\Sigma^{-1}_{\neg m,\neg m})\Sigma_{\neg m,m}\theta_m \,.$$

Proof. By properties of the multivariate Normal distribution, we have that

$$\begin{aligned}&\mathbb{E}_{X_m}[X_m \mid X_{\neg m}]\\ &= \mathbb{E}[X_m] + \Sigma_{m,\neg m}\Sigma^{-1}_{\neg m,\neg m}(X_{\neg m} - \mathbb{E}[X_{\neg m}])\end{aligned}$$

and as a result, following the reasoning above,

$$\begin{aligned}&\mathbb{E}[Y \mid X_{\neg m}]\\ &= (\theta_{\neg m} + (\Sigma_{m,\neg m}\Sigma^{-1}_{\neg m,\neg m})\theta_m)^\top X_{\neg m} + C_m\\ &= (\theta_{\neg m} + \Delta_m)^\top X_{\neg m} + C_m,\end{aligned}$$

where $C_m = \theta_m^\top(\mathbb{E}[X_m] - \Sigma_{m,\neg m}\Sigma^{-1}_{\neg m,\neg m}\mathbb{E}[X_{\neg m}]) + \alpha_m$, which is constant w.r.t. $X_{\neg m}$. □

A.4 Sparsity in optimal model

Proposition (Proposition 2 restated)**.** *Suppose that a covariate $j \in [d]$ is observed under pattern m, i.e., $m_j = 0$, and assume that X_j is non-interactive with every covariate $X_{j'}$ that is missing under m. Then $(\Delta_{\neg m})_j = 0$.*

Proof. Let $\bar{\theta}_{\neg m} = \theta_{\neg m} + \Delta_{\neg m}$. Recall that $S = \Sigma^{-1}$ is the precision matrix for $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and permute the rows and columns of Σ into observed and unobserved parts, such that, without loss of generality, we can write

$$\Sigma = \begin{bmatrix} \Sigma_{\neg m,\neg m} & \Sigma_{\neg m,m} \\ \Sigma^T_{\neg m,m} & \Sigma_{m,m} \end{bmatrix} . \tag{8}$$

Note that by the definition of $\Delta_{\neg m}$ (Proposition 1),

$$\begin{aligned}\Sigma \begin{bmatrix} \Delta_{\neg m} \\ -\bar{\theta}_{\neg m} \end{bmatrix} &= \begin{bmatrix} \Sigma_{\neg m,\neg m} & \Sigma_{\neg m,m} \\ \Sigma^T_{\neg m,m} & \Sigma_{m,m} \end{bmatrix} \begin{bmatrix} \Delta_{\neg m} \\ -\bar{\theta}_{\neg m} \end{bmatrix}\\ &= \begin{bmatrix} 0 \\ g_m \end{bmatrix},\end{aligned}$$

where g_m is a suitable vector. Hence

$$\begin{bmatrix} \Delta_{\neg m} \\ -\bar{\theta}_{\neg m} \end{bmatrix} = \Sigma^{-1} \begin{bmatrix} 0 \\ g_m \end{bmatrix} = S \begin{bmatrix} 0 \\ g_m \end{bmatrix}$$

We conclude the result by noting that $(\Delta_{\neg m})_j$ is zero if in the jth row of S, all entries corresponding to the unobserved part is zero. □

A.5 Consistency in linear-Gaussian DGPs

Theorem (Theorem 1 restated)**.** *Suppose that Condition 1 holds with parameters $(\theta, \{\Delta_{\neg m}\})$ as in Proposition 1, such that, for each covariate j, the number of patterns m for which $m_j = 0$ and $(\Delta_{\neg m})_j = 0$ is strictly larger than the number of patterns m' for which $m'_j = 0$ and $(\Delta_{\neg m'})_j \neq 0$. Then, with $\gamma = 0$ and $\lambda > 0$, the true parameters $(\theta, \{\Delta_{\neg m}\})$ are the unique solution to* (5) *in the large-sample limit, $n \to \infty$.*

Proof. Consider the optimization problem in eq. (5) with $\gamma = 0$. In the large-sample limit ($n \to \infty$), minimizers of the empirical risk over n samples will also minimize the expected risk and, since the outcome is linear-Gaussian, satisfy the constraint in eq. (9). Then, solving (5) is equivalent to solving the following problem:

$$\begin{aligned}\underset{\theta', \{\Delta'_{\neg m}\}}{\text{minimize}} \quad & \sum_m \|\Delta'_{\neg m}\|_1 \\ \text{subject to} \quad & \theta'_{\neg m} + \Delta'_{\neg m} = \theta_{\neg m} + \Delta_{\neg m}, \quad m \in \mathcal{M}\end{aligned} \tag{9}$$

Many parameters (θ', Δ') can satisfy the constraint, due to translational invariance. However, for any value of $\lambda > 0$, regularization in (5) steers the solution towards the one with the smallest norm, $\|\Delta'\|_1$. The reasoning is similar to the argument in the proof of Proposition 3, adding the assumption that the true system is linear-Gaussian. Under the added assumptions of Theorem 1, we can now prove that we also get the correct decomposition.

