Explaining Deep Survival Analysis Models for Heterogenous Data

MORITZ WAGNER

Master Thesis

Studiengang: Statistik mit wirtschafts- und
sozialwissenschaftlicher Ausrichtung
LMU München
Institut für Statistik

Betreuer: Dr. Sebastian Pölsterl, Prof. Dr. Christian Wachinger

München, 05.05.2021

Abstract

The aim of this work is to predict the progressions from mild cognitive impairment (MCI) to Alzheimer's disease (AD) based on heterogeneous data provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack Jr et al., 2008). We consider a slice of the coronal plane of the 3D MRIs of the brain and tabular biomarker data. To pursue this, we leverage state of the art methods in the area of Deep Survival Analysis. While the predictive performances are already promising, it is still a challenging task to integrate them into medical diagnosis systems. This is due to a lack of transparency and interpretability of these algorithms and their predictions (Singh et al., 2020). To enhance interpretability, Shapley values (Shapley, 1953) depict a prominent choice to determine which structures in the brain are responsible for an either accelerated or decelerated disease progression. Shapley values, however, rely on a specified *baseline* against which the considered MRI and its corresponding prediction is compared. To identify a suitable baseline has turned out to be challenging (Sturmfels et al., 2020). The literature refers to that as the *baseline selection problem* (Shih et al., 2020). We argue that the optimal *baseline* must represent a meaningful and contrasting example to the original MRI. If the original MRI contributes to an accelerated/decelerated progression, the *baseline* must contribute to a decelerated/accelerated progression. To ascertain meaningfulness, we require the baseline to represent a realistic sample which differs from the original MRI only in those features that are directly linked to AD progression. The latter criterion prevents us from selecting a sample from the available data, but rather requires to synthetically generate a hypothetical MRI. To pursue this, we rely on the general ideas of image-to-image translation. We propose a novel and unique framework - *the baseline generator* - that allows to uniquely identify an optimal *baseline* for each MRI. While similar methods have already been proposed for binary classification (Bass et al., 2020), our proposed framework is applicable to survival analysis. Within this work it will become evident why this general conceptual transfer is essential. Due to the limited scope of this thesis, we refrain from applying the *baseline generator* framework on the ADNI data. Instead, we identify a unique simulation setting to fully verify the functioning of the established framework. By doing so, we can conclude that the framework fills a non-negligible gap for making survival times predictions - based on unstructured image data - interpretable. We argue that this serves as a decisive step to enhance interpretability of predictions of AD progression.

Contents

1 Introduction **1**

2 Survival analysis **5**
2.1 Basic concepts 5
2.2 Likelihood for (semi-) parametric estimation 6
2.3 Cox-PH model 7
2.4 Multimodal Deep Cox Proportional Hazards model 9
2.5 The orthogonalization trick 10

3 Attribution methods **11**
3.1 Attribution methods - a meta level consideration 11
3.1.1 Definition of meta-level criteria 11
3.1.2 Ranking of meta-level criteria 12
3.1.3 Classes of interpretation methods - a comparison 14
3.2 Requirements on attribution methods 16
3.3 Shapley values 18
3.3.1 Integrated Gradients 19
3.3.2 Sampled Shapley values 20

4 The baseline generator **22**
4.1 The baseline selection problem 22
4.2 The optimal baseline 24
4.3 Identification of the optimal baseline 27
4.4 The baseline generator in the context of current research 32

5 Experiments **35**
5.1 Experimental setup 35
5.2 Experimental strategy 35
5.3 ADNI 36
5.3.1 Data 36
5.3.2 Training 37
5.3.3 Results 38
5.4 Simulations 39
5.4.1 Data 39
5.4.2 Training 42
5.4.3 Results 44

6 Discussion **48**
6.1 Baseline images and interpretability 48
6.2 Generalizability to further survival models 49

6.3 External Validity 53

7 Conclusion **54**

References **55**

Appendices **60**

A Architectural design 60

A.1 ADNI 60

A.2 Simulations 60

B Main hyperparameters 61

C Main results 61

1 Introduction

For people over 70 years of age it is common to suffer from cognitive impairments (Knopman & Petersen, 2014). While some patients merely suffer from mild cognitive impairments (MCI) and therefore are not limited in their everyday activities, a non-negligible share of affected patients develop Alzheimer's disease (AD), the most common form of dementia. In such cases, the cognitive and motor abilities are heavily limited to such a degree that an independent living is impossible (McKhann et al., 2011). While it is understood that MCI represents a pre-dementia stage, it remains challenging to fully understand why some patients remain stable and others progress to dementia. To pursue this, it is crucial to correctly diagnose whether a patient with MCI actually suffers from dementia but has not yet developed to the severe stage or merely shows cognitive impairment symptoms. To correctly diagnose the neurodegenerative disorder, experts rely on biological biomarkers which arguably serve as suitable predictors for the progression of the disease. To enable researchers to identify strong predictors for conversion from MCI to AD, the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack Jr et al., 2008) - a longitudinal study started in 2003 - collects clinical and biomarker data as well as MRI scans of the brain from patients affected by MCI or AD.

A vast amount of studies have already investigated the problem of predicting the conversion from MCI to AD. The progression to AD is often modeled as a binary classification task, where class 0 means, the corresponding patient did not yet convert to AD and 1 otherwise (Moradi et al., 2015; Tong et al., 2016). The length of the observation period is fixed, so that only conversions that took place within a specified time span are considered. For this approach to be valid, two major assumptions must hold. Firstly, it must be assumed that once a patient has remained stable for the fixed time span, she will remain stable and will not convert to AD any time later. This assumption, however, is highly restrictive and if violated implies that the patient has been erroneously considered stable, even though conversion took place. While there exists evidence that some patients remain stable (Clem et al., 2017), there is neither an empirical nor a theoretical justification for arguing that after a certain time stability is guaranteed. Secondly, it must be assumed that any dropouts can be considered random. If this assumptions holds, those patients can simply be ignored. If, however, the dropout is not random w.r.t. progressing to AD, the binary classification approach is not a valid choice. Again, we cannot observe for any dropout whether she converted to AD or remained stable. In summary, the binary classification approach assumes that every patient and the development of her disease is fully observed. If, however, some of the patients can only be partially observed, the validity of the approach breaks.

To avoid to rely on these restrictive assumptions, the need for exactly determining whether

conversion took place or not has to be given up. Given a patient who has not progressed to AD until the end of the study, it is not justified to argue that she will remain stable. By contrast, reasoning about conversion probabilities or conversion time is indeed justified - e.g. a statement that the probability of conversion is comparably low once a certain time has been passed is valid. This type of inference can be obtained by applying standard techniques from the survival analysis literature, where conversion probabilities are predicted for each patient over time. This also allows to reason about conversion probabilities for partially observed observations which is prohibited within binary classification. Besides a theoretical justification for survival analysis, there is also a practical motivation for refraining from binary classification and adhering to survival analysis. So far there is no cure for AD, but there are effective treatments for delaying and decelerating the progression of cognitive and functional decline (Rountree et al., 2013; Yiannopoulou and Papageorgiou, 2013). For the treatments' effectiveness, however, it is crucial to start treatments at an early stage of cognitive impairments (Dubois et al., 2010; Sperling et al., 2011). To pursue this, we must predict conversion times to identify when introducing a treatment is sensible. Binary classification is not suitable, as predictions only indicate whether conversion occurs or not, but not *when* it occurs. There is extensive work on leveraging both tabular clinical data and MRI scans for predicting survival times (e.g. Platero and Tobar, 2020). However, these approaches are mostly based on extracting and engineering hand-crafted features from the raw MRIs which are then fed into a linear survival model. Recent studies have proposed methods that allow the training of a multimodal approach in an end-to-end fashion (Kopper et al., 2020; Nakagawa et al., 2020; Pölsterl et al., 2019). However, instead of training on the raw MRIs of the brain, they use 3D anatomical shape representations (point clouds) or the volume extracted from the 3D MRI. Within this study, we will build upon the approach introduced by Pölsterl et al. (2019), except that we will use 2D slices from the coronal plane of the raw MRIs.

While the predictive performances of the proposed methods are already promising, it is still challenging to integrate them into medical diagnosis systems. The reluctant acceptance of these algorithms by regulators, doctors and patients is mainly driven by the lack of transparency and interpretability of these algorithms and their predictions (Singh et al., 2020). To increase acceptance and trust, a variety of explainability methods have been introduced, the class of attribution methods being one of the most prominent ones (Montavon, 2019). In general, these methods are simple to compute and their results are intelligible. The attribution methods aim at determining an importance score for each pixel in the MRI which are then visualized by *attribution maps*. Thereby, the recipient yields a visual understanding of which regions in the MRIs were the driving forces for the prediction. Besides an intelligible interpretation method, it is evident to also demand robust and unambiguous results derived by the attribution methods. A set of theoretical axioms have been established to flash out the notion of robustness and unambiguity. In

this context, these axioms allow to infer which types of interpretations are admissible for each attribution method (Friedman, 2004). It was shown that Shapley values (Shapley, 1953) represent the unique method that satisfies all axioms and thus allows for the highest degree of interpretability. Shapley values determine the marginal contributions of each pixel to the overall prediction by inherently comparing the original prediction to a baseline prediction. The narrative is as follows: For each pixel, the corresponding baseline value shall simulate *missingness* so that the derived contribution reflects the impact on the prediction if the pixel was missing. In other words, the baseline serves as a reference point against which the original prediction is compared. To identify a baseline that indeed represents *missingness* has turned out to be challenging (Sturmfels et al., 2020). The literature refers to that as the *baseline selection problem* (Shih et al., 2020).

In the domain of predicting AD progression, we can pursue the objective of simulating *missingness* as follows: For an MRI that corresponds to a *sick/healthy* patient, we seek to identify a baseline MRI that represents a *healthy/sick* patient. The structural differences between the original and the baseline MRI reflect *missingness* - either *sickness* or *healthiness* is missing. Thereby, to fully reflect *missingness*, the optimal baseline contributes neither to an accelerated nor to a decelerated AD progression which is fulfilled when the survival model predicts a zero risk score for the baseline MRI. But for the baseline to represent *missingness* is not a binding criterion, since any baseline is potentially optimal as long as it represents a semantically meaningful reference point. Aligned with the objective to predict survival times, we can require the baseline to represent a specific quantile of the predicted survival times - e.g. the median survival time - against which the original MRI is compared. Then, we can infer which pixels contributed to a more decelerated/accelerated AD progression, when the predicted survival time of the MRI is before/beyond the specified quantile. To yield a reliable reference point, we require the baseline MRI to only differ from the original MRI in those features that are responsible for AD progression, while all other domain-unspecific characteristics must remain constant. But since the opposite clinical picture of a patient is not observed, we must synthetically generate the unique baseline MRIs. To pursue this, we propose a framework, *the baseline generator*, which allows to uniquely identify for each MRI a corresponding meaningful baseline. Thereby, we rely on the general concepts of image-to-image translation (e.g. Isola et al., 2017). Previous works have already proposed frameworks to translate MRIs with MCI to MRIs with AD and vice-versa (Bass et al., 2020; Baumgartner et al., 2018). However, these methods are not applicable in survival analysis as they rely on a binary classification setting. Hence, by introducing *the baseline generator*, we are the first to transfer the concept of image-to-image translation from the classification domain to the survival analysis domain. Beyond that, we are the first to discuss the *baseline selection problem* in the context of survival analysis. This is a significant contribution to enhancing interpretability in the field of predicting AD progression.

The remaining is structured as follows: in chapter 2, we will provide a detailed explanation on how AD progression can be explicitly modelled by means of a Cox-PH model (David et al., 1972). This serves as a preliminary to understand how the loss function is derived to train on the unstructured MRIs and the clinical tabular data, jointly. Subsequently, the *orthogonalization trick* is elaborated which is essential to avoid identification issues when training in a multi modal setting. Once the Deep Cox-PH model and the *orthogonalization trick* is introduced, chapter 3 focuses on a theoretic derivation of the most suitable interpretation methods w.r.t. the unstructured MRIs. We will then conclude that Integrated Gradients (Sundararajan et al., 2017) as well as sampled Shapley values (Castro et al., 2009) represent the most suitable choices for our purposes. The chapter concludes with a theoretical definition of the chosen methods. Thus completed, in chapter 4 we will formally introduce *the baseline generator* framework in the following steps. First, we point to the *baseline selection problem* and its practical implications. By understanding the inherent problem, we can establish a set of criteria which constitute an optimal baseline. We can then establish an identification strategy for finding the optimal baseline. It will become apparent that our proposed framework embeds well in the current research of image-to-image translation within the medical domain. This will also complete the theoretical section of this study, leading to chapter 5, which will continue with the experimental section. A brief overview of the experimental setup, as well as the experimental strategy will be given. The first experiments are conducted on the data provided by ADNI. Here, the primary focus is put on whether the *orthogonalization* has any impact on the model performance or on the learned coefficients that correspond to the structured part of the model. The second part of the experimental section focuses on the validity of *the baseline generator* framework. We pursue this by training the baseline generator on simulated data. We first visually assess the quality of the generated baseline images, before we evaluate to what extent the generated images represent an appropriate baseline choice for the attribution methods. The latter is pursued by benchmarking two types of attribution maps: those derived from the generated baseline images with those derived from more naive baseline choices. The discussion in chapter 6 deals with three important considerations. First, we will emphasize the stand-alone importance of the generated baseline images in the context of interpretability. In the second part, we will show that the baseline generator framework is easily transferable to variety of other survival models. Lastly, we will discuss the external validity of the proposed framework and thereby, examine whether the framework is transferable to other relevant domains. Finally, we conclude in chapter 7.

2 Survival analysis

Chapter 1 showed why studying AD progression requires techniques from survival analysis. In what follows, we will first elaborate the basic concepts of survival analysis. This will serve as a preliminary to define a general formula for the likelihood for (semi-) parametric estimations on the basis of which we can derive the likelihood for the Cox-PH model (Cox, 1972). Consequently, we can elaborate the approach proposed by Pölsterl et al. (2019) which allows us to train a Cox-PH model on both the tabular data and the unstructured MRIs jointly (multi modal Deep Cox-PH model). This multi modal approach is possible, as the derived likelihood can be leveraged for gradient descent optimization during training (Faraggi & Simon, 1995). Finally, we elaborate the *orthogonalization* trick from Rügamer et al. (2020) which prevents identification issues that otherwise occur when training a Deep Cox-PH model in a multi modal fashion.

2.1 Basic concepts

Hazard rate describes for each time point t the conditional likelihood that a conversion will take place at time t, given that the patient has not converted until time t:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \geq 0 \tag{1}$$

The conditional modelling of event probabilities is crucial, as it cannot be assumed that the observations during the follow-ups of a patient are independently distributed. Further, the hazard rate is an important component to determine the survival function.

Survival function corresponds to the probability that a patient is not converting to AD prior to or at time t. It indicates the minimum time for the patient to convert.

$$S(t) = 1 - F(t) = P(T > t) \tag{2}$$

where $F(t) = P(T \leq t)$ is the cumulative distribution function of T. The survival function can also be written as the integral of the probability density function $f(t)$

$$S(t) = P(T > t) = \int_t^\infty f(t)\, dt \tag{3}$$

where $f(t)$ approximates the probability that conversion will take place at time t. This is the central function, as it allows to reason about conversion probabilities of patients who are only observed until time t. Recall, while we had to make restrictive assumptions for the partially observed patients when applying a binary classification, we can now explicitly model this problem by applying the survival function. We can compute for

every patient at any time a probability of remaining stable until time point t. The survival function can also be written as a function of the cumulative hazard rate $\Lambda(t)$:

$$S(t) = \exp(-\Lambda(t)) \tag{4}$$

whereby $\Lambda(t)$ is defined as:

$$\Lambda(t) = \int_0^t \lambda(u)\, du \tag{5}$$

which reveals a direct relationship between the hazard rate $\lambda(t)$ and $S(t)$. The relationship between $\lambda(t)$, $\Lambda(t)$ and $S(t)$ can be intuitively interpreted in case of discrete conversion times. If $\Lambda(t)$ is large, it follows that the probability $S(t)$ is close to 0. If $\Lambda(t)$ is large, then the respective hazard rates $\lambda(t)$ for time points $t_1 < .... < t_n < t$ must have been comparably large. In short, at any time until t, the patient was likely to convert but somehow did not. The longer the conversion is prolonged, against the odds, the more likely it becomes that conversion takes place at the next time - $S(t)$ decreases.

Censoring one main argument for not analyzing AD progression within the framework of binary classification was the presence of partially observed data. In survival analysis, every subject is considered censored for whom the conversion has not taken place during the observation period. Hence, we do not fully observe the patient. While there are different types of censoring, we will assume that all partially observed patients are right censored. Of all patients who entered the study with MCI, those who are censored either dropped out or did not convert to AD during the study period.

2.2 Likelihood for (semi-) parametric estimation

Given the basic quantities, it remains to be clarified how the likelihood must be specified, to fully consider censored data. In case of no censoring - all conversions from MCI to AD are observed - the standard likelihood for parametric estimations $L = \prod_{i=1}^{n} f(t_i)$ is applicable. While the likelihood contribution $f(t_i)$ is still valid for all subjects were the conversion was observed, we have to adjust the likelihood contribution for those who are censored. We know, by observation, that for those who are censored, conversion did not take place until $t = C_r$ where C_r represents the time of censoring. Information about conversion beyond C_r are not available and therefore, we can only reason about its probability, which is well defined as the survival function evaluated at the time of censoring $S(C_r)$. From that, we can derive the total likelihood for all observations

$$L \propto \prod_{i \in D} f(t_i) \prod_{i \in R} S(t_i) \tag{6}$$

where set D includes all patients for whom exact conversion times are observed and set R subsumes all right censored patients. We can simplify equation 6, by introducing an indicator function δ_i which is $\delta_i = 1$ when for patient i conversion was observed and $\delta_i = 0$, if the patient i is right censored. Then the total likelihood is defined as

$$L \propto \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} \tag{7}$$

We can also rewrite the probability of event times $f(t_i)$ as a function of the hazard rate $\lambda(t_i)$ and the survival probability $S(t_i)$, which yields a different expression of the likelihood

$$L \propto \prod_{i=1}^{n} \lambda(t_i)^{\delta_i} S(t_i) \tag{8}$$

which will be useful to derive the Cox-PH model in the following chapter.

