# Descriptive Model Building in a low Events-Per-Variable Setting: Identifying Risk-Factors for Dysphagia

To Obtain the Master's Degree
(MSc)

at the Faculty of Digital and Analytical Sciences
of the Paris-Lodron University Salzburg, Austria

submitted by

**Gerrit Hendrik Hüller**

01522522

Supervisor: Univ.-Prof. Dipl.-Math. Dr. Arne Bathke

Department of Artificial Intelligence and Human Interfaces

Mering, July, 2023

# Summary

This thesis aimed to identify risk-factors for dysphagia, a clinically relevant swallowing dysfunction. Due to the clinical routine at the *BG Trauma Center Murnau*, observational data of $N = 403$ patients, who sustained cervical spinal cord injury, were available. Of those, $n_{\text{dysphagia}} = 87$ ($p \approx 22\%$) developed dysphagia. To find influential variables for the binary response variable, domain experts incorporated 7 predictor candidates in a *global model*. The resulting Events-Per-Variable rate of $\text{EPV}_{\text{Global}} \approx 12$ was problematically low for variable or even function selection. Since risk-factors for dysphagia had to be identified nonetheless, a potpourri of different methods was applied to provide as sound results as possible in such low EPV setting.

In order to account for statistical uncertainty and derive robust estimates for the predictor candidates, 1000 crossvalidation runs with fixed dysphagia prevalence for training and test data were performed on multiple models. Variable selection itself was performed by 2 complementing modelling approaches: a) the *global model* was implemented in 4 different logistic regressions which were adjusted to some extent for shrinkage. In b), several state-of-the-art variable selection algorithms were applied and accompanied by stability investigations in the form of variable inclusion and model selection frequencies. Due to 1000 crossvalidation runs, each predictor candidate received a subsampling distribution of estimates, from which the median and the subsampling $\text{CI}_{\text{SS.95\%}}$ were used as robust measures for the variable importance.

Ultimately, descriptive model building identified age and tracheostomy to be the most important predictors, and a high neurological level of injury as well as an anterior surgical approach as potential risk-factors for dysphagia. Despite the successful descriptive modelling, the bad fit indices of the final model indicated missing other influential risk-factors which have to be determined in future research. Nonetheless, the applied potpourri of methods proved itself to be an elaborate but informative approach to perform variable selection in the eye of the low EPV setting.

# Contents

# 1 Introduction

The increase in data volume, complexity, and especially availability over the last decades poses many opportunities for different scientific disciplines. Paradata, automatically and incidentally created data by computer usage, offer new behavioral indicators for personality assessment in Psychology (e.g., Callegaro, 2013; Diedenhofen & Musch, 2017; Hüller, 2022). In genomics, enormous amounts of DNA sequences can be analyzed to find genetic origins of rare diseases (e.g., Stephens et al., 2015). By using health insurance data, studies with a lot of statistical power can be conducted to find risk- and protective factors for public health (e.g., Hulsen et al., 2019; Alliance, 2021, 2020). Even without access to such big data, digitalization in hospitals also allows to easily track a significantly larger number of medical measurements than in the past. As in many fields of research, scientific advances also come along with challenges. Storing and accessing enormous amounts of data is a challenge in itself, but analyzing that data is no less complex.

Big data is not only defined by its volume but also by its complexity, such as e.g., non-tabular or nested data structures as well as ambiguities within the data. Thus, the term "big" not necessarily (but also possibly) refers to a high number of measurement points (i.e., sample size), but can also indicate a large number of features/ variables to choose from (e.g., Chen & Wojcik, 2016). Having such a high dimensional feature space is especially challenging in settings with low sample sizes which is not uncommon in life and medical sciences. As a result, the careful selection of explanatory features plays a pivotal role in statistical model building. This is also known as variable selection.

Before starting to think about variable selection, one should first choose a conceptual modelling approach suitable for the research question at hand: should it be a predictive, explanatory or descriptive model (Shmueli, 2010; Sauerbrei et al., 2020)? Depending on the approach, one focuses either on the predictive performance, parsimony/ understandability, or causality. Explanatory modeling concentrates on causality by testing associations between predictors and the response variable via e.g., regression models and minimizing bias. Causality itself is not explicitly modelled, but argued by theory and specific criteria (e.g., cf. Cox, 2018). Those provide the foundation to model the confounders, predictor(s) and their associations (i.e., functional forms) to the response variable. Consequently, theory and (domain) expertise are of primary importance in this type of modelling. Like the name suggests, predictive modelling aims for most accurate predictions of new or future data. Hence, the focus lies less on parsimony, interpretability or theory, but on the outcome and (minimizing) the expected prediction error. In order to improve predictions, the variance-bias trade-of can favor a reduced variance over "unbiasedness". As such, so called shrinkage methods can be utilized which artificially introduce some (small) bias towards zero to reduce the overall prediction error. Finally, descriptive modeling aims to strike a balance between accurately representing the data structure and achieving simplicity, with reduced dependence on theory or, ideally, without relying on theory altogether (Shmueli, 2010; Sauerbrei et al., 2020; Heinze, Wallisch, & Dunkler, 2018).

Depending on the conceptual approach of model building, a higher or a smaller number of predictors or polynomials are allowed in the model which is consequently deemed more or less complex. For smaller (i.e., less complex) models more conservative (selection) criteria are specified for the variable (or function) selection process, whereas more liberal criteria can be applied to select bigger (i.e., more complex) models. In other words: depending on our chosen approach, our model contains more or less parameters to estimate. To obtain accurate estimates and predictions, it is necessary to have a sample size that is sufficiently large relative to model complexity. With a limited number of measurement points (i.e., a small sample), fewer predictor candidates should be considered simultaneously in variable selection (e.g., Heinze et al., 2018). To restrict the search space for potential predictor candidates, both domain knowledge and variable selection algorithms can be applied. Such approaches usually implement a stepwise procedure, where either an empty model is filled with relevant, or a "full" model is reduced by irrelevant variables (Sauerbrei et al., 2020; Sauerbrei, Royston, & Binder, 2007; Heinze et al., 2018). This thesis aims to examine the strengths and weaknesses of various variable (and function) selection methods in different settings.

Variable selection is closely related to function selection in continuous variables. By introducing (fractional) polynomial terms or splines, also non-linear relationships can be modelled (Sauerbrei et al., 2020; Sauerbrei & Royston, 1999; Royston & Sauerbrei, 2008). Although too often neglected or ignored, many phenomenons may be non-linearly associated (Sauerbrei et al., 2020, 2007; Heinze et al., 2018). In life sciences many models claim linear relationships for the sake of simplicity and interpretability (a.k.a., *"no model is correct but some are useful" - attributed to George Box*). Modelling non-linearity with e.g., a polynomial regression introduces new challenges. Such as an increase in e.g., $age^2$ by 1 year while holding *age* constant, has barely any practical meaning without further investigations (Harrell, 2015a). Additionally, some non-linear associations differ in their polynomial terms, but are almost arbitrary for interpretation (e.g., $age^2$ vs. $age^4$ vs. $age^6$; Simonsohn, 2018). A reasonable compromise might be to model local linearity within the data by an "interrupted" regression. Similar to regression splines, one can approximate the true (non-linear) functional form by applying linear models on local data subsets. Another popular approach to circumvent thinking about non-linearity is dichotomization. Although such approach easily allows to model any type of association, information is lost in this process. Also, the number of categories and their thresholds often are debatable and the resulting step-functions might be implausible (Sauerbrei et al., 2020). In the end, introducing categorization conditions (like in rule learning), splines or (fractional) polynomial terms increases the number of parameters to estimate and therefore contribute to statistical uncertainty in low sample size scenarios (Sauerbrei & Royston, 1999; Sauerbrei et al., 2007, 2020; Royston & Sauerbrei, 2008; Heinze et al., 2018).

Again, all comes down to sample size. Smaller samples increase the uncertainty of estimates and increase the chance of overfitting by modelling sample specific random variation. Simultaneously, variable (and function) selection repeatedly test single predictor candidates regarding their relevance. By testing many candidates, pure chance can lead to the in-/exclusion of predictors and consequently lead to biased (i.e., overestimated) regression coefficients and (underestimated) p-values. The latter is a result of repeated testing which invalidates the nominal confidence levels (Sauerbrei et al., 2020; Heinze et al., 2018). As a countermeasure, Sauerbrei et al. (2020) and Heinze et al. (2018) propose variable selection in combination with stability investigations and shrinkage methods, especially in small sample size situations. The former assesses variable inclusion and model selection frequencies within re- or subsampling methods (e.g., bootstrapping & crossvalidation). The latter covers techniques which introduce bias towards zero on predictors' estimates in order to (partly) remove the overestimation introduced by variable selection. While both papers reference literature that suggests shrinkage methods result in improved model selection in small sample scenarios, a simulation study conducted by Van Calster, van Smeden, De Cock, and Steyerberg (2020) contradicts this finding by demonstrating that shrinkage methods may not necessarily lead to superior models.

In summary, variable (and function) selection is valuable in the process of (exploratory) model building, particularly when numerous potential predictors are available. However, it also poses limitations, especially when dealing with small sample sizes. Furthermore, the impact of shrinkage methods on variable selection in small samples remains a subject of controversy.

The practical part of this thesis aims to identify several risk-factors for the binary outcome dysphagia, a swallowing dysfunction. In a cohort of $N = 403$ individuals who sustained acute traumatic cervical spinal cord injury (SCI), $n_{\text{dysphagia}} = 87(21.59\%)$ patients developed dysphagia. Performing variable or function selection on these observational data is comprised in the descriptive modelling framework. Domain experts identified 7 predictor candidates which are quite a lot with respect to the small sample of $n_{\text{dysphagia}} = 87$ dysphagia patients in order to gain robust and trustworthy results. Therefore, variable selection has to be performed in a so-called low Events-Per-Variable setting which ultimately afflicts model selection with uncertainty. As previously discussed, stability investigations and shrinkage methods are proposed as potential solutions (Heinze et al., 2018). But current research challenges the extent to which shrinkage methods contribute to better performing models (Van Calster et al., 2020). Consequently, this thesis examines the effect of different shrinkage methods on descriptive model selection and the corresponding model performance in addition to ”plain” variable selection.

To address this research question, it is essential to conduct a thorough analysis that fulfills the requirements posed by the available sample. To do so, this thesis first lays theoretical foundations for modelling (non-)linear associations with binary data and introduces different variable and function selection procedures (see section 2.3.6). More concrete, this comprises the introduction of different modelling frameworks, such as generalized linear models and generalized additive models. Further, different variable (and function) selection criteria and corresponding algorithms are presented and discussed regarding their limitations. This is followed by a more detailed introduction to the research question and the available data (cf. s. 3.2). Within the methods section (cf. s.4.3), the sample characteristics are presented and the analysis plan is derived with respect to the research questions and the before mentioned theoretical foundations. Ultimately, the results (cf. s.5) are presented, discussed (cf. s.6.2) and conclusions are derived (cf. s.7).

# 2 Theory

The next sections cover technical definitions and formulas. There, matrices $\mathbf{X}$ are written in capital and bold letters, vectors $\mathbf{x}$ in lower case bold letters, random variables $Y$ in italic capital letters, and single observations $y$ as realisations of those random variables in italic lower case letters. Functions are written in plain (i.e., neither bold nor italic) letters.

## 2.1 Modelling Framework: Regression Models

Due to the later presented research question (cf. s.3.2), this thesis is primarily concerned with variable and function selection in (generalized) linear regression models. But as non-linearity is also discussed in function selection procedures, the following section about modelling frameworks cover both ”simple” linear regression models, generalized regression models (i.e., also logistic regression), and generalized additive models, before covering variable and function selection itself.

In general, regression modelling tests associations between a matrix of observations of measurements $\mathbf{X}$ (i.e., covariates, regressors, predictors, explanatory, or ”independent” variables, IV), and the respective vector of responses $\mathbf{y}$ (outcome, target, regressand, or dependent variable, DV) (Myers & Montgomery, 1997; Shmueli, 2010; Harrell, 2015a, p.103). Let's assume we have a sample of $N$ observations, then our outcome observations $\mathbf{y}$ are $N$ identically and independently distributed realizations of a random variable $Y$ (McCullagh & Nelder, 1989, p.26). In order to introduce generalized linear models (GLMs), McCullagh and Nelder (1989) suggest to call the response $\mathbf{y}$ the *random component* of a linear model. Complementary, there also exist some *systematic component* which explains/predicts our *random component* $\mathbf{y}$. In other words, the *systematic component* comprises a (linear) combination of explanatory variables. As explanatory variables are often measurements and both measurement error and true random variation exist in the state of the measured concept (e.g., human body height during the course of the day), there is not only the true *systematic effect* but also a *random effect*, namely the error term $\boldsymbol{\epsilon}$, in our *systematic component* (see equation (1); McCullagh & Nelder, 1989, p.3).

$$\text{random component} = \text{systematic component} = \text{systematic effect} + \text{random effect} \tag{1}$$

Depending on the modelling approach, one would rather call $\mathbf{x}$ predictors or explanatory variables (Shmueli, 2010; Heinze et al., 2018). Since they denote the same concept and there is no fixed convention, these terms are used synonymously throughout this thesis. The earlier presented components which are aimed to be minimized can be found in Eq. (2), such that the *irreducible error* corresponds to the *random effect*.

$$\text{expected prediction error} = \text{irreducible error} + \text{bias}^2 + \text{variance} \tag{2}$$

### 2.1.1 Pre-requisite: ”Simple” Linear Regression Models

Linear regression models assume a linear relationship between a covariate vector (i.e., the predictors) of length $K$ $\mathbf{x} = (x_1, ..., x_K)$ and the observed continuous outcome $y$ (Hastie, Tibshirani, & Friedman, 2009; Royston & Sauerbrei, 2008; Heinze et al., 2018). The deviation of the observed outcomes $y$ to the expected outcomes $\mu = E(y)$ is assumed to originate from the error term $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with a mean of 0 and variance $\sigma^2$. Such that for a single observation, the linear combination of $K$ explanatory variables with their respective weights $\boldsymbol{\beta}$ and intercept $\beta_0$ are addable as in Eq. (3) (Royston & Sauerbrei, 2008, p.10).

$$y = \mathrm{E}(y) + \epsilon = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \epsilon = \beta_0 + \mathbf{x}\boldsymbol{\beta} + \epsilon = \eta + \epsilon \tag{3}$$

In the terms of McCullagh and Nelder (1989, p.3, 26), we explain our *random component* $y$ by the *systematic component* comprising the sum of the *random effect* $\epsilon$ and the linear predictor $\eta$ in Eq. (4) (i.e., "index" or *systematic effect*; Royston & Sauerbrei, 2008, p.11).

$$\eta = \beta_0 + \mathbf{x}\boldsymbol{\beta} \tag{4}$$

The intercept $\beta_0$ and regression coefficients $\boldsymbol{\beta}$ are usually unknown (model) parameters which can be estimated as $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ in our "estimation equation" Eq. (5) via methods such as ordinary least squares (OLS) or maximum likelihood (ML). Both estimation methods aim to solve an optimization problem and thus find "optimal" estimates which either maximize the (log-)likelihood (ML) or minimize the sum of squared residuals (OLS). Residuals are defined as the deviation between observed and expected values $r_i = y_i - \mu_i$ on the data used for parameter estimation. Data used for parameter estimation are also often called training data. New test data in contrast are unrelated to parameter estimation and either used for model validation or performing predictions on them. The differentiation between data used for parameter estimation and "unrelated" test data is especially important when discussing phenomenons such as overfitting (cf., s. 2.2.2). Since those estimation procedures are not of primary interest in this thesis, we generally call them parameter estimation, model fitting/training or simply optimization procedures. More information can be found in e.g., Royston and Sauerbrei (2008) and Hastie et al. (2009) among the other references mentioned in his section. Regardless of the fitting procedure, Royston and Sauerbrei (2008) suggest to interpret the average residual $\bar{r} = \frac{1}{N}\sum_{i=1}^{N} y_i - \mu_i$ for any given and fixed covariates $\mathbf{x}$ as an estimate for the bias in Eq. (2).

$$\mu = \hat{y} = E(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_K x_K = \hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}} = \hat{\eta} \tag{5}$$

### 2.1.2 Generalized Regression Models - Logistic Regression

As implied by its name, GLMs generalize the ordinary linear model we previously discussed by accommodating various types of response data, including continuous variables as well as binary or count data, among others. Hence, the (simple) linear model is a special case of GLMs. Based on McCullagh and Nelder (1989), GLMs consist of a *random component* $\mu$, a *systematic component* $\eta$ and a *link function* $g(\cdot)$ connecting both (cf. Eq.