Take a solution $\theta^*, \{\Delta^*_{\neg m}\}$ of (9). For simplicity of notation below, let vectors $\theta, \Delta_{\neg m}, \theta^*, \Delta^*_{\neg m}$ always be indexed such that the same index j refers to coefficients corresponding to the same covariate X_j. Next, define $I_j = \{m \mid m_j = 0,\ (\Delta_{\neg m})_j = 0\}$ to be the set of patterns where covariate j is observed and without specialization under the optimal model. Similarly, define $I_j^c = \{m \mid m_j = 0,\ (\Delta_{\neg m})_j \neq 0\}$ to be the set of patterns where covariate j is observed and needs specialization. First, note that

$$\begin{aligned}\sum_m \|\Delta^*_{\neg m}\|_1 &= \sum_j \sum_{m|m_j=0} |(\Delta^*_{\neg m})_j| \\ &= \sum_j \sum_{m\in I_j} |(\Delta^*_{\neg m})_j| + \sum_j \sum_{m\in I_j^c} |(\Delta^*_{\neg m})_j|\,.\end{aligned}$$

For $m \in I_j$, we have $\theta^*_j + (\Delta^*_{\neg m})_j = \theta_j$. Hence

$$\sum_j \sum_{m\in I_j} |(\Delta^*_{\neg m})_j| = \sum_j |I_j||\theta_j - \theta^*_j| \tag{10}$$

For $m \in I_j^c$, we have $\theta^*_j + (\Delta^*_{\neg m})_j = \theta_j + (\Delta_{\neg m})_j$ and hence by the triangle inequality, we have

$$\begin{aligned}\sum_j \sum_{k\in I_j^c} |(\Delta^*_k)_j| \geq \sum_j \sum_{k\in I_j^c} \big(|(\Delta_k)_j| - |\theta_j - \theta^*_j|\big) = \\ \sum_m \|\Delta_{\neg m}\|_1 - \sum_j |I_j^c||\theta_j - \theta^*_j|\end{aligned} \tag{11}$$

We conclude that

$$\begin{aligned}\sum_m \|\Delta^*_{\neg m}\|_1 &\geq \sum_m \|\Delta_{\neg m}\|_1 + \sum_j (|I_j| - |I_j^c|)|\theta_j - \theta^*_j| \\ &\geq \sum_m \|\Delta_{\neg m}\|_1\end{aligned}$$

where the last inequality is by the assumption. This provides the desired result. □

B Experiment details

B.1 Real world data sets

ADNI The compiled data set includes 1337 subjects that were preprocessed by one-hot encoding of the categorical features and standardized for the numeric features. The processed data has 37 features and 20 unique missingness patterns. The label set is quite unbalance showing 1089 patients who do not change from their baseline diagnosis, and 248 do. The regression task targets predicting the result of the cognitive test ADAS13 (Alzheimer's Disease Assessment Scale) at a 2 year follow-up (Mofrad et al. 2021) based on available data at baseline.

SUPPORT The data set contains 9104 subjects represented by 23 unique missingness pattern. The following 10 covariates were selected and standardized: partial pressure of oxygen in the arterial blood (pafi), mean blood pressure, white blood count, albumin, APACHE III respiration score, temperature, heart rate per minute, bilirubin, creatinine, and sodium.

B.2 Details of the baseline methods

We compare to the following baseline methods:

Imputation + Ridge / logistic regression (`Ridge`/`LR`) the data is first imputed (see below) and a ridge or logistic regression is fit on the imputed data. The implementation in SciKit-Learn was used (Pedregosa et al. 2011a). The ridge coefficients are shirked by imposing a penalty on their size. They are a reduced factor of the simple linear regression coefficients and thus never attain zero values but very small values (Tibshirani 1996)

Imputation + Multilayer perceptron (`MLP`): The MLP estimator is based on a single hidden layer of size $\in [10, 20, 30]$ followed by a ReLu activation function and a softmax layer for classification tasks and a linear layer for regressions tasks. As input, the imputed data is concatenated with the missingness mask. The MLP is trained using ADAM (Kingma and Ba 2014), and the learning rate is initialized to constant (0.001) or adaptive. We use the implementation in SciKit-Learn (Pedregosa et al. 2011b).

Pattern submodel (`PSM`): For each pattern of missing data, a linear or logistic regression model is fitted, separately regularized with a ℓ_2 penalty. Following Mercaldo and Blume (2020), for patterns with fewer than $2{*}d$ samples available, a complete-case model (CC) is used. Our implementation of `PSM` is based on a special case of our `SPSM` implementation where regularization is applied over all patterns and not in each pattern separately. To enforce fitting separated submodels for each pattern, we set $\gamma = 1e^8$ and $\lambda = 0$.

XGBoost (`XGB`): XGBoost is an implementation of gradient boosted decision trees. Note, XGBoost supports missing values by default (Chen et al. 2019), where branch directions for missing values are learned during training. A logistic classifier is then fit using XGBClassifier while regression tasks are trained with the XGBRegressor (Pedregosa et al. 2011b). We set the hyperparameters to 100 for the number of estimators used, and fix the learning rate to 1.0. The maximal depth of the trees is $\in [5, 10, 15]$.

Imputation methods and hyperparameters for all methods were selected based on the validation portion of random 64/16/20 training/validation/test splits. Results were averaged over five random splits of the data set. The performance metrics for classification tasks were accuracy and the Area Under the ROC Curve (AUC). For regression tasks, we use the mean squared error (MSE) and the R-square, (R^2) value, representing the proportion of the variance for a dependent variable that's explained by an independent variable, taking values in $[-\infty, 1]$ where negative values represent predictions worse than the mean (Dancer and Tremayne 2005). Confidence intervals at significance level $\alpha = 0.05$ are computed based on the number of test set samples. For accuracy, MSE and R^2 we use a Binomial proportion confidence interval (Fagerland, Lydersen, and Laake 2015) and for AUC we use the classical model of (Hanley and McNeil 1983).

Computing Infrastructure The computations required resources of 4 compute nodes using two Intel Xeon Gold 6130 CPUS with 32 CPU cores and 384 GiB memory (RAM). Moreover, a local disk with the type and size of SSSD 240GB with a local disk, usable area for jobs including 210 GiB was used. Inital experiments are run on a Macbook using macOS Montery with a 2,6 GHz 6-Core Intel Core i7 processor.