2.3 Cox-PH model

The Cox-PH model (Cox, 1972) allows to evaluate the effect of a set of covariates $(\boldsymbol{x_1}^T, ..., \boldsymbol{x_n}^T)^T =: \boldsymbol{X} \in \mathbb{R}^{n \times p}$ on the hazard of conversion at time t by establishing the following structural assumption

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t) \exp(\boldsymbol{X}\boldsymbol{\beta}) \tag{9}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ are the corresponding coefficients. The hazard rate $\lambda(t|\boldsymbol{X})$ depends on the baseline hazard rate $\lambda_0(t)$ which is identical for all subjects and solely depends on the time t. The risk score $\boldsymbol{X}\boldsymbol{\beta}$ is determined by time constant covariates and therefore, remains constant over time. As the data were collected within a longitudinal study, it is reasonable to question the strong assumption of time constant covariate effects. However, this will not be a subject of discussion within the scope of this work. If the assumption holds, however, the Cox-PH model comes with advantageous properties w.r.t. interpreting the effects. We can interpret the coefficients $\boldsymbol{\beta}$ as multiplicative factors on the hazard rate, e.g. if the covariate x_1 increases by one unit, the risk of conversion at time t increases ceteris paribus by a factor of $\exp(\beta_1)$. By transforming the hazard rate $\lambda(t|\boldsymbol{X})$ on the log scale, we can directly interpret the learned coefficients $\boldsymbol{\beta}$. Given the semi-parametric structural assumption, we can now model the total likelihood from

equation 8 by adhering to the idea of profile likelihood. We first fix $\boldsymbol{\beta}$ and consider the likelihood $L(\lambda_0(t), \boldsymbol{\beta})$ as a function of $\lambda_0(t)$ only, which results in the following

$$
\begin{aligned}
L(\lambda_0(t), \boldsymbol{x}|\boldsymbol{\beta}) &= \prod_{i=1}^{n} \lambda(t_i)^{\delta_i} S(t_i) && (10)\\
&= \prod_{i=1}^{n} \lambda_0(t_i)^{\delta_i} \exp(\boldsymbol{x_i}^T \boldsymbol{\beta})^{\delta_i} \exp(-\Lambda_0(t_i) \exp(\boldsymbol{x_i}^T \boldsymbol{\beta})) && (11)\\
&= \Big[\prod_{i=1}^{m} \lambda_0(t_{(i)}) \exp(\boldsymbol{x_{(i)}}^T \boldsymbol{\beta}) \Big] \exp\Big[- \sum_{j=1}^{n} \Lambda_0(t_j) \exp(\boldsymbol{x_j}^T \boldsymbol{\beta}) \Big] && (12)\\
&=: L_\beta(\lambda_0(t)) && (13)
\end{aligned}
$$

where $j = 1, ..., n$ is the set of indeces for the subjects and $i = 1, ..., m$ represent the set of indices for the discrete survival times $0 < t_{(1)} < .. < t_{(m)}$ where t_0 is set to 0. If we assume discrete survival times, we can leverage that the cumulative baseline hazard rate can be written as $\Lambda_0(t_i) = \sum_{t_{(i)} \leq t_i} \lambda_0(t_{(i)})$ which yields

$$
\begin{aligned}
L_\beta(\lambda_{01}, ..., \lambda_{0m}) &= \underbrace{\prod_{i=1}^{m} \exp(\boldsymbol{x_{(i)}}^T \boldsymbol{\beta})}_{=:c} \left[\prod_{i=1}^{m} \lambda_{0i} \right] \exp\left[- \sum_{j=1}^{n} \Lambda_0(t_j) \exp(\boldsymbol{x_j}^T \boldsymbol{\beta}) \right] && (14)\\
&= c \left[\prod_{i=1}^{m} \lambda_{0i} \right] \exp\left[- \sum_{j=1}^{n} \sum_{t_{(i)} \leq t_j} \lambda_{0i} \exp(\boldsymbol{x_j}^T \boldsymbol{\beta}) \right] && (15)\\
&= c \left[\prod_{i=1}^{m} \lambda_{0i} \right] \exp\left[- \sum_{j=1}^{n} \lambda_{0i} \sum_{j \in R_i} \exp(\boldsymbol{x_j}^T \boldsymbol{\beta}) \right] && (16)
\end{aligned}
$$

where R_i is defined as the riskset at time $t_{(i)}$ which includes all subjects who did not yet convert to AD and have not been censored until time $t_{(i)}$. We can now take the derivative of the log profile likelihood $l_\beta(\lambda_{01}, ..., \lambda_{0m}) = \log L_\beta(\lambda_{01}, ..., \lambda_{0m})$ w.r.t. the discrete baseline hazard rates $(\lambda_{01}, ..., \lambda_{0m})$

$$
l_\beta(\lambda_{01}, ..., \lambda_{0m}) = \log(c) + \sum_{i=1}^{m} \log(\lambda_{0i}) - \sum_{i=1}^{m} \sum_{j \in R_i} \exp(\boldsymbol{x_j}^T \boldsymbol{\beta}) \tag{17}
$$

$$
\frac{\partial l_\beta}{\partial \lambda_{0i}} = \frac{1}{\lambda_{0i}} - \sum_{j \in R_i} \mathbf{exp}(\boldsymbol{x_j}^T \boldsymbol{\beta}) \tag{18}
$$

which yields the profile likelihood estimator of $\lambda_{0i} = \frac{1}{\sum_{j \in R_i} \exp(\boldsymbol{x_j}^T \boldsymbol{\beta})}$ which can be plugged back into equation 16 to yield the partial likelihood which only depends on the covariates $\boldsymbol{x}$ and coefficients $\boldsymbol{\beta}$.

$$L(\boldsymbol{\beta}) \propto \prod_{i=1}^{m} \exp(\boldsymbol{x_{(i)}}^T \boldsymbol{\beta}) \prod_{i=1}^{m} \frac{1}{\sum_{j \in R_i} \exp(\boldsymbol{x_j}^T \boldsymbol{\beta})} \propto \prod_{i=1}^{m} \frac{\exp(\boldsymbol{x_{(i)}}^T \boldsymbol{\beta})}{\sum_{j \in R_i} \exp(\boldsymbol{x_j}^T \boldsymbol{\beta})} = PL(\boldsymbol{\beta}) \tag{19}$$

2.4 Multimodal Deep Cox Proportional Hazards model

As already discussed, ADNI not only collects structured, tabular biomarkers and socio-economic data, but also unstructured MRI scans of the brain. Hence, we want to determine the predictive power of the structured and the unstructured data, jointly. To pursue this, we follow the approach proposed by Pölsterl et al. (2019). Except for the fact that they were analyzing point clouds instead of MRIs, the objective of their study is identical to ours. By training a neural network $f_{\boldsymbol{\theta}} : \mathbb{R}^{n \times c \times h \times w} \to \mathbb{R}^{n \times q}$, we aim to learn a meaningful, lower dimensional representation $(\boldsymbol{u_1}^T, ..., \boldsymbol{u_n}^T)^T =: \boldsymbol{U} \in \mathbb{R}^{n \times q}$ of the raw MRI scans $(\boldsymbol{z_1}^T, ..., \boldsymbol{z_n}^T)^T =: \boldsymbol{Z} \in \mathbb{R}^{n \times c \times h \times w}$, which serve as predictors for the Cox-PH model. Then, we can concatenate the p features from the tabular data with the q features from the latent representation to obtain the predictor

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma} \tag{20}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ are the coefficients that correspond to the structured part $\boldsymbol{X}$ and $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_q)$ represent the coefficients corresponding to the unstructured part $\boldsymbol{U}$. As introduced in chapter 2.3, we can now directly model the hazard rate $\lambda(t|\boldsymbol{X}, \boldsymbol{U})$ by

$$\lambda(t|\boldsymbol{X}, \boldsymbol{U}) = \lambda_0(t) \exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma}) = \lambda_0(t) \exp(\boldsymbol{\eta}) \tag{21}$$

Hence, the structural assumption of the Cox-PH model remains unchanged, except for the fact that the predictor is extended by the latent representation $\boldsymbol{U}$. To train the model in an end-to-end fashion, we can minimize the log-likelihood from equation 19 as proposed by Faraggi and Simon (1995) and rewrite such that the linear predictor includes both, the structured part $\boldsymbol{X}$ and the latent representation $\boldsymbol{U}$ from the unstructured part

$$\arg\min_{\Theta} = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{x_i}\boldsymbol{\beta} + f_{\boldsymbol{\theta}}(\boldsymbol{z_i})\boldsymbol{\gamma} - \log \left(\sum_{j \in R_i} \exp(\boldsymbol{x_j}\boldsymbol{\beta} + f_{\boldsymbol{\theta}}(\boldsymbol{z_j})\boldsymbol{\gamma}) \right) \right] \tag{22}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ denotes the set of all parameters, whereby $\boldsymbol{\theta}$ correspond to the learned weights of neural network $f_{\boldsymbol{\theta}}$ to map the raw MRIs to the lower dimensional representation. $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ correspond to the learned weights of the last linear layer of our network after concatenation.

2.5 The orthogonalization trick

We proposed to jointly estimate the effects derived from the structured part and the unstructured part in the last layer of the network. Yet, the learned coefficients that correspond to the unstructured part may overlap with the learned coefficients from the structured part. If that is the case, the effects are not uniquely identifiable - the latent representation $\boldsymbol{U}$ captures effects that are also present in $\boldsymbol{X}$. Rügamer et al. (2020) discussed this identification problem and proposed to constrain the learned latent representation such that it does not overlap with the features from the structured part. They show that it is sufficient to orthogonalize the latent representation $\boldsymbol{U}$ to the space spanned by the linear features $\boldsymbol{X}$. Two vectors that are orthogonal to each other are also independent, which precludes any overlap. To pursue this, we need to learn the projection matrix $\boldsymbol{P_X}$ to then determine the orthogonal complement $\boldsymbol{P_X} := \boldsymbol{I_n} - \boldsymbol{P_X}$. We can then left multiply the latent representation with the orthogonal complement $\tilde{\boldsymbol{U}} = \boldsymbol{P_X U}$ which results in a new predictor $\boldsymbol{\eta_k} = \boldsymbol{X\beta} + \tilde{\boldsymbol{U}}\boldsymbol{\gamma}$. A more detailed discussion, including a proof of the concept, is provided by Rügamer et al. (2020). We know from linear regression that the projection matrix is defined as $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}$, whereby it is necessary to find a numerically stable computation of the projection matrix. Among others, the Demmler-Reinsch orthogonalization represents a valid approach to that (Ruppert et al., 2003). While the orthogonalization ensures uniquely identifiable estimates, it also decisively enhances the interpretability of the results. The learned effects $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are independent and therefore, they can also be interpreted independently. This, has also direct implications on the interpretation of the unstructured part.

3 Attribution methods

The *orthogonalization trick* represents not only a suitable method to make the structured part interpretable, but also allows to consider the challenge of interpreting the unstructured part in isolation. While disentangled, both parts cannot be interpreted in the same manner, as the learned latent representation has no comprehensible semantic meaning. Hence, interpreting its corresponding coefficients is hardly expressive. Yet, to obtain a high degree of interpretability, it is essential to yield a better understanding of the input-output mapping process that corresponds to the unstructured MRIs. A vast amount of methods has been proposed, and yet there is clearly no unique solution that satisfies all needs for interpretability. Rather, the appropriate choice for a method is highly dependent on the specific needs of the applicant.

In the following, we aim to deduce a small subset of the available methods that arguably fit best for our purposes. We will proceed by assessing the validity of different methods based on a set of meta-level criteria. By doing so, we will conclude that the class of attribution methods represents the most appropriate choice for our specific needs. By adhering to the selection process, it will also become apparent why there is no universal best interpretation method. We will then use an axiomatic set of criteria to select the most promising and robust methods within the class of attribution methods. It will become evident why we opt for Integrated Gradients (Sundararajan et al., 2017) and sampled Shapley values (Castro et al., 2009). The chapter will conclude with a more detailed and theoretical definition of the selected methods. A formal definition of the attribution methods is mandatory to understand *the baseline selection problem* discussed in chapter 4.

3.1 Attribution methods - a meta level consideration

As outlined previously, the following will serve as a preliminary to derive a consistent reasoning for choosing a suitable class of interpretation methods. The considered meta-level criteria are entirely derived from Ras et al. (2018) and eventually complemented by further studies. As no method will satisfy all criteria and there is no objective ranking of those criteria, we will elaborate a ranking according to the usefulness with regard to the specific needs of our study and will then opt for the class of interpretation methods that is arguably the best aligned with the derived ranking.

3.1.1 Definition of meta-level criteria

From a high level perspective, Ras et al. (2018) argue that a suitable interpretation method should meet the criteria of high fidelity, high interpretability, high generalizability and high explanatory power.

High fidelity corresponds to the requirement that an interpretation method is able to capture the input-output mapping of a deep neural network. Fong and Vedaldi (2019) refer to this property as faithfulness which is fulfilled when the internal processes (how did the model decide?) as well as the external properties of a model (what has the model learned?) are understood. Both insights are arguably of equal importance, as it sheds light on the question of whether the model learned something meaningful or not.

High interpretability refers to the need that the explanation is unambiguous and of low complexity. It must be comprehensible for a human expert and thereby her limitations must be taken into account. Within the medical domain, the human expert is likely to be a physician with little or no machine learning knowledge. Hence, the need for low complexity is high. Ribeiro et al. (2016) discuss the trade-off between fidelity and interpretability and point out that the complexity of an interpretation increases considerably with increasing fidelity. Further, while a ML expert might be able to classify ambiguous results, as she understands the properties of the underlying method, we probably cannot expect that from a domain expert.

High generalizability requires that the interpretation method should not depend on the choice of the architecture or any other configuration. In the best case, we want to obtain a two stage approach where training the model for the prediction task is independent from the interpretation method. Hence, we can first fully focus on the model performance before making the predictions interpretable. The lower the generalizability, the more entangled these two steps are. A high entanglement may inherently restrict the model's capacity and subsequently its performance for the sake of interpretability.

High explanatory power is given when an interpretation method answers different kinds of relevant questions. Ras et al. (2018) refer to the varying needs of different stakeholders who require different types of interpretations. What this means in concrete terms can ultimately only be clarified if the stakeholders and their needs are clearly defined.

3.1.2 Ranking of meta-level criteria

As already partially indicated, the above defined meta-level criteria cannot be satisfactorily fulfilled simultaneously. For instance, some methods will come with a high degree of interpretability, but will lack in generalizability while others are hard to interpret but have a high explanatory power. As no method will completely outrank the others, we have to establish a ranking of the meta-level criteria. For instance, do we either prefer a method with high interpretability or one with a high degree of generalizability? Since an objective ranking is infeasible we have to determine a set of priorities to then infer a ranking of the criteria.

Model performance vs. interpretability? For our specific use case, the highest priority is model performance. We argue that if the model lacks in performance, it becomes useless and the need for interpretability is pointless. Once the model performance is maximized, we can care about the interpretability of the results.

Recipient of explanation While pursuing the goal of interpretability, it remains to clarify for whom and for which purposes the interpretability has to be assured. As mentioned before, the domain expert is considered to be the most important recipient. The domain expert is arguably also primarily interested in having a well-performing model and only secondarily in its interpretability. In general, she wants to have two main questions answered. What was the prediction and which regions in the MRI were the driving forces for the prediction. Note, if we also considered regulators, we would probably conclude differently, as they would probably focus more on an understanding of the model's internal workings.

Within or post-hoc? We must recall that we prioritize model performance over interpretability. Given a within approach, we would try to directly incorporate the objective of interpretability into the model. For instance, in the field of representation learning, it is often considered useful to learn a disentangled representation (Kumar et al., 2017). Yet, a disentangled representation is often less useful for potential downstream tasks than an entangled representation. In such case, we sacrifice model performance for interpretability and thus violate our highest priority. Yet, Baumgartner et al. (2018) argue that a classifier is likely to ignore features with low discriminative power, if features with large discriminative power are identified. Their reasoning is based on the findings of Shwartz-Ziv and Tishby (2017) who show that during training the mutual information between input and output is minimized and hence, only the most salient features are considered by the model. Therefore, a post-hoc interpretation method will not be able to identify the unconsidered features in retrospective, while a within method might enforce the model to not ignore those features. Yet, we do not share the urge of fully identifying all discriminative features for two reasons. First, it is reasonable to assume that ignoring them has no impact on the model performance. Secondly, we are not interested in interpreting the data but rather the model. By focusing on the latter, we derive a potential validation check for whether the model has learned what we expected. Further, the model might have distilled information unknown to the domain expert which is only revealed when we interpret the model and its predictions, instead of the data. Thus, we advocate for post-hoc interpretation methods.

Global vs. local explanation? A global explanation is mainly concerned about the internal workings of a model - how the model makes decisions and how inputs are processed to produce the outputs. The need for global explanation is of special importance, if the

standard performance measures do not imply trust in the model and therefore fail to ascertain faithfulness (Ribeiro et al., 2016). Many models such as decision trees, sparse linear models or some Bayesian models do inherently come with a certain degree of global intepretability (Lipton, 2018). For more complex models, however, it is very difficult to yield a complete understanding of the model's internal workings (Oh et al., 2019). If we still aim for a high degree of global interpretability, we must either sacrifice an uncomplex interpretability for model performance or vice-versa. Yet, since we do not want to sacrifice one for the other a global explanation is not realisable. By contrast, local methods provide only explanations for one specific observation at a time. They function mostly in a post-hoc fashion and therefore do not have any impact on the model performance. This case-by-case consideration is also probably more in line with the needs of the domain expert. She wants to understand why a specific patient converted or not and not necessarily, how the internal workings of a model are functioning. Again, she requires an unambiguous and uncomplex interpretation which is more satisfied with local methods. Thus we prefer a local explanation over a global explanation.

Based on this set of priorities we can now derive the subjective ranking of the meta-level criteria. We argue that a high degree of interpretability and a high degree of generalizability is paramount compared to fidelity and explanatory power. We argue that interpretability depicts the second priority after model performance which is founded in the recipient's limitations. Even though both criteria seem to interfere with each other, a high performance can still be enforced by both, requiring a high generalizability of the interpretation method and opting for a post-hoc interpretation. By doing so, the objective of interpretability is disentangled from the objective of performance. As fidelity conflicts with interpretability, the former is conceded only a minor importance. Further, Lipton (2018) argues that interpretation can be informative without revealing the inner workings of the model. Within our use-case, it might be sufficient to yield interpretations that allow the domain expert to stress the model's behavior against her own intuition. Further, the need for a high degree of explanatory power is arguably limited, as we merely consider the problem of intepretability from the perspective of a domain expert. Hence, the need for a variety of different explanations is limited.

3.1.3 Classes of interpretation methods - a comparison

We have now determined that we rather prefer methods that assure a high degree of interpretability and generalizability and not necessarily fidelity and explanatory power. Based on that ranking we can compare the following three classes of interpretation methods. The definition of the classes are obtained from Ras et al. (2018) who distinguish between *rule-extraction methods*, *intrinsic methods* and *attribution methods*. This abstraction level is appropriate, as the partitions are mutually exclusive and to a large extent collectively exhaustive.

Rule-extraction methods are mainly concerned with extracting human interpretable decision rules from the learned model. By doing so, we aim at a high level of fidelity as it allows a holistic understanding of the internal workings of the model. However, these methods are more concerned with providing a global instead of a local interpretation. With an increasing number of extracted decision rules, interpretability decreases. While these methods are considered to have a high explanatory power, they mostly lack in generalizability.[1] Although these methods are indeed a valid choice we conclude that, due to the shortcomings in interpretability and generalizability, they are not suitable for our use-case.

Intrinsic methods enhance the degree of interpretability by directly incorporating the objective of interpretability into the modeling process. This can be either done by adapting the loss function (i.e. disentanglement learning with Variational Autoencoders (Mathieu et al., 2019)) or some internal structures of the architecture (e.g. Goudet et al., 2018). While such methods may have some advantageous properties w.r.t. interpretability, they clearly lack in generalizability. However and more importantly, it is also likely that the model's capacity and therefore prediction performance is limited. Since our highest priority is model performance, we argue that intrinsic methods are in general not suitable for our purposes. Note, for the *intrinsic methods* the same reasoning applies as that made us opt for post-hoc instead of within methods.

Attribution methods are concerned with assigning for each feature $i \in \{1, ..., n\}$ of an observation x a contribution score $[\phi_1(x, f), .., \phi_n(x, f)] \in \mathbb{R}^n$ to a model's prediction $f(x)$. This allows us to infer which feature made a contribution to the prediction and which not. Without too much loss of generality, attribution methods can be categorized in perturbation- and gradient-based methods (Ancona et al., 2017, Montavon, 2019). The former subsumes methods that perturb or remove features in order to then determine how the output has changed. Gradient-based methods rely on the calculated gradient of the output w.r.t. the considered feature. The magnitude of the gradient indicates the impact of a small perturbation of the feature on the prediction (Ancona, Ceolini, et al., 2019). In both cases, the magnitude of the attributed value indicates the importance and its sign indicates in which direction the prediction has changed due to the feature. Further, it is standard to visualize the attributions by superimposing the values over the original input which results in *attribution maps*. This approach neglects global interpretability, but allows for a high degree of local interpretability. This is due to the fact that only single observations are considered and no information about the internal workings of the model are revealed. The visualization of the results do ensure a high degree of interpretability, as it directly allows to infer which regions or even features in the image were responsible for the prediction. As discussed above, this is perfectly in line with the

[1] Ras et al. (2018) provides a more detailed discussion of the *rule-extraction methods*.

requirements of the domain expert. Beyond that, the attribution methods are in most cases highly generalizable, as they do not require any alterations on the original model and can be classified as post-hoc interpretation methods. Hence, the attribution method has no impact on the model performance. Despite a considerable lack in fidelity and explanatory power, we still argue that attribution methods depict the most suitable class of interpretation methods for our purposes.

3.2 Requirements on attribution methods

As an intermediate result, we conclude that attribution methods meet our criteria best. Yet, within the class of attribution methods, the appropriateness of the methods varies decisively. To reliably identify the most appropriate methods, the literature has established a set of theoretical axioms which must be satisfied so that a method is considered valid. The raison d'être for these axioms is as follows: It is almost impossible to empirically evaluate whether wrongly determined attributions stem from errors of the model or from a flawed attribution method. Consider a pixel which has assigned a high attribution, but the ground-truth indicates a low attribution. Now we have to find out whether this divergence is due to a flawed attribution method or due to a poorly performing model. In case of the latter, we would actually expect this divergence, as we are interested in interpreting the model and not the ground-truth data. In case of the former, we must conclude that the attribution method is not appropriate. A purely qualitative evaluation of the resulting attribution maps has proven to be deficient and biased (Adebayo et al., 2018) and therefore, some sort of quantitative assessment is recommended. Besides, the interpretability criteria demands not only an uncomplex interpretation but also an unambiguous one. Therefore, it must be clarified in advance which types of interpretations are admissible and which are not by elaborating the implications of each axiom on the interpretation. This will be dealt with in the following paragraphs.

Axiom: Implementation Invariance as decscribed by Sundararajan et al. (2017). If two differently implemented networks are functionally equivalent. i.e. the same outputs are generated, we expect the attribution method to yield identical results. Hence, the results do not depend on the model configuration, but only on inputs, outputs and what the model has learned. Implementation invariance is crucial as it allows for a certain degree of generalizability of the attribution method. Otherwise, the interpretations corresponding to two different but functionally equivalent models would differ.

Axiom: Continuity requires that for two almost identical inputs, the respectively assigned attributions must also be almost identical, e.g. $\phi_i(x) \approx \phi_i(x + \epsilon)$. Hence, the predictions must also be nearly identical. Although this represents a standard criterion it can be shown that some attribution methods do not satisfy it (Ancona, Oztireli, et al.,

2019). If not satisfied, the resulting diverging interpretations are likely to contradict human intuition. For two inputs that are almost identical, we would expect to yield almost identical attribution maps.

Axiom: Linearity as defined in Sundararajan et al. (2017). Given two submodels f_1 and f_2, we obtain attribution values for each of the submodels. Now, if both models are combined to $\alpha f_1 + \beta f_2$, we want the attribution method to allocate the shares of the total attribution depending on its shares. This property is important within the context of multi-modality networks. If we combine the structured with the unstructured part, we expect the attribution method to consistently assign the *true* attributions accordingly. If not satisfied, the attribution method might overemphasize one part which results in misleading conclusions.