(6) & (7)).

$$g(\mu) = \eta \tag{6}$$

$$\mu = g^{-1}(\eta) \tag{7}$$

In Eq. (5) we derived $\mu = \mathrm{E}(y) = \hat{\eta}$, where the link between the *random and systematic component* is the identity function. In the latter, our observed continuous response $y$ is assumed to be normally distributed. GLMs can model the response of any distribution from the exponential family. Such that (non-negative) count responses can be modeled by a logarithmic link, binary responses by e.g., the logit function, or exponential responses by the gamma function. In other words, link functions $g(\cdot)$ restrict the range of $\mu = \mathrm{E}(\mathbf{y})$. By doing so, the visual appearance of our GLM graph might suggest non-linear associations, which is not true. Linearity is provided by the application of the link function. Let $z = g(\mu)$ be our transformation of $\mu$, then we can derive $z = \eta$, which is identical to Eq. (6) and looks similar to our linear statistical model in Eq. (5). Ultimately, the intepretability of the *random or systematic component* is influenced by the selection of the (inverse) link function. The link function itself can take any form as long it is monotonic differentiable and the distribution of the response belongs to the exponential family. For more information on the link function and parameter estimation see McCullagh and Nelder (1989, p.27ff.).

**Logistic Regression** Given that this research thesis focuses on descriptive modeling using observational (retrospective) data and a binary response variable, logistic regression can be utilized (McCullagh & Nelder, 1989, cf. s.4.3.3 on p.111ff.). Logistic regression is a GLM which models a Bernoulli distributed binary response. Hence, $y$ can either take the value 1 with the probability $\mathrm{P}(y = 1) = \pi$ or the value 0 with $\mathrm{P}(y = 0) = 1 - \pi$. While 1 is often associated with an event like "success" or the "caseness" of an outcome (e.g., successfull recovery after a clinical treatment), 0 is its complement (i.e., non-event; Royston & Sauerbrei, 2008, p.12). Although, the observations are only either 1 or 0, we are interested in the expected probability $\pi$ for a "successfull" response, given a set of covariates. Logistic regression examines this association between the (expected) response probability $\mu = \mathrm{E}(\mathbf{y}) = \pi$ and the covariates within the *systematic component* (McCullagh & Nelder, 1989, p.98).

In order to model this association and hence calculate the probability for each $i = 1, ..., N$ observation $\mu_i = \pi_i$, one must estimate adequate (model) parameters from the observed binary responses $y$ and restrict the range of $\mu = E(y)$ between 0 and 1. While there exist several link functions, the logistic (or logit) link function is probably the most common one for this task. In contrast to other link functions, it can model both prospective and retrospective sampled data and enables the elegant interpretation of the transformed response in form of the log-odds = $\log(\frac{\pi}{1-\pi})$ (cf. Eq. (8) & (9); McCullagh & Nelder, 1989, p.108). Consequently, when a single covariate is increased by one unit while holding all other covariates constant, one can either a) calculate the change in the (log-) odds ratio for "success" (i.e., $y = 1$) as shown in Eq. (8), or b) calculate the probability $\pi$ for a specific combination/expression of covariates (cf., Eq. (9)).

$$g(\pi) = \log(\frac{\pi}{1-\pi}) = \eta = \beta_0 + \mathbf{x}\boldsymbol{\beta} \tag{8}$$

$$\pi = g^{-1}(\eta) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{e^{\beta_0+\mathbf{x}\boldsymbol{\beta}}}{1+e^{\beta_0+\mathbf{x}\boldsymbol{\beta}}} \tag{9}$$

A detailed description of the parameter estimation, as well as transformations for the (binomial) distribution in order to achieve attributes like stable variance, symmetry, or additivity are out of the scope of this thesis and can be found in McCullagh and Nelder (1989, Chapter 2 & s.4.2.5 in p.105 respectively). A brief summary regarding parameter estimation would be, that - analogous to other GLMs - fitting can be achieved iteratively by maximizing the log-likelihood via the Newton-Raphson method, such that the maximum log-likelihood function for the model parameters $\boldsymbol{\theta} = (\beta_0, \beta_1, ..., \beta_k)^T$ corresponds to $\ell_{LR}(\boldsymbol{\theta})$ in Eq. (10) (cf. e.g.; Van Calster et al., 2020).

$$\ell_{LR}(\boldsymbol{\theta}) = \sum_{i=1}^{N}\left\{y_i\log\Big(\pi_i(\boldsymbol{\theta})\Big)+\Big(1-y_i\Big)\log\Big(1-\pi_i(\boldsymbol{\theta})\Big)\right\} \tag{10}$$

### 2.1.3 Generalized Additive Models - Modelling Non-Linear Associations

Generalized Additive Models (GAMs; Hastie & Tibshirani, 1990) extend GLMs by introducing smoothed non-linear terms and hence relax the linearity assumption (Sauerbrei et al., 2020; Wood, 2017). Consequently, GAMs are utilized to model more complex associations between continuous covariates and the response. Those smoothed non-linear terms are univariate transformations (i.e., each function takes at most one covariate as input) and are sometimes called basis-function methods (Hastie et al., 2009). In contrast to (fractional) polynomials, which specify the functional form between predictor and response in a (semi-) parametric way, GAMs only define "smooth functions" which can result in various functional forms depending on the fitted values (i.e., in a more non-parametric way; Wood, 2017, p.161).

In GAMs, the originally linear *systematic component* $\eta = \beta_0 + \mathbf{x}\boldsymbol{\beta} = \beta_0 + \sum_{k=1}^{K}\beta_k x_k$

describes the sum over the products between covariate $x_k$ and their corresponding slope $\beta_k$. Hastie and Tibshirani (1990) describe GAMs in the form of $y = \beta_0 + \sum_{g=1}^{G} f_g(x_g)$ where $f(\cdot)$ denote more complex transformations than the simple dot product $x_k\beta_k$. This transformation can either be the identity function, i.e., the covariate $x$ is assumed to have a linear relationship to the outcome, or some other function like splines (Sauerbrei et al., 2020). Now, let's define the additive component as $\gamma = \sum_{g=1}^{G} f_g(x_g)$ and modify our linear predictor $\eta = \eta_0 + \eta_+$ where $\eta_0$ corresponds to the intercept $\beta_0$ and $\eta_+$ to the sum of the linear dot products $\eta_+ = \sum_{k=1}^{K} x_k = \mathbf{x}\boldsymbol{\beta}$. Then, based on (Wood, 2017, p.161) a formal GAM can be written as in Eq. (11), where $\eta$ denotes the sum of all $K$ linear and $\gamma$ the sum of all $G$ transformed parameters. Of course, GAMs can either contain solely linear components $\eta_+$ or solely additive components $\gamma$ in addition to the optional intercept $\eta_o = \beta_0$.

$$g(\mu) = \eta + \gamma = \eta_0 + \eta_+ + \gamma = \beta_0 + \sum_{k=1}^{K} x_k\beta_k + \sum_{g=1}^{G} f_g(x_g) \tag{11}$$

Since GAMs use smoothing functions $f(\cdot)$ to model non-linear relationships, not only the smoothing function itself but also their degree of "smoothness" or "wiggliness" has to be specified (Wood, 2017, p.161). While some types of smoothing functions are discussed later, the extent of smoothness can be controlled with an added penalization term $\lambda$ within optimization process (e.g., ML + penalization term). Such that a high penalty $\lambda \to \infty$ leads to a straight line and $\lambda = 0$ leads to an un-penalized ("wiggly") regression estimate (Wood, 2017, p.168). Details on the fitting procedure as well on determining the smoothing parameter $\lambda$ via generalized cross validation scoring can be found in Wood (2017). Penalization terms in general are discussed in the section about regularized regression models in s.2.2.3.

The transformations $f(\cdot)$ are usually represented as a linear combination of $J$ specific basis functions, such that for a specific transformation $f_g(\cdot)$ on the $g^{\text{th}}$ covariate corresponds to

$$f_g(x_g) = \sum_{j=1}^{J} b_{gj}(x)\beta_{gj} \ , \tag{12}$$

where $b_{gj}(x)$ is the $j^{\text{th}}$ basis function of covariate $x_g$ with its respective slope $\beta_{gj}$. Analogous to Wood (2017, p.162), we could model a 3rd order polynomial on $x_g$ with the basis functions $b_{g1}(x_g) = 1$, $b_{g2}(x_g) = x_g$, $b_{g3}(x_g) = x_g^2$, and $b_{g4}(x_g) = x_g^3$, and its transformation $f_g(x_g) = \beta_{g1} + x_g\beta_{g2} + x_g^2\beta_{g3} + x_g^3\beta_{g4}$.

Nonetheless, using polynomials as basis-functions and hence modelling non-linear relationships parametrically has limitations (Wood, 2017, p.162-164). Polynomials have a tendency to be oversensitive to noise within the data, and consequently perform worse in inter- and extrapolation tasks as they are heavily influenced by single observations. In contrast to polynomial basis-functions, so-called spline smoothing functions are much more often mentioned in the context of GAMS (Sauerbrei et al., 2020; Hastie & Tibshirani, 1990; Wood, 2017; Royston & Sauerbrei, 2008).

**Splines** Spline functions are linear combinations of spline basis functions to model non-linear effects. Wood (2017, in Sauerbrei et al., 2020, p.9) describes spline functions as "a set of piecewise polynomial functions [...] that are joined smoothly at a set of knots spread accross" the range of values within the respective covariate. Since those knots are distributed over a pre-specified range on the covariate, spline functions model the association between two consecutive knots in a "local" and piecewise manner. Basically, the more knots are modelled, the more flexible the function is (Royston & Sauerbrei, 2008, p.204). Different types of splines specify the range on the covariate as well as the functional form of the association between two consecutive knots (Sauerbrei et al., 2020). There are several different sub-types of splines among the more general smoothing splines and regression splines. Detailed explanations about splines can be found in Hastie and Tibshirani (1990, chapter 2.9) or Wood (2017, chapter 5). In order to provide some general idea, how spline regression models (i.e., GAMs with spline functions) work, natural cubic splines (a.k.a., restricted cubic splines) are presented exemplary:

Assume, we have a sequence of $M + 2$ evenly distributed and ordered knots $\xi_{\min} < \xi_1 < ... < \xi_M < \xi_{\max}$ over the range of values within the covariate $x$. This would result in $M$ inner knots and 2 boundary knots $\xi_{\min}$ and $\xi_{\max}$ which are usually but not necessarily placed on the extremes of $x$ (Royston & Sauerbrei, 2008, p.203). Further, $M + 1$ sections between consecutive knots would have to be modelled non-linearly. Each section is modelled with a cubic term $x^3$ and hence, each of those $M + 1$ sections owns a separate regression coefficient. At the knots themselves, the slope and curvature (i.e., the first 2 derivatives) is matching to the neighbouring sections. In other words: the spline function is continuous at the knots. As a *natural* cubic spline model the tails of the spline function (i.e., "outside" of the boundary knots) are restricted to be linear (hence the alternative name *restricted cubic splines;* Royston & Sauerbrei, 2008, p.203). Consequently, the end knots $\xi_{\min}$ and $\xi_{\max}$ have zero second derivatives (Wood, 2017, p.196).

A natural cubic spline consists of several components, which are linearly combined. There is a simple linear regression component $\beta_{(0)} + \beta_{(1)}x$ to model values outside of our spline boundaries and there is the cubic component $C(\cdot)$ (cf. Eq. (15); Gauthier, Wu, &

Gooley, 2020). The cubic component comprises $M$ cubic basis functions $b_m = (x - \xi_m)^3_+$ for the inner knots $\xi_1, ..., \xi_M$, as well as a penalization term $\lambda_m$ weighting the distance to the boundary knots $\xi_{\min}$ and $\xi_{\max}$ in Eq. (13). The "plus function" $(\cdot)_+$ indicates, that only the positive results or 0 are returned (see Eq. (14); Binder, Sauerbrei, & Royston, 2013; Royston & Sauerbrei, 2008).

$$\lambda_m = \frac{\xi_{\max} - \xi_m}{\xi_{\max} - \xi_{\min}} \tag{13}$$

$$(x - \xi)^3_+ = \begin{cases} (x - \xi)^3 & \text{if } x \geq \xi \\ 0 & \text{if } x < \xi \end{cases} \tag{14}$$

$$C_m(x) = \Big[(x - \xi_m)^3_+ - \lambda_m (x - \xi_{\min})^3_+ - (1 - \lambda_m)(x - \xi_{\max})^3_+\Big] \tag{15}$$

Finally, a specific natural cubic spline transformation $f(\cdot)$ for the covariate $x$ can be specified as shown in Eq. (16) (cf.; Royston & Sauerbrei, 2008; Binder et al., 2013; Gauthier et al., 2020). The application of spline functions involves the incorporation of M+2 parameters which need to be fitted for each covariate undergoing the transformation. In multivariable spline modelling, the number of parameters which require estimation rapidly increases.

$$f(x) = \beta_{(0)} + \beta_{(1)} x + \sum_{m=1}^{M} \beta_m C_m(x) \equiv$$

$$f(x) = \beta_{(0)} + \beta_{(1)} x + \sum_{m=1}^{M} \beta_m \Big[(x - \xi_m)^3_+ - \lambda_m (x - \xi_{\min})^3_+ - (1 - \lambda_m)(x - \xi_{\max})^3_+\Big] \tag{16}$$

Whereas spline functions model non-linearity between consecutive knots, fractional polynomials aim for non-linear modelling by applying polynomials with restricted complexity in order to circumvent the limitations of polynomials mentioned earlier.

