RESEARCH ARTICLE

Deep transformation models for functional outcome prediction after acute ischemic stroke

Lisa Herzog[1,2,3] | Lucas Kook[1,2] | Andrea Götschi[1] | Katrin Petermann[1] |
Martin Hänsel[3] | Janne Hamann[3] | Oliver Dürr[4] | Susanne Wegener[3] |
Beate Sick[1,2]

[1]Epidemiology, Biostatistics & Prevention Institute, University of Zürich, Zürich, Switzerland

[2]Institute for Data Analysis and Process Design, Zurich University of Applied Sciences, Winterthur, Switzerland

[3]Department of Neurology, University Hospital Zurich, Zürich, Switzerland

[4]Institute for Optical Systems, Konstanz University of Applied Sciences, Konstanz, Germany

Correspondence
Beate Sick, Epidemiology, Biostatistics, and Prevention Institute, University of Zürich, Hirschengraben 84, CH-8001 Zürich, Switzerland. Institute for Data Analysis and Process Design, Zurich University of Applied Sciences, Technikumstrasse 81, CH-8400 Winterthur, Switzerland.
Email: sick@zhaw.ch

Authors Lisa Herzog and Lucas Kook contributed equally.

Abstract

In many medical applications, interpretable models with high prediction performance are sought. Often, those models are required to handle semistructured data like tabular and image data. We show how to apply deep transformation models (DTMs) for distributional regression that fulfill these requirements. DTMs allow the data analyst to specify (deep) neural networks for different input modalities making them applicable to various research questions. Like statistical models, DTMs can provide interpretable effect estimates while achieving the state-of-the-art prediction performance of deep neural networks. In addition, the construction of ensembles of DTMs that retain model structure and interpretability allows quantifying epistemic and aleatoric uncertainty. In this study, we compare several DTMs, including baseline-adjusted models, trained on a semistructured data set of 407 stroke patients with the aim to predict ordinal functional outcome three months after stroke. We follow statistical principles of model-building to achieve an adequate trade-off between interpretability and flexibility while assessing the relative importance of the involved data modalities. We evaluate the models for an ordinal and dichotomized version of the outcome as used in clinical practice. We show that both tabular clinical and brain imaging data are useful for functional outcome prediction, whereas models based on tabular data only outperform those based on imaging data only. There is no substantial evidence for improved prediction when combining both data modalities. Overall, we highlight that DTMs provide a powerful, interpretable approach to analyzing semistructured data and that they have the potential to support clinical decision-making.

KEYWORDS
deep learning, distributional regression, ordinal regression, transformation models

1 | INTRODUCTION

Although prevention, diagnosis, and treatment of stroke have improved largely, it remains one of the leading causes of long-term disability and death worldwide (Benjamine et al., 2019). Each year, approximately 15 million people experience a stroke, 40% die, and 30% suffer lasting functional disability. To achieve the best possible outcome, patients have to be treated as fast as possible and decisions for or against different treatment options have to be made under immense time pressure. For clinical studies, the patient's functional outcome 3 months after hospital admission is primarily used to assess treatment success. Functional outcome is quantified on the modified Rankin Scale (mRS), an ordinal score comprising seven levels ranging between no symptoms at all (mRS of 0) and death (mRS of 6, Quinn et al., 2009). Often, neurologists are not directly interested in predicting the exact class of the mRS but rather in stratifying the chances of a patient having a favorable (mRS of 0–2) versus unfavorable (mRS of 3–6) functional outcome (Weisscher et al., 2008).

Semistructured data comprise the basis for various decisions in medicine (e.g., in stroke and cancer, Ebisu et al., 1997; Jafari et al., 2018). For instance, when predicting functional outcome in stroke patients, unstructured data such as brain images resulting from computed tomography (CT) or magnetic resonance imaging (MRI) are as important as structured data, such as tabular patient and clinical characteristics (Copen et al., 2011). Different brain imaging modalities provide insight into the extent of tissue injury, the exact location of the stroke lesion, as well as previous brain infarcts. While in clinical practice, information from brain imaging is frequently used for difficult clinical decisions, functional outcome prediction is limited with current image analysis strategies (see Section 1.1). It is currently an open question to what extent the imaging data and tabular data help in reliably predicting functional outcome. In a previous study, Hamann et al. (2021) found no additional benefit for stroke outcome prediction when adding expert-derived image features alongside clinical features. Trustworthy models for outcome prediction relying on data of both modalities are lacking but of high interest to the neurologist to assess the vast amount of complex medical data under immense time pressure.

Recently, machine learning (ML) and deep learning (DL) models, in particular, have proven outstanding prediction results on unstructured data like images. The models are fast, precise, and reproducible when it comes to analyze the large amount of data appearing in daily clinical practice (e.g., Campanella et al., 2019). Nonetheless, there is often distrust in ML-derived predictions, which is mainly due to their "black-box" character (Rudin, 2019). Questions like "How does the model come to its prediction?," "How certain is the model about the prediction?," or "What is the impact of different patient features on the prediction" have to be answered, in order for medical experts to trust the model. Therefore, ML models should not only focus on achieving the most accurate predictions but also on interpretability and uncertainty, that is, the models should be tailored to provide a distributional outcome prediction instead of a point prediction.

We present deep transformation models (DTMs) to analyze semistructured data. DTMs unite classical statistical models with (deep) neural networks, provide distributional outcome predictions, and achieve interpretable model parameters without sacrificing the high prediction performance of DL models. We demonstrate the use of DTMs on data of patients admitted to the hospital due to stroke symptoms. In particular, we present models for predicting a patient's functional outcome measured by the ordinal mRS 3 months after hospital admission that rely on tabular data, brain imaging, or a combination of both. We apply DTMs on a semistructured data set of 407 stroke patients to model the conditional distribution of a patient's functional outcome 3 months after hospital admission. We describe briefly how DTMs can be used to model continuous or censored outcomes, like time-to-event data, which makes them applicable to many different research questions. We discuss how DTMs yield interpretable effect estimates of the different input modalities and how the model arrives at its predictions. Moreover, we highlight baseline-adjusted DTMs conditioning on a patient's prestroke mRS, which is expected to be strongly predictive of outcome. Baseline-adjusted DTMs for un- and semistructured data are novel and of high interest to data analysts working in medical research, in which integrating baseline variables for outcome prediction is a common requirement.

1.1 | Related work

In the following, we describe work related to semistructured distributional regression approaches as DTMs.

1.1.1 | Classical regression models

Classical regression models like logistic regression or Cox proportional hazard models are the standard when analyzing structured data (e.g., tabular features) in medical applications. They are considered highly trustworthy because they are transparent, interpretable, and provide uncertainty measures (e.g., Steyerberg, 2019). However, unstructured data like

images or text cannot directly be analyzed with such models. First, tabular features have to be extracted from the unstructured data to be subsequently analyzed in a regression model—potentially together with other tabular data (e.g., Thiran & Macq, 1996). Yet, this feature engineering step is disconnected from optimizing the model parameters and necessarily discards information, which makes it difficult to know if the engineered features reflect relevant information in the original data well enough.

1.1.2 | Deep neural networks

(Deep) neural networks (DNNs), on the other hand, learn relevant features for a task at hand as a part of the model-fitting process and therefore omit the feature engineering step, whereas they can be trained on structured data, unstructured data, or a combination of both (Goodfellow et al., 2016). For instance, previous work has focused on analyzing combinations of image and tabular data to predict stroke patient outcomes with DNNs. Pinto et al. (2018) used a model consisting of a convolutional neural network (CNN) for the image data where they attach tabular data to the feature vector in the dense part of the CNN. This enables interactions between image and tabular data. Another pilot study for stroke outcome prediction used a combination of a CNN and a dense NN for integrating image and tabular data into one model (Bacchi et al., 2020). However, like the majority of the DNNs, the existing approaches are black box models that do not quantify uncertainty. They lack interpretable model parameters and estimate point predictions like the conditional mean rather than a conditional outcome distribution.

1.1.3 | Distributional regression

Distributional regression focuses on estimating an entire conditional distribution rather than the first conditional moment(s) (Kneib et al., 2021). Therefore, when fitted by empirically optimizing a proper score like the negative log likelihood, a distributional regression model directly quantifies aleatoric uncertainty inherent in the data. To achieve a well fitting distributional regression model, a complex conditional outcome distribution might be required. Generalized linear models (GLMs) are based on members of the exponential family, defined by the first two moments, for modeling the conditional outcome distribution, whereas they provide interpretable model parameters. Generalized additive models for location, scale, and shape (GAMLSS) extend GLMs by allowing to specify all parameters of the assumed outcome (Stasinopoulos & Rigby, 2007). A GAMLSS implementation with flexible specification of the conditional moments of $\mathbb{P}_{Y|X=x}$ using DNNs is, for example, presented in Rügamer et al. (2020). However, these models still require the choice of a parametric family of conditional outcome distributions.

1.1.4 | Transformation models for distributional regression

Transformation models (TMs) are a more recent method for distributional regression, which do not require to prespecify the family of the outcome distribution (Hothorn et al., 2014, 2018). In TMs, the conditional outcome distribution is decomposed into a simple, parameter free, target distribution F_Z (e.g., normal or logistic), and a conditional transformation function $h(y|x)$, such that $F_{Y|X=x}(y) = F_Z(h(y|x))$. More details are given in Section 2. Independent of TMs, normalizing flows were developed in the DL community (Rezende & Mohamed, 2015), which are based on the same idea as TMs. But while normalizing flows solely aim at predicting a flexible (conditional) distribution and constructing the transformation function as a chain of simple transformations, TMs are tailored for interpretable distributional regression models. The construction of the transformation function and the choice of the simple distribution F_Z give rise to extremely flexible TMs for conditional distributions. For instance, Sick et al. (2021) and Baumann et al. (2021) use $F_Z = \Phi$ and predict different outcome distributions with variously flexible transformation functions on commonly used benchmark data sets in DL and demonstrate state-of-the-art prediction performances. Rügamer et al. (2021) use DTMs for time series data by including autoregressive components in the transformation function.