Axiom: Null Player is satisfied if the method attributes always zero to a feature which the function does not depend on. This is crucial as it ensures that irrelevant features are not erroneously considered important. If the ground-truth indicates no importance, but the attribution method assigns a non-zero value, while the axiom is satisfied, we can confidently infer that the model has learned something wrong. Otherwise, if the axiom is not satisfied, we cannot reliably interpret any attribution score.

Axiom: Sensitivity (a) complements the *null player* axiom in cases where an attribution method relies on some baseline value. The axiom states that if the input differs from the baseline only in one feature and the outcome difference is non-zero, then the attribution score for that feature must be non-zero as well. If the axiom is violated, relevant features might be considered irrelevant.

Axiom: Conservation allows for a new dimension of interpretability. If only the above discussed axioms are fulfilled, we can indeed distinguish between relevant and irrelevant features but a direct interpretation of the features' absolute and relative importance is not given. To enable the latter, this axiom must hold. It requires that the returned score of the attribution method shall match the magnitude of the predicted outcome. In short, the attribution method shall correctly distribute the shares of the predicted risk scores to the relevant features, where the size of the share is equivalent to its absolute importance.

Axiom: Completeness extends the *conservation* axiom in cases where the attribution method relies on some baseline values. If the input differs from the baseline only in one feature and the outcome difference is non-zero, then the attribution score for that feature must be non-zero as well. Further, it states that if the function is evaluated at input x and baseline x', the attribution method must fully account for the output difference $f(x) - f(x')$. The *completeness* becomes even more crucial, if the output and hence,

the output differences can be numerically interpreted. While the outcome differences in the context of binary classification might be not semantically meaningful, the outcome difference between two predicted risk scores definitely are. By considering the relative increase/decrease of the risk score, we can evaluate to what extent the AD progression has accelerated/decelerated compared to the baseline.

3.3 Shapley values

It is desirable to apply attribution methods that satisfy these axioms. Yet, most of the methods that can be assigned to the class of attribution methods do not satisfy all of the axioms. In most cases, the important axiom of Completeness or relatively weak axioms such as Sensitivity (a) or Continuity are violated (Ancona, Oztireli, et al., 2019). If that is the case, the expressiveness of the interpretation is decisively limited. We can merely obtain an indication of whether a feature has contributed to the prediction, but by no means a measure of the actual magnitude of contribution.

Friedman (2004) shows that Shapley values (Shapley, 1953) represent the unique method that satisfies all axioms. The Shapley value has its initial motivation from cooperative game theory, where the outcome of a function f is considered to be a total surplus that is generated by a coalition of N players. Thereby, the Shapley value reflects for each player its marginal contribution to the total surplus. In other words, it determines the importance of each single player in isolation. To obtain that, it is not sufficient to compare the total surplus while all players are participating with the total surplus that is generated without the considered player. Instead, to obtain the marginal contribution we have to integrate out all other players. Therefore, we have to average the total surplus differences over all possible coalitions that can be formed with the available players. Thereby, the size of a possible coalition can vary from only one player to all players included. While Shapley values are quite intuitive within economic game theory, they can also be directly translated to machine learning problems. Within our context, the predicted risk score for one observation corresponds to the total surplus, while each player corresponds to a pixel of the input image[2]. The maths behind calculating the average marginal contribution of a feature is rather straightforward. Given a function $f : \mathbb{R}^N \to \mathbb{R}$ and a set of features N, the contribution of feature i is given by:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \Big[f(S \cup \{i\}) - f(S) \Big] \tag{23}$$

where S is a subset of N and $f(S \cup \{i\})$ corresponds to the prediction given subset S and feature i and $f(S)$ corresponds to the prediction given subset S without feature i.

[2]Lundberg and Lee (2017) introduced the idea of transferring Shapley values to machine learning. Further, they showed empirically that the calculated Shapley values agree considerably with human intuition.

Note, again, as we want to integrate out all other features $-i$, it is not sufficient to calculate the prediction difference for only one subset, but rather we have to average over all possible coalitions. Thus, calculating the marginal contribution for each feature becomes computationally infeasible with an increasing number of features, as for N players, there are 2^{N-1} possible coalitions. Therefore, we rather rely on two methods *Integrated Gradients* (Sundararajan et al., 2017) and *Sampled Shapley values* (Castro et al., 2009) that approximate the *true* Shapley values but still maintain its advantageous properties.

3.3.1 Integrated Gradients

We know from the first theoreom of calculus that the difference between the outcome given the explicand $f(x)$ and the outcome given the baseline $f(x')$ can be expressed as

$$f(x) - f(x') = \int_{x'}^{x} \frac{\partial f(x)}{\partial x}\, dx \tag{24}$$

The intuition is as follows. The integral defines a path from the baseline value x' to the explicand x, whereby the calculated gradients reflect the change in outcome for any given point along the path. Hence, the integral of the gradients determines the overall, absolute change in outcome when going from baseline x' to explicand x. It can be shown that the equivalent is achieved when calculating the Integrated Gradient for each observation i, which is defined as follows

$$IG_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha \times (x - x'))}{\partial x_i}\, d\alpha \tag{25}$$

to then sum over all observations i

$$f(x) - f(x') = \sum_{i=1}^{n} IG_i(x) \tag{26}$$

Thereby it becomes evident, that the Integrated Gradients satisfy the *completeness* axiom. Further, as the derivative is taken w.r.t. the i-th feature, it is assured that any changes in the outcome are only assigned to the i-th feature. While we could theoretically show that the Integrated Gradients are equivalent to Aumann-Shapley values (Aumann & Shapley, 2015) and therefore fulfill the desired axioms, we will rather focus on a more intuitive explanation on the equivalence between Shapley values and Integrated Gradients. To pursue this, we will adhere to the illustration provided by Sundararajan and Najmi (2020). The intuition is as follows: The Integrated Gradients define a smooth path between the baseline and the original input. By contrast, equation 23 achieves that by following a discrete path, whereby in each step one more feature is *turned on*. The followed path is defined by the edges of a hypercube where each node represents one feature that is *turned*

on. The Shapley value is then the average over all discrete paths, whereas the Integrated Gradients choose the internal diagonal of the hypercube. To make this computationally feasible, we can approximate equation 25 by taking the sum of gradients from a subset of m points that lay along the path between x and x'

$$IG_i^{approx}(x) = (x_i - x_i') \times \sum_{k=1}^{m} \frac{\partial f(x' + \frac{k}{m} \times (x - x')}{\partial x_i} \times \frac{1}{m} \tag{27}$$

In general, the points k are chosen such that they are equally distributed along the path. The approximation via Integrated Gradients is favorable, as its implementation allows for an simple and fast computation of the features' attributions.

3.3.2 Sampled Shapley values

A more intuitive but also naive approach was introduced by Castro et al. (2009). They leverage the fact that the Shapley value for feature i as defined in equation 23 can also be expressed as

$$\phi_i = \frac{1}{n!} \sum_{O \in \pi(N)} \Big[f(Pre^i(O) \cup i) - f(Pre^i(O)) \Big] \tag{28}$$

where $\pi(N)$ corresponds to an ordered set of all possible permutations with cardinality $n!$ and $Pre^i(O)$ corresponds to the set of features that precede feature i in the respective permutation $O \in \pi(N)$. For instance, for a given permutation $O \in \pi(N)$ with cardinality n, the feature i is placed in k-th position, then $Pre^i(O)$ includes all features $-i$ that are in positions $1, ..., k-1$, while the remaining are left unconsidered. It is straightforward to see that if $\pi(N)$ includes all possible permutations, equation 28 is equivalent to equation 23. To approximate equation 28, we consider a randomly sampled subset of size M of all possible ordered sets $O \in \pi(N)$. The corresponding pseudo algorithm is outlined here:

Algorithm 1: Approximate Shapley values proposed by Castro et al. (2009)

Input: model: f
Input: ordered set with cardinality $N!$: $\pi(N)$
Input: number of samples: M
Output: Shapley values ϕ_i for $\forall i \in N$
Initialize: $\phi_i = 0$ for $\forall i \in N$
for $m = 1 \rightarrow M$ **do**
  Sample $O \in \pi(N)$ with probability $\frac{1}{N!}$
  for $i \in N$ **do**
    Derive $Pre^i(O)$
    Calculate $\phi_i^m = f(Pre^i(O) \cup i) - f(Pre^i(O))$
    $\phi_i = \phi_i + \phi_i^m$
  end
end
$\phi_i = \frac{\phi_i}{M}$ for $\forall i \in N$

The authors further show that the approximation is unbiased and consistent in probability. For a large enough M the approximation converges towards the exact value. Štrumbelj and Kononenko (2014) proposed an extension of this approximation which allows to further reduce the computational complexity. As it will become apparent in chapter 5, we will apply the attribution methods only in a simulation setting. Our concerns are therefore not directed to computational complexity. Hence, we opt for the arguably simplest approximation. If, however, we were to apply the attribution methods on the ADNI data, it is definitely recommended to opt for the method that comes with lowest complexity. For such purposes, Ancona, Oztireli, et al. (2019) proposed an efficient approximation which reduces the computational complexity from $\mathcal{O}(2^N)$ to $\mathcal{O}(KN)$, where K corresponds to the number of sampled coalitions and N to the number of features. This approach, however, comes with one non-negligible shortcoming. It requires to transform the deterministic point estimates of the outputs and activations of the learned model to probabilistic output layers and distributions, respectively. To pursue this, they rely on the concept of *Lightweight Probabilistic Deep Networks* introduced by Gast and Roth (2018). The framework, however, is not applicable for all potential activation functions (e.g. tanh activation). Therefore, this attribution method is limited in its generalizability that ought to be maximized, as discussed in chapter 3.1.2.

4 The baseline generator

The calculation of Shapley values, either via Integrated Gradients or sampled Shapley values, relies on the specification of a baseline value. If the choice of the baseline impacts the interpretability of the derived Shapley values, then it is reasonable to assume that there exist more appropriate and less appropriate baseline choices. This chapter serves to clarify which factors constitute an appropriate baseline and how to identify the baseline as a consequence. We pursue this by adhering to the following three steps. In chapter 4.1, we will discuss the problems if a sub optimal baseline is chosen. Chapter 4.2 establishes a set of criteria that allow to infer what constitutes an optimal baseline. In chapter 4.3, we will then elaborate an identification strategy to obtain a baseline that satisfies all criteria in order to - at least theoretically - solve the discussed problems. Finally, in chapter 4.4 we put the derived *baseline generator* in the context of current research and thereby point towards the strengths of this framework.

For the remaining, it is important to note that we assume an uni modal setting - only the unstructured MRIs are considered while the structured tabular data is left unconsidered. We emphasize that for notational reasons. To consider the MRIs in isolation is justified due to the *orthogonalization trick* (chapter 2.5).

4.1 The baseline selection problem

To thoroughly understand the problem, we have to once again consider Shapley values from the cooperative game theoretical perspective. We permute over all possible coalitions, whereby the size of a coalition varies from one player to all players. When only a subset of players is involved, the remaining players are excluded from the game. In other terms, the remaining players are *missing* for this particular round. In the transfer of the concept of Shapley values from game theory to machine learning, we have to exclude all features that are not included in a current subset. Thereby the following question remains: How do we exclude features? Applying the equivalent by simply removing those features is technically impossible - we cannot just cut out pixels from the image. Hence, we have to find a replacing baseline value that appropriately simulates the case of *missingness*.

To pursue this, the literature often suggests to replace the original features values with zero (Sundararajan et al., 2017). While this can indeed depict a valid choice for tabular data, it does definitely not apply to images. For instance, Jha et al. (2020) state that within a variety of genomic applications the value zero has a biological meaning. Also in the application of grading diabetic retinopathy, the zero value is far away from representing *missingness* (Sayres et al., 2019). In the case of predicting AD progression, the zero value might artificially increase the size of the hippocampus which is arguably a strong

predictor for AD progression.

In view of this problem, Sturmfels et al. (2020) analyze the impact of different baseline value choices w.r.t. the objective of simulating *missingness*. They do so by training a convolutional neural network on the ImageNet dataset (Deng et al., 2009) and applying different baselines such as *Maximum Distance*, *Uniform*, *Gaussian* and *Blurred* as input for Integrated Gradients. They argue that inducing randomness into the determination of a baseline value increases the chance of selecting a baseline that has no semantic meaning. While this argument is potentially valid, inducing randomness is far away from an universally valid approach. The authors provide a qualitative discussion of the different choices, as well as a first attempt at quantitative assessment and thereby admit that they were not able to identify an optimal choice. We argue that their failure is mainly due to a lack of an elaborate, theoretical discussion of the actual problem.

A theoretical discussion of what indeed represents *missingness* has to consider the actual practical implication of choosing a sub optimal baseline. One of the dominant reasons to opt for Shapley values lies in the fact that they theoretically satisfy the axiom of *completeness*. While this holds for any arbitrary choice of the baseline, the interpretation of the Shapley values is decisively influenced by that. To better understand that, we must deviate from the term *missingness*, but rather consider what *missingness* implies for the prediction task. In both contexts, game theory and machine learning, a missing player or feature implies a zero contribution to the outcome. In short, a feature that contributes nothing is equivalent to a feature that does not participate. Identifying a non-contributing baseline yields an unambiguous interpretation - the relative contribution of the feature is equivalent to the absolute contribution. By contrast, if the baseline has a non-zero contribution, we have to interpret the feature's contribution in reference to the baseline's contribution: If we have no understanding why the prediction of the baseline is non-zero, we have no chance to obtain a meaningful interpretation. This problem is unavoidable when we randomly select a baseline as there is no semantic meaning. In the worst case, features could possible be assigned a non-zero attribution value, even though the model did not focus on that. If that is the case, any conclusion is definitely misleading and therefore, wrong. The problem becomes even more severe, if we demand a certain degree of robustness. Given the same data, the same model and the same attribution method, but different baselines, the resulting Shapley values differ (Merrick & Taly, 2020). Due to the fact that the reference point which is determined by the baseline varies, we can conclude that there is an urgent need for identifying a unique and meaningful baseline that leaves no room for ambiguities.

4.2 The optimal baseline

When we apply attribution methods or more specifically, when we calculate Shapley values, it is not advisable to entirely rely on its axiomatic properties. In the previous chapter, we examined why an ill-considered baseline choice is likely to break the entire interpretability of the results, despite the axioms being satisfied. This happens if the baseline produces a non-zero outcome, while its reasons are not understood. If we do not know what the baseline represents, we have no chance to obtain a meaningful interpretation where the baseline serves as a reference point. Merrick and Taly (2020) argue that, for this specific reason, the baseline choice should not be considered merely as an implementation detail, but rather as "a first-class argument to the framework" (Merrick and Taly, 2020, p.10). For them the optimal baseline must allow to yield a meaningful contrasting explanation. In short, we require a baseline which serves as a meaningful reference to the original outcome.

For illustration, imagine an MRI of the brain for which the model predicts a positive risk score - the MRI reveals structures that lead to an accelerated AD progression. Now, we aim to understand which pixels were responsible for the prediction and what was the share of contribution, respectively. To answer that, a proper reference point represents an image that is considered healthy. Then, we can assign outcome differences to those pixels which differ between input and reference point, while the pixels that remain constant cannot have made any contribution to the prediction of the risk score. To obtain this contrasting explanation, it is key to formulate the corresponding contrasting question which should be answered by means of the baseline. In our case, a valid and expressive contrasting question could be formulated as follows:

> Given an MRI of the brain for which the model predicted a positive/negative risk score, how is the corresponding baseline defined so that the model predicts a negative/positive risk score?

This discussion brings us one step closer to answering the contrasting question, but does not guarantee an identification of the optimal baseline. Recall, we require a baseline which only differs from the input in those pixels that are, according to the trained model, relevant for predicting the risk score. In fact, the contrasting question ensures that a baseline is identified which contrasts to a healthy/sick input image, however, it does not ensure that irrelevant pixels remain unchanged. If the latter is not guaranteed, the total contribution (outcome difference) is not only shared among the relevant pixels but also among the irrelevant pixels. Thus, the interpretation of the assigned attributions is deteriorated. Further, it was argued that different reference points result in different attributions. As answering the contrasting question does not provide us with a unique reference point, the reference explanation is not unique either.

To avoid this shortcoming, we define three criteria that ensure a robust and unique identification of the optimal baseline. To pursue this, we adhere to the set of criteria established by Shih et al. (2020).

Criterion 1: The baseline belongs to the target domain (w.r.t. the model) If the model predicts a positive risk score for the input, we require that the baseline results in a negative risk score and vice versa. Then, we can contrast the explanation for a sick brain with a healthy brain and vice versa. The resulting interpretation is quite intuitive as we can now evaluate which structural characteristics were responsible for an accelerated/decelerated AD progression.

Criterion 2: The baseline is a realistic sample We require the baseline to represent a realistic sample, as it allows to effectively understand structural differences between the target and the baseline image. In addition to a quantitative assessment of outcome differences, we can also yield a domain-specific qualitative assessment. If the baseline did not represent a realistic sample, we could again not comprehend why the baseline belongs to the target domain and hence the reference point is not meaningful.

Criterion 3: The baseline is close to the input This criterion ensures that only structural changes that are linked to AD are captured. All other characteristics visible in the MRI must stay constant. If we did not require to meet this criteria, we could simply select a sample from the training data that belongs to the target domain and compare it to the sample of interest. However, this implies that we compare the sick/healthy brain of person A, with a healthy/sick brain of person B. Rather, we want to compare the sick/healthy brain of person A with a hypothetical healthy/sick brain of person A. If that criterion is not satisfied, the baseline image would also differ in domain unspecific characteristics so that the attribution method assigns non-zero scores to structures (features) that were irrelevant for predicting AD progression.

Figure 1 serves as an illustration for an intuitive explanation on how the optimal baseline is identified and how the fulfillment of the criteria enforce the optimal solution. The blue areas represent the latent space covered by realistic samples, whereas the red areas represent the latent space covered by observed training samples. The decision boundary in the latent space represents all instances where the model predicts a zero risk score. To the right all instances correspond to a negative risk score and to the left all instances correspond to a positive risk score. Observations that are farther away from the decision boundary correspond to a more positive/negative risk score. For a given input, three potential baselines are illustrated. The yellow rectangle represents a baseline that is closest to the input but does not represent a realistic sample. Therefore it is not a valid choice. The green rectangle (MDTS = mimimum distance training sample) lies within

the target domain and indeed represents a realistic image. However, as it is a sample from the training data, it does not only differ in domain-specific characteristics from the input image. Therefore it is not a valid choice, either. The red point represents the only baseline which satisfies all three criteria and therefore represents the unique optimal baseline. The baseline lies within the area of realistic samples, it lies in the target domain and - if those two criteria are fulfilled - is closest to the input.

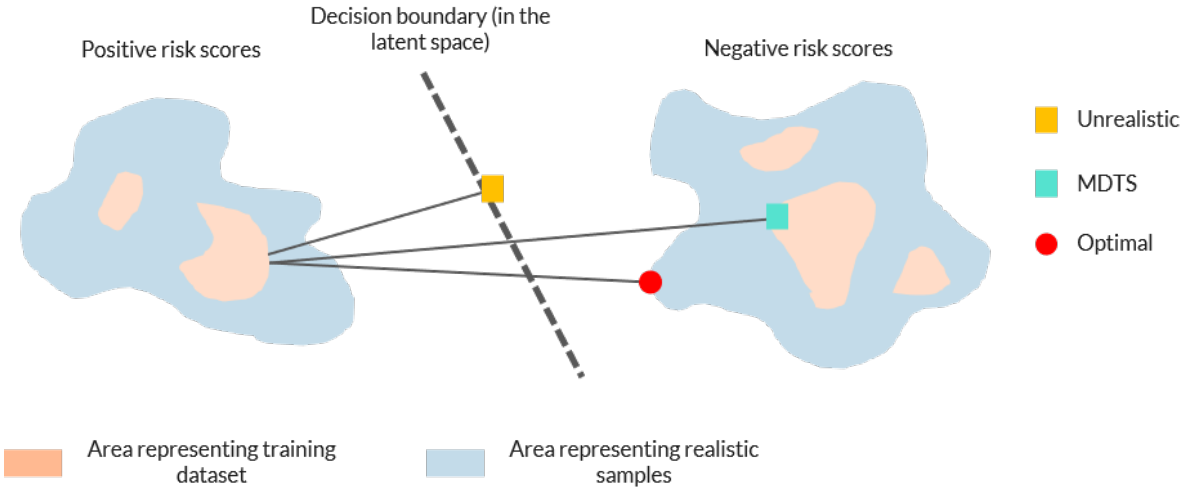
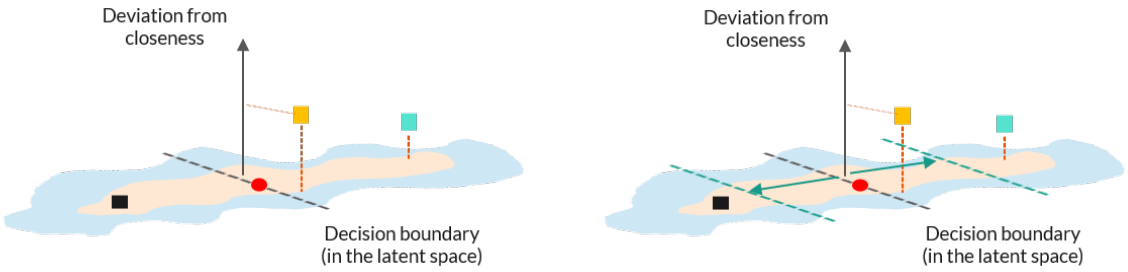


Figure 1: Illustration of a discontinuous latent space and the learned decision boundary of the survival model.