C Additional experimental results

C.1 Simulation results

Results for synthetic data with missingness Setting B (pattern-dependent) and Setting C (MCAR) can be found in Figures 5 and 6, respectively.

C.2 Results for ADNI and SUPPORT

A figure illustrating the performance on SUPPORT with varying data set size is given in Figure 7. Table 4 presents the MSE score

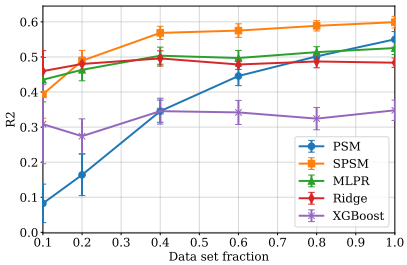

Figure 5: Performance on simulated data Setting B. Error bars indicate standard deviation over 5 random data splits. The complete data set has $n = 10000$ samples.

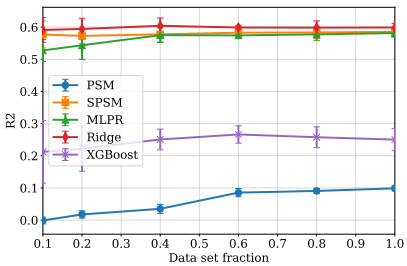

Figure 6: Performance on simulated data Setting C (MCAR). Error bars indicate standard deviation over 5 random data splits. The complete data set has $n = 10000$ samples.

Linear Methods	**MSE**
ADNI	
Ridge, I_0	0.36 (0.26, 0.46)
XGB, I_μ	0.60 (0.48, 0.74)
MLP, I_μ	0.37 (0.27, 0.47)
PSM	0.50 (0.38, 0.62)
SPSM	0.35 (0.25, 0.45)
SUPPORT	
Ridge,I_0	0.61 (0.56, 0.66)
XGB, I_μ	0.69 (0.63, 0.75)
MLP, I_0	0.44 (0.39, 0.48)
PSM	0.47 (0.42, 0.52)
SPSM	0.47 (0.42, 0.51)

Table 4: Experimental results of regression methods for ADNI and SUPPORT data set.

as an additional performance metric for the regression tasks using ADNI and SUPPORT data. For the MAR setting in the SUPPORT data, we present the results for classification and regression tasks in Table 6 and Table 5. Moreover, the full table of pattern 4 non-zero coefficients with the corresponding missing features is displayed in Table 7.

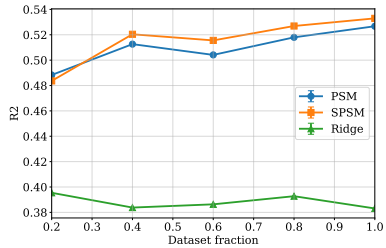

Figure 7: Performance on SUPPORT data for a regression task. Error bars indicate standard deviation over 5 random subsamples of the data.

Regressions	R^2	**MSE**
SUPPORT		
Ridge, I_0	0.38 (0.34, 0.41)	0.62 (0.57, 0.67)
XGB, I_μ	0.27 (0.23, 0.31)	0.73 (0.67, 0.78)
MLP, I_0	0.55 (0.52, 0.58)	0.45 (0.40, 0.49)
PSM	0.51 (0.48, 0.54)	0.49 (0.44, 0.53)
SPSM	0.52 (0.49, 0.58)	0.47 (0.42, 0.51)

Table 5: Experimental results of regression methods for SUPPORT data set MAR.

C.3 HOUSING data

The Ames Housing data set (HOUSING) (De Cock 2011) was compiled by Dean De Cock for use in data science education. The data set describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data set contains 2930 observations and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involved in assessing home values. In this study we used a subset of the features 27 features to describe the main characteristics of a house. Examples of features included are measurements about the land ('LotFrontage', 'LotArea', 'LotShape', 'LandContour', 'LandSlope'), the 'Neighborhood', and 'HouseStyle', when the house was build ('YearBuilt'), or remodeled ('YearRemodAdd'). Moreover, features describing the outside of the house ('RoofStyle', 'Foundation'), technical equipment ('Heating', 'CentralAir', 'Electrical', 'KitchenAbvGr', 'Functional', 'Fireplaces', 'GarageType', 'GarageCars', 'PoolArea', 'Fence', 'MiscFeature'), and information about previous house selling prices and conditions ('MoSold', 'YrSold', 'SaleType', 'SaleCondition'). The numeric features

Classifiers	**AUC**	**Accuracy**
SUPPORT		
LR, I_0	0.82 (0.80, 0.84)	0.75 (0.72, 0.78)
XGB, I_0	0.83 (0.81, 0.85)	0.76 (0.74, 0.78)
MLP, I_0	0.85 (0.84, 0.87)	0.78 (0.76, 0.81)
PSM	0.83 (0.81, 0.85)	0.78 (0.74, 0.80)
SPSM	0.83 (0.81, 0.85)	0.76 (0.73, 0.80)

Table 6: Experimental results of classifiers for SUPPORT data with MAR.