1.1.5 | (Deep) transformation models for ordinal outcomes

The main application of this article features an ordinal outcome (mRS). Models for the conditional distribution of an ordinal outcome given covariates like the proportional odds logistic regression model have been studied in statistics for several
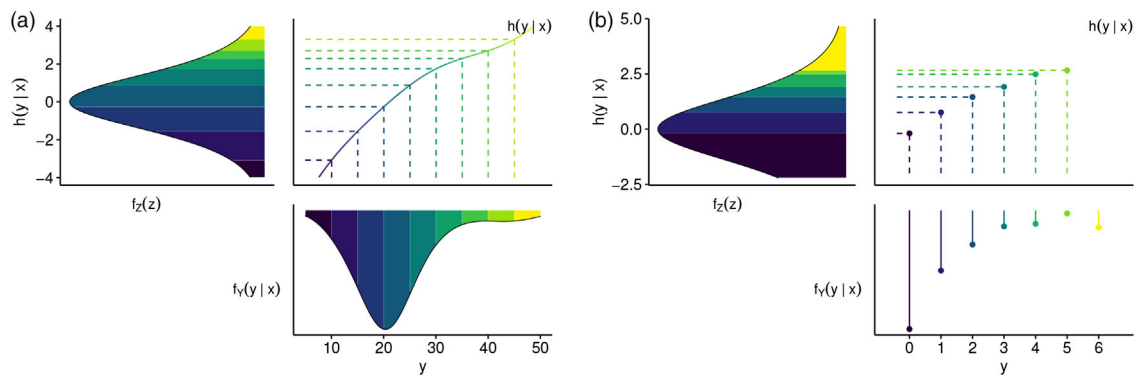
FIGURE 1 TMs for continuous (a) or ordinal categorical outcome (b). The lower right part of each panel shows the conditional density of Y given $\boldsymbol{x}$, which is mapped onto the density of the latent variable Z (see upper left part in each panel). The transformation is done via a monotone transformation function h (upper right part). This transformation function can be continuous (a) or discrete (b).

decades (McCullagh, 1980). Baseline-adjusted proportional odds models have been described from a TM perspective in Buri et al. (2020). However, only recently a special DTM, focusing on ordinal neural network TMs (ONTRAMS), has been developed in DL and applied to several publicly available (nonmedical) data sets (Kook & Herzog et al., 2022b). However, DTMs were not yet applied in the context of stroke.

1.1.6 | Transformation ensembles

Ensembling in terms of aggregating the predictions of multiple models to improve prediction performance is commonly seen in practical applications. In the field of DL, ensembling often means aggregating the predicted probabilities of a few DNNs that possess the same architecture and are trained on the same data after random initialization (Lakshminarayanan et al., 2017). These deep ensembles are not only used to achieve more accurate predictions but also to quantify epistemic uncertainty by means of the variation of the different predictions. However, the special structure and the interpretability of deep TMs are in general lost after aggregating them via deep ensembling. Kook et al. (2022a) recently developed transformation ensembles that aggregate DTMs on the scale of the transformation function preserving structure and interpretability (see Section 2).

  This article is organized as follows. Section 2 presents detailed background on distributional regression models with semistructured data and the experimental setup including model evaluation. Results are presented in Section 3. We end with a discussion of the various types of questions that may be answered by deep distributional regression models like DTMs in Section 4.

2 | METHODS

In the following, we briefly introduce TMs that are used to integrate semistructured data, model highly flexible conditional outcome distributions, and provide interpretable model parameters. Since our application features an ordinal outcome, we will pay special attention to this case.

2.1 | Distributional regression with transformation models

In TMs, the problem of estimating the potentially complex conditional outcome distribution of $(Y|X = x)$ is approached by learning a parameterized monotone transformation $h(y|\boldsymbol{x}; \boldsymbol{\theta})$ that maps between the distributions of $(Y|X = x)$ and the latent variable Z. The distribution of the latent variable Z (with log-concave density) has to be defined a priori. Usually, a parameter-free distribution, such as the standard Gaussian or logistic distribution, is chosen (see Figure 1).

  The parameters $\boldsymbol{\theta}$ in $h(y|\boldsymbol{x}; \boldsymbol{\theta})$ determine the functional form of the transformation function and thus the corresponding conditional outcome distribution (we drop $\boldsymbol{\theta}$ in the following to simplify notation). The parameters are fitted via maximum

likelihood, that is, by minimizing the negative log-likelihood (NLL),

$$\text{NLL} = -\frac{1}{n} \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}), \tag{1}$$

where ℓ_i is the log-likelihood contribution of the ith training observation.

In case of a continuous outcome, the likelihood contribution of an exact observation $(y, \boldsymbol{x})$ is given by the value of the conditional density at the observed outcome $f_{Y|X=x}(y)$ that can be determined via f_Z and h by using the change of variables formula $f_{Y|X=x}(y) = f_Z(h(y|\boldsymbol{x})) \cdot h'(y|\boldsymbol{x})$. The transformation function h is a smooth function (see Figure 1a) that can be modeled via a basis expansion with basis functions $\boldsymbol{a}(\cdot)$, yielding $\boldsymbol{a}(y)^\top \boldsymbol{\theta}$. A common choice for $\boldsymbol{a}(\cdot)$ are polynomials in Bernstein form $\mathbf{a}_{\text{Bs},P}(\cdot)$ of order P. Here, the required monotonicity of h can be easily guaranteed via linear constraints on the parameters $\boldsymbol{\theta}$ (Hothorn et al., 2018). Complex dependence on the input $\boldsymbol{x}$ can be achieved by controlling $\boldsymbol{\theta}(\boldsymbol{x})$ via a deep NN.

If a continuous observation is censored, which often occurs in (but is not limited to) survival data, the outcome is measured as an interval $y \in (\underline{y}, \bar{y}]$ and the likelihood contribution can be derived from the cumulative distribution function, as $F_Z(h(\bar{y}|\boldsymbol{x})) - F_Z(h(\underline{y}|\boldsymbol{x}))$.

For an ordered categorical outcome, the discrete monotone increasing transformation function h maps the observed outcome classes y_k to the conditional cut points $h(y_k|\boldsymbol{x})$, $k = 1, \dots, K-1$ of the latent variable Z, as illustrated in Figure 1(b). This allows to view the ordinal outcome as a result of an underlying continuous latent variable Z with interval-censored observations. The likelihood contribution of an observation $(y_k, \boldsymbol{x})$, given by the probability p_k for the observed class y_k, can correspondingly be determined by the area under f_Z between the cut points $h(y_k|\boldsymbol{x})$ and $h(y_{k-1}|\boldsymbol{x})$ and is computed as $p_k = F_Z(h(y_k|\boldsymbol{x})) - F_Z(h(y_{k-1}|\boldsymbol{x}))$. If dummy-encoding is used for y_k, that is, the class k is encoded by a vector $\boldsymbol{a}(y)$ of length K that holds a *one* at position k and *zeros* elsewhere, then h is given by $h(y_k|\boldsymbol{x}) = \boldsymbol{a}(y)^\top \boldsymbol{\theta}(\boldsymbol{x})$ with $\boldsymbol{\theta}(\boldsymbol{x})$ being constrained to $\theta_1(\boldsymbol{x}) \leq \theta_2(\boldsymbol{x}) \leq \cdots \leq \theta_K(\boldsymbol{x}) = +\infty$.

2.1.1 | Interpretability in transformation models

To achieve the same interpretability as in commonly used regression models, such as proportional hazard or proportional odds models, the flexibility of h needs to be restricted. This can be done by decomposing h in a baseline transformation (intercept function) h_0 that does not depend on the input data and one or several shift terms $h(y|\boldsymbol{x}) = h_0(y) - \text{shift}(\boldsymbol{x})$. In such a shift model, h_0 determines the shape of the transformation function h and only the shift terms depend on $\boldsymbol{x}$, moving h up and down (see Figure 1). A particularly simple example is a linear shift model of some tabular input data x_j, $j \in 1, \dots, J$, which looks as follows for a continuous outcome $h(y|\boldsymbol{x}) = h_0(y) - \boldsymbol{x}^\top \boldsymbol{\beta}$. Depending on the chosen distribution for Z, the parameters $\boldsymbol{\beta}$ have a straightforward interpretation. A summary of commonly used distributions for F_Z and the corresponding interpretational scales is given in Siegfried and Hothorn (2020).