Even though figure 1 illustrates the concurrence of the three criteria rather well, the schematic depiction is not accurate. In a survival analysis setting, it is reasonable to assume that the latent space is continuous and not discontinuous. In this case, the decision boundary is likely to intersect both the red and the blue area. Then, however, the illustration in figure 1 is not suitable anymore to reliably identify the optimal baseline. Every baseline that lies infinitesimally close to the decision boundary is equally favorable - at least on a visual level. They represent a realistic sample and lie in the target domain. But we cannot distinguish between a baseline that violates the *closeness* criterion and a baseline that satisfies that criterion. The indistinguishability is due to the fact that the latent space merely captures the domain-specific characteristics but not the domain-unspecific characteristics upon which we evaluate to what extent the *closeness* criterion is satisfied. To circumvent this limitation of figure 1, we additionally introduce figure 2. There, the assumption of a discontinuous latent space is voided. In order to display the *closeness* criterion, we add a second dimension - *deviation from closeness* - which captures to what extent the baseline differs from the the original input in the domain-unspecific characteristics. The degree of deviation is indicated by the dashed lines in red, as shown for the yellow and the blue rectangles, respectively. By doing so, we can conclude that the red circle is favorable over the yellow rectangle, even though both are infinitesimally close to the decision boundary and therefore correspond to a zero risk score prediction (see figure 2a).

By acknowledging the presence of a continuous latent space, we can narrow down the conditions of criteria 1 to the extent that the baseline not only lies within the target domain but also that the baseline fully represents the pre-specified reference point which is determined by the decision boundary. We prefer an exact and unambiguous baseline in order to obtain a more robust interpretation. Only by tightening criterion 1, we can control to what extent the baseline deviates from the optimal reference point. Otherwise, a certain degree of variability is induced which prohibits a unique and robust identification of the baseline which again is essential for an unambiguous interpretation. Further, by tightening the criterion 1 we achieved to identify a baseline that indeed represents *missingness* even though we refrained from explicitly looking for it. In the domain of predicting AD progression *missingness* corresponds to the case where the observed structures in the MRI contribute neither to accelerated nor to decelerated progression. Hence, the baseline image makes no contribution to the overall predicted risk score. Put differently, for a sick MRI *sickness* is missing, while for a healthy MRI *healthiness* is missing.



(a) Unique decision boundary **(b)** Multiple decision boundaries

Figure 2: Illustration of a discontinuous latent space and the learned decision boundary of the survival model.

It is also important to note that this identification strategy allows to yield further reference points, apart from the zero risk score. Then, however, the baseline identification can no longer be motivated by the objective to represent *missingness*. Yet, we argued that as long as the criteria are satisfied and the baseline is semantically meaningful, any reference point is potentially suitable. Hence, we can map any original MRI to any desired reference point, as long as the reference point is meaningful. This consideration will be resumed in chapter 4.3. On a visual level, this corresponds to a right or left shift of the decision boundary in the latent space as depicted in figure 2b. The dashed lines in green represent two possible reference point choices, respectively. In this case the optimal baseline lies on the new decision boundary.

4.3 Identification of the optimal baseline

Given the elaborated criteria, we still have to clarify how the optimal baseline can be identified. As the third criteria must hold, it is not valid to select a sample from the

training data as baseline, but instead we must synthetically generate a unique baseline for each data instance, respectively. To achieve this, the class of generative modeling depicts a valid choice. To meet the criteria of generating a baseline that belongs to the target domain, but also maintains all domain-unspecific characteristics, the family of image-to-image translation networks first introduced by Isola et al. (2017) is suitable. This family is only applicable, if the main goal is to map an input image to a desired output image which fully reflects our goal of baseline generation. While there is a variety of image-to-image translation networks that hypothetically could yield the desired results, we opted for StarGAN (Choi et al., 2018), because it allows to directly incorporate the pre-trained survival model as a discriminator.

The StarGAN consists of three parts which are illustrated in figure 3. The generator $G(\boldsymbol{x}, \boldsymbol{d}) : \mathbb{R}^{n\times(c+d)\times h\times w} \rightarrow \mathbb{R}^{n\times c\times h\times w}$ generates realistic baseline images that belong to the respective target domain. Given the input tuple $(\boldsymbol{x}, \boldsymbol{d})$, the generator G produces a baseline image $\tilde{\boldsymbol{x}}$ that only differs from the input image in the domain-specific characteristics and thereby represents a realistic image. The one-hot encoded vector $\boldsymbol{d}$ indicates into which domain the generator G has to map the original image. The discriminator $D(\boldsymbol{x}) : \mathbb{R}^{n\times c\times h\times w} \rightarrow [0, 1]$ learns to distinguish between real and fake images which enforces the generator G to indeed produce realistic images. Lastly, the survival model $S(\boldsymbol{x}) : \mathbb{R}^{n\times c\times h\times w} \rightarrow \mathbb{R}$ judges whether the generated baseline image belongs to the target domain or not. Hence, the survival model S provides feedback to the generator G to what extent the generated images lie within the target distribution.

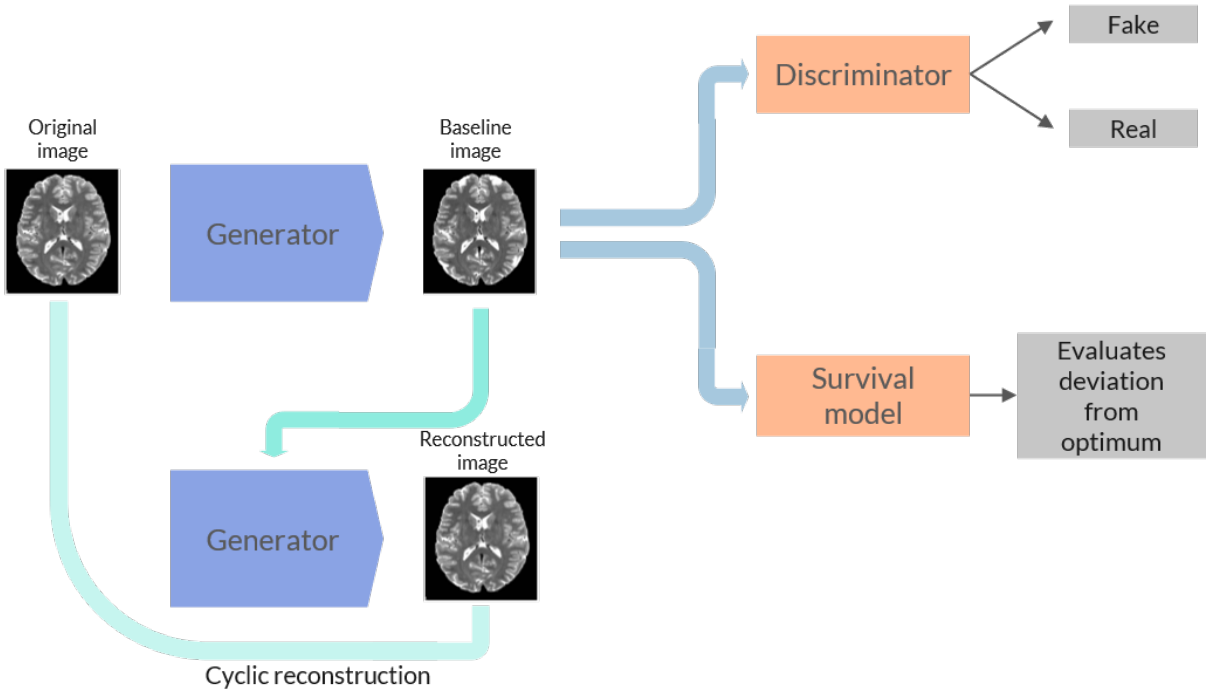


Figure 3: Illustration of StarGAN framework in the setting of survival times prediction

How the three models interact during training can be understood if we consider the corresponding loss functions for optimization. In the following, the adversarial loss $\mathcal{L}_{adv}$,the domain loss $\mathcal{L}_{cls}$, the reconstruction loss $\mathcal{L}_{rec}$ and the gradient penalty $\mathcal{L}_{gp}$ are discussed.

Adversarial loss is adopted to ensure that the generator G produces realistic images

$$\mathcal{L}_{adv} = \mathbb{E}_{\boldsymbol{x}}[logD(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x},\boldsymbol{d}}[log(1 - D(G(\boldsymbol{x}, \boldsymbol{d}))] \tag{29}$$

where G produces the baseline images given the tuple $(\boldsymbol{x}, \boldsymbol{d})$ and the discriminator distinguishes between real and fake images. The generator G aims to minimize L_{adv} while the discriminator D aims to maximize the objective. Note, the adversarial loss does not enforce to generate baseline images that belong to desired target domain, but merely enforces realistic baseline images.

Domain loss enforces the generator G to produce images that belong to the target domain. In a classification task setting, this can be pursued by the following loss function

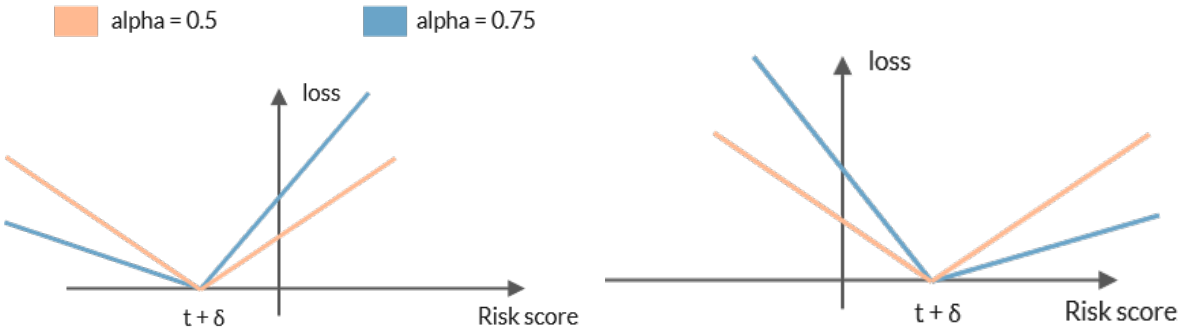
$$\mathcal{L}_{cls} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{d}}[-log(D_{cls}(\boldsymbol{d}|G(\boldsymbol{x}, \boldsymbol{d})))] \tag{30}$$

where D_{cls} represents a second discriminator which learns to classify the images correctly. The generator G tries to minimize the objective so that the discriminator classifies the synthetic image to the desired target class d. In the standard StarGAN setting, it is assumed that the labels are known, while such labels do not exist in the survival times setting. Therefore, it needs to be determined how the labels can be obtained. As illustrated in figure 1, we could define a fixed decision boundary (threshold) where the predicted risk score that corresponds to the MRI is 0. In one domain (class), all corresponding MRIs contribute to an accelerated AD progression, while in the other domain (class) all corresponding MRIs contribute to a decelerated AD progression. Then, we can define $d = \{0, 1\}$ where d takes 0, when the target domain represents a negative risk score and 1, vice-versa. Now we can similarly adapt the domain loss in equation 30. However, this only enforces the generator G to produce baseline images that lie in the target domain with a high confidence. Yet, we aim to find the baseline that lies on the decision boundary in the latent space (see figure 2). To pursue this, we can apply the following quantile loss

$$\mathcal{L}_{surv} = (1 - \alpha) \sum_{\boldsymbol{y} > \boldsymbol{d}\tau} \|\boldsymbol{y} - (\tau + \delta)\| + \alpha \sum_{\boldsymbol{y} \leq \boldsymbol{d}\tau} \|\boldsymbol{y} - (\tau + \delta)\| \tag{31}$$

where τ represents a threshold which is considered to be optimal at $\tau = 0$. Any deviation

of the predicted risk scores $\boldsymbol{y}$ from the optimal threshold τ is penalized. To further increase the confidence that the baseline image lies in the corresponding target domain, we add a small residual to the threshold $\tau+\delta$, whereby $\delta > 0$ if $d = 1$ and $\delta < 0$ if $d = -1$. To additionally strengthen the penalization when the target domain is not fulfilled, the parameter $\alpha \in [0, 1]$ can be set to $\alpha > 0.5$. In short, deviations of $\boldsymbol{y}$ from τ are less penalized when $\boldsymbol{y}$ lies in the target domain. Figure 4 illustrates the loss function when the target domain lies in the range of negative (see figure 4a) and positive risk scores (see figure 4b), respectively. Note, the illustration assumes the L1-norm, whereby any reasonable distance norm is applicable.



(a) Target domain: negative risk scores

(b) Target domain: positive risk scores

Figure 4: Illustration of the domain loss for two different target domains, respectively

Even though the derived domain loss is applicable in theory, it can be questioned whether the optimal threshold $\tau = 0$ is in fact a suitable choice in practice. The problem becomes apparent when we consider the loss function for optimizing the DeepCoxPH model (equation 22). While the loss function enforces the survival model S to learn the correct ordering of the survival times, it does not directly encourage the survival model S to learn the exact risk scores. Hence, there is no guarantee that the predicted risk scores are aligned with the ground truth. In the worst case, the decision boundary where $\tau = 0$ is not covered by the latent space and therefore, determining a reference point that satisfies $\tau = 0$ does not represent a realistic sample. To circumvent this, we have to identify a more robust choice of the threshold τ. We require a threshold τ that is in line with the ordering objective of the survival model S. One applicable choice for the threshold τ is determined by the median survival time prediction which can be derived as follows. We know that the survival function $S(t)$ can be expressed as

$$S(t) = \exp(-\Lambda_0(t) \exp(\boldsymbol{x}^T \boldsymbol{\beta})) \tag{32}$$

where the risk score, and therefore the threshold τ, can be written as

$$\tau = \boldsymbol{x}^T \boldsymbol{\beta} = \log\Big(- \frac{\log S(t)}{\Lambda_0(t)} \Big) \tag{33}$$

If we aim for a threshold that corresponds to the median survival time, we set $S(t) = 0.5$ which yields

$$\tau = \log\Big(-\frac{\log 0.5}{\Lambda_0(t)}\Big) \tag{34}$$

Thereby the target domain $d = -1$ relates to all observations that had an event beyond the median survival time and $d = 1$, if the observation had an event before the median survival time. As illustrated in figure 2b, the median survival time threshold merely corresponds to either a left or right shift of the decision boundary so that the validity of the loss function remains. This alternative derivation of the threshold τ comes with two advantages. Firstly, it does not rely on an exact prediction of the risk scores, but only on the correct ordering which is better aligned with the general evaluation of the survival times prediction. Secondly, the choice of the threshold τ is not limited to one unique value. While $\tau = 0$ depicts the only semantic meaningful threshold for the risk score related derivation of the threshold τ, the choice of the survival time quantile is not bound to the median survival time. In fact, there might be use-cases where a smaller or larger quantile is more sensible. Such cases are subject to discussion in chapter 6.3. This flexibility of the threshold choice allows for extended interpretations of the baseline images which will be discussed in chapter 6.1.

Reconstruction loss causes the generator G to produce baseline images that only differ in domain-specific characteristics, but not in the domain-unspecific ones. Both, the domain loss $\mathcal{L}_{surv}$ and the adversarial loss $\mathcal{L}_{adv}$ do not guarantee the preservation of domain-unspecific characteristics, unless we adopt the cyclic reconstruction loss from Choi et al. (2018)

$$\mathcal{L}_{rec} = \mathbb{E}_{\boldsymbol{d},\boldsymbol{d'},\boldsymbol{x}\in\boldsymbol{d'}}[||\boldsymbol{x} - G(G(\boldsymbol{x},\boldsymbol{d}),\boldsymbol{d'})||_1] \tag{35}$$

which takes the point-wise L1-difference between the generated image and the original image. Note, we obtain the generated image $G(G(\boldsymbol{x},\boldsymbol{d}),\boldsymbol{d'})$ by applying the generator G twice. First, we produce a synthetic image that belongs to the target domain d and then we map the synthetic image back to the original domain d' of the input image x.

Gradient penalty corresponds to an additional penalization term for the discriminator D. Training a GAN often suffers from non-convergence or mode collapse which can occur if the discriminator D overfits, while the generator G still outputs low quality images (Arjovsky et al., 2017). Then, the gradients of the generator G diminish and hence, the generator G is prevented from learning. To circumvent the problem of vanishing gradients, Gulrajani et al. (2017) proposed the method of gradient penalty which is

defined as follows

$$\mathcal{L}_{gp} = \mathbb{E}_{\hat{\boldsymbol{x}} \sim \mathbb{P}_{\hat{\boldsymbol{x}}}} \Big[(\|\Delta_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_2 - 1)^2 \Big] \tag{36}$$

where $\hat{\boldsymbol{x}}$ is defined as

$$\hat{\boldsymbol{x}} = \gamma \tilde{\boldsymbol{x}} + (1 - \gamma)\boldsymbol{x} \tag{37}$$

where $\tilde{\boldsymbol{x}}$ and $\boldsymbol{x}$ correspond to a batch of generated images and a batch of original images, respectively and γ is uniformly sampled with $0 \leq \gamma \leq 1$. It has been empirically shown that the penalization term $\mathcal{L}_{gp}$ contributes considerably to the stability of the GAN training (Gulrajani et al., 2017). Note that, adding the gradient penalty $\mathcal{L}_{gp}$ has no impact on the general framework, but will only improve training.

Given the defined objective functions $\mathcal{L}_{adv}$, $\mathcal{L}_{rec}$, $\mathcal{L}_{surv}$ and $\mathcal{L}_{gp}$, we can now define the objective function for the discriminator D and generator G, respectively. The discriminator tries to minimize the following objective function

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{gp}\mathcal{L}_{gp} \tag{38}$$

while the generator G tries to minimize

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{surv}\mathcal{L}_{surv} + \lambda_{rec}\mathcal{L}_{rec} \tag{39}$$

The hyper-parameters λ_{gp}, λ_{surv} and λ_{rec} control the relative importance of the gradient penalty, the domain loss and the reconstruction loss, respectively and in comparison to the adversarial loss. As desired, the survival model S is not further optimized during the training, as we aim to explain the predictions made by the survival method S in a post-hoc fashion. By including the survival model S during optimization, the prediction performance may be negatively influenced[3].

4.4 The baseline generator in the context of current research

The objective to translate MRIs of the brain with AD to MCI and vice-versa has already been extensively studied. Baumgartner et al. (2018) pursued an approach, called VA-GAN, where 3D MRIs of the brain with AD were translated to MCI. They pursue this by learning an *effective disease map* which captures the class specific characteristics that distinguish the MRI with AD from a corresponding MRI with MCI. They pass an

[3]The problem of within interpretation methods has been thoroughly discussed in chapter 3.1.2

original image x from the baseline class $c = 0$ (AD) through a generator which outputs the disease map $M(x)$. The learned disease map is then added to the original MRI $y = x + M(x)$, where y belongs then to the target class $c = 1$ (MCI). Similar to our approach, the generated and original MRI shall only differ in the domain-specific characteristics while all other characteristics remain constant. However, their approach comes with a major drawback. As the framework only allows to translate AD to MCI and not vice versa, it is required to know the class labels a *priori*. To circumvent this, Bass et al. (2020) introduced an approach, called VAE-GAN, which also leverages the idea of image-to-image translation to learn *effective disease maps*. By slightly modifying the approach from Lee et al. (2018), they encode each class specific MRI in a class irrelevant and class relevant latent representation, respectively. The class relevant encoding is then used for classification, while a cross-combination of the class specific encodings is passed to a generator which translates the original MRI to the target domain. Then, taking the point-wise difference between the original MRI and the generated MRI yields the *effective disease map*. While this approach does not require to know the class labels a *priori*, it is still not applicable for our purposes for one specific reason. In chapter 3, we argued that we want to explain the model in a post-hoc fashion. Within their approach, however, the classification model is trained simultaneously with the generator framework and therefore, a within approach has been chosen. A flexible adaption to a post-hoc setting seems not feasible. Beyond that, both methods require a binary classification setting. As discussed in chapter 1, this depicts a poor framework when studying progression from MCI to AD. Although Bass et al. (2021) have extended their original approach by also allowing for regression tasks, this framework still requires to distinguish between sick and healthy MRIs a *priori*[4].

To sum up, there exists a variety of elaborate approaches that seek to identify the disease-specific characteristics and thereby yield similar outputs as our derived attribution maps. Yet, we still argue that our proposed *baseline generator* framework is superior. From a pure theoretical perspective, our framework is applicable to survival models which clearly depicts a more appropriate choice to model AD progression. Further, the generated baseline images are inputs for axiomatic verified attribution methods. Thus, the admissibility of interpretations are theoretically founded. This is not given with the learned *effective disease maps*. Lastly, our framework allows for interpretations that are more revealing. In our case, both the baseline's prediction and the prediction corresponding to the original MRI have a numeric semantic meaning. Hence, the prediction difference can be interpreted and as the *completeness* axiom is satisfied, the prediction difference is fully captured by the assigned attribution scores. By contrast, the predicted logits from binary classification are hardly meaningful as they only indicate the confidence of the model's

[4]The regression tasks are apparently only applied to translate within the domain of ages. Thus, the framework allows to translate any *young* MRI to any *old* MRI and vice versa. Yet, the fact that the need to distinguish between MRIs with MCI and AD a *priori* prohibits a transfer to survival analysis

prediction. Hence, the prediction differences are also not that revealing. As we not only require the baseline to contrast an MRI that corresponds to an accelerated/decelerated AD progression, but also require a baseline that reflects the pre-specified quantile, we know exactly what the reference point represents. Hence, we can reliably assess to what extent the original MRI contributes to a more accelerated or decelerated AD progression compared to the identified baseline. As our framework provides for all original MRIs reference points that represent the median survival time, we can justifiably also compare the resulting attribution maps between observations. By contrast, the current methods do only enforce to translate the respective MRIs to their contrasting clinical picture and thereby it is not assured that the reference point is consistent for all MRIs. Given two MRIs with MCI, one baseline represents AD distinctly, while the other baseline represents AD only weakly. Then, for two (almost) identical MRIs with the same clinical picture, the *effective disease maps* differ because the reference points are not identical. If that is the case, the reliability of the resulting interpretations are questionable.