No. of missingness pattern	Feature	Δ_m	θ	$\theta + \Delta_m$
Missing features in pattern 0: None				
0	Age	-0.038	0.121	0.082
	EDUCAT	0.014	-0.005	0.009
	APOE4	0.046	-0.010	0.035
	FDG	-0.032	-0.039	-0.071
	ABETA	0.027	-0.000	0.027
	LDELTOTAL	0.051	-0.391	-0.340
	Entorhinal	0.007	-0.131	-0.124
	ICV	0.013	0.093	0.106
	Diagnose MCI	0.078	-0.139	-0.061
	GEN Female	-0.054	0.003	-0.050
	GEN Male	0.000	0.062	0.062
	Not Hisp/ Latino	0.047	-0.114	-0.067
	Married	0.115	-0.159	-0.044
Missing features in pattern 1: FDG				
1	Age	-0.052	0.121	0.069
Missing features in pattern 4: ABETA, TAU and PTAU at baseline (bl)				
4	Age	-0.140	0.121	-0.019
	FDG	-0.090	-0.039	-0.129
	Whole Brain	0.000	-0.045	-0.044
	Fusiform	0.016	0.021	0.037
	ICV	0.001	0.093	0.094
Missing features in pattern 10: FDG, ABETA (bl), TAU (bl), PTAU (bl)				
10	APOE4	0.038	-0.010	0.027

Table 7: Full table showing Δ_m in the regression task using SPSM for ADNI.

where standardized and the categorical ones are one-hot-encoded during preprocessing. The HOUSING data set shows 15 different missingness patterns. An exploratory analysis has shown that the house sale prices are somehow skewed, which means that there is a large amount of asymmetry. The mean of the characteristics is greater than the median, showing that most houses were sold for less than the average price. In the classification predictions, we look if the sale prices for a house are above or below the median, while for regression tasks we predict the sale price for a house.

We report the results of the HOUSING data set in Table 9. In classification, on average a high performance over all models, whereas the best performing one, `XGB` achieves an AUC of 0.96 and an accuracy of 0.91. `SPSM` achieves only slightly lower prediction power of 0.95 AUC and 0.88 accuracies than `XGB`. While `LR`, `XGB` and `MLP` depend on mean or zero imputation, `PSM`and `SPSM` are able to achieve comparable results without adding bias to their prediction with high confidence on average. For the HOUSING regression, the validation power suggested $\gamma = 10, \lambda = 100$ for `SPSM`, resulting in an R^2 of 0.64 and an MSE of 0.39. This result is better than for `PSM`(R^2 of 0.58 and MSE of 0.46) and thus demonstrates the benefit of coefficient sharing in `SPSM` compared to no sharing. Although `Ridge`, and `MLP` perform better the differences are only marginal to `SPSM`. The best performing model is the black-box method of `XGB` achieving an R^2 of 0.76 and MSE of 0.27 indicating the non-linearity of the data set.

Pattern number	Number of subjects per pattern	R^2
0	119	0.64 (0.53, 0.75)
1	30	0.30 (-0.10, 0.55)
6	27	0.71 (0.50, 0.92)
10	28	0.71 (0.50, 0.91)
others	≤ 13	undefined or insignificant

Table 8: A minimum sample size is required for `SPSM` to maintain predictive performance

Figure 8: Heatmap visualizing the tradeoff between interpretability and prediction power including different hyperparameter values for γ and λ, expressed by the R^2 using SUPPORT data. Each cell is indicating a γ,λ combination, e.g. 1,100 represents 1 = γ and 100 = λ.

C.4 Analysis of interpretability

By enforcing sparsity in pattern specialization, we ensure that the resulting subset of features is reduced to relevant differences which will foster interpretability for domain experts; `SPSM` allows for more straight-forward reasoning about the similarity between submodels and the effects of missingness. Lipton et al. (2016) provides qualitative design criteria to address model properties and techniques thought to confer interpretability. We will show that `SPSM` satisfies some form of transparency by asking, i.e., *how does the model work?*. As stated in (Lipton et al. 2016), transparency is the absence of opacity or black-boxness meaning that the mechanism by which the model works is understood by a human in some way. We evaluate transparency at the level of the entire model (simulatability), at the level of the individual components (e.g., parameters) (decomposability), and at the level of the training algorithm (algorithmic transparency). First, simulatability refers to contemplating the entire model at once and is satisfied in `SPSM` by it's nature of a sparse linear model, as produced by lasso regression (Tibshirani 1996). Moreover, we claim that `SPSM` is small and simple (Rudin 2019), in that we allow a human to take the input data along with the parameters of the model and perform in a reasonable amount of time all the computations necessary to make a prediction in order to fully understand a model. The aspect of decomposabilty (Lipton et al. 2016) can be satisfied by using tabular data where features are intuitively meaningful. To that end, we use two real-world tabular data sets in the experiments and present the coefficient values for input features in Table 3. More-

Housing

Classification	**AUC**	**Accuracy**
LR, I_μ	0.96 (0.94, 0.98)	0.90 (0.85, 0.95)
XGB, I_0	0.96 (0.94, 0,98)	0.91 (0.87, 0.96)
MLP, I_0	0.96 (0.93, 0.98)	0.90 (0.85, 0.94)
PSM	0.93 (0.90, 0.96)	0.88 (0.83, 0.93)
SPSM	0.95 (0.92, 0.97)	0.88 (0.83, 0.94)
Regression	R^2	**MSE**
Ridge, I_μ	0.68 (0.62, 0.75)	0.35 (0.25, 0.44)
XGB, I_0	0.76 (0.70, 0.81)	0.27 (0.18,0.35)
MLP, I_0	0.64 (0.58, 0.71)	0.39 (0.29, 0.49)
PSM	0.58 (0.50, 0.65)	0.46 (0.35, 0.57)
SPSM	0.64 (0.57, 0.71)	0.39 (0.29, 0.49)

Table 9: Experimental results of classification and regression methods for HOUSING data set.

over, one can choose to display the coefficients in a standardized or non-standardized way to provide even better insights. The comprehension of the coefficients depends also on domain knowledge. Finally, algorithmic transparency is given in SPSM, since in linear models, we understand the shape of the error surface and have some confidence that training will converge to a unique solution, even for previously unseen test data. Additionally, Henelius, Puolamäki, and Ukkonen (2017) claims that knowing interactions between two or more attributes makes a model more interpretable. SPSM shows in $\theta + \Delta$ the coefficient specialization between the main model and the pattern-specific model and therefore reveals associations between attributes.