When choosing, for example, the minimum extreme value distribution for Z, that is, $F_Z(z) = 1 - \exp(-\exp(z))$, the parameters β_j, $j = 1, \dots, J$ can be interpreted as log hazard ratios. A well-known example is the proportional hazard model that is often used for survival analysis, where the bounded continuous outcome is a survival time. Survival analysis poses additional challenges. For instance, usually, not all patients experience the event of interest during follow-up, leading to (right-) censoring with $y \in (\underline{y}, +\infty)$, which can be easily handled in TMs, as described above.
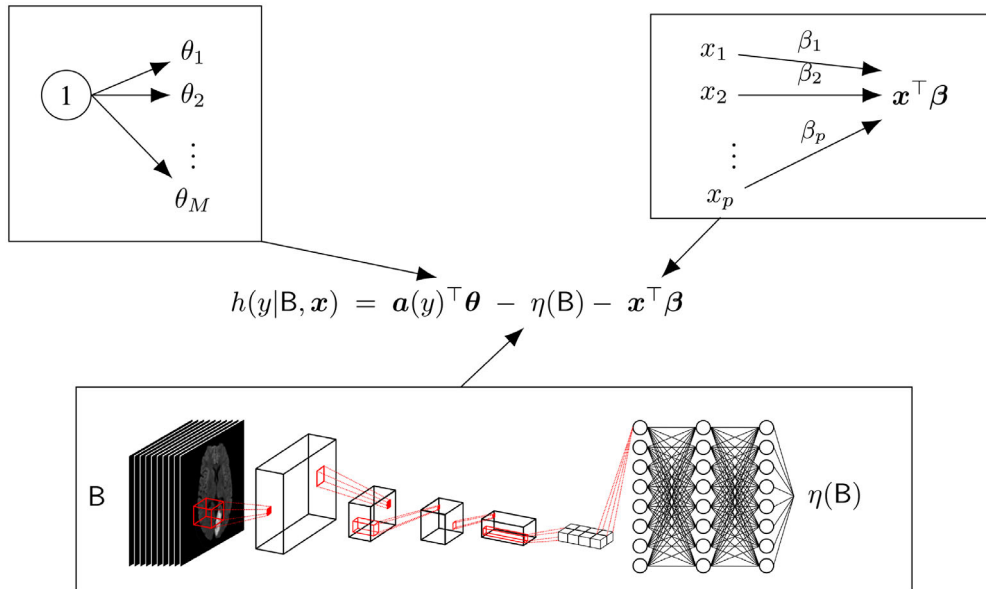
2.1.2 | Semistructured regression

In semistructured regression, the problem is to combine both structured data, for example, tabular features $\boldsymbol{x}$, and unstructured data, for example, images B, in one single model. This can be realized with NNs, which take both structured and unstructured data as input and control the parameters of h (see Figure 2). Depending on the architecture of the NNs, more or less flexible models can be described.

The most flexible model is achieved, if h depends in complex manner on all inputs corresponding to a complex intercept model with

$$h(y|\text{B}, \boldsymbol{x}) = \boldsymbol{a}(y)^\top \boldsymbol{\theta}(\text{B}, \boldsymbol{x}), \tag{2}$$

$$h(y|\mathrm{B},\boldsymbol{x}) = \boldsymbol{a}(y)^\top\boldsymbol{\theta} - \eta(\mathrm{B}) - \boldsymbol{x}^\top\boldsymbol{\beta}$$

FIGURE 2 An SI-CS$_\mathrm{B}$-LS$_{\boldsymbol{x}}$ DTM with simple intercepts, not depending on the input, a linear shift term in the tabular input data $\boldsymbol{x}$ and a complex shift term in the images B. All additive components of h are controlled by NNs, that is, shallow dense NNs without hidden layer for SI and LS$_{\boldsymbol{x}}$ and a three-dimensional CNN for CS$_\mathrm{B}$, which are jointly fitted by minimizing the NLL via stochastic gradient decent.

where an NN controls $\theta(\cdot)$ depending on imaging data (B) and tabular data ($\boldsymbol{x}$) and thus potentially allowing for interactions between B and $\boldsymbol{x}$. Without restricting $\theta(\cdot)$ in any way besides being monotone increasing, maximal flexibility is achieved. Most often in biostatistics, a shift model is assumed for h (i.e., a proportionality assumption is made) and no interactions between the input data are allowed. In this scenario, the model simplifies to

$$h(y|\mathrm{B},\boldsymbol{x}) = \boldsymbol{a}(y)^\top\boldsymbol{\theta} - \eta(\mathrm{B}) - \beta(\boldsymbol{x}), \tag{3}$$

where β and η are controlled by two separate NNs and are interpretable, for example, as log odds ratios if the logistic distribution $F_Z(z) = \mathrm{expit}(z) = \frac{1}{1+\exp(-z)}$ is chosen. If a linear effect is assumed for each tabular feature, and the effect of each feature should be interpretable as log odds ratio, then further simplifications have to be made by using a linear shift term for the tabular data

$$h(y|\mathrm{B},\boldsymbol{x}) = \boldsymbol{a}(y)^\top\boldsymbol{\theta} - \eta(\mathrm{B}) - \boldsymbol{x}^\top\boldsymbol{\beta}. \tag{4}$$

Such a model with simple intercept (SI) $\boldsymbol{a}(y)^\top\boldsymbol{\theta}$, linear shift $\boldsymbol{x}^\top\boldsymbol{\beta}$, and complex shift $\eta(\mathrm{B})$ term is depicted in Figure 2 and referred to as SI-CS$_\mathrm{B}$-LS$_{\boldsymbol{x}}$ in this work.

In general, the primary goal is to develop a model with adequate prediction performance. Usually, simpler (i.e., fewer parameter) and more interpretable models are preferred over black boxes. Only if the more complex model yields a substantial improvement in terms of prediction performance, the more complex model should be preferred. We investigate the ramifications of model selection in Section 3.

2.1.3 | Transformation ensembles

We construct transformation ensembles of DTMs that are fitted on the same data but with different random initialization. Transformation ensembles average the predicted transformation functions of the DTMs, which preserves the model structure and interpretability, improves prediction performance, and allows to quantify epistemic uncertainty (Kook et al., 2022a).
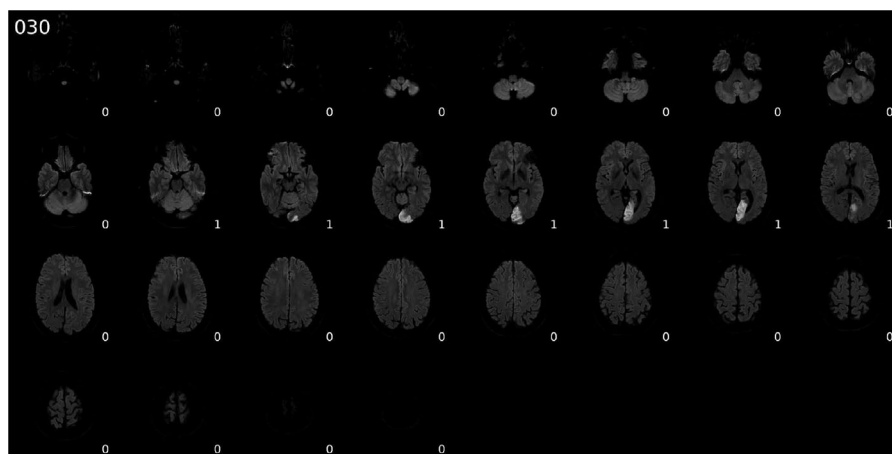
FIGURE 3 A pseudo-3D diffusion-weighted image of an example stroke patient. 2D slices where a stroke lesion is visible are labeled with a 1 and 0 otherwise. Each patient is represented by 28 diffusion-weighted images (DWIs) after preprocessing. Ischemic stroke lesions appear as hyperintense signal on one or multiple images of a sequence.

2.2 | Data

Our cohort consists of 407 patients who are either diagnosed with ischemic stroke (295 patients) or transient ischemic attack (TIA, 112 patients). As opposed to stroke, TIA causes only temporary stroke symptoms and no permanent brain damage. The cohort was collected retrospectively. All patients were admitted to the University Hospital of Zurich between 2014 and 2018 and had MRI records in the acute phase. Ethical approval for the study was obtained from the Cantonal Ethics Committee Zurich (KEK-ZH-No. 2014-0304).

In this study, we use the stroke patient's brain imaging and tabular baseline data for functional outcome prediction. Diffusion-weighted images (DWIs) represent brain pathology in a 3D manner as ordered sequences of multiple 2D images per patient. On DWIs, stroke lesions appear as hyperintense signals, typically on multiple, subsequent images in the sequence (see Figure 3). They give valuable insight into stroke location and severity. TIA patients show no visible lesion on DWIs. All collected DWIs were recorded within 3 days after hospital admission. After preprocessing, each 3D image is of dimension $128 \times 128 \times 28$ with zero mean and unit variance (see Figure 3). We consider baseline covariates, that is, patient characteristics including age and sex, risk factors including hypertension, prior stroke, smoking, atrial fibrillation, coronary heart disease (CHD), prior TIA, diabetes, and hypercholesterolemia, the National Institutes of Health Stroke Scale at baseline (NIHSS at BL) highlighting stroke symptom severity as an ordinal sum score with 42 levels, and the pre-stroke mRS (mRS at BL) informing about the patient's functional disability before stroke. All factor variables are dummy encoded and all other tabular features are standardized to make the magnitude of estimated parameters comparable.

The outcome of interest is the ordinal mRS, which consists of seven levels: 0 = no symptoms at all, 1 = no significant disability despite symptoms, 2 = slight disability, 3 = moderate disability, 4 = moderately severe disability, 5 = severe disability, 6 = death (Grotta et al., 2016). In our cohort of 407 patients, we observed the following $n_k, k = 1, \ldots, K$ for the $K = 7$ outcome classes: $n_0 = 184$ (45.2%), $n_1 = 88$ (21.6%), $n_2 = 60$ (14.7%), $n_3 = 25$ (6.1%), $n_4 = 20$ (4.9%), $n_5 = 5$ (1.2%), $n_6 = 25$ (6.1%). Figure B.1 in the Appendix shows the distribution of predictors stratified by the outcome among all 407 patients. Since in clinical practice, the neurologists are often primarily interested in the patient's chance for a favorable (mRS ≤ 2, $n_f = 332$, 81.6%) versus unfavorable (mRS > 2, $n_u = 75$, 18.4%) outcome (Weisscher et al., 2008), we additionally considered the binary mRS.