**Fractional Polynomials** Fractional polynomials (FP) are an extension of power transformations of covariates (e.g., Sauerbrei et al., 2020; Royston & Sauerbrei, 2008). Although their relationship to GAMs is not explicitly addressed in Hastie and Tibshirani (1990), Wood (2017) or in Sauerbrei et al. (2020), one can derive from Binder et al. (2013) that FP - as transformations - can be modelled via $f(\cdot)$ in GAMs.

While splines model non-linearity locally between consecutive knots, FP model non-linearity over the whole range of the covariate (Sauerbrei et al., 2020). Contrary to full polynomial basis functions, fractional polynomials apply a power transformation on a covariate $x_g$ from a restricted set of powers $p \in S$ with $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, where for $p = 0 : x^0 := \log(x)$. For most use cases, a fractional polynomial transformation of degree 1 (FP1; cf. Eq. (17)) or of degree 2 (FP2; cf. Eq. (18)) is considered sufficient (Sauerbrei et al., 2020; Binder et al., 2013; Royston & Sauerbrei, 2008).

$$f(x) = \beta x^p \tag{17}$$

$$f(x) = \beta_1 x^{p_1} + \beta_2 x^{p_2} \tag{18}$$

If $p_1 = p_2 = p$, the FP2 model is defined as $f(x) = \beta_1 x^p + \beta_2 x^p + \log(x)$. With the beforementioned set of restricted powers, FP1 defines 8 and FP2 36 different models which can be fitted with maximum likelihood estimation (for more details see chapter 4 in; Royston & Sauerbrei, 2008). FP and the choice of its power(s) $p$ are usually applied in the context of variable and function selection, which is covered in multivariable model building with FP (MFP; see section 2.3.4 or chapter 6 in; Royston & Sauerbrei, 2008). Although splines are more flexible regarding functional forms, Sauerbrei et al. (2020) highlights the easier interpretability and the fewer number of parameters which have to be estimated in FP, while still covering a wide range of different functional forms.

Within the framework of GAMs, linear and non-linear associations can be modeled with any response variable, as long as its distribution belongs to the exponential family. Additional to GLMs and GAMs, spline transformations and fractional polynomials were presented as more or less flexible and interpretable methods to model non-linear relationships. Up to now, model fitting is assumed to result in the estimation of the ”best possible” parameters. In the following sections, we first discuss what this optimization procedure means with respect to variance and bias. Further, the terms overfitting and shrinkage are introduced with their countermeasure, namely shrinkage methods. Post-estimation shrinkage factors (PESF) and penalized likelihood optimization are two procedures presented for deriving shrinkage factors. Despite dealing with overfitting, the application of shrinkage factors can also be employed to handle multicollinearity or to perform variable selection.

Section 2.2 first introduces the Events-Per-Variable (EPV) as a heuristic to determine model complexity before differentiating between bias, variance, overfitting, and shrinkage. Afterwards, post-estimation shrinkage factors and penalized likelihood methods like ridge and Lasso regression, as well as Firth's correction for rare-events in logistic regression are presented to lay foundations for state-of-the-art (SOTA) variable and function selection approaches and its challenges.

## 2.2 Sample Size, Bias, Variance, & Shrinkage

### 2.2.1 Events-Per-Variable (EPV)

Events-Per-Variable (EPV) describe the ratio between sample size and the number of parameters (excluding the intercept) which have to be estimated in the model (Van Calster et al., 2020; Heinze et al., 2018; Royston & Sauerbrei, 2008). In case of dichotomous outcome variables, the number of measurement points in the smallest group is used to calculate the EPV. I.e., in a logistic regression model for a binary response variable, $\text{EPV} = \frac{\min(n_1,n_0)}{|\boldsymbol{\theta}|-1}$, with $|\boldsymbol{\theta}| :=$ number of model parameters. Of course, non-linear transformations of predictors possess their own parameters in a regression model and hence decrease the EPV when applied (Heinze et al., 2018).

The lower the EPV rate, the less information (i.e., measurement points) are available for the estimation of each parameter. The corresponding higher uncertainty in the estimations depends not only on the sample size (i.e., EPV), but also on the correlation structure (i.e., multicollinearity), and effect size of predictors (Courvoisier, Combescure, Agoritsas, Gayet-Ageron, & Perneger, 2011). Hence, popular rule of thumbs for regression modelling, such as EPV $> 10$ (Harrell, Lee, Califf, Pryor, & Rosati, 1984 in Heinze et al., 2018) or the more recent EPV $> 15$ (Harrell, 2015b) are oversimplifications. Regardless, EPV are an easily accessible heuristic for defining model complexity (Heinze et al., 2018).

### 2.2.2 Bias, Variance and Overfitting

In Eq. (2) of section 2.1 which discusses regression methods, we derived that the expected prediction error comprises the sum of an irreducible error, (squared) bias, and variance. Assuming a true mean of our outcome variable exists, the irreducible error describes the variance around the true mean and can not be extinct, unless there exists no *random effect* $\epsilon$, i.e., variation $\sigma_\epsilon^2 = 0$ within the data. The (squared) bias reflects the (systematic) average deviation from our predicted values to their true mean. Variance is defined by the (unsystematic) average squared dispersion of our predicted values and their mean (Hastie et al., 2009, p.223). The latter is the result from estimating the parameters on a sample, instead of the whole population. High bias therefore represents a systematic (prediction) error and results from misspecifying the applied model such as e.g., not including *all* relevant predictors or modelling a linear association when it is indeed a quadratic one (Shmueli, 2010). Low bias on the other hand indicates, that the model is a good approximation to "reality" and therefore generalizes well on new data. Low variance reflects accurate and therefore representative predictions.

**Overfitting** The parameters of a statistical model comprise both a *random effect* $\epsilon$ and the *systematic effect* which reflects the true association. A model "overfits", if its estimated parameters not only capture the *systematic effect*, but also (training) sample inherent irreducible error (a.k.a., "noise"). Such noise can not be found in other samples from the same population. In overfitted models, the prediction error is increased since it expects training data specific sample characteristics which do not exist in new (i.e., test) data from the same population. Consequently, overfitted models do not generalize well on new data. The risk for overfitting increases with model complexity, i.e., with increasing number of model parameters respective to the (trainings) sample size (e.g.; Van Calster et al., 2020).

**Bias-Variance Tradeoff** The bias-variance tradeoff describes the balance between overfitting and generalizability. Assuming there is some amount of irreducible error, a good model would explain a lot of variation within the training data and is generalizable on new data. Thus, there would be minimal residuals on training and minimal prediction error on testing data. As mentioned before, the chance of overfitting increases with model complexity. Simultaneously, we can not hope for a good model if we do not include all relevant predictors for a specific outcome. Consequently, we have to balance between a "good" approximation of reality (low bias) which goes along with the risk of overfitting, and generalizability in form of accurate and correct predictions on new data (Briscoe & Feldman, 2011; Neal et al., 2019). For the latter, previous literature states that the bias-variance tradeoff is true for kernel regression and splines, but is only partially applicable to some types of neural networks, which can both minimize bias and variance. Although, more research is needed in this area, Neal et al. (2019) also support the statement that this tradeoff applies in many cases. Building on Shmueli (2010), explanatory models aim for minimal bias in order to get the most precise and hence un-biased estimates for the theory represented by the model. In predictive modelling, the minimization of the prediction error on new data is the primary objective. Although both bias and variance are ought to be minimized here, sometimes intentional model misspecification (i.e., bias) is applied in order to have a lower prediction error (Shmueli, 2010).

### 2.2.3 Shrinkage: Regularization vs. Post-Estimation Shrinkage Factors

The prediction error on new test data is typically greater than the ”prediction” error on training data, primarily due to the potential for overfitting and the inherent nature of parameter estimation on samples. The phenomenon shrinkage occurs if there is a systematic ”over-estimation” bias in the predictions.

**The Phenomenon Shrinkage** According to Heinze et al. (2018), the phenomenon shrinkage occurs when the observed outcomes are closer to their mean than their model prediction. This happens if the regression coefficients are ”too optimistic/extreme” and thus overestimate the predictors' effect sizes. Ultimately, shrinkage aggravates due to overfitting. In order to counter such over-optimistic predictions and consequently an increased prediction error, one can multiply shrinkage factors to the parameters (e.g.; Sauerbrei et al., 2020; Van Calster et al., 2020; Heinze et al., 2018; Dunkler, Sauerbrei, & Heinze, 2016). The corresponding procedures are called shrinkage methods.

**Shrinkage Methods as a Countermeasure** Shrinkage methods counteract the phenomenon shrinkage by introducing a (small) bias towards zero on the regression coefficients which were estimated too optimistic due to the optimization procedure or overfitting. On average, the application of shrinkage factors leads to lower coefficient variability, lower prediction error, and hence better predictive performance (e.g.; Dunkler et al., 2016; Heinze et al., 2018; Sauerbrei et al., 2020; Van Calster et al., 2020).

There are different types of shrinkage factors which can be applied to a model. Dunkler et al. (2016) present global (i.e., uniform), parameterwise, and joint shrinkage factors. Uniform shrinkage calculates one shrinkage factor for all predictors in the model. Parameterwise shrinkage factors in contrast are estimated for and applied to each single parameter in the model. In case of semantic (i.e., theoretically) similar or even correlated covariates, parameterwise shrinkage might be inappropriate because it can treat similar covariates severely different. Dunkler et al. (2016) propose so-called joint shrinkage factors which combine the benefits of both global and parameterwise shrinkage factors: similar covariates receive uniform shrinkage, while dissimilar covariates receive parameterwise shrinkage. The authors suggest joint shrinkage factors for transformed covariates like splines or FP which consequently yield common information and are therefore correlated.

Since variable selection procedures also suffer from parameter overestimation, shrinkage procedures can further stabilize variable selection procedures, especially in low EPV settings (Heinze et al., 2018). Recent research suggests that (penalized likelihood) shrinkage procedures in low EPV settings partially apply too much shrinkage, increase parameter variance, and increase the prediction error (Van Calster et al., 2020). More details are discussed in section 2.3.

The next paragraphs cover two shrinkage methods. In general, shrinkage factors can either be determined by post-estimation or by penalized likelihood (i.e., regularization) procedures.

**Post-Estimation Shrinkage Factors (PESF)** As the name implies, post-estimation shrinkage factors (PESF) are determined after model fitting. Dunkler et al. (2016) summarized two methods for the estimation of either global, parameterwise, or joint PESF: leave-one-out cross-validation (i.e., jackknifing) or an approximation based on DFBETA residuals.

In order to convey the basic ideas behind the calculation of PESF with cross-validation, the three-step procedure for global PESF estimation of Verweij and Van Houwelingen (1993, in Dunkler et al., 2016) is presented. Further details on parameterwise and joint PESF estimation, as well as their approximation with DFBETA residuals can be found in Dunkler et al. (2016).

Assumed we have a sample size of $N$ observations and a regression model with covariates $\mathbf{X}$ and outcome $\mathbf{y}$. 1) the parameters $\boldsymbol{\beta}$ of the model get re-estimated $N$ times, where each time one individual observation $i$ gets excluded (i.e., "leave-one-out" cross-validation). This approach results in $N$ estimations for each parameter, where a single observation $i$ is excluded: namely $\hat{\boldsymbol{\beta}}^{(-i)}$ for $i = 1, ..., N$. 2) we can calculate the prediction for the excluded $i^{\text{th}}$ individual with its respective covariate row vector $\mathbf{x}_i$ in the form of $\hat{\eta}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}^{(-i)}$, that the $i^{\text{th}}$ individual had no influence on the parameter estimation $\hat{\boldsymbol{\beta}}^{(-i)}$. 3) a new regression model is introduced and fitted with the un-influenced predictions $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, ..., \hat{\eta}_i, ..., \hat{\eta}_N)$ as a single predictor (i.e., covariate) vector with its slope $c$ for the outcome $y$. The coefficient $c$ is then the global shrinkage factor which can be applied on the original model parameters. To calculate parameterwise or joint PESF, steps 2) and 3) have to be modified for each (set of) covariate(s). Such that the predictions in 2) are performed for each of the $K$ covariates $x_k$. Ultimately, the new model in 3) contains $K$ predictors $\hat{\eta}_k$ and their respective slopes $c_k$, which are in return the parameterwise or joint PESF.

Although the jackknifing approach yields precise PESF estimations, it is quite computational intensive. Especially, when re- or subsampling approaches are applied in combination with iterative fitting procedures like in GLMs or GAMs. The DFBETA procedure utilizes information matrices which are available in such iterative fitting procedures to approximate PESF. In small samples, where the single observations are more influential than in large samples, DFBETA approximations are slightly closer to 1 than their jackknifing counterparts. Consequently, DFBETA might yield underestimated PESF in small samples, but PESF converge in both procedures for bigger samples (Dunkler et al., 2016).

Alternatively to estimating shrinkage factors after model fitting, model fitting itself can be augmented in order to bias parameters towards zero. This can be achieved by penalized likelihood procedures.

**Penalized Likelihood: Ridge and Lasso Regression** Penalized likelihood procedures, which are also called regularization methods, shrink the size of the parameters within the maximum likelihood estimation via a penalization term.

For a given log-likelihood function $\ell(\boldsymbol{\theta})$ with $K$ parameters in $\boldsymbol{\theta}$, like the one for logistic regression in Eq. (10), a multiple $\lambda$ of the penalty term is subtracted from the "ordinary" log-likelihood of the parameters. If $\lambda = 0$, there is no penalty term and it is hence an "ordinary" maximum likelihood estimation. Equivalent to GAMs "wiggliness" factor in section 2.1.3, $\lambda \rightarrow \infty$ indicates infinite penalty which results in the arithmetic mean as the most "(log-) likely" model. $\lambda$ can be determined via cross-validation and grid-search (Van Calster et al., 2020).