Paper III

Learning replacement variables in interpretable rule-based models

Lena Stempfle, Fredrik D. Johansson

Manuscript in preparation

Learning replacement variables in interpretable rule-based models

Lena Stempfle, Fredrik D. Johansson

Abstract

Rule models are favored in many prediction tasks due to their interpretation using natural language and their simple presentation. When learned from data, they can provide high predictive performance, on par with more complex models. However, in the presence of incomplete input data during test time, standard rule models' predictions are undefined or ambiguous. In this work, we consider learning compact yet accurate rule models with missing values at both training and test time, based on the notion of replacement variables. We propose a method called `MINTY` which learns rules in the form of disjunctions between variables that act as replacements for each other when one or more is missing. This results in a sparse linear rule model that naturally allows trade-off between interpretability and goodness of fit while being sensitive to missing values at test time. We demonstrate the concept of `MINTY` in preliminary experiments and compare the predictive performance to baselines.

1 Introduction

Rule-based models, such as risk scores, rule lists, and linear rule models, are favored in prediction problems and domains where interpretability is a concern (Margot and Luta, 2021; Wei et al., 2019; Fürnkranz et al., 2012). For example, clinical scoring systems are defined using a small number of rules with associated points that add up to a score, indicating, e.g., the risk of mortality for a patient (Knaus et al., 1991). In the same domains, it is common for some variables used in rules to be unobserved at the time of prediction, due to varying tool availability, examination protocols, or heterogeneous data sources Madden et al. (2016). Despite this, most rule-based models lack built-in principled ways for making predictions with missing values.

Approaches to prediction with incomplete data, include imputation (Rubin, 1976), Bayesian modeling, fallback default rules (Chen and Guestrin, 2016), weighted estimating equations Ibrahim et al. (2005) and prediction with missingness indicators (Le Morvan et al., 2020). Drawbacks of existing methods are that they are specific to a non-interpretable model class or that they reduce the interpretability of rule-based models by relying on auxiliary models which may not be interpretable (imputation, estimation weighting) or on parameters associated with missingness itself (default rules, missingness indicators) (Stempfle and Johansson, 2022).

If there is redundancy in the covariates set, where two variables have similar associations to the outcome, we may not need to observe both of them to predict accurately. Instead, redundant variables A and B could be used as replacements for each other when one of them is missing: "If A is not available, use B", or "If B is not available, use A".

Below we show an example of rules which illustrate how replacement variables can be used in the context of linear rule models with binary covariates; if at least one variable in each rule is observed and active, the prediction is the same whether other variables in the rule are missing.

$$\begin{aligned}\textbf{prediction} = &\ \text{Coefficient}_1(\text{Variable}_1 \text{ OR } \text{Variable}_2)\\ &+ \text{Coefficient}_2(\text{Variable}_3 \text{ OR } \text{Variable}_4)\\ &+ \text{Coefficient}_3(\text{Variable}_5)\end{aligned}$$

Using replacement variables also avoids direct dependence on imputation or missingness indicators.

In this project, we aim to *learn replacement variables for missing values at test time using a rule-base interpretable model.* Replacements for unobserved variables should be learned during the training phase and then retrieved at test time. We propose a new methodology, `MINTY` which utilizes replacement variables, defined by disjunctions of literals, in generalized linear rule models. Replacement variables indicate which features can be alternatively used in situations where the original feature was not measured. In addition, they ensure a comparable predictive power to their original counterpart.

Contributions. Our contributions can be summarized as follows: 1) We propose an algorithm for optimizing generalized linear rule models to learn disjunctions of replacement variables as its rules. This technique, which we call `MINTY`, does not rely on imputation or missingness indicators. 2) We optimize `MINTY` using iterated constrained optimization by

adding conditions on the missingness of replacement variables to the column generation approach of Wei et al. (2019). 3) We perform empirical experiments comparing `MINTY` to baseline models on synthetic data and illustrate that our formulation achieves comparable prediction performance but benefits from the flexibility to handle missing values at test time and provide interpretable results for domain experts.

2 Prediction with missing values at test time

We consider a supervised learning problem of predicting a continuous outcome of interest $Y \in \mathbb{R}$ based on a vector of d input features $X = [X_1, ..., X_d]^\top$. In our setting, features may be missing both at training time and at test time, as indicated by a random *missingness mask* $M = [M_1, ..., M_d]^\top \in \{0,1\}^d$ such that $M_j = 0$ if X_j is observed and $M_j = 1$ otherwise. We let $\tilde{X} \in (\mathbb{R} \cup \{\texttt{NA}\})^d$ indicate the partially observed feature vector, with `NA` indicating missingness.

We are given a training data set of samples (x_i, m_i, y_i) for $i = 1, \ldots, n$, drawn i.i.d. from a distribution p, with $x_i = [x_{i1}, ...x_{id}]^\top$ the feature vector of sample i with missing values, and m_i, y_i defined analogously. For convenience, we let $\mathbf{X} \in (\mathbb{R} \cup \{\texttt{NA}\})^{n \times d}, \mathbf{M} \in \{0,1\}^{n \times d}, \mathbf{Y} \in \mathbb{R}^{1 \times d}$ denote data matrices of features, missingness masks and outcomes for all observations, respectively.

We assume that all features X_j represent logical literals, taking values in $\{0,1\}$, where $X_{ij} = 1$ represents that literal j is true for observation i. For instance, in a health care example, feature j may represent the literal $Age \geq 70$ and a patient i that is 73 years old would have the observation $x_{ij} = 1$. There are standard ways to transform observations of continuous and categorical variables to such a representation, such as discretization by quantiles and dichotomization, see e.g., Rucker et al. (2015).