2.3 | Experimental setup

2.3.1 | Models

We compare models with varying degrees of interpretability and flexibility for ordinal mRS prediction (see Table 1). The goal is to obtain a model that achieves the highest possible prediction power while being adequately interpretable. In all

TABLE 1 Summary of all models used for binary and ordinal functional outcome prediction in the stroke data. If applicable, the transformation function is given. The model names are combinations of the components (simple/complex intercept/shift), the subscript indicates which modality enters which component, where, for example, LS_x indicates that the tabular data are the input. SI, simple intercept; CI, complex intercept; LS, linear shift; CS, complex shift. Note that LS_x includes all predictors, including prestroke mRS, whereas LS_{mRS} contains prestroke mRS only.

Outcome	Input data	Model name	Transformation function
Binary mRS	Images only	CI_B-Binary	$\theta(B)$
Ordinal mRS	None	SI	θ_k
	Tabular only	SI-LS_x	$\theta_k - \boldsymbol{x}^\top\boldsymbol{\beta}$
	Tabular only	SI-CS_{age}-$LS_{\bar{x}}$	$\theta_k - \gamma(\boldsymbol{x}_{age}) - \boldsymbol{x}_{-age}^\top\boldsymbol{\beta}$
	Images only	SI-CS_B	$\theta_k - \eta(B)$
	Images + tabular	SI-CS_B-LS_x	$\theta_k - \eta(B) - \boldsymbol{x}^\top\boldsymbol{\beta}$
	Images + prestroke mRS	CI_B-LS_{mRS}	$\theta_k(B) - \boldsymbol{x}_{mRS}^\top\boldsymbol{\beta}$
	Images + tabular	CI_B-LS_x	$\theta_k(B) - \boldsymbol{x}^\top\boldsymbol{\beta}$
	Tabular only	GAM	$\theta_k - \gamma(age) - \boldsymbol{x}_{-age}^\top\boldsymbol{\beta}$
	Images only	CI_B	

models, we choose $F_Z(z) = \text{expit}(z)$, such that shift parameters in h can be interpreted as log odds ratios. By comparing models based on tabular data, image data, and a combination of both, we assess if tabular and image data carry complementary information and which of the two contains more information for outcome prediction. As a baseline, we consider performance metrics of an unconditional model, which takes no input data and hence consists of an SI only. This model predicts the prevalence of each outcome class. To assess binary mRS prediction, we consider the outcome as censored and sum up the predicted probabilities of the respective ordinal model. The probability for a favorable outcome is the sum across the probabilities for classes 0–2, the probability for unfavorable outcome is the sum across the probabilities for classes 3–6.

We define an image-only model that is fitted using the binary mRS (CI_B-Binary in Table 1). This dichotomized version of the mRS can be viewed as a censored version of the ordinal mRS and can therefore be directly compared to all models fitted on the ordinal scale (see Table 1). We fit the CI_B-Binary model primarily as a benchmark for the performance of the models that are trained for the ordinal but evaluated for the binary mRS.

The most interpretable model for the ordinal mRS is a linear proportional odds model based on all tabular features. It consists of an SI and a linear shift in $\boldsymbol{x}$ (SI-LS_x). The SI-CS_{age}-$LS_{\bar{x}}$ model allows the outcome to depend on age in a nonlinear way by estimating a potentially complex and continuous log odds-ratio function x_{age}. We additionally fit models depending on image data only (SI-CS_B, CI_B) and on a combination of image and tabular data (SI-CS_B-LS_x and CI_B-LS_x models). Integrating the images as complex intercept (CI_B) rather than as complex shift term (CS_B) allows to increase model complexity further. In the image model CI_B-LS_{mRS}, we additionally adjust for the prestroke mRS to achieve a fairer comparison between image-data-only and tabular-data-only.

2.3.2 | Implementation

SI and linear shift terms for tabular features are modeled with fully connected NNs without hidden layers. A fully connected NN with multiple hidden layers is used to integrate age as complex shift term. The complex intercept and complex shift terms for the images are modeled with a 3D CNN. In all models, the number of output nodes is equal to six (since the mRS has seven levels) in NNs for intercept terms and equal to one in NNs for shift terms. The last layer activation is always linear and no bias terms are used.

All models are trained by minimizing the negative log-likelihood (see Equation (1)) using the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 5×10^{-5} and a batch size of six. Augmentation of the image data is used to prevent overfitting. In addition, we use early stopping, that is, we select the model weights from the epoch that shows the smallest NLL on the validation data. More details on NN architectures, hyperparameters, augmentation procedure, and software are given in Appendix A.

2.3.3 │ Training and evaluation

We randomly split the data six times into a train (80%), validation (10%), and test set (10%). This results in six fits for each model type (see Table 1) that allows us to assess the variation of the achieved test performance. For all models, which include the image data as input, we perform transformation ensembling (Kook et al., 2022a). For that, we train five models on the same data in each split. CNNs controlling the image term in the model are initialized randomly. Additional SI and LS_x terms are initialized with the corresponding parameters of the $SI\text{-}LS_x$ model fitted on the same split. This results in an ensemble model (constructed from five members) in each of the six splits for each model type.

2.3.4 │ Performance measures

All models are mainly evaluated with proper scoring rules (Gneiting & Raftery, 2007). The score we consider primarily for model comparison is the test NLL (Good, 1952, a.k.a. log-score). We further assess the Brier score for the binary outcome. For the ordinal functional outcome, we calculate the ranked probability score (RPS) as an additional proper score (Bröcker & Smith, 2007). As measures of discriminatory ability, we compute area under the receiver operating characteristic (ROC) curve (AUC) and accuracy for binary outcomes and quadratic weighted Cohen's κ (QWK) for the ordinal outcome (Steyerberg, 2019). We construct 95% bootstrap confidence intervals by taking $B = 1000$ bootstrap samples of size n_{test} of test predictions (e.g., NLL contributions) for each of the $S = 6$ random splits of the data, by computing the 2.5th, 50th, and 97.5th percentile of the B bootstrap metrics averaged over the S splits.

3 │ RESULTS AND DISCUSSION

We first present results for predicting and discriminating binary and ordinal mRS. Then, we discuss how to interpret linear and nonlinear model components.

3.1 │ Binary mRS prediction

The test performance and calibration plots of all models from Table 1 evaluated for the binary mRS are summarized in Figure 4. We first compare models that only include the image modality and only differ in the number of classes (CI_B-Binary trained with two vs. CI_B trained with seven classes). The CI_B-Binary model shows a worse average performance and a higher variability in predictions across the six random splits compared to the CI_B. This highlights the importance for training with all available class levels rather than with a dichotomized version of the outcome—whenever possible. The average performance of the CI_B-Binary is similar to that of the unconditional model (SI), indicating that the model has primarily learned the class frequencies. Decreasing model complexity by modeling the image data with a complex shift ($SI\text{-}CS_B$) rather than with a complex intercept (CI_B) leads to a comparable performance. Both models, $SI\text{-}CS_B$ and CI_B, achieve a better average prediction performance than the unconditional model (SI), indicating that the image data contain information for mRS binary prediction.

The most interpretable model based on tabular features only ($SI\text{-}LS_x$) shows a better prediction performance than all models based on image data only (CI_B, CI_B-Binary, $SI\text{-}CS_B$) in terms of NLL and Brier Score. Like the models based on image data only, the $SI\text{-}LS_x$ outperforms the unconditional model (SI, Figure 4). This indicates that not only image but also tabular data are useful for binary mRS prediction. For a fairer comparison of image-data-only versus tabular-data-only, we adjust for prestroke mRS in the most flexible image model. In this comparison, the baseline-adjusted model ($CI_B\text{-}LS_{mRS}$) shows a performance similar to the unadjusted model (CI_B).