The difference between Lasso and ridge regression comes down to the penalty term used: Lasso, the "least absolute shrinkage (and) selection operator", uses the L1 norm for penalization. Due summing over $K$ absolute parameters, Lasso estimation can set parameters exactly to zero which corresponds to variable selection. In contrast, ridge regression uses the L2 penalty which is the sum over the squared parameters. Ridge regression therefore penalizes extreme slopes more, but it can not set parameters exactly to zero. Therefore no variable selection is comprised in the "basic" ridge regression. The likelihood functions of both Lasso and ridge regression are presented in Eq. (19).

$$\begin{aligned} \ell(\boldsymbol{\theta})_{\text{Lasso}} &= \ell(\boldsymbol{\theta}) - \lambda \sum_{k=1}^{K} |\beta_k| \\ \ell(\boldsymbol{\theta})_{\text{ridge}} &= \ell(\boldsymbol{\theta}) - \lambda \sum_{k=1}^{K} \beta_k^2 \end{aligned} \tag{19}$$

In order to penalize all variables in the same fashion, standardizing the variables (e.g., by z-standardization) is necessary. Otherwise, variables with intrinsic higher values (e.g., weight in grams) would have more influence than variables with lower intrinsic values (e.g., weight in kilograms; Sauerbrei et al., 2020).

More details on Lasso and ridge, general penalized maximum likelihood estimation, as well as further developments like adaptive Lasso or Elastic Net can be found in Schaefer, Roi, and Wolfe (1984), Le Cessie and Houwelingen (1992), Hastie et al. (2009, chapter

3.4), Heinze et al. (2018), Van Calster et al. (2020), or Sauerbrei et al. (2020).

Further, Van Calster et al. (2020) found, that among other methods Lasso and ridge regression have the tendency to shrink the parameters too much on average. They propose that this potentially undesirable effect originates in the cross-validation grid-search of the hyperparameter $\lambda$, since the cross-validation is usually performed on a smaller sample than the one used for model fitting.

**Firth's Correction for Rare Events** Firth's procedure also introduces a penalty term to the maximum likelihood function $\ell(\boldsymbol{\theta})$, where $I(\boldsymbol{\theta})$ represents the Fisher information matrix (cf. Eq. (20); Firth, 1993; Van Calster et al., 2020). The primary application of Firth's regression is to counter perfect separation in logistic regression, but with its penalization term it also introduces shrinkage. Furthermore, it is also known to benefit a model if the prevalence of the outcome is low (e.g.; Puhr, Heinze, Nold, Lusa, & Geroldinger, 2017; Van Calster et al., 2020). Although the correction by Firth (1993) solves the problem in maximum likelihood estimation when perfect discrimination in the outcome is present, it leads to a bias in the predicted probability. Puhr et al. (2017) assume, that this bias is negligent with very low or very high prevalences in the outcome, but propose correction approaches for medium prevalences. Those modifications comprise a correction of the intercept or the inclusion of an additional covariate in order to further improve the predictions from a Firth's regression model.

$$\ell(\boldsymbol{\theta})_{\text{Firth}} = \ell(\boldsymbol{\theta}) + \frac{1}{2}\log|I(\boldsymbol{\theta})| \tag{20}$$

Van Calster et al. (2020) found that Firth's correction led to less shrinkage than other shrinkage procedures like Lasso and ridge regression. Further, Firth's correction was the only shrinkage approach which led to a reduction in variability but no "shrinkage overshoot".

While the first section 2.1 presented frameworks to model (non-) linear associations with any type of response variable from the exponential distribution, section 2.2 introduced the influence of sample size and model complexity on model quality. Since variable and function selection aim for building high quality models, the problems arising from overfitting exacerbate and severely harm model quality. Overfitting results in too optimistic parameter estimations, which in return can make covariates appear more important than they really are. As this apparent importance influences variable and function selection, the wrong conclusions might be derived, whereby the resulting model is mis-specified and consequently heavily biased.

Section 2.3 covers approaches and algorithms to perform variable and function selection, addresses arising problems, and examine model quality based on the previously discussed methods.

## 2.3 Variable & Function Selection

The selection of relevant variables and – in case of continuous variables – the selection of their respective functional forms is essential in model building (Sauerbrei et al., 2020). Over the last decades several procedures were introduced to determine the influence of variables based on a defined evidential level (e.g., the nominal significance level; Sauerbrei et al., 2020).

Digitalization and the resulting possibilities of data measurement and storage provides researchers in various fields to examine an increased data volume, far surpassing what was available a century ago. Consequently, variable selection gained popularity especially in observational studies in order to determine the importance of predictors and confounders (c.f. section 2.3.1). Nowadays, many statistical software packages provide variable selection procedures and therefore facilitate their application. This gives rise to several problems: ignorance regarding inflated nominal significance levels, variance and bias in combination with missing stability and sensitivity analyses, as well as general miscomprehension of the validity of variable selection procedures (e.g., univariate selection) lead to questionable results (e.g.; Heinze & Dunkler, 2017; Heinze et al., 2018; Sauerbrei et al., 2020).

Since there are many pitfalls in variable (and function) selection, the following section provides an overview of procedures which are currently deemed as state-of-the art (SOTA) and which are not recommended.

### 2.3.1 Domain Knowledge & Directed Acyclic Graphs (DAGs)

The most common variable selection procedure in research is probably the selection by background knowledge. The ultimate goal is to fully determine all key features with their (assumed) functional forms solely by domain knowledge. Usually, both a domain expert and a methodological affine person cooperate to integrate background knowledge in the study and analysis design. Heinze et al. (2018) describe a two-step process, where domain experts first pre-select some variables which are deemed as potentially influential. Then the methodologist guides a discussion about model specifications, such as variable types (i.e., discrete or continuous), functional forms, or (causal) dependencies. This process can be visually assisted by the use of directed acyclic graphs (DAGs). Finally, this process should lead to a (working) set of influential variables for a specific outcome which can be formalized in a *global model*.

**DAGs** In DAGs, causally associated variables are represented as vertices which are connected by directed edges (i.e., arrows) in the direction of the association. Three types of associated triplets can be differentiated: confounders, mediators and colliders (Rohrer, 2018; Heinze et al., 2018; Sauerbrei et al., 2020).

*Confounders* are variables which influence both a predictor as well as the outcome. Hence, their directed associations lead to both predictor and outcome; $predictor \longleftarrow confounder \longrightarrow outcome$. Confounders should be included in the model, since their influence can be accounted for and thus leads to adjusted effect estimates between predictor and outcome. Although including confounders in the model decreases bias, their inclusion might lead to increased variance in case of small effect sizes. Consequently, Heinze et al. (2018) propose that confounders should only be included if their effect on both predictor or outcome is big enough. The influence and therefore importance of confounders can be tested by change-in-estimate criteria (cf. s. 2.3.2).

*Mediators* are causally affected by the predictor, but are also the causal origin of the outcome. They ”explain” the association between the (original) predictor and the outcome; $predictor \longrightarrow mediator \longrightarrow outcome$. Mediators are of special interest in order to identify and dissolve e.g., spurious correlations, in mediation analyses. Apart from that, mediators should not be included in a *global model*, since the inclusion of both predictor and mediator does not explain more variance but leads to multicollinearity and hence model instability. Thus, if a mediator is mistakenly included in the model, associations may disappear unintentionally.

*Colliders* are variables which are both affected by predictor and outcome, although predictor and outcome are deemed unrelated; $predictor \longrightarrow collider \longleftarrow outcome$. Colliders should also be excluded from the model since their inclusion might introduce spurious associations between predictor and outcome.

In sum, variable (pre-) selection based on domain knowledge primarily aims to identify a (working) set of predictors and influential confounders for a specified outcome. In case of explanatory models (cf.; Shmueli, 2010), the formalisation of causal influences in order to avoid bias is a prerequisite and (visually) benefits from the usage of DAGs (Heinze et al., 2018). Generally, a holistic representation of a DAG with all causally related variables might be ambitious, if not even impossible in some (observational) research questions. Especially in research questions where due to unclear associations additional variable (and function) selection procedures are needed. Despite this limitation, the process of integrating assumptions and domain expertise in a DAG often leads to better understanding, and ultimately supports analyses.

From a hypothesis testing perspective, the background knowledge approach corresponds to the “typical” confirmatory hypothesis testing, if the selection criteria are based on a (strong) theory. In contrast, the “automated algorithms”, presented in the following, resemble exploratory hypothesis testing. Corresponding to exploratory analyses, ”exploratory variable selection” are accompanied by problems like inflated significance levels, variance and bias as mentioned earlier.

### 2.3.2 Selection Criteria for Selection Algorithms

In general, ”automated” variable (and function) selection algorithms need decision rules when to in-/exclude variables or their transformations. These decision rules can be assigned to either information, significance, or change-in-estimate criteria and are outlined in the following paragraphs.

**Significance Criteria** Significance criteria basically utilize the respective $p$-value of a hypothesis test to determine the in- or exclusion of a variable (or function transformation). Such that, the in-/ exclusion of a single variable can be tested via e.g., a likelihood ratio test for two competing models: One model includes the currently tested variable $x_k$ with its slope $\beta_k$, and the competing model assumes $\beta_k = 0$ and therefore excludes $x_k$. In case of iterative fitting procedures like in GLMs or GAMs, this hypothesis test can either be approximated by the score-test for forward selection or the Wald-test for backward elimination (cf. section 2.3.3; Heinze et al., 2018).

Typically, variable selection algorithms are performed on a set of variables where single ones are iteratively tested regarding their importance. Since this test procedure comprises a large number of tested models which are not pre-specified, the resulting $p$-values are generally inflated due to multiple testing. Consequently, those $p$-values are not only underestimated, but also solely reflect the relevance of a specific variable given the particular set of other adjusted variables within this model (Heinze et al., 2018).

**Information Criteria** In contrast to the hypothesis testing on a single variable with significance criteria, information criteria compare models of arbitrary sized differences. This comes in handy, if a single best model should be chosen from a set of competing models resulting from e.g., a best subset variable selection (cf. s. 2.3.3). If the model differences reflect the in- or exclusion of a single variable, information criteria can be directly translated to significance criteria i.e., $p$-values.

Generally speaking, the model fit increases with the number of included covariates regardless of whether those are truly associated with the outcome(cf. s. 2.2; or Heinze et al., 2018). Hence, ”bigger” models would be automatically preferred over smaller ones, even if they are probably overfitted or uninterpretable. Information criteria penalize model sizes. Consequently, only the inclusion of truly relevant variables in a model lead to an improvement in the information criteria. In case of the inclusion of unrelated and hence irrelevant variables, the information criteria worsens.

Two often applied information criteria are Akaike's information criterion (AIC) and Schwartz's bayesian information criterion (BIC). There exist different formulations of AIC and BIC (”smaller vs. bigger is better”). Regardless of their form, their penalty term is important for variable and/or model selection. For $K$ covariates and a sample size of $N$, the AIC adds a penalization of $2K$, and the BIC of $\log(N)K$ to a multiple of the model log-likelihood. Consequently, the BIC penalizes a big model (i.e., with many covariates) heavier than the AIC. The BIC was originally developed under the assumption that among all possible/detectable models, there exist the true data generating model. Hence, it is assumed that all causally influential variables are available for variable selection and the BIC therefore can detect a set of true predictors. Since this assumption might be unrealistic, especially in life sciences, information criteria are generally used to select single models from a set of competing ones Heinze et al. (2018).

While the AIC selects a large(r) number of (weak) predictors with increasing sample size, the BIC gets more and more strict. The corresponding $p$-values reflect this: As mentioned earlier, AIC and BIC can be translated to $p$-values for hierarchically nested models differing by one variable. Such that the AIC corresponds to the significance criterion of $p \leq .157$. And for the sample sizes of $N_1 = 100$ and $N_2 = 400$, the BIC corresponds to the significance criteria $p_1 = .032$ and $p_2 = .014$ (Sauerbrei et al., 2020). Consequently, the AIC is a more liberal and the BIC a conservative criterion for variable and/or model selection. In settings with $\text{EPV}_{\text{global}} \leq 100$, Heinze et al. (2018) recommends the application of the AIC or even corresponding higher significance criteria.

In sum, information criteria are primarily useful in selecting competing models, while single variable (and function) selection traces back to significance criteria.

**Change-In-Estimate Criteria** The change-in-estimate criteria examine the change of model parameter estimates when a single variable is in- or excluded. Such that, the inclusion of an important confounder leads to ”significantly” adjusted slopes of the other predictors in the model. ”Significance” of a change-in-estimate is defined as a bigger absolute change in any slope than a predefined threshold such as e.g., $\tau = 0.05$ . The change-in-estimate criterion therefore can support the detection of confounders (Heinze et al., 2018).

In a simulation study Dunkler, Plischke, Leffondré, and Heinze (2014) showed, that the exclusion of a truly relevant variable from a model led to a significant change-in-estimate. Vice versa, the exclusion of an irrelevant variable led to an insignificant change-in-estimate. The authors propose a combination of significance and change-in-estimate selection criteria to derive the augmented backward elimination algorithm which is presented in s. 2.3.3.

### 2.3.3 Variable Selection Algorithms

All presented variable (and function) selection algorithms in the following are based on an implementation of the previously discussed criteria. The lower the threshold of the criterion (i.e., $p \leq .05$ vs. $p \leq .01$), the fewer variabels are selected, and hence the smaller the model. Analogous to the "no free lunch" theorem (NFL; e.g., Wolpert & Macready, 1997), there is also no single best variable (and function) selection algorithm. Each has its advantages and limitations in a specific setting, given a specified task (e.g., creating a prediction model), the number of variables to choose from, their correlation structure, their effect sizes, and sample size (e.g.; Sauerbrei et al., 2020; Heinze et al., 2018). Regardless of the algorithm, variable (and function) selection is related to exploratory analyses, whereby the nominal significance level often can not be met, if not explicitely controlled for. The resulting uncertainty should be examined by accompanying stability analyses. Further, Heinze et al. (2018) recommend not to process "well-known" influential variables in the variable selection process, since the resulting instability of such approaches might result in the exclusion of such well-known variables. In order to test their functional form or their importance comparative to other selected models, one still can apply model comparisons based on information criteria. The algorithms presented in the following are applicable when limited or no previous knowledge is available.

**Univariate Selection** In this procedure, for each predictor candidate the bivariate association (e.g., correlation) to the outcome is calculated separately. All variables with significant results are deemed important and hence included in the "final" model. A common miscomprehension is, that a significant bivariate association is a necessary condition for being significant in an (adjusted) multivariate model. This is not the case, and adjusted effects can go in any direction. Consequently, univariate variable selection should be avoided, since it provides no relevant information about the importance of variables and therefore negatively affects model stability (Heinze & Dunkler, 2017; Heinze et al., 2018; Sauerbrei et al., 2020).