Our goal is to predict Y *under missingness* M in X using functions $f : (\mathbb{R} \cup \{\texttt{NA}\})^d \to \mathbb{R}$, with minimum risk with respect to the squared loss on p,

$$\min_f R(f), \text{ where } R(f) \coloneqq \mathbb{E}_{\tilde{X},Y \sim p}[(f(\tilde{X}) - Y)^2] \,. \quad (1)$$

We assume that features and their missingness have the same distribution at test time as during training.

A common strategy for learning and prediction with missing values is to impute unobserved variables based on observed ones Rubin (1976), and proceed as if no values were missing in the first place. However, when missingness itself depends on unobserved values—variables are missing not-at-random (MNAR)—this strategy is suboptimal, in general, Jamshidian and Mata (2007).

In our setting, under the assumption that Y has centered, additive noise, $Y = g(X, M) + \epsilon$ where $\mathbb{E}[\epsilon] = 0$, the Bayes-optimal predictor of Y is $f^* = \mathbb{E}[Y = 1 \mid \tilde{X}, M]$ Morvan et al. (2021), which depends directly on the missingness mask M itself. However, when interpretability is wanted, letting models include features such as "Age is missing" may be undesirable. Next, we propose a rule-based interpretable model which is sensitive to test-time missingness but avoids direct dependence on missingness indicators.

3 Methodology

We propose `MINTY`, a linear rule model for learning replacement variables when values expected by the model may be unobserved at test time. Our goal is to obtain small, interpretable models with high predictive performance when inputs are incomplete. We first describe the model class and then show how we solve the regression task using constrained optimization.

`MINTY` is a generalized linear rule model (Wei et al., 2019) with three main components:

1. *Rule definitions* $z_{\cdot k} \in \{0,1\}^d$, for rules $k = 1, ..., K$, which define logical rules in terms of d features (literals)
2. *Rule activations* $a_{ik} \in \{0,1\}$, which indicate whether individual $i = 1, ..., n$ satisfies rule k
3. *Rule coefficients*, $\beta = [\beta_1, ..., \beta_K]^\top \in \mathbb{R}^K$, where β_k relates rule k to the predicted outcome. By letting rule 1 always be true, β_1 takes the role of an intercept.

`MINTY` handles missing values by making predictions based on rules formed as *disjunctions* of literals, such as "(Age $>$ 20) or (Female)". If the value of "Age" is missing, the rule depends only on the value of "Female". If none of the features in a disjunction is observed, the rule is inactive—acting like zero-imputation. To prevent this from happening, at training time, we require that, for each observation and each rule, at least one literal is observed. We formalize the `MINTY` model as follows.

Given an observation x_i, with missingnenss, let $\bar{x}_i$ denote its zero-imputation, $\bar{x}_i = x_i \mathbb{1}[x_i \neq \texttt{NA}]$. Then, define the activation of rule k for x_i to be,

$$a_{ik} = \bigwedge_{j=1}^{d} z_{jk} \bar{x}_{ij} = \max_{j \in [d]} z_{jk} \bar{x}_{ij}$$

where $z_{jk} = 1$ indicates that literal (feature) j is included in disjuction (rule) k. Given such activations, the prediction for an input x_i is made as

$\hat{y}_i = \sum_{k \in S} \beta_k^\top a_{ik}$, where S denotes the set of disjunctions under consideration, defined by indicators z_{jk}. We aim to find both a set of rules S and coefficients β that minimize the regularized empirical risk,

$$\min_{\beta, S} \frac{1}{n} \sum_{i=1}^{n} (\sum_{k \in S} \beta_k a_{ik} - y_i)^2 + \sum_k \lambda_k |\beta_k|, \qquad (2)$$

with an ℓ_1-penalty $\lambda_k |\beta_k$ for including rule k. By choosing λ_k, we can control the number and size of rules used by the model.

When generating the rule set, we restrict S to only include rules where at least one of the variables in each rule k is measured for each subject i.

3.1 Optimization

By letting S be the set of all possible disjunctions $\mathcal{K} = \{0, 1\}^d$, our learning problem (2) reduces to a LASSO problem with known solvers, but with a number of rules and coefficients growing exponentially in d. Even for moderate-size problems, it would be intractable to enumerate all of them. Instead, we follow the column-generation strategy by Wei et al. (2019), which intelligently searches the space of disjunctions and builds up $S \subseteq \mathcal{K}$ incrementally.

The idea is to first solve a restricted problem with a small set of candidate rules S_0, in our case just the intercept rule, and then iterative adding new candidates based on the optimal dual solution of the restricted problem. A rule is selected based on the marginal benefit (or partial derivative) of introducing it to the restricted problem. If the partial derivative for the most promising column is non-negative, the procedure terminates. We modify their approach by requiring that each added rule has at least one observed feature for each observation in the training set.

Given a current set of disjunctions S and coefficients β, a new rule is added to S by finding a disjunction that can explain the largest part of the residual of the current model, $\mathbf{R} = \mathbf{A}\beta - \mathbf{Y}$, where $\mathbf{A} = [a_{1\cdot}, \ldots, a_{n\cdot}]^\top$ is the matrix of rule assignments for all observations $i = 1, \ldots, n$ in the training set. Wei et al. (2019) show that such a rule z may be found by solving the following optimization problem for both signs of the first term in the objective.

$$\begin{aligned} \underset{\substack{a \in \{0,1\}^n \\ z \in \{0,1\}^d}}{\text{minimize}} \quad & \pm \frac{1}{2n} \sum_{i=1}^{n} r_i a_i + \lambda_0 + \lambda_1 \sum_{j=1}^{d} z_j \\ \text{subject to} \quad & a_{i,k} = \sum_{k=1}^{K} \max(x_{ij} z_j) \\ & \forall i, k : \sum_j (1 - M_{ij}) z_j \geq 1 \end{aligned} \qquad (3)$$

The first constraint in (3) makes sure that rule activations a_{ik} correspond to a disjunction of literals x_{ij} as indicated by z_j. For the second constraint, we require that, for all rules, at least one of the included literals $j : z_j = 1$ is observed for every individual i. To find an approximate solution to (2), we start with a subset S_0 of rules, solve (2) with respect to β for this set, and compute the residual $\mathbf{R}$ for the current model. Then, repeatedly, a single rule is added to S based on maximizing its correlation with the residual $\mathbf{R}$, solving (3), and the coefficients β are refit. When no rule can be found with a negative solution to (3), the algorithm terminates and the coefficients β are refit one last time.