The semistructured models incorporating both image and tabular data ($CI_B\text{-}LS_x$ and $SI\text{-}CS_B\text{-}LS_x$) achieve a similar or slightly better average performance than the model including tabular data only ($SI\text{-}LS_x$, see Figure 4). $CI_B\text{-}LS_x$ does not assume proportional odds for the image modality and outperforms $SI\text{-}CS_B\text{-}LS_x$ on some splits. The latter assumes proportional odds for both tabular and image data. Overall, there is no convincing evidence that combining tabular and imaging data in a $CI_B\text{-}LS_x$ model improves binary mRS prediction. The added image information increases variability in prediction performance.
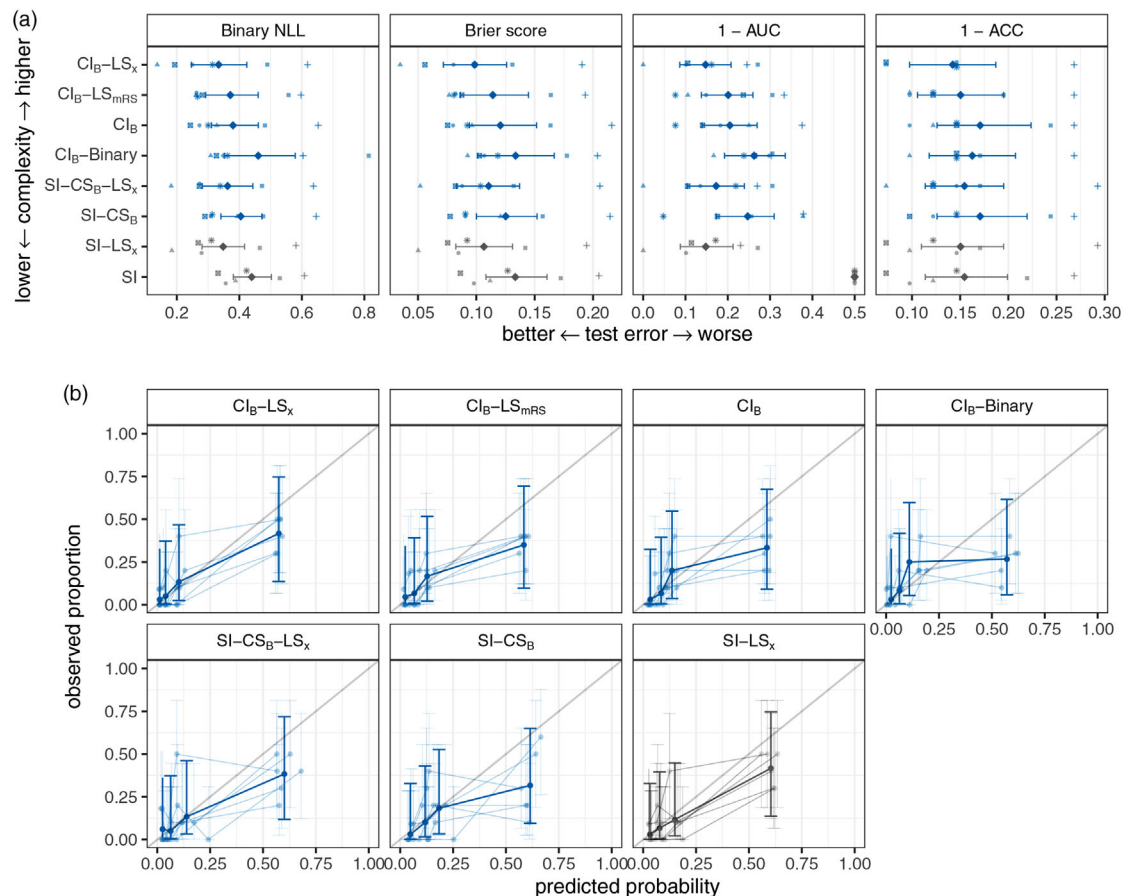
FIGURE 4 Test error (a) and calibration plots (b) of transformation ensembles (blue) and reference models (gray) evaluated for the binary mRS outcome (mRS 0–2 vs. mRS 3–6). In (a), the test performance is quantified in terms of negative log-likelihood (NLL), Brier score, discrimination error (1 − AUC), and classification error (1 − ACC). From the test performance in the six random splits (indicated by different symbols), we consider the average test error and 95% bootstrap (B = 1000) confidence intervals. For the calibration plots in (b), the predicted probabilities are split at the 0.25, 0.5, and 0.75 empirical quantiles to produce the four bins for which the average predicted probabilities and the observed proportion of an unfavorable outcome are computed. The confidence interval is plotted at the midpoint of the respective bin. Average calibration across all six random splits is shown as thick line, whereas the calibration of the single splits is shown as thin lines.

Scores highlighting discriminatory ability of the models (AUC and accuracy) show similar results. Slight differences in the ranking of models are possible because these measures are improper scoring rules (Gneiting & Raftery, 2007). Note that SI has no discriminatory ability (AUC = 0.5) because it always predicts the most frequent class (mRS 0). The relative test performance to the benchmark SI-LS$_x$ model (i.e., the differences in performance within splits) can be found in Appendix B.2.

Well-calibrated predictions are hard to achieve for highly imbalanced outcomes. The calibration plots in Figure 4 show no substantial evidence for miscalibration. However, all models seem to slightly overpredict the probability for an unfavorable outcome. This effect is most pronounced in the models based on image data only (CI$_B$, CI$_B$-Binary). The semistructured and tabular data-only models show a slightly better calibration.

3.2 | Ordinal mRS prediction

Figure 5 summarizes the test performance and calibration plots for all models in Table 1 trained and evaluated for the ordinal mRS.

As in the binary case, the models based on image (CI$_B$, SI-CS$_B$) and tabular data only (SI-LS$_x$) show better average prediction performances in terms of NLL, RPS, and QWK than the unconditional model (SI). And again, the most inter-
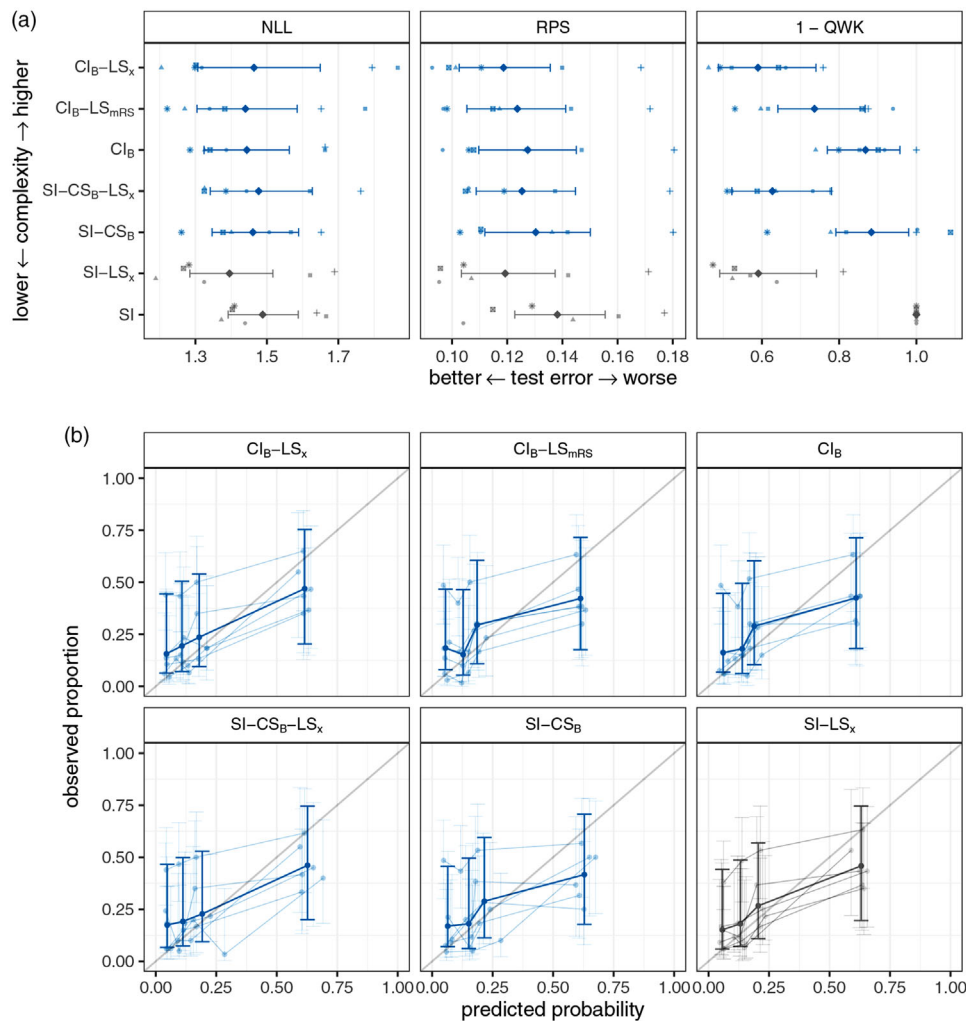
FIGURE 5 Test error (a) and calibration plots (b) of transformation ensembles (blue) and reference models (gray) evaluated for the ordinal mRS outcome. In (a), test error is quantified in terms of negative log-likelihood (NLL), ranked probability score (RPS), and discrimination error $(1 - \text{QWK})$. The average test error and 95% bootstrap $(B = 1000)$ confidence intervals are depicted for six random splits of the data (indicated by the different symbols). For the calibration plots in (b), the predicted probabilities are split at the 0.25, 0.5, and 0.75 empirical quantiles to produce the four bins for which the average predicted probabilities and the observed proportion of an unfavorable outcome are computed. The confidence interval is plotted at the midpoint of the respective bin. Average calibration across all six random splits is shown as thick line, whereas the calibration of the single splits is shown as thin lines. The 95% confidence intervals are averaged across classes and splits.

pretable model based on tabular data only (SI-LS_x) outperforms the more flexible black box image-only models, indicating that tabular features contain more information for ordinal mRS prediction than the images (at the available sample with only 407 patients). As in the binary case, we find no substantial evidence that using tabular and image data together in a semistructured model $(\text{CI}_B\text{-LS}_x, \text{CI}_B\text{-LS}_{mRS}, \text{ or SI-CS}_B\text{-LS}_x)$ improves average test performance compared to SI-LS_x (see Figure B.3).

In terms of calibration (Figure 5b), we again observe that all models overpredicted the probability for an unfavorable outcome.