**Best Subset Selection** This is a brute force algorithm which creates all $2^K$ possible models for $K$ covariates. Then, these models can be compared regarding their AIC or BIC. This approach is only feasible for a small set of variables from which the best fitting set is ought to be selected. Although bigger variable sets are processable with modern computation capacity, the sheer number of resulting models increases the risk for selecting ”spurious predictors” (Sauerbrei et al., 2020). Although Sauerbrei et al. (2020) report of no known literature on implementing additional function selection in the best subset algorithms, the resulting number of competing models would multiply and hence exacerbate the depicted problem. Overall, these authors do not recommend best subset selection, since other selection procedures like penalized likelihood methods or backward elimination are available.

**Penalized Likelihood Selection** Penalized likelihood selection applies shrinkage methods which allow the penalty terms to result in slopes equal to zero like in Lasso (cf. Eq. 19). Since Lasso shrinks all variables equally, it tends to shrink influential variables more than less influential variables. The result is, that Lasso tends to include more variables with weak effects than other selection algorithms. Hence, Lasso in its original form rather corresponds to a variable ”screening” than variable selection procedure. Especially in existence of multicollinearity, Lasso fails to select all relevant (correlated) variables (Sauerbrei et al., 2020). To address these problems, advanced methods like adaptive Lasso or Elastic Net can be used. The former introduces weights according to the effect size of a variable and hence favors more influential variables by shrinking them less, while shrinking variables with smaller effects more strongly towards zero. Elastic Net both uses the L1 (from Lasso) and L2 (from ridge) penalization to reach the same goal.

Overall, (Sauerbrei et al., 2020; Heinze et al., 2018) recommend Lasso only in very low EPV settings (EPV $\leq 10$), since Lasso in its original form contains less parameters to estimate than its advancements. In higher EPV settings (EPV $> 10$), one can either apply the before mentioned more complex shrinkage methods or algorithms which were explicitly developed for variable selection.

**Backward Elimination/Forward Selection** Backward elimination (BE) and forward selection (FS) are two variable selection algorithms, which are typically implemented with a significance criterion.

**Forward Selection (FS)** starts with an empty model and then uses the score-test to select and include the variable with the highest bivariate fit (i.e., lowest $p$-value) which reaches the predefined significance criterion. In each further step, FS uses the score test to determine the significance of one to the model added variable, while simultaneously accounting for the already included variables. If one or more significant additions to the model exist FS ”greedily” selects the ”most significant” one (i.e., with the lowest $p$-value) and adds it to the model. If there are no significant candidate variables anymore, the algorithm stops.

**Backward Elimination (BE)** in contrast starts with a predefined *global model*. On each included variable the Wald-test is performed on. And among all insignificant variables, the least significant (i.e., highest $p$-value) gets excluded. Then, the resulting reduced model gets re-estimated and again, if there exist insignificant variables, the least significant gets excluded. BE terminates, if all included variables are significant.

In most research use-cases, Heinze et al. (2018) recommends BE over FS, since the starting *global model* already makes sense and adjusts for multiple effects. In high dimensional variable selection settings, no *global model* can be built. Then FS is still applicable. Both algorithms can be combined in different ways: if there exist one well-known *global model*, FS can be used to find further influential predictors among variables which are not included in the model yet. Another combination is the so-called "stepwise" procedure.

**Stepwise BE/FS** The stepwise procedure is a modification for FS and BE and introduces additional BE or FS steps respectively (Heinze et al., 2018).

Such that, after the FS inclusion of a new variable, all variables in the model are tested by a BE Wald-test to determine their significance. In case of insignificant ones, the least significant variable gets excluded. In the subsequent FS step, also previously excluded variables are again considered/ tested for inclusion.

Stepwise BE works vice versa: After each BE exclusion step, all not included variables are then score-tested for significance. Among all significant variables, the most significant (i.e., lowest $p$-value) gets again included to the model.

Both stepwise FS and BE terminate, if either a counter is reached (in case of in-/exclusion cycling) or all variables in the model are significant and all variables not in the model are insignificant.

**Augmented Backward Elimination** The augmented backward elimination (ABE) algorithm extends BE and its intrinsic significance criterion with an additional change-in-estimate criterion. Originally proposed as "purposeful selection" (Bursac, Gauss, Williams, & Hosmer, 2008; Hosmer, Lemeshow, & Sturdivant, 2013), Dunkler et al. (2014) modified the algorithm by standardizing the change-in-estimate criterion, such that it is independent from the variable scalings.

ABE performs classic BE steps with an additional test for a change-in-estimate after the exclusion of a variable. If the exclusion of a variable results in a change-of-estimate equal or bigger than a predefined threshold $\tau$, the (originally excluded) variable is deemed important to be controlled for and re-enters the model. ABE stops, if all variables in the model are either significant or their exclusion lead to a change-in-estimate equal or bigger than $\tau$ in any other included variable.

With a fixed significance criterion, the additional change-in-estimate criterion makes the ABE algorithm somewhat more liberal than the BE algorithm. Consequently, ABE tends to include more variables than BE and therefore to have less biased parameters. ABE and purposeful selection are reported to determine not only influential predictors, but also confounders (Dunkler et al., 2014; Heinze et al., 2018; Sauerbrei et al., 2020).

**Feature Ordering (by) Conditional Independence - FOCI** The FOCI algorithm by Azadkia, Chatterjee, and Matloff (2021) is a rank-based non-parametric forward selection algorithm. Thus, it starts with an empty model and includes in each step a single variable which maximizes its measure. FOCI tests the conditional (in-) dependence of two or more variables by a generalization of partial $R^2$ and stops when its measure achieves a non-positive value the first time.

Due to its non-parametric nature, FOCI has the advantage to also detect non-linear associations or even interaction terms. On the other hand, non-parametric measures potentially have lower statistical power as they do not use all (parametrically assumed) information available and thus benefit from bigger sample sizes than their parametric counterparts (e.g., Prajapati, Dunne, & Armstrong, 2010).

### 2.3.4 Function Selection Procedure

Similar to variable selection, function selection procedures aim for determining the importance of transformed continuous variables. Such that the model fit with a transformed covariate $f(x)$ is compared to its un-transformed $x$ and with a model, where $x$ is excluded.

**A priori assumed linearity** The most naive approach is, to simply assume linear associations between all variables and the outcome. Although this approach less likely results in overfitting, assuming solely linear associations is often an oversimplification and therefore results in increased bias instead. As always, such restrictions on the transformations (i.e., here identity transformations) have to be carefully discussed. Especially, in low EPV settings where even variable selection alone is deemed problematic, additional function selection with more complex transformations like FP or splines might exacerbate problems. Hence, assuming linearity might be the best approach possible in some settings, in order to reach somewhat stable results. A compromise approach would utilize often observed local linearity in generally non-linear associations by applying e.g., an interrupted regression. For a more detailed discussion, see Sauerbrei et al. (2020), Heinze et al. (2018), and Binder et al. (2013).

**Categorization** According to Sauerbrei et al. (2020), the categorization of continuous variables is a popular but highly problematic procedure for deriving easy interpretable results. In this approach, continuous data are dichotomized in two (or more) categories by one (or more) thresholds. It is arguably a very flexible approach to model non-linearity and receive easily interpretable results such as risk-ratios, Odds ratios, $\chi^2$ or mean comparisons. But the specifications of the thresholds themselves, as well as the underlying step function result in interpretation problems, inflated type I errors (if not corrected), and are generally accompanied by a loss of information. Specifying the threshold by background knowledge is somewhat arbitrary and debatable in the eye of domain experts. Deriving such thresholds from the data, a loss in generalizability is expectable due to the increased risk for overfitting and excessive type I error if not adjusted. Further, dichotomizing continuous data results in an information loss, since resulting bins can only approximate the true association. In case of a true polynomial association, the dichotomization into two groups does not capture the true association at all. But one usually does not know, how many groups/bins are necessary to minimize the loss in information and somewhat approximate the true association. Additional to that, the interpretation of such step-functions might also be problematic: if two very similar individuals are assigned two different bins, their odds for contrafactual outcomes might differ substantially despite the similarity of their covariate vectors.

Overall, Sauerbrei et al. (2020) strongly recommend not to apply categorization if possible and instead treating and modelling continuous data as such.

**FP & Splines** As introduced in the GAM section 2.1.3, fractional polynomial (FP) and spline transformations can be applied to approximate non-linear associations. These procedures have been combined with existing variable selection algorithms like backward elimination (BE) to select not only influential variables, but also their most influential transformation.

In the **multivariable fractional polynomial (MFP)** approach, the BE algorithm is extended with significance tests for FP terms of predefined complexity. Hence, before performing MFP one has to specify the most complex FP model allowed. Basically, each continuous covariate is therefore not only tested regarding the significance of the linear association, but also regarding the significance of its non-linear FP1, FP2, ... association. The best fitting transformation is then kept in the model. The MFP implements a closed test procedure in order to control for a inflated type I error due to multiple testing. A detailed depiction of this algorithm can be found in Binder et al. (2013) and Royston and Sauerbrei (2008, chpt. 6). Analogous to the MFP, Royston and Sauerbrei also provide similarly working variable and function selection algorithms for cubic regression splines in the **multivariable regression spline (MVRS)** and for cubic smoothing splines in the **multivariable smoothing splines (MVSS)** algorithms (Royston & Sauerbrei, 2008, 2007).

There also exist several other function selection approaches, which are mentioned in Sauerbrei et al. (2020) but are out of the scope of this thesis. Generally, there is few literature comparing these approaches, especially MFP with its spline based alternatives. For instance, Binder et al. (2013) compared MFP with algorithms implementing restricted cubic splines or penalized splines. Results indicate, that low sample sizes (i.e., low EPV rates) and very noisy data (i.e., $R^2 \leq .2$) generally result in bad models regarding variable and function selection, irrespective of model complexity (i.e., the type of non-linear modelling). In such settings, linear models performed equally bad and can therefore be preferred. In settings with a medium amount of information available (i.e., $R^2 \approx .5$), MFP results in better performing models than its spline alternatives. This might be due to the smaller number of parameters which have to be estimated and thus less risk of overfitting for FP, in comparison to spline alternatives. Simultaneously, splines are more flexible and therefore can model associations FP2 (and FP1) can not. But when a large amount of information is available (i.e., $R^2 \approx .8$) all non-linear approaches perform equally well in the simulation study of Binder et al. (2013). Ultimately, these authors favor MFP over its spline alternatives, since FP are easier to implement, interpret, and produced better results with a limited amount of information available.

### 2.3.5 SOTA: Recommendations for Variable (and Function-) Selection

Different types of variable and function selection approaches were proposed in the last section. The preferred use-cases depend on the modelling approach, the Events-Per-Variable (EPV) rate, pre-existing domain knowledge, as well as the correlation structure among and the effect sizes of predictor candidates, and therefore the information available (i.e., how noisy the data are) (Sauerbrei et al., 2020; Heinze et al., 2018; Courvoisier et al., 2011; Shmueli, 2010).

Generally, in Heinze et al. (2018) it is argued that BE is the best variable selection procedure, since BE requires a predefined *global model* which is assumed to be unbiased. Results from BE resemble those based on AIC and BIC procedures (with comparable significance levels), BE is easy to use and extendable for finding confounders and performing additional function selection depending on your own demands (Sauerbrei et al., 2020; Heinze et al., 2018; Royston & Sauerbrei, 2008). No explicit recommendations were found regarding stepwise extensions of FS and BE. Due to the nature of the selection procedure, stepwise BE might lead to bigger models than BE, and stepwise FS to smaller models than FS. Although (stepwise) FS is generally not recommended, contrary to BE it is still applicable on very high dimensional data. Although Lasso in its original form comes with some disadvantages, it is an adequate "screening" tool and possess adaptations (e.g., Elastic Net) that are better suited for variable selection. Univariate selection is generally not recommended.

When a *global model* is available and $\text{EPV}_{\text{global}} \leq 10$, Heinze et al. (2018) advise against any type of variable selection. Instead, they suggest to perform shrinkage methods like ridge regression on the global model and emphasize caution in the interpretation of the model. When $10 < \text{EPV}_{\text{global}} \leq 25$, variable selection on predictors with unknown associations should either be performed with penalized likelihood methods such as Lasso, or algorithms like e.g., ABE should be accompanied by PESF (cf. Dunkler et al., 2016, 2014). Neither Sauerbrei et al. (2020), nor Heinze et al. (2018) have mentioned a clear recommendation for function selection. But since Binder et al. (2013) found little benefit of function selection in low EPV and noisy settings, I would argue to perform additional function selection only in settings with $\text{EPV} > 100$, where model selection can be accompanied by the usage of BIC. Ultimately, Heinze et al. (2018) emphasizes, that variable selection should only be performed on a set of variables with unknown effects - hence, not performed on variables with known strong effects. Since EPV are only a heuristic, the recommendations from Heinze et al. (2018) can be loosened when the predictor candidates are (mostly) independent from each other. If the predictor candidates are correlated (e.g., by introducing interaction terms), the EPV heuristic need to be tightened. Only in settings with $\text{EPV} \gg 50$, the authors start trusting in the asymptotic properties of selection algorithms, where less caution is needed when interpreting the results of variable selection algorithms.

In either case and setting, Sauerbrei et al. (2020) and Heinze et al. (2018) strongly recommend to perform stability analyses whenever variable (and function) selection is performed.

### 2.3.6 Stability Investigations

Due to the exploratory character of variable (and function) selection procedures, bias and uncertainty are introduced. The resulting stability issues should therefore be examined by re- or subsampling approaches (Heinze et al., 2018).

**Resampling: Bootstrapping** The non parametric bootstrap draws $l$ samples of size $N$ with replacement from the original sample. Hence, single observations can occur multiple times in a single the bootstrap sample. On the long run such as for e.g., $l =$

1000 bootstrap samples, this method reduces the influence of outliers, since frequently occurring ”average” observations are more likely to be drawn than rare outliers.

**Subsampling: Crossvalidation** In crossvalidation, the data set is randomly divided in a training and a testing data set at a fixed partition rate. Thus, a single duplet of training and test data each contain unique observations and are non-overlapping. Similar to bootstrapping, this procedure also can be repeated for e.g., $l = 1000$ times. While training data are used for model fitting, testing data can be used to validate the model performance (e.g., the prediction error) on new data. The splitting procedure can be modified so that the resulting data sets meet some pre-specified properties, such as equal size or fixed prevalence of a binary outcome.

Usually, several hundreds to thousands of subsets or resamples are drawn at random, whereby ”natural” variation in the resulting bootstrap, or training and test data can be simulated. By estimating the model or performing variable (and function-) selection in each of those several hundreds to thousand subsets, one gets a distribution of estimates for each model parameter. From those distributions robust measures such as the median estimate or the 2.5 and 97.5 percentiles corresponding to a 95% confidence interval ($\mathrm{CI}_{95\%}$) can be derived.