5 Experiments

5.1 Experimental setup

For conducting the experiments, we were provided with a single GPU from LRZ. Further, we published the entire code on GitHub[5] to enable a certain degree of reproducibility of the conducted experiments. Further instructions on how the experiments can be reproduced can be found on the GitHub repository itself.

5.2 Experimental strategy

Our experimental strategy is subdivided into two phases. In the first phase, we train survival models on the ADNI data (chapter 5.3). Thereby we aim to understand to what extent the multi-modal approach (tabular data *and* MRIs) can enhance performance in comparison to the uni-modal approaches (tabular data *or* MRIs). Thereby, we also seek to understand to what extent the orthogonalization (chapter 2.5) impacts the performance of the survival model. If the performance of the model with orthogonalization does not decline decisively, we can conclude that we obtained a higher degree of interpretability without any sacrifice in performance. Beyond mere performance considerations, we will also investigate whether the orthogonalization has any impact on the estimated linear weights that correspond to the structured part. To pursue both, we will benchmark four models. First, we train a simple linear Cox-PH model (Cox, 1972) on the tabular data only. Secondly, we train a Deep Cox-PH model on the MRI data only and lastly, we train two multi modal Deep Cox-PH models with and without orthogonalization, respectively.

In the second phase of our experiments (chapter 5.4), we will focus on the applicability of the baseline generator framework (chapter 4). To pursue this, we refrain from applying the framework on the ADNI data, but rather aim for identifying a simulation setting which allows to fully understand the internal workings of this framework. By doing so, we can show that the framework indeed generates baseline images that satisfy the criteria which were established in chapter 4.2. To further emphasize the necessity of the baseline framework, we evaluate the attribution maps derived from Integrated Gradients (Sundararajan et al., 2017) and sampled Shapley values (Castro et al., 2009). It will become evident that the attribution maps provide only sensible insights when applied on the generated baselines, while arbitrary baseline choices deteriorate interpretability to a non-negligible extent.

[5]This GitHub repository entails the full codebase to run all experiments. To do so please follow the instructions in the Readme. [Link: https://github.com/MoritzWag/DeepSurvival]

5.3 ADNI

5.3.1 Data

In the first part of the experiments, we use the data provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack Jr et al., 2008) which represents a longitudinal study that started in 2003. The initiative aims at identifying strong predictors for the conversion from MCI to AD, by collecting clinical and biomarker data as well as MRI scans of the brain. For our purpose, we selected 795 subjects with MCI at the entry of the study and at least one follow-up visit. To reduce computational complexity, we take one slice from each raw 3-dimensional MRI to obtain a 2-dimensional representation of the brain. The slices are taken from the coronal plane, as this arguably represents the hippocampus the best. There seems to be some evidence that the size of the hippocampus is a strong predictor for AD progression (Goukasian et al., 2019). By following this approach, we reduce the size of one MRI from (128, 160, 128) to (160, 128). Note, for each MRI scan, we took the same slice from the coronal plane which might imply that for some instances the hippocampus is better represented than for others. To what extent a fairly naive choice of the slices impacts the performance of predicting survival times is left for future research. With respect to pre-processing the MRI scans, we normalized the data from a range of 0 to 255 to 0 to 1. Besides that we refrained from applying any other pre-processing on the image data.

The MRIs are further complemented by tabular clinical data which include the level of education, age and sex. Beyond that, we consider relevant biomarker data such as: FDG-PET, AV45-PET, APOE4, levels of beta amyloid 42 peptides (Aβ42), total tau protein (T-tau), and Tau phosphorylated at threonine 181 (p-Tau). All covariates are normalized between 0 and 1, to obtain a higher degree of stability during training. Regarding the biomarker data, we had to cope with missing data. We set the covariate's value to 0, for those where no information was available. To control for missingness, we additionally included missingness indicators into our model. We augment the clinical data by entirely following the approach suggested by Pölsterl et al. (2019). We account for non-linear effects of the covariate age, by applying a natural B-spline expansion with four degrees of freedom and additionally include an interaction term between age and gender. The categorical variable education was encoded with an orthogonal polynomial coding. We replicate the pre-processing steps suggested by Pölsterl et al. (2019), in order to assess more reliably how much information for predicting AD progression is contained in the 2d-slices compared to the used point clouds. Note, however, a final comparison is still invalid as the data selection process is not identical.

In total, 795 observations were available for training, validation and testing. We split training and test data into 90% and 10% shares, whereby we further exclude 20% of

the training data for validation purposes. This results in 572 instances for training, 143 instances for validation and 80 instances for testing. Splitting the data was repeated 5 times with different splits to obtain cross-validated results.

5.3.2 Training

We trained a ResNet (He et al., 2016) with two residual blocks and two residual bottleneck blocks on the MRI 2d-slices. The latent representation is then concatenated with the tabular clinical information to train one last linear layer for predicting survival times, jointly. We adhere to the approach from Pölsterl et al. (2019) which was first introduced by Cheng et al. (2016). To ensure distinct interpretability, we also replace the simple concatentation with an orthogonalization of the unstructured latent representation on the tabular data, as discussed in chapter 2.5. A detailed description of the architectural design can be found in table 1.

We trained the model for 150 epochs using AdamW (Loshchilov & Hutter, 2017) and weight decay. While the architectural design of the ResNet was not subject to tuning, we tuned the size of the latent representation of the unstructured part with Hyperband (Li et al., 2017). The learning rate, weight decay and the scheduler gamma were also considered for tuning. In total, we tuned the model for 24h on the validation data, without cross-validation. Furthermore, we initialized the weights of the ResNet with Glorot initialization (Glorot & Bengio, 2010) and the linear weights that correspond to the tabular data were pre-trained with a linear Cox-PH model (David et al., 1972). As the liner Cox-PH model already yielded good performance results, we decided to exclude the last linear layer from weight decay. An overview of the hyperparameter specifications is given in table 6a.

The performance was evaluated with Harrell's concordance index (c-index), which evaluates whether the ordering of the predicted survival times is concordant with the observed survival times (Harrell et al., 1982). A c-index of 50% corresponds to random guessing by the model, while a c-index of 100% states that the model has perfectly learned the ordering of the observed survival times. While we can evaluate the model's discriminative power, the c-index does not indicate how accurately the model is able to predict exact survival times. Further, we compare the performance of the multi-modal approach with the performance of two baseline models. The first baseline model was trained on the MRI scans of the brain only, while the second baseline model was trained on the tabular data only. This allows us to draw three conclusions. Firstly, we can judge which modality contributes the most to predicting survival times and secondly, we can assess whether the multi-modal approach can boost performance compared to the uni-modal baseline results. Lastly, we compare the multi-modal model with and without orthogonalization to understand whether the latter has any negative impact on the performance results.

5.3.3 Results

Figure 5 illustrates the performance of the multi-modal models (with and without orthogonalization) and the baseline models. It reveals that the tabular data with a median c-index of 74.76% are stronger predictors than the 2D-slices of the MRIs with a median c-index of 63.38%. When we combine the tabular data with the MRIs we slightly outperform the linear model with a median c-index of 76.20% when no orthogonalization is applied and a median c-index of 77.00% with orthogonalization. The results are robust over all 5 splits and comparable with the results derived from Pölsterl et al. (2019). Interestingly, the multi-modal model with orthogonalization consistently outperforms all other models. Therefore, we can conclude that no performance was sacrificed for the sake of interpretability.

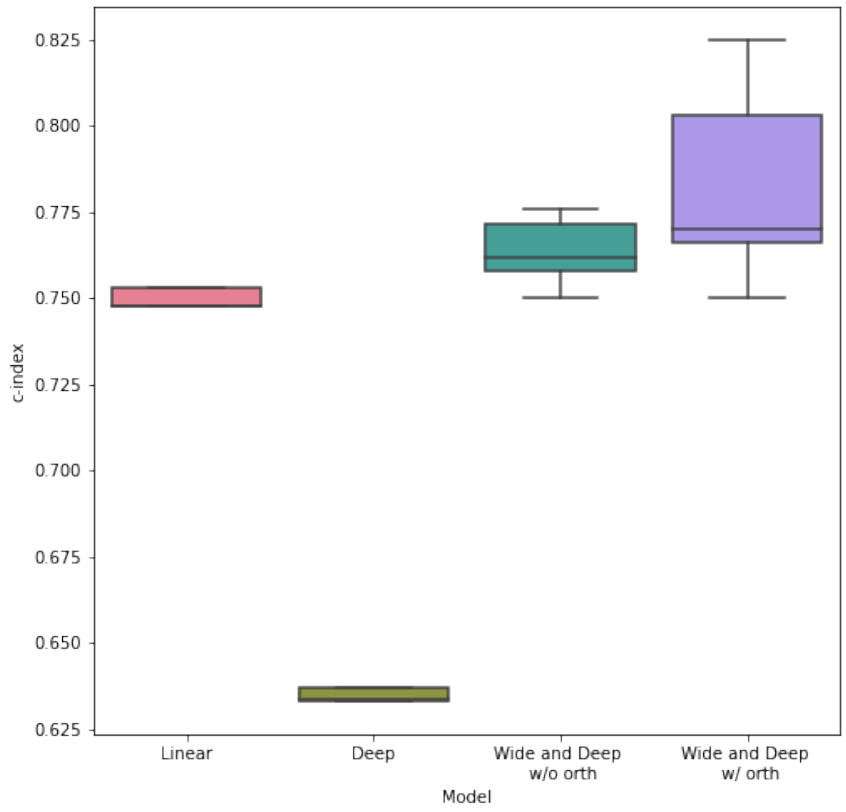


Figure 5: Performance of the different models across five random splits of the data. w/o orth: multi-modal model is trained without orthogonalization. w/ orth: multi-modal model is trained with orthogonalization

Beyond that, we compare and interpret the coefficients of the linear model with the linear part of the multi-modal model with and without orthogonalization in figure 6. The coefficients can be interpreted either as multiplicative factors on the hazard rate or directly on the log scale. The coefficients are shown across all five splits. A negative coefficient

implies that the corresponding feature contributes to a decelerated AD progression, while a positive coefficient implies contribution to an accelerated AD progression. We observe that the signs of the coefficients are in line with those reported by Pölsterl et al. (2019). Interestingly, the pre-trained coefficients remained stable while training the multi-modal framework. This contradicts the findings from Pölsterl et al. (2019), who observed a shrinkage of the coefficients towards 0 when training the multi-modal model with pre-trained weights. We argue that the stable coefficients are due to the fact that we excluded the coefficients from weight decay. The results further suggest that the orthogonalization does not impact the results and therefore, we can finally conclude that interpretability is achieved without any loss in performance.

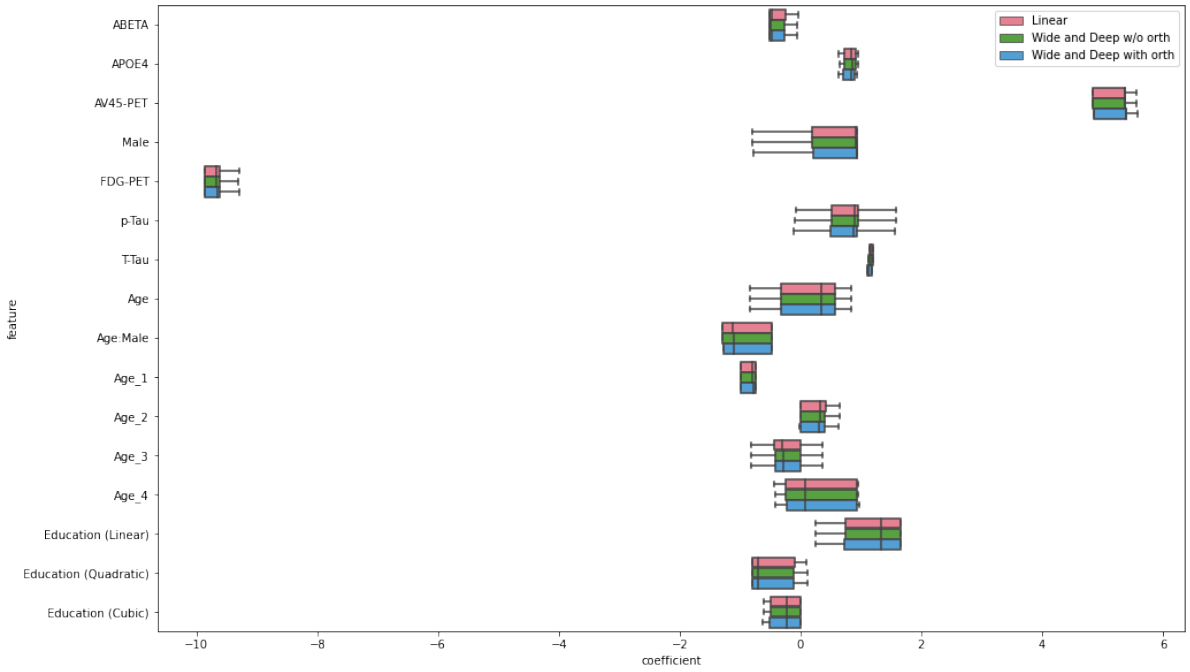


Figure 6: Comparison of the learned coefficients corresponding to the tabular clinical features. w/o orth: multi-modal model is trained without orthogonalization. w/ orth: multi-modal model is trained with orthogonalization.

5.4 Simulations

5.4.1 Data

To simulate survival times, we have to follow a two step approach. In a linear regression model, the response variable is directly associated with the covariates, the coefficients and the error terms. However, within the setting of a Cox-PH model, we associate the simulated terms to the hazard rates, so that we have to translate the hazard rates to survival times (R. Bender et al., 2005). Even though, both terms are direct expressions of

each other, the algorithm requires survival times instead of hazard rates. In what follows, we will make use of a general formula that specifies the relation between hazard rates and survival times, derived by R. Bender et al. (2005). Further, the practical implementation of the theoretical concepts is fully reproduced from the blog post published by Pölsterl (2019).

We know, that the survival function is associated with the cumulative baseline hazard $\Lambda_0(t)$ and the risk score $\boldsymbol{\eta} = \boldsymbol{x}^T\boldsymbol{\beta}$ as follows

$$S(t|\boldsymbol{x}) = \exp[-\Lambda_0(t) \times \exp(\boldsymbol{\eta})] \tag{40}$$

where

$$\Lambda_0(t) = \int_0^t \lambda_0(u)\, du \tag{41}$$

Furthermore, we can use the relation of the survival function and the distribution function to yield

$$F(t|\boldsymbol{x}) = 1 - \exp[-\Lambda_0(t) \times \exp(\boldsymbol{\eta})] \tag{42}$$

Now, we can leverage the idea of inverse transform sampling (Devroye, 2006), where we can generate random numbers from any probability distribution by using its inverse cumulative distribution F^{-1}. The probability integral transform states that given a random variable X with continuous distribution function F_X, the random variable $U = F_X(X)$ follows a uniform distribution on $[0, 1]$ (Angus, 1994). Then, the random variable $F_X^{-1}(U)$ follows the same distribution as X. Further, it is straightforward to see that $(1 - U)$ follows equivalently a uniform distribution on $[0, 1]$. Given the distribution function in equation 42, we can now write

$$U = \exp[-\Lambda_0(t) \times \exp(\boldsymbol{\eta})] \sim \text{Uni}[0, 1] \tag{43}$$

which yields

$$t = \Lambda_0^{-1}[-\log(U) \times \exp(\boldsymbol{\eta})] \tag{44}$$

Given structural assumptions about $\Lambda_0(t)$ and a specification of the predictors $\boldsymbol{\eta}$, we can now sample from a uniform distribution function to generate survival times t. In what follows, we assume that the survival times follow an exponential distribution $t \sim \exp(\lambda)$

with scale parameter $\lambda \in \mathbb{R}^+$ for which the inverse cumulative baseline hazard is defined as (see R. Bender et al., 2005)

$$\Lambda_0^{-1}(t) = \lambda^{-1}t \tag{45}$$

we can now insert equation 45 into equation 44 to yield

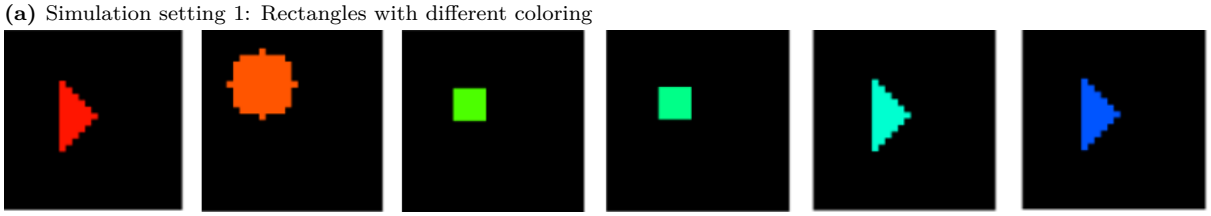
$$t = \lambda^{-1}[-\log(U) \times \exp(-\boldsymbol{\eta})] = -\frac{\log(U)}{\lambda \times \exp(\boldsymbol{\eta})} \tag{46}$$

Here, it is important to note, that assuming the survival times to be exponentially distributed implies that the baseline hazard rate $\lambda_0(t)$ is constant over time. Yet, we argue that this depicts no shortcoming, as we are merely interested in identifying a suitable simulation setting for stressing the validity of baseline generator framework. λ is chosen such that the median survival time is 20 days. Further, we randomly sample censored survival times from a uniform distribution which results in an approximate 45% of censored survival times which allows us to determine the event indicator e where $e = 1$ if an event is observed and $e = 0$ if an observation is censored (see Pölsterl, 2019).

At this point, we still have to determine an appropriate choice for simulating the risk scores $\boldsymbol{\eta} = \boldsymbol{x}^T\boldsymbol{\beta}$. To do this, our major focus is on identifying a setting which allows to fully stress the functioning of the baseline generation framework. To stress whether the framework is able to generate baseline images that satisfy the established criteria, we rely on two different settings. In the first setting, we generate images with colored rectangles on a black background. For each generated image, the positioning as well as the size of the rectangle is equivalent. Hence, the color of the rectangle which is solid represents the only varying factor and uniquely determines the risk score. During first experiments, we witnessed that training with RGB channels makes training the baseline generator considerably more complex. To simplify, we did not train the model on RGB images but on HSL images. Thereby, for different coloring of the rectangles, we only need to vary the hue (H) channel while holding the saturation (S) and lightness (L) channel constant. Hence, for each image, we randomly sample a hue value between 0 and 255 while fixing the saturation at 50 and the lightness at 100. The color palette reaches from dark red $H = 0$ to dark blue $H = 255$. The hue channel is then standardized on [-1, 1] and the only predictor for the risk scores. Hence, the simulated risk scores are equivalently in the range of [-1, 1]. Therefore, an image with a dark red rectangle corresponds to a strongly negative risk score and an image with a dark blue rectangle corresponds to a strongly positive risk score. Note that we refrained from artificially inducing a multi modal setting, as doing so would not enhance the ability to validate the *baseline generator* framework. Recall, we aim to identify for each original image a

baseline image which represents the median survival time. Hence, we expect the baseline generator to generate images that only differ from the original images in the coloring of the rectangles. Further, the color must correspond to the median survival time and as the color is the only varying factor, the generated baseline images should ideally be identical. However, within this simulation, we are limited in the ability to stress whether the baseline generator indeed outputs synthetic images that only differ from the original images in the domain-specific characteristics (coloring of the rectangle). To cope with this limitation, we also consider the second simulation setting. The second simulation is equivalent to the first, with the exception that now different geometric shapes with varying locations are introduced. In short, the rectangles are further complemented by triangles and circles, whereby the coloring of the geometric shape still uniquely determines the risk score. Now, we expect the baseline generator to output generated images that keep location and shape of the geometric figure while only changing the coloring. Thereby, all generated baseline images should represent the same coloring with shapes that are equivalent to the corresponding original image. Figure 7 shows samples of the simulated images, for both the first simulation setting (see figure 7a) and the second simulation setting (see figure 7b).

(a) Simulation setting 1: Rectangles with different coloring



(b) Simulation setting 2: Different geometric shapes with different coloring

Figure 7: Illustration of the simulated images for both settings. The images are ordered w.r.t. their associated risk scores. The leftmost image corresponds to a strong negative risk score and the rightmost to a strong positive risk score. Both figures depict merely an extract of the sampled colors.

In total, 1000 samples were available for training, validation and testing. We split training, validation and test data into 70%, 20% and 10% shares. This results in 700 instances for training, 200 instances for validation and 100 instances for testing.