Overall, we cannot draw a definitive conclusion about which data modality (tabular or image data) is more useful for functional outcome prediction and if adding image to tabular data aids mRS prediction. The confidence intervals overlap largely and average test performance is similar. In particular, this can be attributed to the small sample size. In Appendix B.3, we conclude that collecting more data could further enhance performance. When we artificially reduce sample size via subsampling and refit all models, we find no evidence of plateauing prediction performance. However,
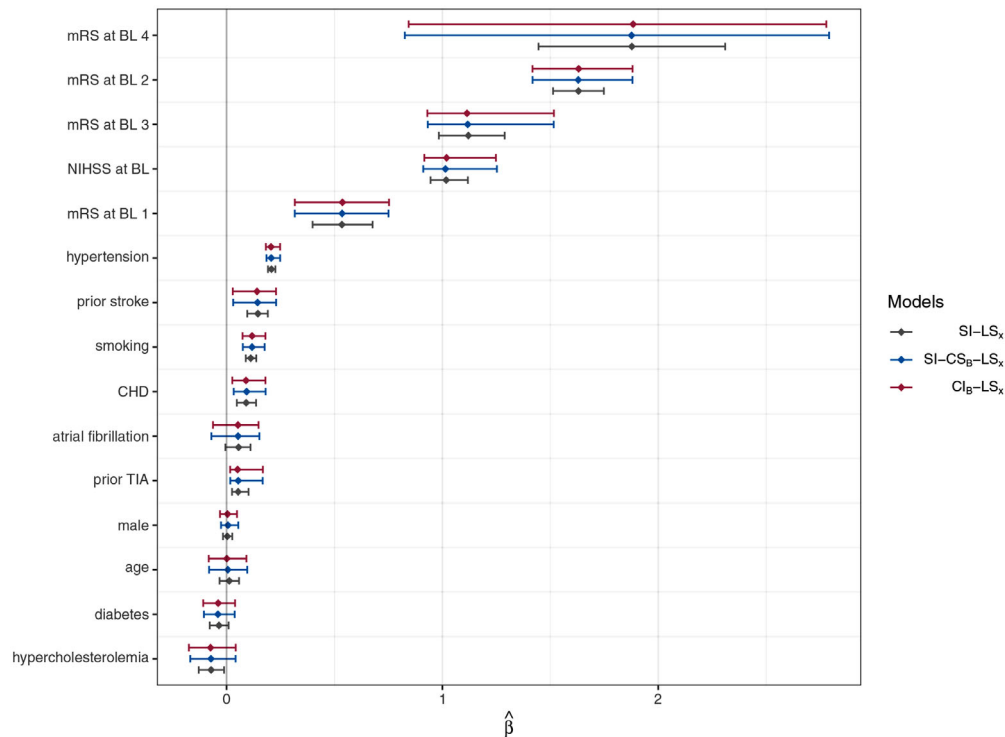
FIGURE 6 Pooled log odds ratios ($\hat{\beta}$) across all six random splits and 95% bootstrap ($B = 1000$) confidence intervals for all models with linearly included tabular features (see Section 2.2). With the exception of age and NIHSS, all features are categorical and the plot shows log odds-ratios with respect to the reference level (note that the largest observed prestroke mRS is 4). The coefficients are based on standardized features and sorted in increasing order.

no differential increase in prediction performance is observed for the tabular-data-only model compared to the most complex DTM.

## 3.3 | Interpretation of model parameters

Figure 6 visualizes the effect sizes of the tabular features in the linear shift terms LS_x of different models.

Because the logistic distribution is chosen for Z, the coefficients $\boldsymbol{\beta}$ in the linear shift term are interpretable as log odds ratios. Comparing tabular-data-only models with semistructured models shows that adjusting for the images (in CI_B-LS_x or SI-CS_B-LS_x) changes the $\hat{\boldsymbol{\beta}}$ estimates only slightly. The log odds ratios are comparable across all variables because the variables are standardized. Thus, the effect sizes reflect a change in log-odds of a worse outcome for a one-standard-deviation increase in the respective variable. Accordingly, Figure 6 shows that the strongest prognostic factors are the prestroke mRS and NIHSS on admission. This is expected when predicting 3 months mRS. The prestroke mRS captures functional disability of a patient before stroke, whereas NIHSS measures stroke severity on admission. This additionally emphasizes the importance for being able to adjust for prestroke mRS.

Similar to both linear and complex shift terms, complex intercepts of categorical predictors are directly interpretable. Here, cumulative baseline log-odds of the outcome are estimated for each stratum of the predictor. Thus, differences in the complex intercepts can be interpreted as class-specific log-odds ratios (Buri et al., 2020). For continuous predictors or images, this simple interpretation is lost to an extent that depends on the complexity of the neural network component that is modeling the complex intercept term.

Alongside interpretation, quantifying uncertainty in both predictions and parameters is of high importance, but generally difficult to achieve in DL models (Wilson & Izmailov, 2020). The use of transformation ensembles and random splits allows uncertainty quantification for the coefficient estimates in terms of bootstrap confidence intervals. This way, both aleatoric and epistemic uncertainty are captured. Note that the model coefficients of the five (members) times six (splits) are repeatedly sampled and that the models are not additionally refitted to obtain the 95% confidence intervals.

FIGURE 7 The smooth log-odds function for age fitted by a GAM (a generalized additive proportional odds model using the `mgcv` package), depicted as blue solid line along with pointwise 95% confidence band (dashed lines), and by a DTM (SI-LS$_{\bar{x}}$-CS$_{age}$) fitted on 50 bootstrap samples, depicted as gray lines. In addition, the linear effect of an SI-LS$_x$ is displayed (red line). Although both models, GAM and DTM, allow for a nonlinear effect in age, there is no evidence against a linear effect of age.
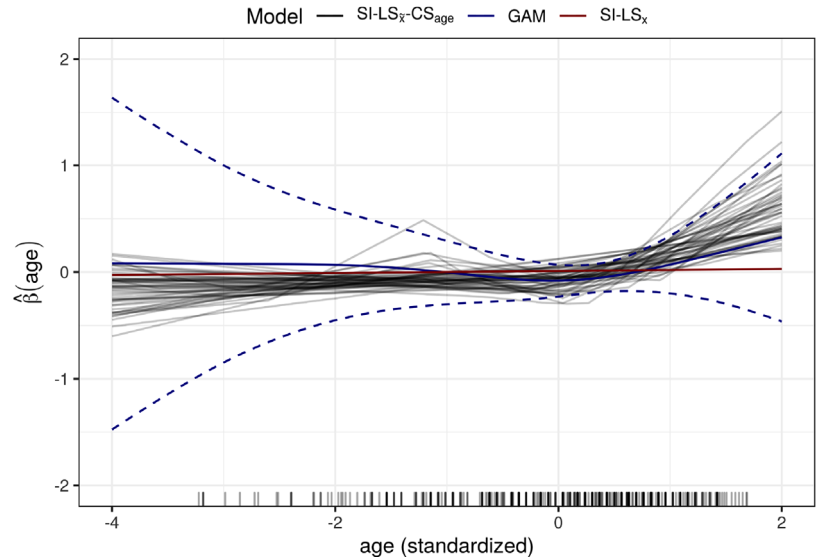
To investigate if assuming a linear age effect is appropriate, we evaluate models including the age effect with a flexible function, $\theta_k - \gamma(x_{age}) - \boldsymbol{x}_{-age}^{\top}\boldsymbol{\beta}$. We show the results of a GAM (a generalized additive proportional odds model) and a DTM (SI-LS$_x$-CS$_{age}$) that depict the estimated age effect function as shown in Figure 7. The GAM and the DTM agree in the functional form of the effect, which is constant up to a standardized age of 0.5 (corresponding to an age of 75 years) and then increases the odds for a worse outcome. However, there is no evidence against a linear effect when we consider the pointwise confidence band for the GAM. Note how the GAM enforces smoothness of the estimated function, whereas the neural network produces a piecewise linear estimate.

4 | SUMMARY AND OUTLOOK

DTMs provide a novel and flexible way to integrate multimodal data for interpretable prediction models for various kinds of outcomes. We demonstrate the potential of DTMs on a semistructured data set with an ordinal outcome (mRS) describing the functional disability of stroke patients 3 months after hospital admission. We discuss how the best trade-off between interpretability and flexibility can be achieved. In essence, we follow the top-down model approach to model building for TMs (Hothorn, 2018).

By investigating the interpretable model parameters, we judge the relative importance of the predictors and show that in a baseline-adjusted DTMs, the baseline mRS is the variable with the most relevant predictive effect. We also investigate the question, which input modality is most important for functional outcome prediction and whether predictive performance in terms of NLL and calibration can be improved by including both tabular and imaging data. While for binary mRS prediction, models seemed to slightly benefit from the addition of brain imaging data, this is not observed for ordinal mRS prediction. In general, a definitive judgment on whether the images contain information to aid mRS prediction cannot be made. This is because all results have to be interpreted conditional on (i) the small sample size and (ii) limited computation time for joint hyperparameter tuning.

When artificially increasing the sample size up to the available 407 patients, there is no evidence for differential performance gain of the semistructured over tabular-data-only models. However, extrapolating these results to larger sample sizes is in general extremely difficult.

In general, DNNs (including DTMs) are difficult to train with limited sample size and require a carefully chosen optimization procedure. For instance, transfer learning in terms of adapting the weights of a CNN that is already trained on a different data set by retraining it with the data of interest potentially improves predictive performance even with smaller sample sizes. However, methods for transferring the weights of well-known 2D CNN architectures to their 3D counterparts did not improve predictive performance in our application (results not shown). In general, it is difficult to access weights of trained 3D CNNs to then fine-tune the models.

For ordinal functional outcome prediction in our cohort, the model SI-LS$_x$ seemed to be most appropriate when including tabular features only and modeling them as linear effects. Here, classical statistical inference provides uncertainty

measures (confidence intervals) and the model is fully interpretable. Using semistructured models, including tabular and brain imaging data, improved binary mRS prediction to some extent. However, including images as a complex intercept or complex shift reduced interpretability of the model and increased variability.

TMs also work naturally for other kinds of outcomes, such as survival times, which often feature censored observations (e.g., Hothorn et al., 2018). Because the dichotomized mRS could be viewed as a censored version of the ordinal mRS, the very same models, trained on ordinal outcomes, can also be used for different dichotimizations (or binnings) of the ordinal outcome, without the need to refit the models on the binned outcome (Lohse et al., 2017).

In summary, being able to fit distributional regression models with complex outcome types and multimodal input data and following statistical principles for model building opens up vast areas of applications. Especially in medicine, these models have the potential to aid decision-making, because of their state-of-the-art prediction performance and transparency.

CONFLICT OF INTEREST
The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT
The dataset cannot be made available at this moment.

OPEN RESEARCH BADGES
This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to data confidentiality issues and their computational complexity.