In combination with post-estimation shrinkage factors (PESF), those sub- or resampling distributions of the estimates can be used to examine the bias and variance of a model, when comparing them to the (unbiased) parameter estimates from the *global model*. Further, re- and subsampling variable inclusion frequencies or model selection frequencies can also be tracked as a measure for (un-)certainty of the selected set of variables, functions and thus models. In case of correlated predictors, Wallisch, Dunkler, Rauch, Bin, and Heinze (2020) recommend the application of variable inclusion and model selection frequencies within a subsampling approach, such as $l = 1000$ crossvalidation runs.

Over the last sections, a vast foundation about regression modelling, variable and function selection was laid. With this theoretical knowledge, the research question and its challenges are presented and discussed in the next section. Ultimately, the analysis plan is derived.

# 3 Practical Research Question: Descriptive Model Building - Identifying Risk-Factors for Dysphagia

## 3.1 Overview

Dysphagia, a swallowing dysfunction, negatively affects life quality, increases the risk of pulmonary complications, and ultimately leads to mortality after a preceding stroke or traumatic cervical spinal cord injury (SCI; Wolf & Meiners, 2003; Chaw, Shem, Castillo, Wong, & Chang, 2012; Hayashi et al., 2017; Iruthayarajah et al., 2018; Shem, Wong, Dirlikov, & Castillo, 2019). For individuals sustaining a cervical SCI, dysphagia, despite its severe primary and secondary complications, is still under-diagnosed, and no specific treatment has been established yet (Hayashi et al., 2020). Early diagnoses are a necessary prerequisite for adequate patient-centered treatment by speech pathologists in order to at least prevent secondary complications like aspiration pneumonia (Chaw et al., 2012).

Determining influential risk and protective factors would inform clinical decision-making and therefore benefit patients' health and recovery. But due to methodological weaknesses and small sample sizes in previous studies, the prevalence as well as such risk and protective factors are yet not fully determined or still critically discussed. Based on the descriptive modelling framework of Shmueli (2010), this thesis aims for determining such influential variables for the binary response dysphagia. Variable selection is performed on a dataset consisting of $N = 403$ patients with cervical SCI, of whom some developed dysphagia in the course of their treatment at *BG Trauma Center Murnau*. In cooperation with the *Spinal Cord Injury Center of BG Trauma Center Murnau* in Germany, the *Institute of Molecular Regenerative Medicine* and the *Spinal Cord Injury and Tissue Regeneration Center Salzburg (SCI-TReCS)* of the Paracelsus Medical University, as well as the *Department of Artificial Intelligence and Human Interfaces* of the Paris Lodron University Salzburg in Austria, this thesis complements the research paper of Meissner et al. (2023).

Based on the approach presented in section 2.3.1, experts in the fields of speech and language therapy, trauma surgery, and medical science, as well as myself identified 7 potential predictor candidates for dysphagia which are incorporated in a *global model* (Eq. (21)). The assumed associations of this domain knowledge based model are depicted in the DAG (cf. s. 2.3.1) in Fig. 1. The variable descriptions are presented in a list below. These predictor candidates belong to a large pool of measurements. But due to the low dysphagia prevalence, the resulting $\text{EPV}_{\text{global}} \approx 12$ for those 7 predictor candidates is already on the lower border, for which Heinze et al. (2018) deem variable selection still as feasible. In the following, the *global model* and its variables are shortly introduced.

## 3.2 Building a Global Model for Dysphagia

The *global model* is formally expressed in Eq. (21) and based on the DAG in Fig. 1. Here one can see, that a high neurological level of injury (NLI.High) is an assumed confounder, both influencing the predictor tracheostomy and the outcome dysphagia positively. Among the predictors, the total motor score (TMS) is assumed to be negatively associated with dysphagia, while age is assumed to be positively related. The type of surgery for tracheotomized patients (Tracheostomy:T.Surgery), as well as the number of fusioned segments (OP:No.Fusions) and the surgical access (OP:OP.Ventral) among operated individuals are defined as moderating variables. These are assumed to strengthen the association between the outcome and the main effect of receiving a tracheostomy or a surgery (OP).

Here, a higher number of fusioned vertebrae and a frontally (i.e., ventrally) accessed surgery, as well as a surgical tracheostomy are expected to result in higher odds for developing dysphagia. Those moderators are causally related to their main effect counterpart. The type of surgery for tracheostomy (T.Surgery) only matters if the patient receives a tracheostomy at all. Similarly, the surgical access (OP.Ventral) and the number of fusioned vertebrae (No.Fusions) only matter in patients who had undergone surgery (OP). Consequently, to model both not-operated as well as different types of operated patients, one needs those interaction terms to prevent missing data (i.e., *NA*s).

$$\begin{aligned} \text{Dysphagia} \sim \text{TMS} + \text{NLI.High} + \text{Age} + [\text{OP:OP.Ventral} + \text{OP:No.Fusions}] + \\ [\text{Tracheostomy} + \text{Tracheostomy:T.Surgery}] \end{aligned} \tag{21}$$

### Variable Description

- **Dysphagia**: a swallowing disorder; the binary response variable
- **TMS**: the (continuous) Total Motor Score (TMS) of the International Standards for Classification of Spinal Cord Injury Motor Score (ISNCSCI); used as an inverted proxy for injury severity (i.e., higher score = less severe injury)
- **NLI.High**: dichotomized Neurological Level of Injury (NLI); NLI.High = 1 indicates a high neurological level of injury on the vertebrae C1-C4, NLI.High = 0 a lower neurological level of injury (i.e., on C5 and lower)
- **Age**: the patient's age at the time of the accident

- **OP**: binary indicator if patient had a surgery (i.e., OP = 1: ”yes”)
- **OP.Ventral**: dichotomized surgery access variable; OP.Ventral = 1 indicates either solely ventrally accessed surgery or a combined ventral and dorsal access, OP.Ventral = 0 indicates solely dorsal access
- **No.Fusions**: the (continuous) number of a patient’s vertebrae which were surgically fused together
- **Tracheostomy**: binary indicator if a patient received a tracheostomy (i.e., Tracheostomy = 1: ”yes”)
- **T.Surgery**: binary indicator how the tracheostomy was applied; T.Surgery = 1 indicates a surgical tracheostomy, while T.Surgery = 0 indicates a dilatative tracheostomy

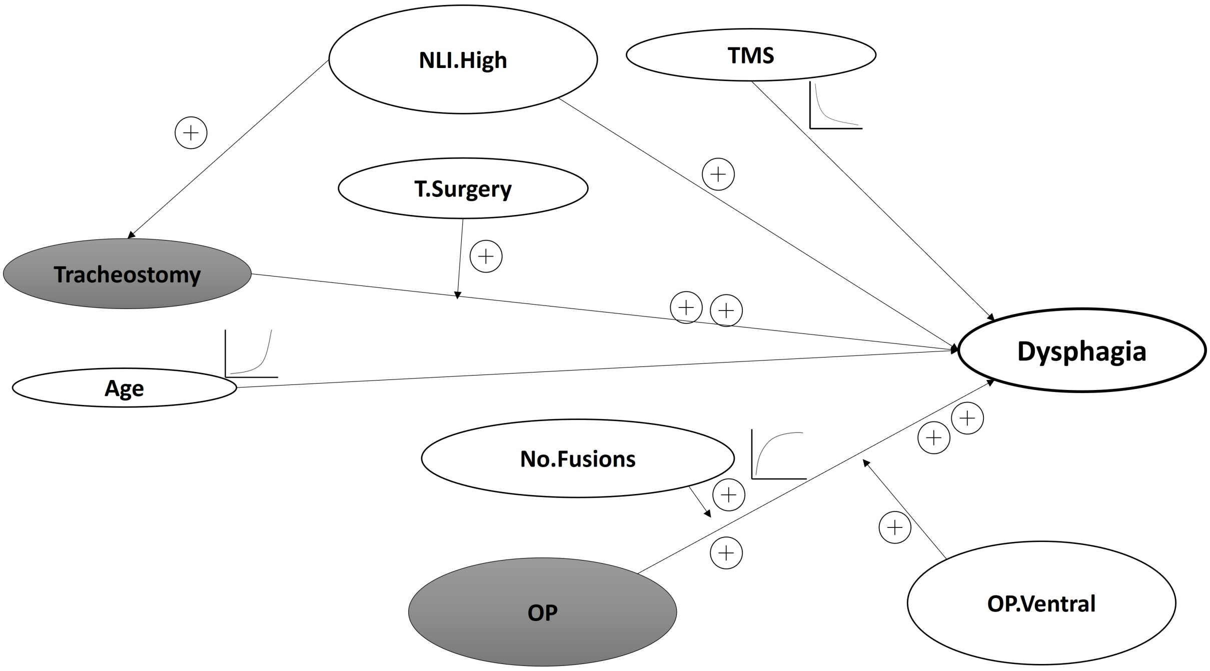

Figure 1: Identified predictor candidates by domain experts. In grey: main effects for the moderation terms. The assumed moderators (i.e., T.Surgery, No.Fusions, & OP.Ventral) are conditional to their main effect variables, as they represent subsets of patients with a specific treatment. The signs in circles indicate the assumed relationships between the binary predictor candidates and the response dysphagia. The assumed relationships of the continuous predictor candidates are depicted by a small mock-up graph.

Those interaction terms and their main effects can be interpreted as different subgroups of patients. For the interaction doublet of the binary tracheostomy and T.Surgery, there exist the "no tracheostomy" (Tracheostomy = 0, T.Surgery = NA), the "dilatative tracheostomy" (Tracheostomy = 1, T.Surgery = 0) and the "surgical tracheostomy" (Tracheostomy = 1, T.Surgery = 1) groups. Analogously, there are the "no surgery"(OP = 0, OP.Ventral = NA), the "dorsal surgery" (OP = 1, OP.Ventral = 0) and the "ventral surgery" (OP = 1, OP.Ventral = 1) groups.

Since there are severely different group sizes between (not) operated individuals stratified for dysphagia and the $\text{EPV}_{\text{global}}$ rate is already very low, the experts chose to exclude the main effect OP and solely include the interaction terms OP:OP.Ventral and OP:No.Fusions. Consequently, the "no surgery" and "dorsal surgery" groups are combined to a single "no or dorsal surgery" group, and operated individuals without fusioned segments are not explicitly modelled. Although $\sim 26\%$ of observations are affected by this, the more balanced group sizes and therefore increased stability outweigh the information loss. The same reasoning was applied to the commonly modeled (confounding) variable sex, since it also was heavily imbalanced and additionally suffers from a potential sampling bias (e.g., men tend to show more risky behavior which increases the risk for cervical SCI).

In sum: due to the low prevalence of dysphagia cases and thus low $\text{EPV}_{\text{global}}$, dependencies among predictors, and imbalanced patient groups, the *global model* was pruned to comprise 6 predictors and 1 confounder for the binary response. As $\text{EPV}_{\text{global}} \approx 12$, the amount of information available is probably low or medium at best. Although, these factors do not provide a good foundation for model building, influential variables for dysphagia still have to be found in order to improve treatment of affected patients. Thus, variable selection but no function selection should be performed on this *global model* (Heinze et al., 2018; Sauerbrei et al., 2020). Further, due to the low to medium information available and the assumed monotonic relationships, linear models should suffice for approximating the influence of the continuous predictor candidates (Binder et al., 2013).

The low $\text{EPV}_{\text{global}}$, as well as the dependencies among predictors make variable selection prone to errors. Consequently, stability analyses are absolutely necessary. Further, the dependencies introduced by interactions and the potentially low amount of information available increase the risk of overfitting. Thus, shrinkage methods should be applied to account for these tendencies. In order to not only test this *global model* but also to identify the influence of those pre-selected variables, variable selection algorithms such as (augmented) BE or FS by FOCI should be applied to maximize the knowledge gain about risk and protective factors for dysphagia.

In the next section, the sample available will be presented and the analysis plan meeting all previously mentioned conditions is derived.

# 4 Methods

## 4.1 Sample

For this thesis, the data from the observational study of Meissner et al. (2023) were used. There were 407 datasets from patients with cervical SCI who were treated in the level I trauma center *BG Trauma Center Murnau* in Germany. All those patients agreed to participate in the *European Multicenter Study about Spinal Cord Injury (EMSCI)* registry.

*BG Trauma Center Murnau* has implemented a standardized assessment of swallowing function in 2013. Therefore, routinely collected data from patients who were treated with acute SCI between 2013 and 2022 were included. Patients who sustained traumatic brain injury, and patients with a previously diagnosed dysphagia due to pre-existing neurological diseases were excluded. In this study cohort, dysphagia was diagnosed if either clinical examination using Daniel's clinical screening assessment 21 , fiber optic endoscopic evaluation of swallowing (FEES), or videofluoroscopy (VFSS) by speech and language therapists indicated the disease.

Of the original 407 observations, 4 had to be excluded due to missing values in the variables included in the *global model* (cf. Eq. (21)). Ultimately, a data set of $N = 403$ observations was available, and contained $n_{\text{dysphagia}} = 87(21.59\%)$ dysphagia cases. Thus, the 7 predictor candidates of the *global model* result in an $\text{EPV}_{\text{global}} = 12.43$.

## 4.2 Analysis Plan

In order to model the (non-)manifestation of dysphagia among patients with cervical SCI, the analysis plan in Fig. 2 is split into three stages. Stage A) covers the prevalence estimation of dysphagia and the sample partitioning process for the crossvalidation. Within each cross-validation run, B) comprises model building with shrinkage methods and variable selection algorithms and is accompanied by performance measurements and stability analyses. Finally, in C) earlier derived models are compared against each other and the best one(s) get(s) selected.

### 4.2.1 A) Estimating the Prevalence of Dysphagia & Data Partition

Due to variable selection and the low $\text{EPV}_{\text{global}}$, stability analyses are generally indicated (Heinze et al., 2018). Since the interaction terms violate the assumption of independence among the predictors of our *global model*, the stability analyses were applied via subsampling in form of a 1000 crossvalidations as suggested by Wallisch et al. (2020). The $l = 1000$ is an arbitrary value, which is assumed to be high enough to provide robust estimates. The overall dataset with $N = 403$ is partitioned into training and test data for each crossvalidation run. Such that $\text{EPV}_{\text{global training}} > 10$ and both data sets have approximately the dysphagia prevalence estimated by the median of 1000 bootstrap samples. Fixing the prevalence in each sample is important, since different prevalences in training and test data lead to biased estimates and thus to a loss in model accuracy. Further, if one does not fix the prevalence in the $l = 1000$ randomly drawn training subsamples, the occurrence of dysphagia free samples is possible which ultimately terminates the fitting procedure and therefore artificially limits the crossvalidation procedure.