4 Experiments

We evaluate the proposed `MINTY` model[1] on synthetic data, aiming to answer two main questions: How does the accuracy of `MINTY` compare to baseline models; How do replacement variables affect performance and interpretation?

Experimental Setup In the column generation subproblem, to find values for β, given rule definitions S, we use the `LASSO` implementation in scikit-learn (Buitinck et al., 2013) with covariate weighting to achieve variable-specific regularization strength. We iteratively add variables to S by optimizing (3) using Gurobi, a general-purpose optimization solver Gurobi Optimization, LLC (2023).

The objective function regularizes each rule $z_{\cdot k}$ with strength $\lambda_k = \lambda_0 + \lambda_1 \|z_{\cdot k}\|_0$, penalizing high numbers of unmeasured literals per rule. The values of λ_0 range within $[10^{-5}, 0.5]$ and we set $\lambda_1 = 0.2 \cdot \lambda_0$.

We run two `MINTY` variations, where one includes missingness in the input data and uses zero-imputation, therefore learning disjunctions but disregarding the missingness constraint in (3).

Baseline models We compare the two `MINTY` variations, to baseline algorithms in a regression task with synthesized data. The baselines are ℓ_1-regularized linear regression models (`LASSO`) applied to data imputed with either zero imputation or single imputation by chained equations, implemented as the IterativeImputer in scikit-learn (Buitinck et al., 2013), but without posterior sampling.

Data We use simulated data to illustrate the process of generating replacement variables focusing on predictive performance and interpretablity. We sample

[1] Code to reproduce experiments is available at `https://github.com/Healthy-AI/minty`.

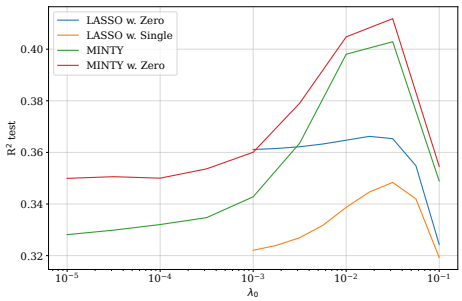

Figure 1: Performance on simulated data. The full data set has $n = 200$ samples and was generated over 100 turns.

$n \times d$ independent binary input features such that $X_{ij} \sim \text{Bernoulli}(p)$, with $p = 0.3$. The outcome Y is given by a disjunctive linear rule model of S, without noise (see Table 1). We limit our experiments to the missing-at-random (MAR) mechanism given by Mayer et al. (2019) for now. To give the rules meaningful descriptions we constructed column names based on the features available in the Alzheimer's Disease Neuroimaging Initiative (ADNI)[2] database.

4.1 Results

We review our results for the regression task for predictive performance and comment on the interpretablity aspects of rule-based models.

In Figure 1 we compare predictive performance in terms of the coefficient of determination (R^2) on the test set, averaged across multiple draws of synthetic data, against the regularization strength indicated by the λ_0. We have fit four models namely `MINTY`, `MINTY` with zero imputation, and `LASSO` models with either zero or single imputation by chained equations as baseline models. We find that the `MINTY` models perform better than the baseline models since they most likely benefit from their nonlinear function class. A potential reason why `MINTY` with zero-imputation performs better than the `MINTY` including missing values is that the missingness constraint in (3) is always inactive for $\texttt{MINTY}_{zero}$. This leads to the conclusion that $\texttt{MINTY}_{zero}$ has more rules in the model class to choose from. The constraint to improve interpretablity seems to introduce some cost on `MINTY` when learning rules. Further investigation is needed and we plan to run more experiments with more data, perform hyperparameter tuning, and will investigate different missingness mechanisms.

[2] `http://adni.loni.usc.edu`

Table 1: Customized rule sets for predictions based on the ground true rule set S

Rules	Coeff.
(Age≥70) OR (Sex= $female$)	+2
(Heart rate≥120BPM) OR (Education_{low})	+3
(Education_{low}) OR (Prior AD diagnosis)	+2
(Age≥70) OR (Heart rate≥120BPM)	−5
Intercept	+0

Table 2: Learned rules sets and the corresponding coefficients

Rules	Coeff.
(Age≥70) OR (Heart rate≥120BPM) (Education_{low}) OR (Prior stroke) OR (Prior AD diagnosis)	−1.96 +1.56
Intercept	+1.54
Prediction	−0.41

Customized rule sets Next, we present descriptions for individual instances and describe the rules relevant to them while the coefficients sum up to the prediction made by the model (Table 1). In the context of the experimental data, this means that variables that are not measured are removed from the rules, and the coefficients of the rules that become equal due to the removed variables are summed up. We implement this by adding a parameter *only_active_rules* indicating that the description contains only features that are "1", i.e., apply to that instance. The simple representation as in Table 1 supports domain experts, such as clinicians to make use of `MINTY` in their decision-making.