5.4.2 Training

For training on the simulated data, we chose a simple three layer CNN to predict survival times. Note, its configuration was not subject to tuning, as the prediction task proved uncomplex. The architectural designs of the discriminator and the generator to train the *baseline generator* were also not subject to tuning, mainly because no valid objective to

tune upon could be identified. Instead, we adhered to the architectural design proposed by Choi et al. (2018). The generator takes as inputs a batch of real images and a one-hot encoded vector which defines the target domain. In the first part, we used two down-sampling layers with instance normalization (Ulyanov et al., 2016) and ReLU activation (Xu et al., 2015) which was followed by six residual bottleneck blocks. The last part of the generator consists of one up-sampling layer with deconvolution and one last convolution layer to obtain an output that matches the shape of the input images (see table 3). The discriminator consists of four convolutional layers with LeakyReLU activation (Xu et al. (2015)). In the standard GAN setting, the discriminator outputs one probability score for whether the image being fake or real. Instead, we follow the typical approach used in image-to-image translation setting where the standard discriminator is replaced with a PatchGAN (see Zhu et al., 2017 or Isola et al., 2017). Thereby, the PatchGAN outputs a patch P of size $N \times N$ where P_{ij} indicates whether the patch ij in the image is real or fake (see table 4).

Regarding the main hyperparameters, we distinguish between those belonging to the survival model (see table 6b) and those that belong to the baseline generator framework (see table 7). The survival model was trained for 30 epochs. To enhance stability of the training, we added an additional tanh activation layer after the last linear layer. By doing so, we did not observe any deterioration in prediction performance, possibly due to the simplicity of the prediction task. By contrast, it provides considerably more stability w.r.t. the range of predicted risk scores. When the tanh activation was not included, we observed that given the same data, model and training configurations, the predicted risk scores varied decisively while the overall prediction performance remained constant. As the objective function (see equation 22) does not enforce the model to predict the exact risk scores but only the right ordering of the risk scores, a shift or scaling of the range of predicted risk scores has no impact on the performance. Yet, leaving the stochasticity unconsidered would have serious implications for the training of the baseline generator. The threshold τ for the domain loss $\mathcal{L}_{surv}$ would vary to a non-negligible extent which would impede a robust identification of the loss weights. Thus, for one run, a given configuration may work well, while another run on the same configuration may perform poorly.

The baseline generator's specific set of hyperparameters that were subject to changes included the number of steps trained, the learning rates for the discriminator and the generator, the loss weights (λ_{rec}, λ_{surv}, λ_{rec}) as well as the linear rampup length of λ_{surv}. Further, we observed that the parameter α for balancing the domain loss λ_{surv} had a decisive impact on training and was set for both simulations to 0.6. The tolerance δ for deviating from the threshold τ was set for both experiments to 0.001, as we specifically did not want to allow strong deviations from the optimum. The generator and discriminator

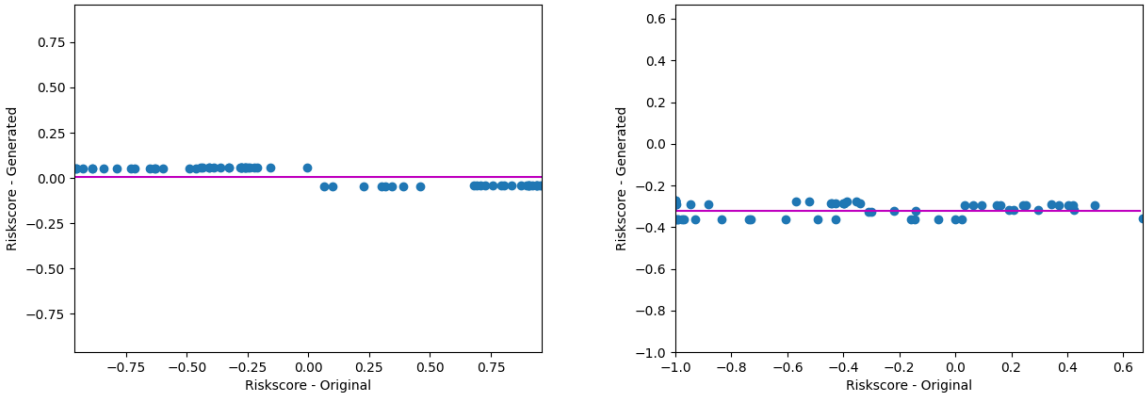
were trained for a total of 20,000 steps with equal learning rates while the generator parameters were updated every fifth parameter update of the discriminator. Again, we followed the training approach proposed by Choi et al. (2018). The loss weights λ_{rec}, λ_{surv}, λ_{gp} were determined by visually evaluating the quality of the generated baseline images and by an evaluation on how well the generated baseline images represent the specified reference point. As the desired degree of stable results could not be assured to a full extent, we logged the results on the test data every 500 steps. After the training was completed, we visually assessed on the validation data at which step the generator yielded the most satisfying results. The remaining discussions of the results are then based on the test data.

5.4.3 Results

The discussion of the results is conducted as follows: We evaluate the quality of the generated baseline images. Given satisfactory results, we then evaluate to what extent the generated baseline images serve as a better baseline choice compared to the zero baseline and a colored baseline where the color corresponds to the color of the geometric figure. It will become evident that the generated images are more appropriate from both a quantitative and a qualitative perspective. Thus completed, we will discuss whether the baseline images themselves might in fact allow for a more expressive visual interpretation of what the model has learned compared to the attribution maps.

Baseline generation Figure 10a and figure 10b show a sample of the baseline generator results for the first simulation setting. Each figure shows a sample of original images (right column) and their corresponding generated baseline images (left column). To visually assess the quality of the baseline images, we have to assess to what extent the three defined criteria (see chapter 4.2) are satisfied. Firstly, the baseline images must represent realistic samples. The background of the generated baseline images is black and the geometric shapes and their location are also in line with those from the original images, while their coloring is part of the original color palette. Thus, we conclude that the requirement for realistic samples is satisfied. Secondly, we require the baseline images to be close to their corresponding original images. Again, they must only differ from the original images in the domain-specific characteristics while all other characteristics remain constant. As discussed, to fully stress whether this criterion is satisfied, the first simulation setting is only of limited suitability. Therefore, we consider the baseline generation results corresponding to the second simulation setting (see figure 11a and figure 11b). We observe that only the coloring of the geometric shapes has changed, while the shape, its location and the background color remains unchanged. Therefore we can conclude that only the domain-specific characteristic of the original images was subject to changes while generating the baseline images. To assess whether the generated baselines belong to the specified target domains, figure 8 provides more revealing insights. We plot

the predicted risk scores that correspond to the real images (x-axis) against the predicted risk scores which correspond to the generated baseline images (y-axis). The horizontal magenta line depicts the threshold which equals the risk score that corresponds to the median survival time. To conclude that the third criteria is satisfied, we expect that the points in the negative range of the x-axis are slightly above the horizontal line and the points in the positive range of the x-axis are slightly below the horizontal line. In short, for observations for which a prolonged survival time was predicted, we seek a reference point that corresponds to a survival time slightly before the median survival time and vice versa. For the first simulation (see figure 8a) we observe that in fact the results are in line with our expectations so that we can conclude that the third criteria is satisfied. Given the second simulation setting (see figure 8b) the results are less distinct. Indeed, the predicted risk scores that correspond to the generated baseline images are centered around the magenta line, however, the clear pattern from the first simulation is not observed. Yet, we find that this is not due to a malfunctioning of the baseline generator but rather because of some unexpected behavior of the survival model. Hence, we conclude that at least to some extent the third criteria is satisfied here as well. These findings can be further confirmed by looking at the generated baseline images (see figures 10, 11). For all generated baseline images, the geometric shapes have equal coloring which represents the color that corresponds to the median survival time. Hence, the baseline generator understood to generate a reference point that corresponds to the median survival time, it must only change the coloring of the geometric shapes which must then be identical for all instances.



(a) Simulation setting 1: Rectangles with different coloring **(b)** Simulation setting 2: Different geometric shapes with different coloring

Figure 8: Scatterplot of original risk scores (x-axis) and generated risk scores (y-axis). The magenta horizontal line depicts the risk score that corresponds to the median survival time.

Figures 12, 13 and 14 illustrate the resulting attribution maps based on the chosen baselines and attribution methods. Each figure illustrates the attribution maps for four different samples of the test data. Thereby we can directly compare the attribution maps that

depend on different baseline choices. Note, the derived attribution maps based on the sampled Shapley values seem to provide visually more distinct results. Yet, as expected, the Integrated Gradients and sampled Shapley values lead to the same results.

Attribution maps: Zero baseline At first glance, the attribution maps based on the zero baseline yield meaningful results. The shapes and the position of the rectangles are well covered so that one understands which region the model focused on. Yet, independently from the input image, the pixel-wise attributions are positive (red colored) except for the red and orange colored rectangles. For that to happen, the zero baseline image must correspond to a strongly negative prediction so that almost every input image corresponds relatively to a more positive prediction. This finding contradicts the premature assumption that the zero baseline corresponds to a zero prediction (Sundararajan et al., 2017). If the reference point was semantically meaningful, an interpretation beyond a spatial one would be still feasible. As, however, this reference point is not semantically meaningful this is impossible. In fact, the zero baseline yields a negative risk score because the hue channel is set to 0 which corresponds to the most negative risk group ($H = 0$ for a red geometric shape). We conclude that even within a strongly simplified simulation setting, it can be shown that the zero baseline choice can result in misleading interpretations. If one interprets the results naively, she would always conclude that most rectangles make a positive contribution to the overall risk score. This is clearly wrong.

Attribution maps: Colored baseline By contrast, the attribution maps based on the colored baseline are less intuitive. The results suggest that the model did not spatially focus on the rectangle but rather on its surroundings. Indeed, this could be meaningful if the shape or location of the geometric figure was the predictor. As, however, the coloring of the geometric figure is the decisive factor, the attribution maps do not allow for any sensible interpretation. Beyond that, the pixel-wise attributions and their values are also hardly meaningful. While the zero baseline is a quite common choice, the colored baseline choice is admittedly arbitrary. Yet, it illustrates well how an ill-considered choice can deteriorate the entire interpretability of the attribution maps.

Attribution maps: Generated baseline Similarly to the zero baseline, the generated baselines result in attribution maps that allow a spatial interpretation of the results. More importantly, however, by using a semantically meaningful baseline, the pixel-wise attributions can be interpreted in a concise and reliable manner. We know that the baselines represent the median survival time. Therefore, we expect that for any attribution map with positive pixel-wise attributions (red) the original image corresponds to an instance for which the predicted survival time is beyond the median survival time. Similarly, we expect that for any negative attributions (blue) the predicted survival time of the original image is before the median survival time. We observe for images with dark green or

blue rectangles, the attributions are positive (red), while for images with orange or red rectangles the attributions are negative (blue). Further, we observe that the thickness of the attributions increases, the more the predicted survival time deviates from the median survival time. Therefore, we can conclude that the generated baseline images allow for a reliable interpretation that goes beyond a spatial one. Thus, the generated baseline is clearly superior to the remaining baseline choices.

In chapter 3.2, it became evident that the completeness axiom represents the ultimate axiom an attribution method should satisfy. According to the literature, the axiom is fulfilled when the attribution method fully captures the prediction difference between the baseline and the input. We argued, however, in chapter 4.1, that the completeness axiom is only favorable, if the baseline prediction is semantically meaningful. To illustrate that, we selected four different samples for which to report both the actual corresponding predictions and the differences between the prediction and the prediction that corresponds to a respective baseline choice (see table 9). Further, table 9 enlists for each sample and each baseline the sum of attributions generated by Integrated Gradients and sampled Shapley values, respectively. It reveals that independently of the input and baseline, both attribution methods do fully capture the prediction differences. Consequently, to fulfill the theoretical axiom of completeness, the chosen baseline represents no critical factor. This finding illustrates that even though theoretically no assumption is violated, there is no guarantee that the practical implications are meaningful. The predictions that correspond to the zero baseline or the colored baseline can simply not be understood and therefore the completeness axiom does not help to yield more interpretable attribution maps.

	Delta			**Integrated gradient**			**Shapley value**		
Prediction (1)	Zero (2)	Colored (3)	Generated (4)	Zero (5)	Colored (6)	Generated (7)	Zero (8)	Colored (9)	Generated (10)
-0.96	-0.14	-0.14	-0.98	-0.14	-0.14	-0.98	-0.14	-0.14	-0.98
0.94	1.76	0.97	0.94	1.76	0.97	0.94	1.76	0.97	0.94
-0.52	0.3	0.3	-0.54	0.3	0.3	-0.54	0.3	0.3	-0.54
0.76	1.58	0.92	0.75	1.58	0.92	0.75	1.58	0.92	0.75

Figure 9: The table illustrates that for any sample of the test data, the completeness axiom is satisfied independently of the baseline choice or attribution method. Column (1): prediction corresponding to an original image. Columns (2) - (4): reflect the prediction difference between the prediction of the original image and the prediction corresponding to the respective baseline. Columns (5) - (7): correspond to the sum of attributions derived via Integrated Gradients for each baseline, respectively. Columns (8) - (10): correspond to the sum of attributions derived via sampled Shapley values for each baseline, respectively.

This whole work was devoted to emphasize the importance of a proper identification of the baseline. Yet, the generated baselines are still considered as a mere input to the attribution methods. However, we argue that the baseline itself provides a considerable degree of interpretability. The baseline always serves as a semantically meaningful reference point against which we compare the actual prediction. Thereby we pursue to answer

two fundamental questions. Firstly, which structures in the images must change and secondly, how must the structures change to yield the reference point which corresponds in our case to the median survival time. In fact, the attribution maps answer the first question, but on closer consideration do not provide any insights for reliably answering the second. For instance, given the two simulation settings, the attribution maps do only allow for a spatial interpretation: we understand which structures must change. However, we only understand the coloring of the geometric shapes as the domain-specific characteristic, if we directly look at the generated baseline images. Even if we can infer from the attribution maps that the color represents the domain specific characteristic, we still do not know which color represents the reference point. Hence, we conclude despite an attribution method satisfying all theoretical axioms and a properly identified baseline, the attribution maps only provide limited insights on what the model has learned. A more detailed discussion of the limitations of attribution maps and the potential of the baseline images will be given in chapter 6.1.

6 Discussion

6.1 Baseline images and interpretability

While we primarily focused on the necessity of a proper identification of the baseline as input for the attribution methods, we also pointed towards the stand alone importance of the generated baselines. While attribution maps allow for a spatial interpretation, the generated baseline images do provide insights on *what* is different. This finding is further supported by Narayanaswamy et al. (2020) who argue that the attribution methods merely provide a spatial support by indicating where the model looked, but do not provide any insights on how the structures must change to yield the reference point. This, however, becomes necessary if the model distills information unknown to the domain expert, as she might be interested in a visual understanding of how MRIs differ. As this cannot be provided by the attribution methods, Narayanaswamy et al. (2020) conclude that attribution methods are useful for validation but lack in the capability of exploration.

Narayanaswamy et al. (2020) also argue that it might be revealing to not only understand the structural differences between input and reference point, but to also consider the interpolations between them. With our proposed framework, we can implement that. Recall, our framework relies on the StarGAN (Choi et al., 2018) which allows to translate an image to a variety of different domains. So far, we only considered an image translation from the source domain (before/beyond median survival time) to the respective target domain (beyond/before median survival time). Thereby, the threshold which corresponds to the median survival time uniquely identifies the boundary between source and target domain. Yet, our baseline generator framework is neither restricted to one target domain, nor to the median survival time as threshold choice. Hence, we can directly and simulta-

neously translate an original image towards a reference point that represents any quantile of the survival times. For instance, we can visually evaluate how an image must change so that it corresponds to the 25%-quantile, the median quantile, and the 75%-quantile or any other quantile of the survival times. By doing so, we can then understand how gradual changes in the input affect the predictions. This in return yields a complete picture of what the model has learned. Therefore, we can potentially yield a more sophisticated understanding on how small structural changes in the brain affect the progression of AD.

The conjecture that the generated baseline images provides a better explanation than the attribution maps is further confirmed by Jeyakumar et al. (2020). In their study, the authors evaluate which explanation method is preferred by end-users who have no explicit expertise in machine learning. They trained a simple CNN on Cifar10 (Krizhevsky et al., n.d.) and applied a variety of explanation methods on the predictions. Among others, the explanation methods included Shapley values, saliency maps (Simonyan et al., 2013), LIME (Ribeiro et al., 2016) and explanation-by-example (Caruana et al., 1999). The authors then asked the users which explanation method they preferred. They found that approximately 90% of the users preferred the explanation-by-example method. The authors found that the prioritization was due to the intuitive and semantically meaningful results - the nearest training examples represented the considered input and the model's decision well. By contrast, the remaining methods often gave fairly unintuitive results. For instance, for an image with an airplane depicted, the remaining methods marked the background sky as important, while the users would rather expect to mark the airplane itself as important. Again, by obtaining a direct reference point, we can reliably distinguish between the domain-specific and the domain-unspecific characteristics which is not always guaranteed with attribution maps.

Note, however, that a definite evaluation of whether the baseline images or the attribution maps provide better insights is impossible but rather depends on the preferences of the domain expert and also on the domain. In case of the predicting AD progression the resulting attribution maps might be less ambiguous and the domain expert might be capable to infer from the attribution maps *what* the structural changes actually are. In this case, the baseline images would probably not provide too much new information. Yet, regardless of the amount of additional information provided by the baseline images, we can still conclude that the baseline images can at least clarify interpretations. Hence, we recommend a holistic view where both outputs are considered jointly.

6.2 Generalizability to further survival models

The discussed results confirmed the theoretically derived advantageous properties of the *baseline generator* framework. It was shown that the generated baselines represent the only valid baseline choice for the considered attribution methods (among the considered

baseline choices). Further, in chapter 6.1 it was emphasized that the generated baselines do not only serve as inputs for the attribution methods, but are also key to yield a thorough understanding of what the model has learned. In what follows, we will stress whether the *baseline generator* framework can be transferred to a variety of other survival models that might be in some cases more appropriate than the Cox-PH model. To pursue this we consider the additive Cox model, the piecewise exponential models (PEM) (Cox, 1972) as well as the piecewise additive mixed models (PAMM) (A. Bender et al., 2018).

The additive Cox-PH model differs from the Cox-PH model to the extent that it allows for nonlinear and time-variant effects of the covariates, so that the hazard can be written as

$$\lambda(t, \boldsymbol{x}) = \lambda_0(t) \exp(\sum_{p=1}^{P} \boldsymbol{x_p}(t)\beta_p + \sum_{l=1}^{L} f_l(\boldsymbol{x_l}(t))) \tag{47}$$

where the covariates $\boldsymbol{x}(t)$ now depend on time t and $\boldsymbol{\beta}$ capture the linear effects and the function $f(.)$ captures the non-linear effects of the covariates $\boldsymbol{x}(t)$ on the hazard which is often defined via basis representations. From a practical perspective, we have to prepare the data set in a longitudinal format, whereby the number of observations for each subject depends on the length of the observation. In this case, the applicability of the baseline generator framework breaks, as a specific covariate has varying effects on the hazard and therefore on the survival time. Therefore, we cannot uniquely identify baseline values for all covariates $\boldsymbol{x}(t)$ that represent the median survival time.

Yet, it remains to be clarified if we can maintain the applicability of the framework, if we make the restrictive assumption that only the structured part has time-varying effects on the hazard. Then, we can rewrite equation 47 as

$$\lambda(t, \boldsymbol{x}) = \lambda_0(t) \exp(\sum_{p=1}^{P} \boldsymbol{x_p}(t)\beta_p + \sum_{q=1}^{Q} \boldsymbol{u_q}\gamma_q + \sum_{l=1}^{L} f_l(\boldsymbol{x_l}(t))) \tag{48}$$

where the covariates $\boldsymbol{u}$ correspond to the latent representation of the unstructured part and the covariates $\boldsymbol{x}$ to the structured part. The coefficients $\boldsymbol{\gamma}$ capture only time-constant effects and are therefore independent from time t. Then, the survival function $S(t)$ is given by

$$S(t) = \exp\Big(- \Lambda_0(t) \exp\Big(\sum_{p=1}^{P} \boldsymbol{x_p}(t)\beta_p + \sum_{q=1}^{Q} \boldsymbol{u_q}\gamma_q + \sum_{l=1}^{L} f_l(\boldsymbol{x_l}(t)\Big)\Big) \tag{49}$$

To identify the threshold that corresponds to the median survival time $S(t) = 0.5$, we

can rearrange equation 49 as follows

$$0.5 = \exp\Big(- \Lambda_0(t) \exp\Big(\sum_{p=1}^{P} \boldsymbol{x_p}(t)\beta_p + \sum_{q=1}^{Q} \boldsymbol{u_q}\gamma_q + \sum_{l=1}^{L} f_l(\boldsymbol{x_l}(t)\Big)\Big) \tag{50}$$

$$\log 0.5 = -\Lambda_0(t) \exp\Big(\sum_{p=1}^{P} \boldsymbol{x_p}(t)\beta_p + \sum_{q=1}^{Q} \boldsymbol{u_q}\gamma_q + \sum_{l=1}^{L} f_l(\boldsymbol{x_l}(t)\Big) \tag{51}$$

$$\tau = \sum_{q=1}^{Q} \boldsymbol{u_q}\gamma_q = \log\Big(- \frac{\log 0.5}{\Lambda_0(t) \exp(\sum_{p=1}^{P} \boldsymbol{x_p}(t)\beta_p + \sum_{l=1}^{L} f_l(\boldsymbol{x_l}(t))}\Big) \tag{52}$$

Still, however, the survival time depends on time varying effects and thus a unique identification is not possible. Therefore, we have to consider the time-constant unstructured part in isolation which reduces equation 52 to

$$\tau = \sum_{q=1}^{Q} \boldsymbol{u_q}\gamma_q = \log\Big(- \frac{\log 0.5}{\Lambda_0(t)}\Big) \tag{53}$$

If the effects from the unstructured part $\boldsymbol{u}$ and the structured part $\boldsymbol{x(t)}$ are independent, this approach is valid. We can still uniquely identify a threshold τ that corresponds to the median survival time. Note, however, that the median survival time does not refer to the overall predicted survival times anymore. Instead, it refers only to the effects that stem from the unstructured part. If, however, the effects are not independent, we have to apply the *orthogonalization trick*. Then, however, the orthogonalized unstructured latent representation $\tilde{\boldsymbol{u}}$ is not constant over all time intervals anymore as it depends on the space spanned by the time-varying structured part $\boldsymbol{x}(t)$. To still maintain applicability, we could additionally assume that it is sufficient to only consider the risk scores $\sum_{q=1}^{Q} \tilde{\boldsymbol{u}}_{\boldsymbol{q}}\gamma_q$ where $\tilde{\boldsymbol{u}}_{\boldsymbol{q}}$ corresponds to the latent representation of each subject at entry of the study. Then, we can assure that for each subject we obtain a unique latent representation that does not depend on the time t. To what extent such an assumption is too restrictive cannot be evaluated a *priori*.