### 4.2.2 B) Modelling & Variable Selection

Model building and variable selection is performed on each of the $l = 1000$ crossvalidation training samples. Thus, each parameter estimate of a regression model receives an individual subsampling distribution. Simultaneously, variable inclusion as well as model selection frequencies can be derived from the (automated) variable selection procedures which ultimately serve as the stability analyses.

**Knowledge Based Model Building (left path):** With regard to the regression modelling, the *global model* is implemented in two types of purely linear logistic regression models. Due to the low EPV, the assumed monotonic relationships (cf. Fig. 1), and the potentially noisy data, such linearity assumptions should be a sufficient approximation to the true functional forms (Binder et al., 2013; Heinze et al., 2018; Sauerbrei et al., 2020).

One model type applies a "naive" (i.e., normal) logistic regression (NLR), because it is the available and easy to use standard approach to model a binary response. This approach is accompanied by the application of post-estimation shrinkage factors (PESF; Dunkler et al., 2016) and thus results in three comparative models: "raw" NLR, a NLR model with a global PESF (g-NLR), and a NLR model with joint PESF (j-NLR) corresponding to the parenthesis grouping in Eq. (21).

The other model type is a logistic regression with Firth's correction (FLR) which belongs to the penalized likelihood methods and therefore intrinsically applies shrinkage to its parameter estimates (Van Calster et al., 2020). Usually, Firth's correction is used to circumvent problems arising from perfectly separable (i.e., predictable) data which are often associated with rare-event outcomes (Firth, 1993; Puhr et al., 2017). Although it is debatable, if the manifestation dysphagia in cervical SCI patients with $p_{\text{dysphagia}} \approx 22\%$

counts as a rare-event, randomly drawing $l = 1000$ training samples is possible to result in perfect separable data for fitting. Further, since variable selection should be performed, ridge regression is not indicated and Lasso solely "scans" the variables (Heinze et al., 2018). Another contraindication is that Lasso introduces more if not even too much shrinkage on influential predictor candidates and correspondingly less shrinkage on uninfluential variables, wherefore adaptive Lasso or Elastic-Net would be better alternatives (Van Calster et al., 2020; Sauerbrei et al., 2020). But in contrast to Firth's correction, both Elastic-Net and adaptive Lasso further introduce (hyper-)parameters, which again reduce the already low EPV if adequately accounted for.

In sum, model building results in four comparative regression models (i.e., FLR, "raw" NLR, g-NLR, j-NLR), their performance measures on training and test data, as well as the respective estimate distributions for each model parameter. From these distributions the subsampling median and the subsampling 95% CI (i.e., $\text{CI}_{\text{SS.95\%}}$) are used as robust measures for the importance of the predictor candidates in the *global model*.

**"Automated" Variable Selection (right path):** Complementary to the before mentioned theory based model building, "automated" variable selection algorithms are applied with a significance criterion of $\alpha = .157$. Based on Heinze et al. (2018), Backward Elimination (BE, s. 2.3.3) is preferred over most other variable selection approaches. By excluding insignificant predictor candidates from the model, shrinkage is introduced by setting their coefficients to zero. Thus, the "raw" NLR as well as the FLR model are subjected to the BE algorithm. Since confounding variables are influential on both other predictors and the outcome, controlling solely for significance might introduce too much shrinkage and therefore bias. Hence, augmented BE (ABE; cf. s. 2.3.3; Dunkler et al., 2014) with a change-in-estimate threshold of $\tau = .05$ is additionally subjected to the "raw" NLR model, but due to availability problems in the respective R package not the FLR model. Such availability problems are the same reason that the well-known risk-factor age is processed in the variable selection algorithms, except for the ABE. In the latter, age is specified as a "passive" predictor candidate and thus is always included in the ABE-NLR model (Dunkler et al., 2014). Although, the inclusion of well-known predictors in the variable selection process is not recommended (Heinze et al., 2018), the model comparison after the earlier described subsampling approach allows for a stability investigation between the ABE-NLR model where age is surely included and the BE-NLR, BE-FLR, and FOCI model, where age is only selected if enough support is provided.

Also the new Feature Ordering (by) Conditional Independence (FOCI; Azadkia & Chatterjee, 2021) algorithm is used for variable selection. As a forward selection algorithm, FOCI starts from an empty model and selects the most influential variables from an enlarged pool of predictor candidates, which can be found in appendix A. This pool of variables was pre-defined by earlier mentioned experts and lists variables which are at least temporally associated with the (non-)manifestation of dysphagia. In contrast to knowledge based BE and ABE approaches, FOCI is independent of the *global model*, assumption free (i.e., indifferent to functional forms), and therefore represents a new comparative and non-parametric variable selection algorithm (Azadkia et al., 2021).

Ultimately, the "automated" variable selection procedures results in a FOCI model, an ABE reduced NLR model (ABE-NLR), a BE reduced NLR model (BE-NLR) and a BE reduced FLR model (BE-FLR) for each of the $l = 1000$ crossvalidations. Variable inclusion and model selection frequencies are tracked for each of those four variable selection models and are used as stability analyses in part C) for model selection and comparison.

Such that the variable inclusion frequencies of FOCI and all *global model* based variable selection algorithms are presented stratified and summarized. Model selection frequencies are presented for the top 3 models of each variable selection approach.

### 4.2.3 C) Model Selection & Comparison

Ultimately, the in B) derived and presented median coefficients and their $\mathrm{CI}_{\mathrm{SS.95\%}}$ of the four theory based models (i.e., "raw", NLR, g-NLR, j-NLR, & FLR), as well as the "automated" variable selection models (i.e., ABE-NLR, BE-NLR, BE-FLR, & FOCI) are used to derive one or multiple comparative "final" models. In case of multiple comparative models, the best fitting one is determined by a $\chi^2$ test regarding their deviance.

## 4.3 Software

All statistical analyses were conducted in R (R Core Team, 2023, version 4.2.3; running under: Windows 10 x64). Following packages were used: *abe* (Blagus & Babic, 2017), *data.table* (Dowle et al., 2023), *effectsize* (Ben-Shachar et al., 2023), *FOCI* (Azadkia et al., 2021), *ggmosaic* (Jeppson, Hofmann, Cook, & Wickham, 2021), *gridExtra* (Auguie & Antonov, 2017), *logistf* (Heinze, Ploner, Dunkler, Southworth, & Jiricka, 2022), *pacman* (Rinker et al., 2019), *psych* (Revelle, 2023), *qad* (Kasper, Griessenberger, Junker, Petzel, & Trutschnig, 2022), *shrink* (Dunkler & Heinze, 2022), and *tidyverse* (Wickham & RStudio, 2023).

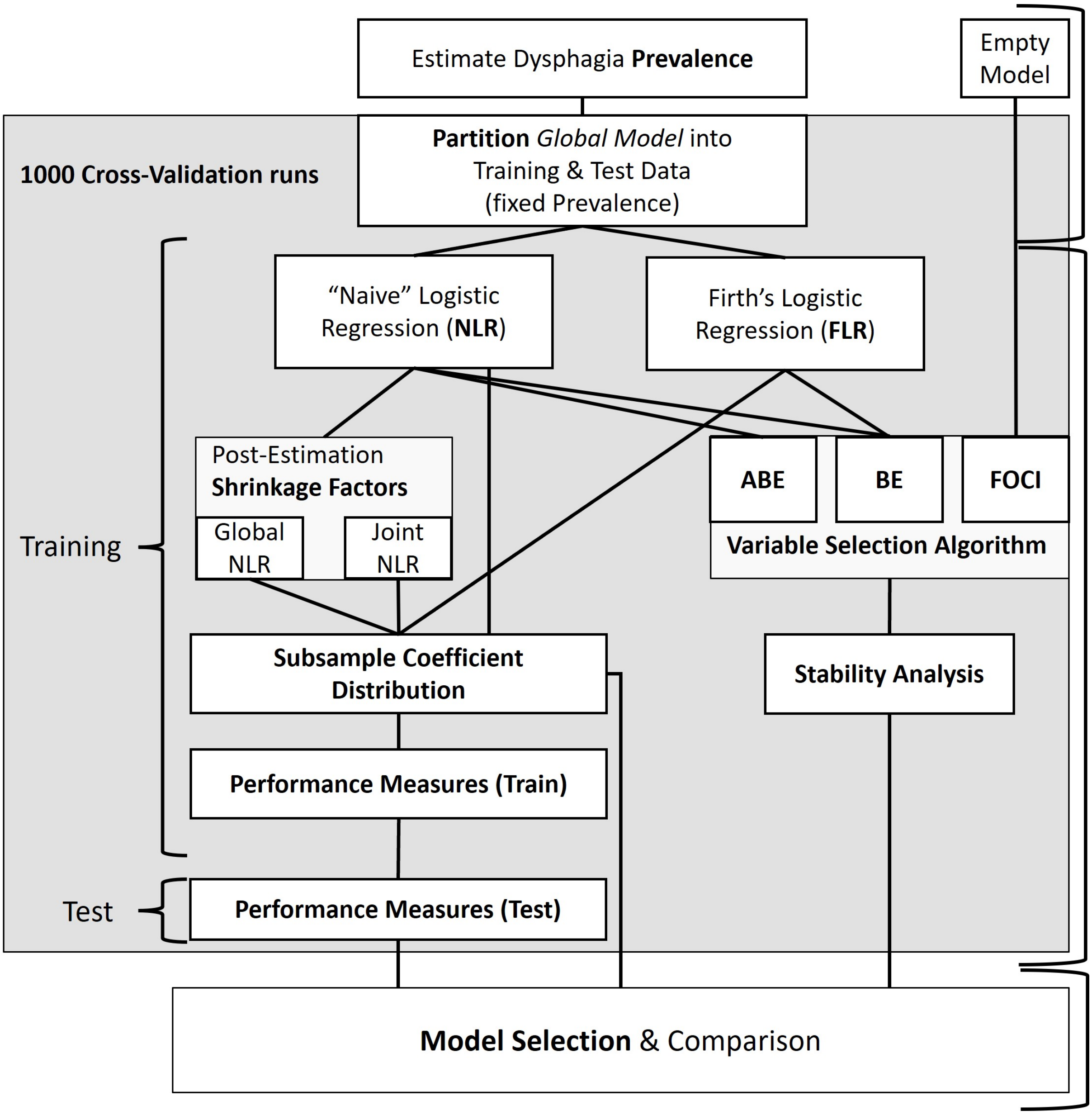
Estimate Dysphagia **Prevalence**
Empty Model
**1000 Cross-Validation runs**
**Partition** *Global Model* into Training & Test Data (fixed Prevalence)
"Naive" Logistic Regression (**NLR**)
Firth's Logistic Regression (**FLR**)
Post-Estimation **Shrinkage Factors**
Global NLR
Joint NLR
ABE
BE
FOCI
Variable Selection Algorithm
Training
Subsample Coefficient Distribution
Stability Analysis
Performance Measures (Train)
Test
Performance Measures (Test)
**Model Selection** & Comparison

Figure 2: This figure shows the analysis plan regarding variable and model selection. Within the variable selection part, BE corresponds to the "normal" Backward Elimination, ABE to the augmented BE, and FOCI to the Feature Ordering by Conditional Independence algorithm.

| **Variables** | **Mean** | **SD** | **Median** | **Minimum** | **Maximum** | **IQR** |
|---|---|---|---|---|---|---|
| Dysphagia* | 0.22 | 0.41 | 0 | 0 | 1 | 0 |
| TMS | 47.42 | 32.83 | 46 | 0 | 100 | 61.5 |
| NLI.High* | 0.59 | 0.49 | 1 | 0 | 1 | 1 |
| Age | 52.84 | 19.52 | 54 | 11 | 87 | 31 |
| OP:OP.Ventral* | 0.74 | 0.44 | 1 | 0 | 1 | 1 |
| OP:No.Fusions | 2.17 | 1.79 | 2 | 0 | 10 | 2 |
| Tracheostomy* | 0.45 | 0.5 | 0 | 0 | 1 | 1 |
| Tracheostomy:T.Surgery* | 0.13 | 0.34 | 0 | 0 | 1 | 0 |

Table 1: Descriptive statistics for the variables of the *global model*. Binary variables are marked by an asterisk *, for which the arithmetic mean represents the probability of the event 1.

# 5 Results

## 5.1 Descriptives & Bivariate Relationships

Within the sample of $N = 403$, there were 330 (81.89%) men and 73 (18.11%) women. Altogether, 87 (21.59%) patients had dysphagia, of whom 76 were males and 11 females. The youngest patient with a cervical SCI was 11 years old and the oldest 87. On average the patients were 52.84 ($\mathrm{Q}_1 = 38, \mathrm{Mdn} = 54, \mathrm{Q}_3 = 69$) years old. Of all 363 patients who received a surgery, 65 were operated dorsally, 210 ventrally, and 88 both dorsally and ventrally. Further descriptives can be found in Tab. 1.

The bivariate dependency plot in Fig. 3 shows different measures adjusted for the types of variables used. Histograms are shown on the diagonal. In case of two binary variables a mosaic plot is accompanied by the $p$-value of a $\chi^2$ test, the Odds Ratio (OR) and an effect transformation to Pearson's $r$ (i.e., $r^*$ with the asterisk indicating the approximation by an effectsize transformation). In case of one binary and one continuous variable, the binary one is used as a grouping variable and the resulting violin and box plots are accompanied by Cohen's $d$, the $p$-value for the Man-Whitney $U$ test and both the effectsize transformation to OR$^*$ and $r^*$. Ultimately for two continuous variables, the maximum non-parametric and asymmetric $q$ measure and its adjusted $p$-value are reported (Junker, Griessenberger, & Trutschnig, 2019). The variables are assigned to the x and y axis, such that the $q$ measure is maximized and thus that the x variable explains variable y better than vice versa. In addition, also Pearson's $r$, its respective $p$-value, and its effectsize transformation OR$^*$ are reported.

The first line in Fig. 3 shows the bivariate associations between the predictor candidates and the outcome. Of those only one, namely the number of fusioned vertebrae (OP:No.Fusions), showed an insignificant association to the response. The direction and strength of those other associations are examined in the logistic regression models. Apart from those relationships to the outcome, several associations between the predictor candidates can be found, and thus indicate potential problems with multicollinearity.