In addition, we show an example of learned rules and compare them to the ground truth rules (Table 1) from the generated data. We can interpret the results in Table 2 by saying that the rule *(Age≥70) OR (Heart rate≥120BPM)* has a coefficient of −1.96, while the true coefficient for this rule is also negative but higher in value (-5) (see last row of rules in Table 2). The model also learns a rule including three literals and a coefficient of +1.56. There is no exact matching rule in the ground truth but $education_{low}$ *or prior stroke occur* together with a coefficient of +2, being somewhat close to the true coefficient.

5 Related work

In this work, we focused on developing an interpretable rule-based model when aiming to predict with missingness at test time by identifying variables that may act as replacements for each other. The literature on interpretable machine learning models con-

tains other methods where learning replacements variables might be beneficial for prediction with test-time missingness. For example, risk scores provide a way to add interpretability in applications with domain-specific constraints (Ustun and Rudin, 2019). They comprise simple logistic-linear models but depend on imputation for dealing with missing data. The tree-based model XGBoost Chen and Guestrin (2016) offers an alternative by assigning default paths through the trees for missing variables.

6 Discussion and Conclusion

In this work, we studied prediction with missingness in situations where groups of variables are correlated and have similar relationships to the outcome of interest, but one or more may be missing at training and test time. We proposed the rule-based interpretable model `MINTY` for learning replacement variables for linear combinations of decision rules. Initial empirical results show that `MINTY` achieves comparable predictive performance to the baseline models and allows us to reason about predictions given incomplete data sets. If meaningful variables are present, they can be evaluated by a domain expert for intuitiveness as a replacement.

There are some limitations that suggest interesting future work. First, more experiments are needed to generalize our results, including several baseline models and data sets. Next, finding features with similar prediction power is only one way to find replacements. Therefore, we could introduce conceptual closeness e.g. similar way to retrieve a measurement using feature concepts associations by domain knowledge into the constraints (Lage and Doshi-Velez, 2020). Moreover, we will investigate the model behavior when no substitute can be found for rare but highly predictive covariates, and describe the properties of such a situation. Furthermore, if only a few features are observed or the model is uncertain in its prediction, it should not make a prediction, but refer to a standard rule that applies to many individuals, or refer to one individual. The latter is referred to as "learning to defer" (Mozannar et al., 2022) and should be applied in especially sensitive areas, such as health care, where it is necessary that no false sense of security is conveyed when the model is uncertain. Finally, if there is high variance in the missingness mechanism when the training data grows, it becomes increasingly unlikely to find a small group of replacement variables among which at least one is always measured. To remedy this, we want to examine combining our approach with imputation, but relying on it as little as possible. As a compromise, instead of having the constraint that every rule be measured for every training sample, we can minimize the number of rules where a variable is not measured.

Acknowledgements This work was supported in part by WASP (Wallenberg AI, Autonomous Systems and Software Program) funded by the Knut and Alice Wallenberg foundation.

Societal impact Our work aims to contribute to the efforts in machine learning for healthcare, where a major goal is to accelerate the deployment of machine learning algorithms to clinical settings.

References

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Johannes Fürnkranz, Dragan Gamberger, and Nada Lavrač. *Foundations of rule learning*. Springer Science & Business Media, 2012.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023. URL `https://www.gurobi.com`.

Joseph G Ibrahim, Ming-Hui Chen, Stuart R Lipsitz, and Amy H Herring. Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100 (469):332–346, 2005.

Mortaza Jamshidian and Matthew Mata. Advances in analysis of mean and covariance structure when data are incomplete. In *Handbook of latent variable and related models*, pages 21–44. Elsevier, 2007.

William A Knaus, Douglas P Wagner, Elizabeth A Draper, Jack E Zimmerman, Marilyn Bergner, Paulo G Bastos, Carl A Sirio, Donald J Murphy, Ted Lotring, Anne Damiano, et al. The apache iii prognostic system: risk prediction of hospital mortality for critically iii hospitalized adults. *Chest*, 100 (6):1619–1636, 1991.

Isaac Lage and Finale Doshi-Velez. Learning interpretable concept-based models with human feedback. *arXiv preprint arXiv:2012.02898*, 2020.

Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, and Gaël Varoquaux. Neumiss networks: differentiable programming for supervised learning with missing values. *Advances in Neural Information Processing Systems*, 33:5980–5990, 2020.

Jeanne M Madden, Matthew D Lakoma, Donna Rusinak, Christine Y Lu, and Stephen B Soumerai. Missing clinical and behavioral health data in a large electronic health record (ehr) system. *Journal of the American Medical Informatics Association*, 23(6):1143–1149, 2016.

Vincent Margot and George Luta. A new method to compare the interpretability of rule-based algorithms. *AI*, 2(4):621–635, 2021.

Imke Mayer, Aude Sportisse, Julie Josse, Nicholas Tierney, and Nathalie Vialaneix. R-miss-tastic: a unified platform for missing values methods and workflows, 2019. URL `https://arxiv.org/abs/1908.04822`.

Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. What's a good imputation to predict with missing values?, 2021. URL `https://arxiv.org/abs/2106.00311`.

Hussein Mozannar, Arvind Satyanarayan, and David Sontag. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5323–5331, 2022.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Derek D Rucker, Blakeley B McShane, and Kristopher J Preacher. A researcher's guide to regression, discretization, and median splits of continuous variables. *Journal of Consumer Psychology*, 25(4): 666–678, 2015.

Lena Stempfle and Fredrik Johansson. Sharing pattern submodels for prediction with missing values. *arXiv preprint arXiv:2206.11161*, 2022.

Berk Ustun and Cynthia Rudin. Learning optimized risk scores. *J. Mach. Learn. Res.*, 20(150):1–75, 2019.

Dennis Wei, Sanjeeb Dash, Tian Gao, and Oktay Gunluk. Generalized linear rule models. In *International Conference on Machine Learning*, pages 6687–6696. PMLR, 2019.