In what follows, we can show that the same reasoning applies for a PAMM. The PAMM as described by Kopper et al. (2020) is specified as

$$\lambda(t|\boldsymbol{x}) = \exp(f(\boldsymbol{x}(t), t)) \tag{54}$$

which describes the hazard λ at time $t \in T$, conditional on a vector $\boldsymbol{x}(t) \in \mathbb{R}^p$ which can include time-varying as well as time-constant covariates. The function $f(.)$ specifies the effect of the features $\boldsymbol{x}(t)$ on the hazard. If we omit the dependence on t, equation 54 reduces to the Cox-PH model. If t is not omitted, equation 54 is approximated

via piecewise constant hazards. To pursue this, we decompose time t into intervals $\alpha_0, \alpha_1, ..., \alpha_m$:

$$(0 = \alpha_0, \alpha_1], (\alpha_1, \alpha_2], ..., (\alpha_{t-1}, \alpha_t], ..., (\alpha_{m-1}, \alpha_m], (\alpha_m, \infty) \tag{55}$$

whereby we assume for each interval a piecewise constant baseline hazard:

$$\lambda_0(t) = \lambda_k \text{ for } t \in I_k = (\alpha_{k-1}, \alpha_k] \tag{56}$$

To estimate the piecewise baseline hazards, we have to transform the data such that each row corresponds to one time interval $(\alpha_{k-1}, \alpha_k], k = 1, ..., K$. Without going more into detail [6], the hazard in its explicit form can then be written as

$$\lambda_k = \exp(\log \lambda_0(t) + \sum_{p=1}^{P} \boldsymbol{x_{k,p}} \beta_p + \sum_{l=1}^{L} f_l(\boldsymbol{x_{k,l}})) \tag{57}$$

where $\log \lambda_0(t)$ corresponds to the baseline hazard and $\boldsymbol{\beta}$ captures the linear effects and $f_l(\boldsymbol{x_{k,l}})$ the univariate, non-linear covariate effects and is in general defined via basis representations. The modelling approach differs from the additive Cox-PH model by estimating a hazard for each interval $k = 1, ..., K$ separately. If we assume again that the effects derived from the unstructured part are time constant, we can rewrite equation 57 as

$$\lambda_k = \exp(\log \lambda_0(t) + \sum_{p=1}^{P} \boldsymbol{x_{k,p}} \beta_p + \sum_{q=1}^{Q} \boldsymbol{u_q} \gamma_q + \sum_{l=1}^{L} f_l(\boldsymbol{x_{k,l}})) \tag{58}$$

where the effects γ_q are constant for each time interval $k = 1, ..., K$. Now, we can derive the survival function $S(t)$ as follows

$$S(t) = \exp(-\int_0^t \lambda(s|x)\, ds) \tag{59}$$

$$S(t) = \exp(-\sum_{k=1}^{M} \lambda_k) \tag{60}$$

whereby the integral can be simplified to a summation term as we have discrete time interval and M defines the last interval that entails time point t. Plugging equation 58 into equation 60 then yields

[6] Kopper et al. (2020) provides a more detailed explanation of PAMs.

$$S(t) = \exp(-\sum_{k=1}^{M} \exp(\log \lambda_0(t) + \sum_{p=1}^{P} \boldsymbol{x_{k,p}}\beta_p + \sum_{q=1}^{Q} \boldsymbol{u_q}\gamma_q + \sum_{l=1}^{L} f_l(\boldsymbol{x_{k,l}}))) \tag{61}$$

Again, the survival time still depends on the time-constants and time varying effects. Hence, identifying the threshold that corresponds to the median survival time $S(t) = 0.5$ is still not feasible. If we adhere to the same reasoning as above, we can determine the median survival time equivalently to equation 53. Again, to what extent this is justified must be considered on a case by case basis. The same applies to the PEM, as we merely omit the basis representations $f(.)$ from the equation.

To transfer the *baseline generator* to a variety of other survival models, we have to induce some restrictive assumptions. First, the effects that correspond to the unstructured part must be time constant. Second, we must assume that the *orthogonalization trick* does not alter the latent representation of the unstructured part too much. Then, it might be reasonable to take the orthogonalized latent representation that corresponds to the time at study entry of each subject. To what extent the assumption of time-constant effects is justified, must be clarified for each application individually. If, however, these assumptions are justified, the *baseline generator* is arguably strong as the complexity of the structured part is not restricted.

6.3 External Validity

The external validity of our proposed framework depicts another strong property. In chapter 6.1, it was discussed that the framework is not restricted to the median survival time as a reference point, but rather the reference point can be defined by any quantile of the survival times. This comes with the advantage that the framework is applicable to domains where the median survival time might be a sub optimal choice. For instance, Miao et al. (2018) predicted the hospital mortality for patients with heart failure and showed that approximately 15% of the patients die within the first 40 days, while afterwards the hazard for dying approaches almost 0. In this case, the median survival time as threshold would clearly depict a poor choice, as during the whole study period the total share of observed events was considerably below the 50%. Hence, in this case, it would be advisable to set the threshold to e.g. the 15%-quantile. Widodo and Yang (2011) studied the degradation of machines within the context of survival analysis. The authors observed that the hazard for failure remains quite low for a long period of time and after a certain time the survival probability decreases drastically. While more than 50% of the observed machines failed during the study period, it would still be not advisable to choose the median survival time as threshold. A more insightful threshold corresponds to the time when the steep decline in survival probabilities occurred. To sum up, independently from domain and the general structure of the survival probabilities, our framework is

arguably robust and therefore, suitable to many applications. This is not only true due to the flexible threshold choices, but also because of the generalizability to a variety of survival models, as discussed in chapter 6.2.

7 Conclusion

This work established the *baseline generator* framework to address the baseline selection problem (Shih et al., 2020) in the context of survival analysis and proposed a solution for it. A theoretical discussion and an empirical verification of the framework provided detailed insights on its functioning. The raison d'être of this framework was further strengthened in chapter 6. There, we pointed towards further advantageous properties of this framework that were not covered in this work.

While the framework is robust in theory, its applicability to more complex use cases must be still verified. While the simulation settings indeed proved that the framework is capable of working, one still has to acknowledge the simplicity of the task. For instance, the location and shape of the geometric shapes are in fact domain-unspecific, but they are not observation-specific. Every shape and location is observed multiple times in the data. It is probably justified to argue that the task would be more complex if the domain-unspecific characteristics were in fact unique to each observation, respectively. Yet, to what extent this would impact the performance of the *baseline generator* cannot be determined a *priori*. Hence, it is essential to test this framework on real data where the domain-unspecific characteristics are in fact unique to each observation. Therefore within the next steps the framework must be applied on the ADNI data. This, however, would have exceeded the scope of this work and therefore we leave this open for future research. If the framework enables to yield robust results on the ADNI data, we could conclude to have made a significant contribution to make predictions of AD progression interpretable.

Yet, applying the *baseline generator* framework on the simulated data was an important exercise. By knowing the ground-truth, we were able to reliably assess the proposed framework. Reliably assessing the proposed framework becomes impossible when applying the framework on the ADNI data as the ground-truth is unknown.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*.

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2017). Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*.

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2019). Gradient-based attribution methods. *Explainable ai: Interpreting, explaining and visualizing deep learning* (pp. 169–191). Springer.

Ancona, M., Oztireli, C., & Gross, M. (2019). Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. *International Conference on Machine Learning*, 272–281.

Angus, J. E. (1994). The probability integral transform and related results. *SIAM review*, *36*(4), 652–654.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. *International conference on machine learning*, 214–223.

Aumann, R. J., & Shapley, L. S. (2015). *Values of non-atomic games*. Princeton University Press.

Bass, C., da Silva, M., Sudre, C., Tudosiu, P.-D., Smith, S., & Robinson, E. (2020). Icam: Interpretable classification via disentangled representations and feature attribution mapping. *arXiv preprint arXiv:2006.08287*.

Bass, C., da Silva, M., Sudre, C., Williams, L. Z., Tudosiu, P.-D., Alfaro-Almagro, F., Fitzgibbon, S. P., Glasser, M. F., Smith, S. M., & Robinson, E. C. (2021). Icam-reg: Interpretable classification and regression with feature attribution for mapping neurological phenotypes in individual scans. *arXiv preprint arXiv:2103.02561*.

Baumgartner, C. F., Koch, L. M., Tezcan, K. C., Ang, J. X., & Konukoglu, E. (2018). Visual feature attribution using wasserstein gans. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8309–8319.

Bender, A., Groll, A., & Scheipl, F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, *18*(3-4), 299–321.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, *24*(11), 1713–1723.

Caruana, R., Kangarloo, H., Dionisio, J. D., Sinha, U., & Johnson, D. (1999). Case-based explanation of non-case-based learning methods. *Proceedings of the AMIA Symposium*, 212.

Castro, J., Gómez, D., & Tejada, J. (2009). Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, *36*(5), 1726–1730.

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., et al. (2016). Wide & deep learning for recommender systems. *Proceedings of the 1st workshop on deep learning for recommender systems*, 7–10.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Clem, M. A., Holliday, R. P., Pandya, S., Hynan, L. S., Lacritz, L. H., & Woon, F. L. (2017). Predictors that a diagnosis of mild cognitive impairment will remain stable 3 years later. *Cognitive and behavioral neurology: official journal of the Society for Behavioral and Cognitive Neurology*, *30*(1), 8.

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

David, C. R. et al. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, *34*(2), 187–220.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Devroye, L. (2006). Nonuniform random variate generation. *Handbooks in operations research and management science*, *13*, 83–121.

Dubois, B., Feldman, H. H., Jacova, C., Cummings, J. L., DeKosky, S. T., Barberger-Gateau, P., Delacourte, A., Frisoni, G., Fox, N. C., Galasko, D., et al. (2010). Revising the definition of alzheimer's disease: A new lexicon. *The Lancet Neurology*, *9*(11), 1118–1127.

Faraggi, D., & Simon, R. (1995). A neural network model for survival data. *Statistics in medicine*, *14*(1), 73–82.

Fong, R., & Vedaldi, A. (2019). Explanations for attributing deep neural network predictions. *Explainable ai: Interpreting, explaining and visualizing deep learning* (pp. 149–167). Springer.

Friedman, E. J. (2004). Paths and consistency in additive cost sharing. *International Journal of Game Theory*, *32*(4), 501–518.

Gast, J., & Roth, S. (2018). Lightweight probabilistic deep networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3369–3378.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256.

Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D., & Sebag, M. (2018). Learning functional causal models with generative neural networks. *Explainable and interpretable models in computer vision and machine learning* (pp. 39–80). Springer.

Goukasian, N., Porat, S., Blanken, A., Avila, D., Zlatev, D., Hurtz, S., Hwang, K. S., Pierce, J., Joshi, S. H., Woo, E., et al. (2019). Cognitive correlates of hippocampal atrophy and ventricular enlargement in adults with or without mild cognitive impairment. *Dementia and geriatric cognitive disorders extra*, *9*(2), 281–293.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, *247*(18), 2543–2546.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., et al. (2008). The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *27*(4), 685–691.

Jeyakumar, J. V., Noor, J., Cheng, Y.-H., Garcia, L., & Srivastava, M. (2020). How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, *33*.

Jha, A., Aicher, J. K., Gazzara, M. R., Singh, D., & Barash, Y. (2020). Enhanced integrated gradients: Improving interpretability of deep learning models using splicing codes as a case study. *Genome biology*, *21*(1), 1–22.

Knopman, D. S., & Petersen, R. C. (2014). Mild cognitive impairment and mild dementia: A clinical perspective. *Mayo Clinic Proceedings*, *89*(10), 1452–1459.

Kopper, P., Pölsterl, S., Wachinger, C., Bischl, B., Bender, A., & Rügamer, D. (2020). Semi-structured deep piecewise exponential models. *arXiv preprint arXiv:2011.05824*.

Krizhevsky, A., Nair, V., & Hinton, G. (n.d.). Cifar-10 (canadian institute for advanced research). http://www.cs.toronto.edu/~kriz/cifar.html

Kumar, A., Sattigeri, P., & Balakrishnan, A. (2017). Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*.

Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., & Yang, M.-H. (2018). Diverse image-to-image translation via disentangled representations. *Proceedings of the European conference on computer vision (ECCV)*, 35–51.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, *18*(1), 6765–6816.

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.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Mathieu, E., Rainforth, T., Siddharth, N., & Teh, Y. W. (2019). Disentangling disentanglement in variational autoencoders. *International Conference on Machine Learning*, 4402–4412.

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. (2011). 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.

Merrick, L., & Taly, A. (2020). The explanation game: Explaining machine learning models using shapley values. *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 17–38.

Miao, F., Cai, Y.-P., Zhang, Y.-X., Fan, X.-M., & Li, Y. (2018). Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest. *IEEE Access*, *6*, 7244–7253.

Montavon, G. (2019). Gradient-based vs. propagation-based explanations: An axiomatic comparison. *Explainable ai: Interpreting, explaining and visualizing deep learning* (pp. 253–265). Springer.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., Initiative, A. D. N., et al. (2015). Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects. *Neuroimage*, *104*, 398–412.

Nakagawa, T., Ishida, M., Naito, J., Nagai, A., Yamaguchi, S., Onoda, K., & Initiative, A. D. N. (2020). Prediction of conversion to alzheimer's disease using deep survival analysis of mri images. *Brain communications*, *2*(1), fcaa057.

Narayanaswamy, A., Venugopalan, S., Webster, D. R., Peng, L., Corrado, G. S., Ruamviboonsuk, P., Bavishi, P., Brenner, M., Nelson, P. C., & Varadarajan, A. V. (2020). Scientific discovery by generating counterfactuals using image translation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 273–283.

Oh, S. J., Schiele, B., & Fritz, M. (2019). Towards reverse-engineering black-box neural networks. *Explainable ai: Interpreting, explaining and visualizing deep learning* (pp. 121–144). Springer.

Platero, C., & Tobar, M. C. (2020). Longitudinal survival analysis and two-group comparison for predicting the progression of mild cognitive impairment to alzheimer's disease. *Journal of Neuroscience Methods*, *341*, 108698.

Pölsterl, S. (2019). *Survival analysis for deep learning*. Retrieved April 1, 2021, from https://k-d-w.org/blog/2019/07/survival-analysis-for-deep-learning/

Pölsterl, S., Sarasua, I., Gutiérrez-Becker, B., & Wachinger, C. (2019). A wide and deep neural network for survival analysis from anatomical shape and tabular clinical data. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 453–464.

Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. *Explainable and interpretable models in computer vision and machine learning* (pp. 19–36). Springer.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Rountree, S. D., Atri, A., Lopez, O. L., & Doody, R. S. (2013). Effectiveness of antidementia drugs in delaying alzheimer's disease progression. *Alzheimer's & Dementia*, *9*(3), 338–345.

Rügamer, D., Kolb, C., & Klein, N. (2020). A unifying network architecture for semi-structured deep distributional learning. *arXiv preprint arXiv:2002.05777*.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge university press.

Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., et al. (2019). Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*, *126*(4), 552–564.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, *2*(28), 307–317.

Shih, S.-M., Tien, P.-J., & Karnin, Z. (2020). Ganmex: One-vs-one attributions using gan-based model explainability. *arXiv preprint arXiv:2011.06015*.

Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *Journal of Imaging*, *6*(6), 52.

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. (2011). 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.

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, *41*(3), 647–665.

Sturmfels, P., Lundberg, S., & Lee, S.-I. (2020). Visualizing the impact of feature attribution baselines. *Distill*, *5*(1), e22.

Sundararajan, M., & Najmi, A. (2020). The many shapley values for model explanation. *International Conference on Machine Learning*, 9269–9278.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*, 3319–3328.

Tong, T., Gao, Q., Guerrero, R., Ledig, C., Chen, L., Rueckert, D., Initiative, A. D. N., et al. (2016). A novel grading biomarker for the prediction of conversion from mild cognitive impairment to alzheimer's disease. *IEEE Transactions on Biomedical Engineering*, *64*(1), 155–165.

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Widodo, A., & Yang, B.-S. (2011). Application of relevance vector machine and survival probability to machine degradation assessment. *Expert Systems with Applications*, *38*(3), 2592–2599.

Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Yiannopoulou, K. G., & Papageorgiou, S. G. (2013). Current and future treatments for alzheimer's disease. *Therapeutic advances in neurological disorders*, *6*(1), 19–33.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Appendices

A Architectural design

A.1 ADNI

Layer	Layer information
Input: $160 \times 128 \times 1$	
Conv Block	5×5 conv, 32 BN & ReLU, stride 2, padding 2
Residual Block	3×3 conv, 32 BN & LR, stride 1, padding 1
Conv Block	3×3 conv, 64 BN & ReLU, stride 2, padding 1
Residual Block	3×3 conv, 64 BN & LR, stride 1, padding 1
Conv Block	3×3 conv, 128 BN & ReLU, stride 2, padding 1
Residual Bottleneck Block	3×3 conv, 128 BN & LR, stride 1, padding 1, BF 64
Conv Block	3×3 conv, 256 BN & ReLU, stride 2, padding 1
Residual Bottleneck Block	3×3 conv, 256 BN & LR, stride 1, padding 1, BF 64
Conv Block	1×1 conv, 4 BN & ReLU, stride 1, padding 0
Linear Layer	FC $320 \rightarrow 60$, ReLU

Table 1: Survival model architecture (ADNI). BN: Batch Normalization, LR: LeakyReLU, FC: Fully connected, BF: bottleneck filters for the Residual Bottleneck Block. The Residual Block consists of two convolutional layer with batch normalization and ReLU activation and a skip connection. The Residual Bottleneck Block consists of three convolutional layer with batch normalization and ReLU activation and a skip connection. Leaky ReLU activation was used with a negative slope of 0.01.

A.2 Simulations

Layer	Layer information
Input: $28 \times 28 \times 3$	
Conv Block	5×5 conv, 6 ReLU & MP(kernel=2), stride 1, padding 0
Conv Block	5×5 conv, 12 ReLU & MP(kernel=2), stride 1, padding 0
Linear Layer	FC $192 \rightarrow 120$, ReLU
Linear Layer	FC $120 \rightarrow 84$, ReLU
Linear Layer	FC $84 \rightarrow 10$

Table 2: Survival model architecture (Simulation). FC: Fully connected, MP: Max Pooling layer with kernel size 2.

Part	Layer	Layer information
Down-sampling	Conv Block	7×7 conv, 64 IN & ReLU, stride 1, padding 1
	Conv Block	5×5 conv, 128 IN & ReLU, stride 2, padding 1
Bottleneck	Residual Block	3×3 conv, 128 IN & ReLU, stride 1, padding 1
	Residual Block	3×3 conv, 128 IN & ReLU, stride 1, padding 1
	Residual Block	3×3 conv, 128 IN & ReLU, stride 1, padding 1
	Residual Block	3×3 conv, 128 IN & ReLU, stride 1, padding 1
	Residual Block	3×3 conv, 128 IN & ReLU, stride 1, padding 1
	Residual Block	3×3 conv, 128 IN & ReLU, stride 1, padding 1
Up-sampling	Deconv Block	3×3 upconv, 64 IN & ReLU, stride 1, padding 2
	Conv Block	5×5 conv, 3 ReLU, stride 1, padding 3

Table 3: Generator architecture (Simulation). IN: Instance Normalization. For all layers, we use instance normalization except the last one. The instance normalization Residual Blocks consists of two convolutional blocks and a skip connection.

Layer	Layer information
Conv Block	4×4 conv, 64 LR, stride 2, padding 1
Conv Block	4×4 conv, 128 LR, stride 2, padding 1
Conv Block	4×4 conv, 256 LR, stride 2, padding 1
Conv Block	3×3 conv, 256 LR, stride 1, padding 0

Table 4: Discriminator architecture (Simulation). LR: Leaky ReLU. Leaky ReLU activation was used with a negative slope of 0.01.