ORCID
Lucas Kook https://orcid.org/0000-0002-7546-7356

REFERENCES
Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. http://tensorflow.org/

Bacchi, S., Zerner, T., Oakden-Rayner, L., Kleinig, T., Patel, S., & Jannes, J. (2020). Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes: A pilot study. *Acadamic Radiology*, *27*, 19–23.

Baumann, P. F. M., Hothorn, T., & Rügamer, D. (2021). Deep conditional transformation models. In *Machine Learning and Knowledge Discovery in Databases. Research Track* (pp. 3–18). Springer International Publishing.

Benjamine, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Chamberlain, A. M., Chang, A. R., Cheng, S., Das, S. R., Delling, F. N., Djousse, L., Elkind, M. S. V., Ferguson, J. F., Fornage, M., Jordan, L. C., Khan, S. S., Kissela, B. M., Knutson, K. L., … Virani, S. S. (2019). Heart Disease and Stroke Statistics – 2019 Update: A report from the American Heart Association. *Circulation*, *139*(10), 56–528.

Bröcker, J., & Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, *22*(2), 382–388.

Buri, M., Curt, A., Steeves, J., & Hothorn, T. (2020). Baseline-adjusted proportional odds models for the quantification of treatment effects in trials with ordinal sum score outcomes. *BMC Medical Research Methodology*, *20*(1), 1–14.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Silva, V. W. K., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, *25*(8), 1301–1309.

Chollet, F. (2015). Keras. https://keras.io

Copen, W. A., Schaefer, P. W., & Wu, O. (2011). Mr perfusion imaging in acute ischemic stroke. *Neuroimaging Clinics of North America*, *21*, 259–283.

Ebisu, T., Tanaka, C., Umeda, M., Kitamura, M., Fukunaga, M., Aoki, I., Sato, H., Higuchi, T., Naruse, S., Horikawa, Y., & Ueda, S. (1997). Hemorrhagic and nonhemorrhagic stroke: Diagnosis with diffusion-weighted and t2-weighted echo-planar mr imaging. *Radiology*, *203*(3), 823–828.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *14*(1), 107–114.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Grotta, J. C., Albers, G. W., Broderick, J. P., Kasner, S. E., Lo, E. H., Mendelow, A. D., Sacco, R. L., & Wong, L. K. (2016). *Stroke: Pathophysiology, diagnosis, and management*, 6th ed Elsevier.

Hamann & Herzog, J. L., Wehrli, C., Dobrocky, T., Bink, A., Piccirelli, M., Panos, L., Kaesmacher, J., Fischer, U., Stippich, C., Luft, J., Gralla, A. R., Arnold, M., Wiest, R., Sick, B., & Wegener, S. (2021). Machine-learning based outcome prediction in stroke patients with middle cerebral artery-m1 occlusions and early thrombectomy. *European Journal of Neurology*, *28*, 1234–1243.

Hothorn, T. (2018). Top-down transformation choice. *Statistical Modelling*, *18*(3–4), 274–298.

Hothorn, T. (2020). `tram`*: Transformation models*. R package version 0.5-1. https://CRAN.R-project.org/package=tram

Hothorn, T., Kneib, T., & Bühlmann, P. (2014). Conditional transformation models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *76*(1), 3–27.

Hothorn, T., Möst, L., & Bühlmann, P. (2018). Most Likely Transformations. *Scandinavian Journal of Statistics*, *45*(1), 110–134.

Jafari, S. H., Saadatpour, Z., Salmaninejad, A., Momeni, F., Mokhtari, M., Nahand, J. S., Rahmati, M., Mirzaei, H., & Kianmehr, M. (2018). Breast cancer diagnosis: Imaging techniques and biochemical markers. *Journal of Cellular Physiology*, *233*(7), 5200–5213.

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR. https://arxiv.org/abs/1412.6980v9

Kneib, T., Silbersdorff, A., & Säfken, B. (2021). Rage against the mean–a review of distributional regression approaches. *Econometrics and Statistics*. https://doi.org/10.1016/j.ecosta.2021.07.006

Kook, L., Götschi, A., Baumann, P. F., Hothorn, T., & Sick, B. (2022a). Deep interpretable ensembles. *arXiv preprint*.

Kook, L., Herzog, L., Hothorn, T., Dürr, O., & Sick, B. (2022b). Deep and interpretable regression models for ordinal outcomes. *Pattern Recognition*, *122*, 108263.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

Lohse, T., Rohrmann, S., Faeh, D., & Hothorn, T. (2017). Continuous outcome logistic regression for analyzing body mass index distributions. *F1000Research*, *6*, 1933.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *42*(2), 109–127.

Pinto, A., McKinley, R., Alves, V., Wiest, R., Silva, C. A., & Reyes, M. (2018). Stroke lesion outcome prediction based on mri imaging combined with clinical information. *Frontiers in Neurology*, *9*, 1060.

Quinn, T. J., Dawson, J., Walters, M., & Lees, K. R. (2009). Reliability of the modified Rankin scale: A systematic review. *Stroke*, *40*(10), 3393–3395.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rezende, D., & Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning* (pp. 1530–1538). PMLR. https://proceedings.mlr.press/v37/rezende15.html

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

Rügamer, D., Baumann, P. F., Kneib, T., & Hothorn, T. (2021). Transforming autoregression: Interpretable and expressive time series forecast. *arXiv preprint at arXiv:2110.08248*. https://arxiv.org/abs/2110.08248

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

Sick, B., Hothorn, T., & Durr, O. (2021). Deep transformation models: Tackling complex regression problems with neural network based transformation models. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE.

Siegfried, S., & Hothorn, T. (2020). Count transformation models. *Methods in Ecology and Evolution*, *11*(7), 818–827.

Stasinopoulos, D. M., & Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, *23*(7), 1–46.

Steyerberg, E. W. (2019). *Clinical prediction models*. Springer.

Thiran, J. P., & Macq, B. (1996). Morphological feature extraction for the classification of digital images of cancerous tissues. *IEEE Transactions on Biomedical Engineering*, *43*, 1011–1020.

Weisscher, N., Vermeulen, M., Roos, Y. B., & De Haan, R. J. (2008). What should be defined as good outcome in stroke trials; a modified Rankin score of 0–1 or 0–2? *Journal of Neurology*, *255*(6), 867–874.

Wilson, A. G., & Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2002.08791

Wood, S. N. (2017). *Generalized additive models: An introduction with R*, 2nd ed. Chapman and Hall/CRC.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Herzog, L., Kook, L., Götschi, A., Petermann, K., Hänsel, M., Hamann, J., Dürr, O., Wegener, S., & Sick, B. (2022). Deep transformation models for functional outcome prediction after acute ischemic stroke. *Biometrical Journal*, 1–18. https://doi.org/10.1002/bimj.202100379

APPENDIX A: COMPUTATIONAL DETAILS

For reproducibility, all code is made publicly available on GitHub https://www.github.com/LucasKookUZH/dtm-usz-stroke.

A.1 | Neural network architectures

The SI terms are modeled with a fully connected single-layer NN without hidden layers and linear activation. No bias term is used. The number of output nodes is always equal to the number of classes minus one, whereas the input is a vector of ones.

The linear shift terms for the tabular data are modeled with fully connected NNs without hidden layers and a linear function as activation. No bias term is used. The number of input units is equal to the number of tabular features, whereas the number of units in the last layer is equal to one.

The complex shift term for the variable age is modeled with a fully connected NN with two hidden layers with 16 units each, ReLU activation function, and L_2 regularization. The number of units in the last layer is equal to one and the activation function in this layer is linear.

The complex intercept and complex shift terms for the image data are modeled with a 3D CNN. The convolutional part of the 3D CNN consists of four convolutional blocks including a convolutional layer with filter size $3 \times 3 \times 3$ and a max pooling layer of size $2 \times 2 \times 2$. The first two layers use 32 filters, the following two use 64 filters. The subsequent fully connected part consists of two fully connected layers with 128 filters each, which are separated by a dropout layer with dropout rate 0.3. The activation function in all layers, expect the last one, is the ReLU nonlinearity. In case the image data are included as complex intercept term, the number of units in the last layer is equal to the number of classes minus one, that is, one when we predict the binary mRS and 6 when we predict the ordinal mRS. When integrated as complex shift term, the number of units in the last layer is equal to one. The activation function in the last layer is linear.

A.2 | Training

All models are fitted with stochastic gradient descent using the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 5×10^{-5} and a batch size of six. We then use the model with the best performance on the validation data in terms of NLL. For all experiments, the 3D images were augmented in x- and y-direction prior to every epoch using the parameters in Table A.1.

All models are implemented in R 4.1.2 (R Core Team, 2020). The models are written in `Keras` based on `TensorFlow` backend using `TensorFlow` version 2.2.0 (Chollet et al., 2015; Abadi et al., 2015) and trained on a GPU. Linear proportional odds models and generalized additive proportional odds models are fitted using `tram::Polr` (Hothorn, 2020) and `mgcv::gam` (Wood, 2017), respectively.

TABLE A.1 Parameter values for augmentation

Parameter	Value
Rotation range	20
Width shift range	0.2
Height shift range	0.2
Shear range	0.15
Zoom range	0.15
Fill	Nearest

APPENDIX B: ADDITIONAL RESULTS

Here, we present descriptive statistics and additional results.

B.1 | Baseline characteristics

Figure B.1 shows the distribution of predictors stratified by the outcome (90 day mRS) in the stroke data set.