B Main hyperparameters

(a) Hyperparameters (ADNI)

Parameter	Values
MRIs only	
Batch size	256
Latent space dimension	60
Optimizer	AdamW
AdamW: weight decay	13.13
AdamW: learning rate	0.0000335
AdamW: scheduler gamma	0.99456
Number of epochs	150
MRIs + tabular data	
Batch size	256
Latent space dimension	60
Optimizer	AdamW
AdamW: weight decay	2.21404
AdamW: learning rate	0.0001329
AdamW: scheduler gamma	0.990049
Number of epochs	150

(b) Hyperparameters (Simulation)

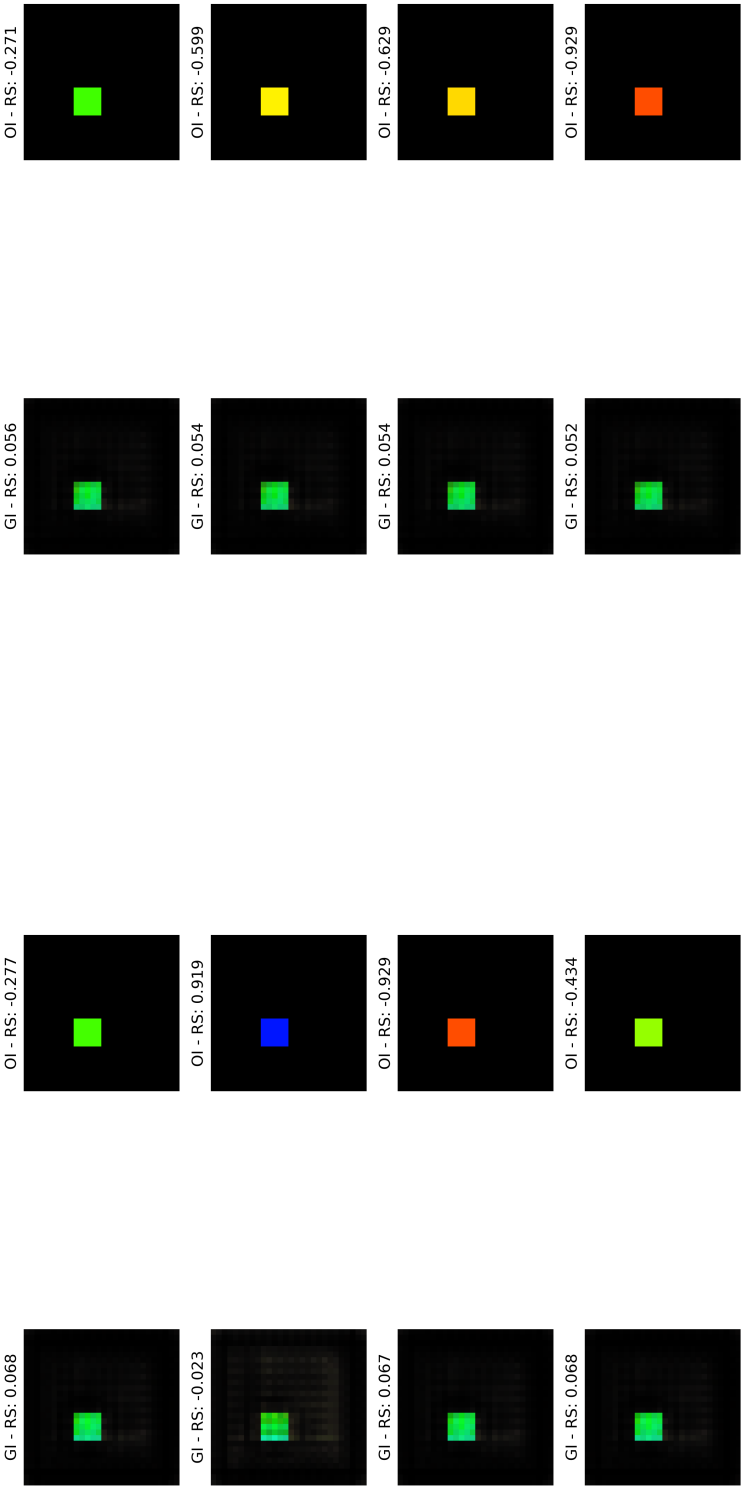
Parameter	Values
Batch size	64
Latent space dimension	10
Optimizer	AdamW
AdamW: weight decay	0.0
AdamW: learning rate	0.01
AdamW: scheduler gamma	0.95
Number of epochs	30

Table 5: Survival models: main hyperparameters. For models that were trained on the ADNI data: We distinguish between the configuration used for training on the MRIs only and the configuration used for training on the MRIs and tabular data jointly.

Model	Parameter	Value (SIM1)	Value (SIM2)
Discriminator	optimizer	Adam	Adam
	Adam: learning rate	0.0001	0.0001
	λ_{gp}	10.0	10.0
	number of steps trained	20,000	20,000
Generator	optimizer	Adam	Adam
	Adam: learning rate	0.0001	0.0001
	λ_{rec}	2.0	2.0
	λ_{surv}	200.0	500.0
	λ_{surv}: linear rampup length	10,000	10,000
	$\mathcal{L}_{surv}$: α	0.6	0.6
	$\mathcal{L}_{surv}$: δ	0.001	0.001
	number of steps trained	20,000	20,000

Table 7: Baseline generator: main hyperparameters. SIM1: Simulation setting 1 (colored rectangles); SIM2: Simulation setting 2 (colored geometric figures)

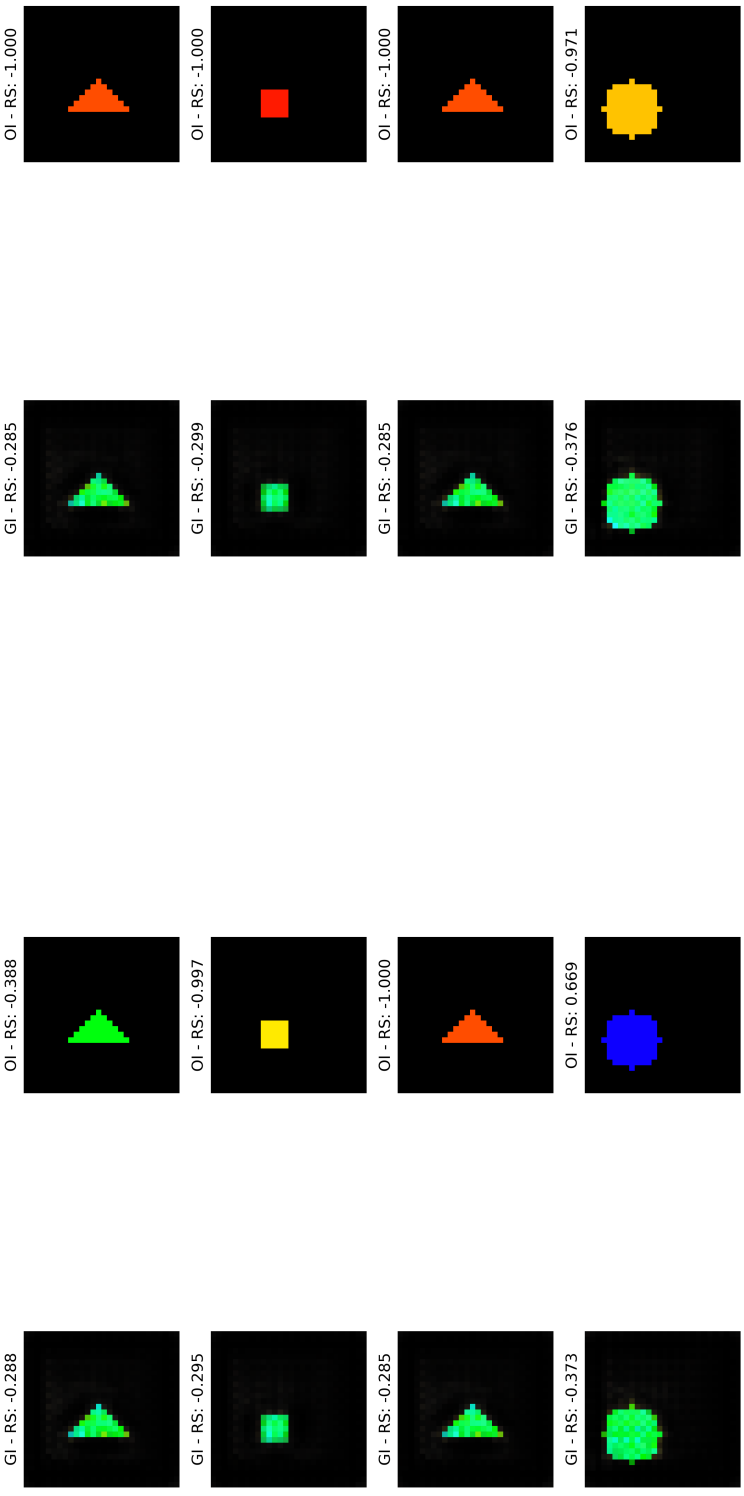
C Main results



(a) Simulation setting 1: Sample 1

(b) Simulation setting 1: Sample 2

Figure 10: Generated baseline images after training the generator for 6499 steps.



(a) Simulation setting 2: Sample 1

(b) Simulation setting 2: Sample 2

Figure 11: Generated baseline images after training the generator for 15499 steps.

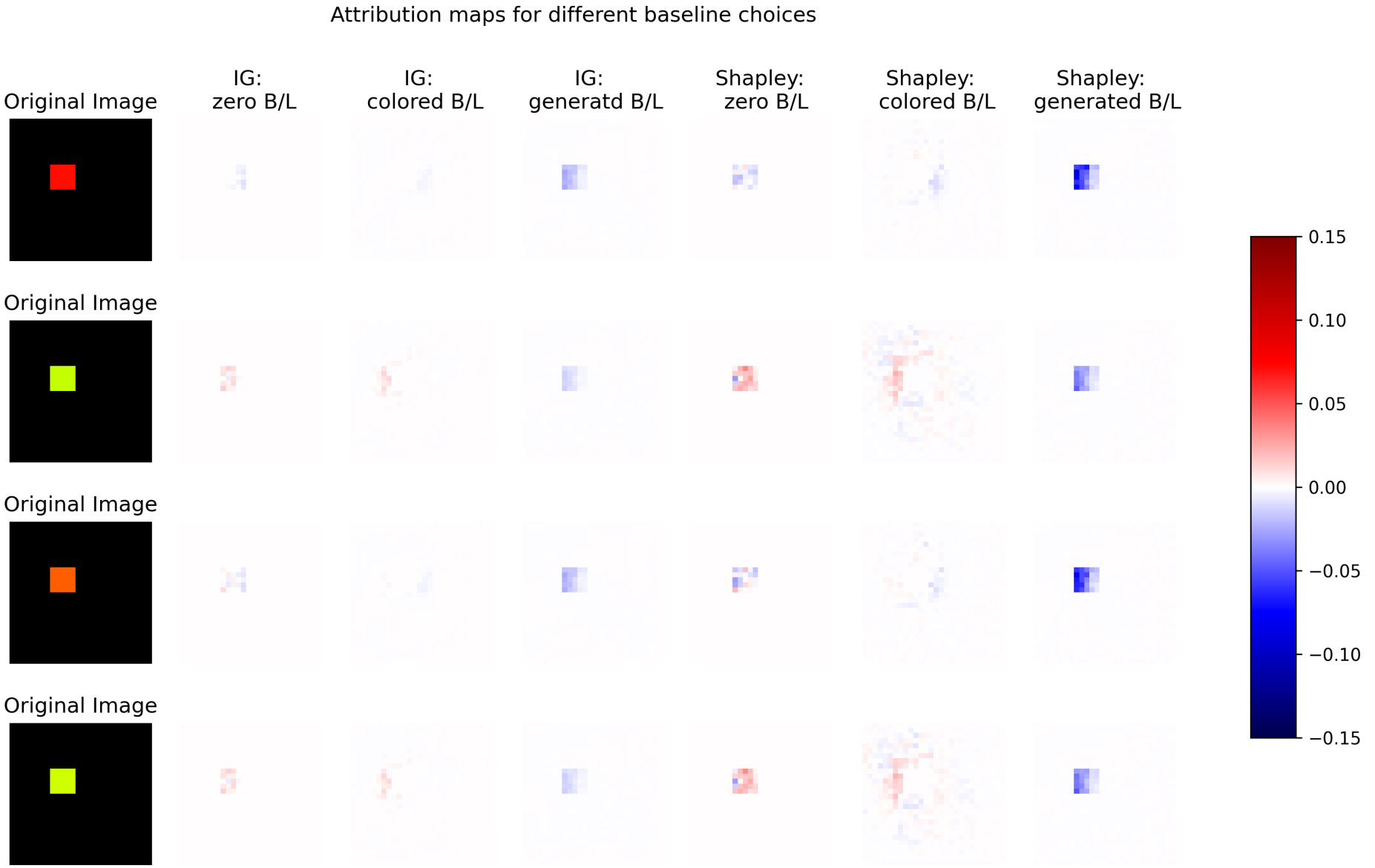


Figure 12: Simulation setting 1, sample 1: Attribution maps for different baseline choices. First three columns correspond to the attribution maps derived with Integrated Gradients. Last three columns corresponds to the attribution maps derived with sampled Shapley values. B/L: baseline. The colorbar depicts the scale of the assigned attribution scores.

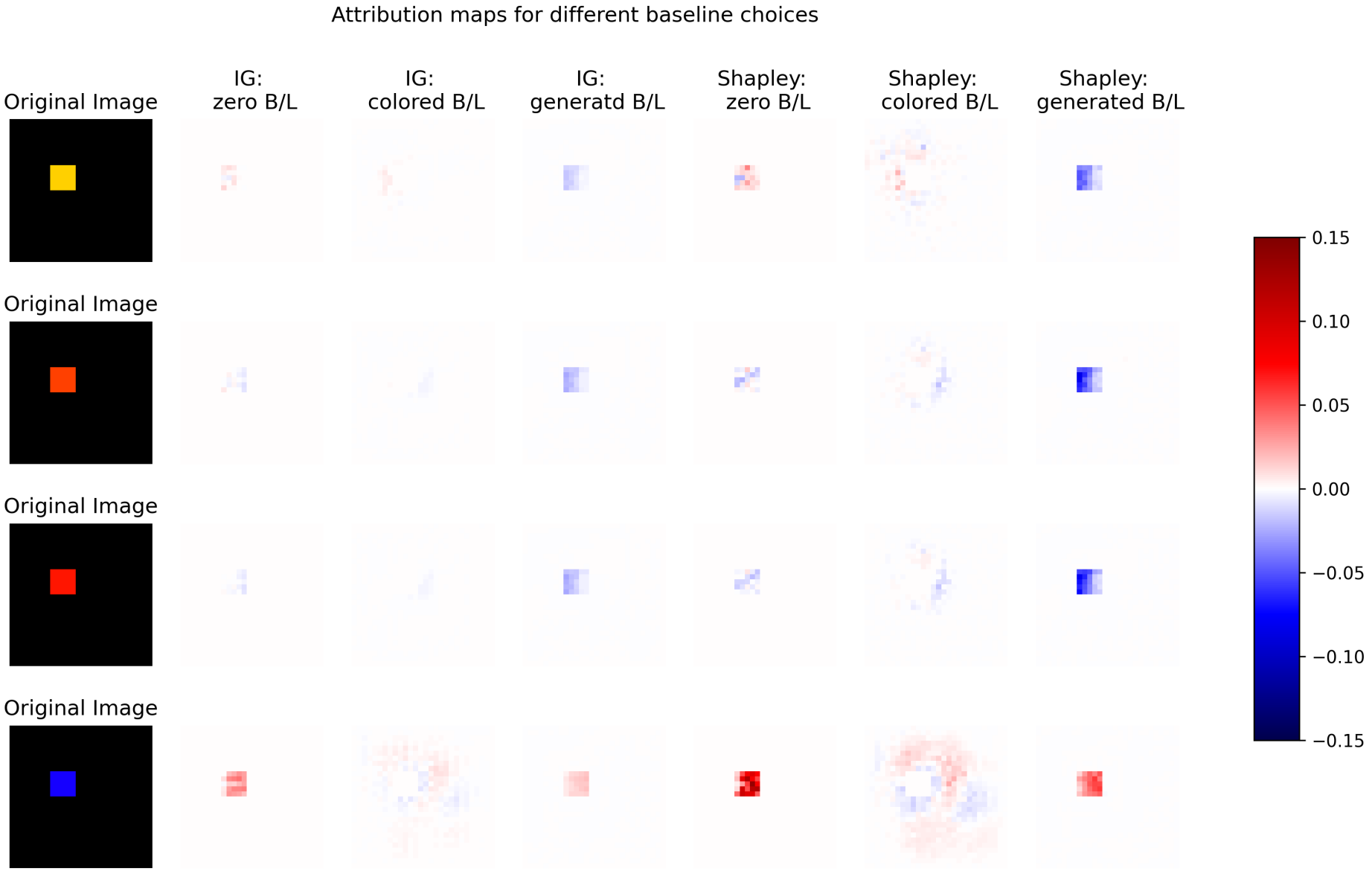


Figure 13: Simulation setting 1, sample 2: Attribution maps for different baseline choices. First three columns correspond to the attribution maps derived with Integrated Gradients. Last three columns corresponds to the attribution maps derived with sampled Shapley values. B/L: baseline. The colorbar depicts the scale of the assigned attribution scores.

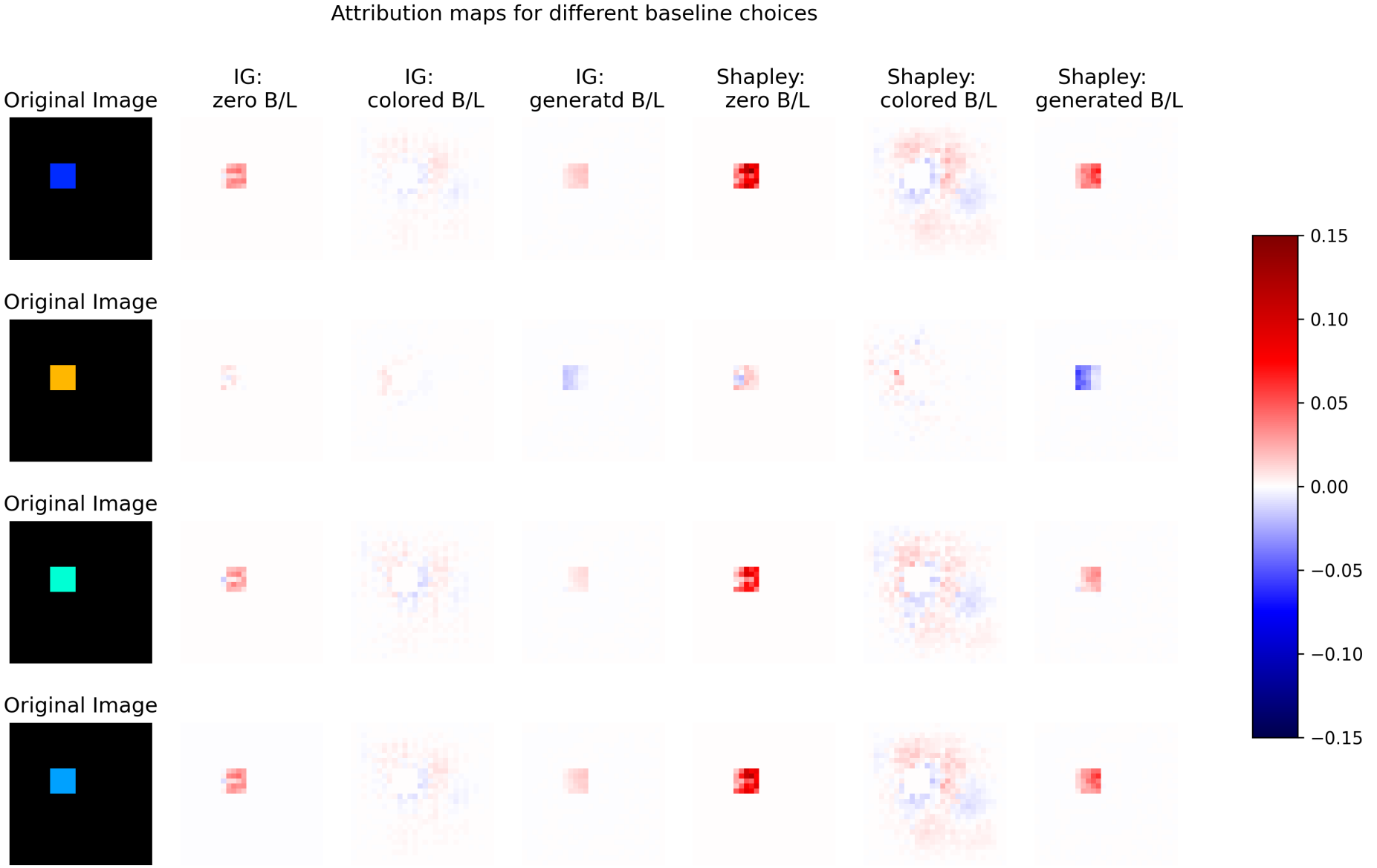


Figure 14: Simulation setting 1, sample 3: Attribution maps for different baseline choices. First three columns correspond to the attribution maps derived with Integrated Gradients. Last three columns corresponds to the attribution maps derived with sampled Shapley values. B/L: baseline. The colorbar depicts the scale of the assigned attribution scores.

Declaration of originality

I hereby declare that this thesis represents my original work and that I have used no other sources except as noted by citations. I have clearly referenced in accordance with departmental requirements. Additionally, I confirm that this work has not been previously or concurrently used for other courses. I confirm that I understand that my work may be electronically checked for plagiarism and appreciate that any false claim in respect of this work will result in disciplinary action.

Munich, 5th of May, 2021

Moritz Wagner