B.2 | Test errors relative to reference model

Figures B.2 and B.3 show the test errors relative to the reference SI-LS$_x$ model evaluated on the binary and ordinal mRS, respectively. After removing the between-split variance, none of the semistructured models improve significantly upon the performance of the SI-LS$_x$ model. Since the SI-LS$_x$ performance was not bootstrapped (the constant split-wise mean was subtracted within split), there is no variance in the average AUC and QWK (because the SI model does not have any discriminatory ability, that is, AUC = 0.5 and QWK = 0) across splits for the unconditional SI model.
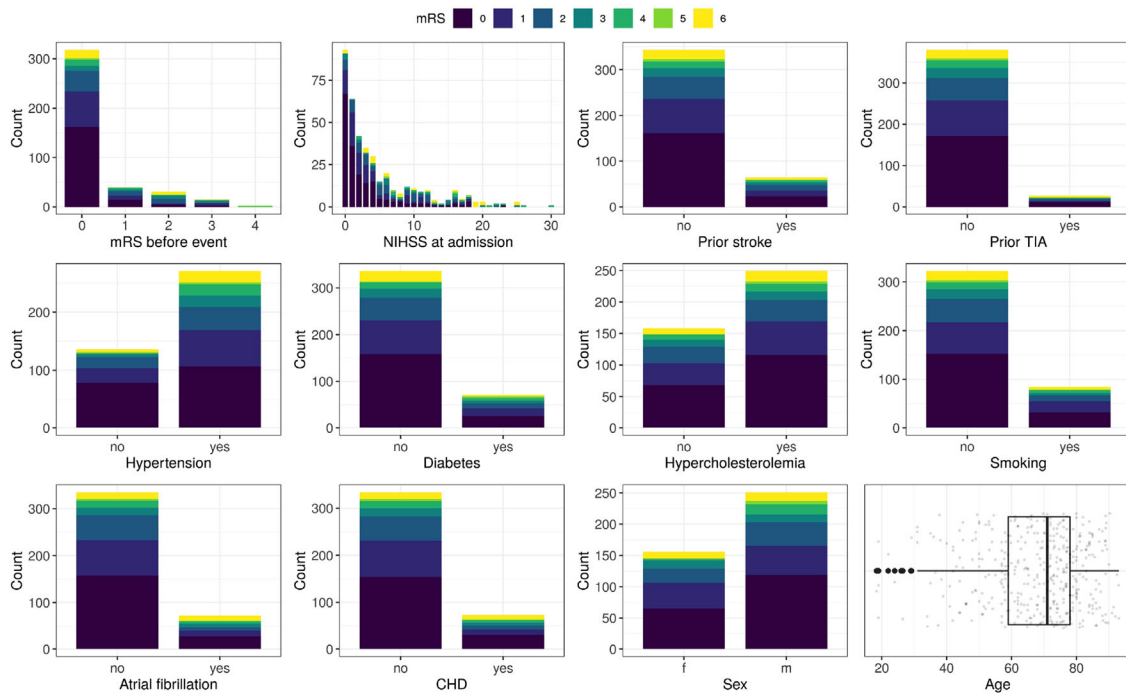

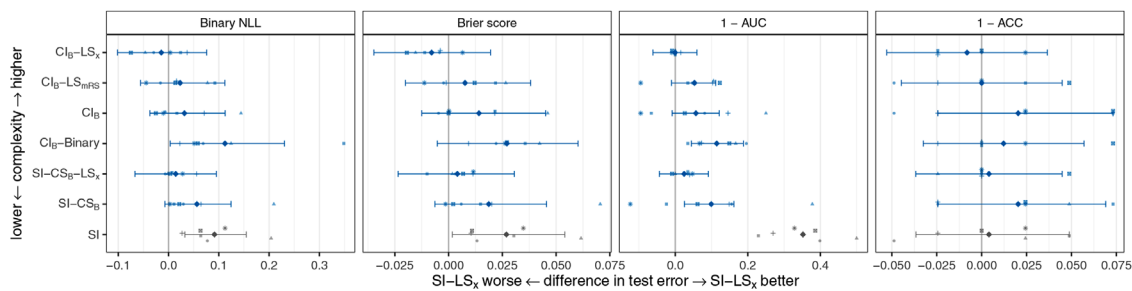**FIGURE B.1** Baseline characteristics stratified by 90 day mRS in the stroke data set

FIGURE B.2 Test error of transformation ensemble models (blue) and reference model (gray) evaluated for the binary mRS outcome (mRS 0–2 vs. mRS 3–6) relative to the test error of the benchmark SI-LS$_x$ model (i.e., a difference of 0 indicates the same performance as the SI-LS$_x$ model). The test error is quantified in terms of binary negative log-likelihood (NLL), Brier score, 1− AUC, and classification error (1 − ACC). The average test error and 95% bootstrap ($B = 1000$) confidence intervals (CI) are depicted for six random splits (indicated by different symbols). The CIs are calculated by substracting the fixed SI-LS$_x$ performance per split.
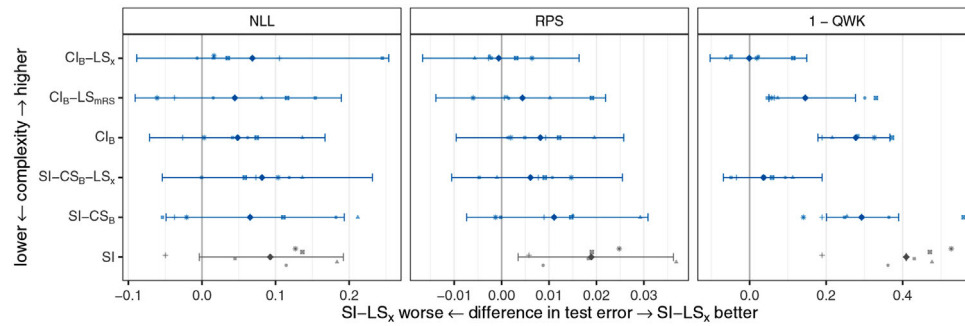
FIGURE B.3 Test error of transformation ensemble models (blue) and reference model (gray) evaluated for the ordinal mRS outcome relative to the test error of the benchmark SI-LS$_x$ model (i.e., a difference of 0 indicates the same performance as the SI-LS$_x$ model). The test error is quantified in terms of negative log-likelihood (NLL), ranked probability score (RPS), and discrimination error (1 − QWK). The average test error and 95% bootstrap ($B = 1000$) confidence intervals (CI) are depicted for six random splits (indicated by different symbols). The CIs are calculated by substracting the fixed SI-LS$_x$ performance per split.
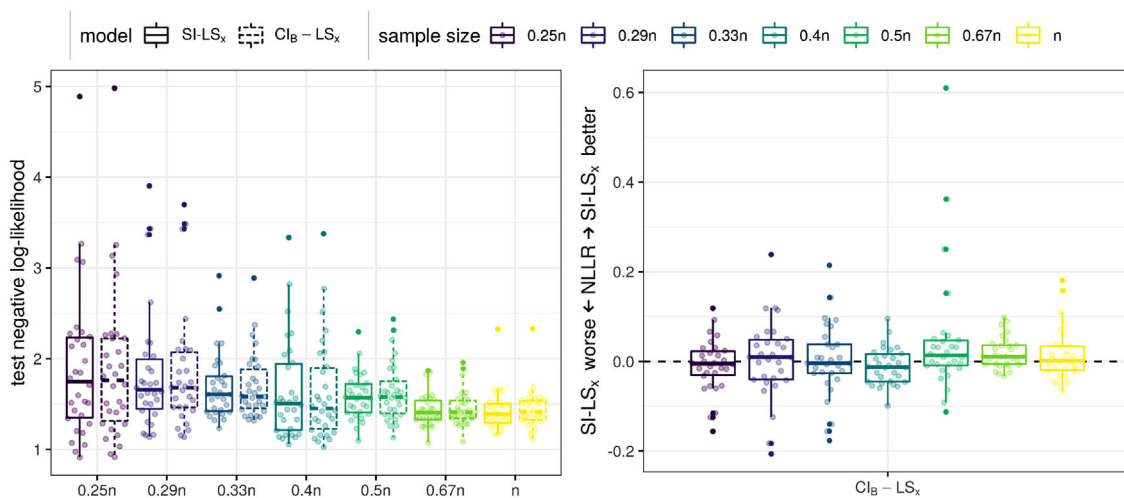
FIGURE B.4 Test performance versus sample size achieved by a subsampling experiment. The semistructured CI$_B$-LS$_x$ model and the proportional odds model SI-LS$_x$ are compared. Both models are fitted to 30 random sub-samples of seven different sample sizes of the original sample size ($n = 407$) and the test NLL is recorded. Both models benefit from increasing sample size. The right panel displays the differences in NLL within split for a given sample size (i.e., the negative log-likelihood ratio NLLR).

B.3 | Sample size

DL typically requires thousands of training images to excel at prediction performance over conventional statistical models (Goodfellow et al., 2016). However, our cohort, like most medical data sets, contained much fewer observations ($n = 407$). To study if collecting more data was a promising approach to enhance the model performance, we artificially reduced sample size by subsampling and refitted the models (see Figure B.4).

In this experiment, the sample size is artificially reduced via subsampling of varying sizes, and then, the prediction performance was measured on a hold-out set of the reduced data set. Subsampling is repeated for seven sample sizes and then 30 train/validation/test splits (with a ratio of 8:1:1) are used for fitting and evaluating the semistructured CI$_B$-LS$_x$ and tabular-data-only model SI-LS$_x$. We observe that the prediction performance, that is, the test NLL, improves for both models with increasing sample size, indicating that the performance may further increase with increasing sample size (left panel of Figure B.4). Directly comparing the performance of both models for the individual splits suggests no evidence that adding the image information to the model that contained the tabular data as input improves prediction performance. The negative log-likelihood ratio fluctuates around zero and no trend with increasing sample size is observable (right panel in Figure B.4).