Since a surgical tracheostomy (Tracheostomy:T.Surgery) covers a subset of patients among those who received a tracheostomy, there is a deterministic dependency between both variables which is also expressed by the infinite Odds Ratio and its ”missing” ($NaN$) effectsize transformation to Pearson's $r^*$. Surprisingly, the interaction terms of ventral surgery (OP:OP.Ventral) and the number of fusioned vertebrae (OP:No.Fusions) are seemingly unrelated to each other, despite their equal origin as individuals had to receive a surgery (OP) in the first place to have a ventral surgery or get their vertebrae fusioned. Despite those two interaction duplets, there are several other associated variable pairs. The total motor score (TMS) is a general proxy for a patient's injury severity. A high TMS corresponds to less severe injuries. As one expects, a high neurological level of injury (NLI.High), a ventral surgery (OP:OP.Ventral), a higher number of fusioned vertebrae (OP:No.Fusions), and a (surgical) tracheostomy (Tracheostomy & Tracheostomy:T.Surgery) are all associated with more severe injuries and thus lower total motor scores. The surprising positive linear association between age and total motor score suggests that older individuals are less severely injured. This can be explained by the clinical observation, that the cause for cervical SCI in younger patients often is (extreme) sport or a traffic accident, while older patients tend to receive cervical SCI after e.g., falling down the stairs. Further, since a tracheostomy is applied if ventilation is needed and thus resulting from more severe injuries, also tracheostomy is associated with all other variables. Such that tracheostomy is associated with a high neurological level of injury, a minimal lower age (although this effect is relatively small and the same argumentation as before is applicable), a ventral surgery, and a higher number of fusioned vertebrae. Finally, ventral surgery is associated with slightly younger patients. Thus, younger patients more often have types of injuries which need ventral surgery, compared to older patients. These associations and potential forms of multicollinearity are later examined by the variance inflation factor (VIF) in Tab. 3.

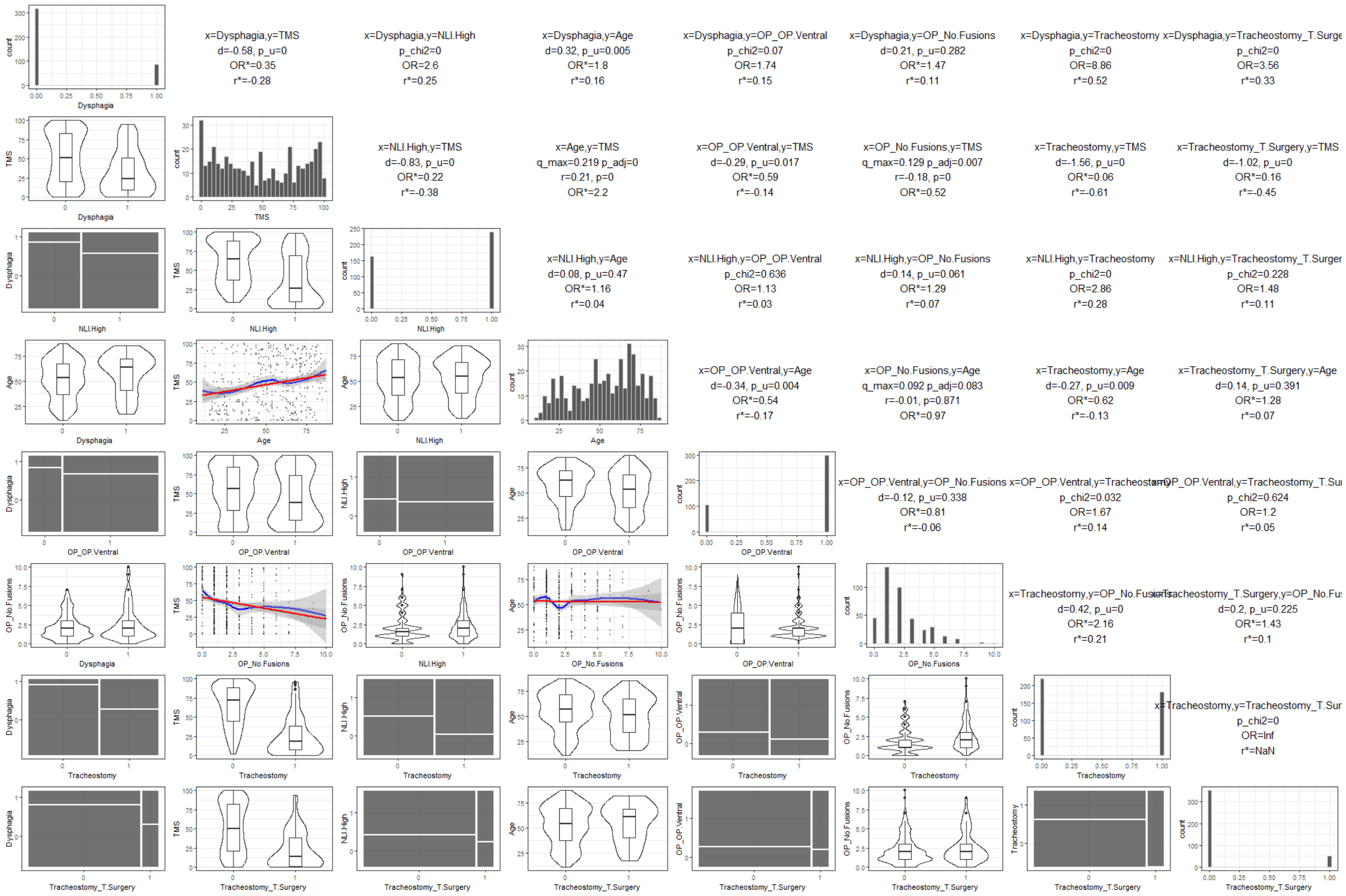

Figure 3: Dependency plot with Cohen's $d$, Odd's Ratio, Pearson's $r$, the maximum (copula based) $q$ measure with its adjusted $p$-value and either $p$-values for $r$ or the Man-Whitney $U$ test. The asterik * indicates an effect size transformation and therefore an approximation. The scatter plots show linear regression lines in red and non-parametric loess-smoothed regression lines in blue. The grey tube around the loess-line indicates its respective 95% CI.

## 5.2 A) Dysphagia Prevalence & Data Partition

In order to fix the dysphagia prevalence in all 1000 training and test samples, first the prevalence is estimated via 1000 bootstrap samples. The resulting median prevalence and its CI was $p_{\text{Boot.Mdn}} = 0.22, \text{CI}_{\text{Boot.95\%}} = [0.18, 0.26]$.

Since the $\text{EPV}_{\text{Global Training}}$ should be at least 10, if not even higher, we define the data partition rate to be 90%/10%. Thus, leading to training samples of $N_{\text{Training}} = 363$ with 79 (21.76%) dysphagia cases and to test samples of $N_{\text{Test}} = 40$ with 8 (20%) dysphagia cases for each of the cross-validation runs. Finally, the EPV of the *global model* on the training data is fixed with $\text{EPV}_{\text{Global Training}} = 11.29$.

## 5.3 B) Model Building and stability Analyses

In sum there are 8 models which were tested in 1000 crossvalidation runs. The first 4 models are a naive logistic regression (”raw” NLR), a logistic regression with Firth's correction (FLR), and two models with post-estimation shrinkage factors (PESF) calculated with DFBETA residuals. One model with a global PESF (g-NLR) and one with parameterwise PESF for the total motor score (TMS), age, a high neurological level of injury (NLI.High), and joint PESF for the surgery duplets (OP:No.Fusions, OP:OP.ventral) as well as the tracheostomy duplets (Tracheostomy, Tracheostomy:T.Surgery). Those duplets correspond the the parentheses in Eq. (21) of the *global model*. Additional to these, also variable selection algorithms are applied to the NLR and FLR models with a significance selection criterion of $\alpha = .157$. Such that augmented Backward Elimination (ABE) with a change-in-estimate threshold of $\tau = .05$ is solely applied to the naive logistic regression model (ABE-NLR) and ”common” Backward Elimination (BE) to both naive and Firth's logistic regression (BE-NLR, BE-FLR). Complementary, FOCI was applied as a modern, non-parametric forward selection algorithm. FOCI was the only algorithm which was not based on the 7 predictor candidates in the *global model*. Instead, it chose from a pool of 11 variables: the 7 of the *global model* and additional 4 which can be found in Appendix A.

Each of those 8 models provided (shrunken) regression coefficients and in case of nonselected variables coefficients of 0. Due to the 1000 crossvalidation runs, each parameter estimate belonged to a subsampling distribution, of which the median and the 2.5% and 97.5% percentiles (i.e., $\text{CI}_{\text{SS.95\%}}$) were derived and used as robust measures for their effect. The median regression coefficients and their CI can be found in table 2 and figure 4 for all predictor candidates. As follows, first the results except the ones from the FOCI algorithm are presented, since it behaved somewhat differently. In essence, one can see that the application of shrinkage methods like applying PESF, penalized likelihood or performing variable selection leads to smaller parameter estimates. Thus, the ”raw” naive logistic regression model (raw NLR) showed on average the biggest (un-shrunken) slopes. On the other end, the application of joint PESF (j-NLR) and FOCI introduced the biggest shrinkage on the slopes, leading to more ”insignificant” predictor candidates having slopes near zero or containing zero in their 95% subsampling confidence interval $\mathrm{CI}_{\mathrm{SS}.95\%}$.

When examining the histograms in Fig. 5 one sees, that most subsampling distributions of the shrinkage models are uni-modal which supports the robustness of the derived effects. But applying joint PESF led to instability (i.e., a bi-modal distribution) in both TMS and tracheostomy. Such bi-modal distributions can also be seen in variable selection coefficient distributions, where the variable was not selected in some of the crossvalidation runs. Strong predictors such as age and tracheostomy mostly showed uni-modal and ”normally” looking coefficient distributions. While weaker predictors showed non-normally distributed effects or even bi-modal distributions, especially among the variable selection algorithms.

Overall, none of those models performed really well: The maximum (adjusted) likelihood ratio $R^2$ (or also McFadden’s pseudo $R^2$; Long, 1997; Menard, 2000) was $R^2_{\mathrm{McFadden}} = .23[.21, .26], R^2_{\mathrm{McFadden\,adj.}} = .19[.16, .22]$ which indicates the absence of other important explanatory variables for dysphagia. Further, the prediction accuracy on test data was at $80\%[70\%, 88\%]$ which is basically not better than chance in a sample with $p \approx 20\%$ dysphagia cases. This is also reflected in the fairly high specificity ($\geq 91\%$) and the simultaneously very low sensitivity ($\leq 25\%$). Thus, all models identified healthy individuals arguably well, but in case of a predicted dysphagia case the probability to be indeed a dysphagia patient (i.e., the precision/ positive predictive value) was around 50% [0%, 100%] in almost all models. It has to be remarked, that predicted dysphagia cases were assigned when their predicted probabilities were $\geq 50\%$. This threshold was arbitrary and **not** optimized by grid search or crossvalidation, due to the low EPV and since we aimed for identifying risk-factors instead of building a performant prediction model.

Regarding the importance of single variables, the total motor score (TMS), the number of fusioned vertebrae (OP:No.Fusions), and the type of tracheostomy (Tracheostomy:T.Surgery) showed no significant effect, since their median slopes were near zero and their $\mathrm{CI}_{\mathrm{SS}.95\%}$ included zero in all models. In contrast, age and tracheostomy showed robust and significant effects in all models except the one from the FOCI algorithm. Over most models, age showed a very narrow effect of around $b = 0.03[0.02, 0.04], \mathrm{OR} = 1.03[1.02, 1.04]$. Thus, when holding all other variables constant, the odds that a 40 year old patient has dysphagia is at $\mathrm{OR} = 3.32$, while the odds that a 50 year old patient has dysphagia are already at $\mathrm{OR} = 4.48$. Likewise, when looking at the models except FOCI and holding all other variables constant, receiving a tracheostomy is associated with a at minimum $\mathrm{OR} = 6.36[3.97, 16.61]$ ($b = 1.85[1.38, 2.81]$) times higher chance of having dysphagia, than not receiving a tracheostomy. A high neurological level of injury (NLI.High) and a ventral surgery (OP:OP.Ventral) showed more ambiguous effects. Both predictor candidates showed similar big slopes and mostly $\mathrm{CI}_{\mathrm{SS}.95\%}$ which did not contain zero, except the FOCI model and the model with joint PESF (j-NLR). In this case, the parameterwise shrinkage factor shrunk the slopes that low, that even their $\mathrm{CI}_{\mathrm{SS}.95\%}$ included zero afterwards. For predictor candidates of the *global model*, multicollinearity was neglible since all $\mathrm{VIF} \leq 1.85$ (see Tab. 3).

Complementary to these single effect sizes per model, they were averaged over all models per crossvalidation run. From the resulting subsampling distribution of arithmetic means for each predictor candidate, the median and respective $\mathrm{CI}_{\mathrm{SS}.95\%}$ were again derived. Such that, the median average regression coefficient of the total motor score was $\mathrm{TMS} = 0[0, 0]$, $\mathrm{NLI.High} = 0.45[0.13, 0.66]$, $\mathrm{Age} = 0.03[0.02, 0.03]$, $\mathrm{OP{:}OP.Ventral} = 0.43[0.11, 0.66]$, $\mathrm{OP{:}No.Fusions} = 0.02[0, 0.04]$, $\mathrm{Tracheostomy} = 2.11[1.78, 2.44]$, and $\mathrm{Tracheostomy{:}T.Surgery} = 0.08[-0.04, 0.27]$. Thus, age, tracheostomy, a ventral operation (OP:OP.Ventral) and a high neurological level of injury (NLI.High) showed on average over all models a significant effect and did not include 0 in their $\mathrm{CI}_{\mathrm{SS}.95\%}$. Further, age and tracheostomy excluded 0 from their CI in all models (cf. Tab. 2) and thus showed the most robust effect.

These main findings are also reflected in the stability analyses. All BE-based variable selection algorithms selected age and tracheostomy in 100% of the crossvalidation runs (see variable inclusion frequencies in Tab. 4). This is not surprising for the ABE-NLR model, since it was specified to always include age. But both BE-NLR and BE-FLR showed convergent validity, as they not only selected age equally frequent but also reported equal effect sizes and $\mathrm{CI}_{\mathrm{SS}.95\%}$. The high neurological level of injury and ventral surgery both were selected in between 68.5% and 88.3%. Insignificant predictor candidates were selected in less than 20%. Again with exception of the FOCI model, the model selection frequencies in Tab. 5 also strongly support the importance of age and tracheostomy, and the ambiguous influence of a high neurological level of injury and a ventral surgery. Generally, the most often selected models of the different algorithms have clearly higher frequencies than the 2nd or 3rd best models. Ultimately, the variable selection models based on the *global model* unanimously identify the combination of age, tracheostomy, a high neurological level of injury and a ventral surgery as the important predictors in 44.8-62.4% of the crossvalidation runs.

The results from the FOCI algorithm have to be presented and discussed separately, since FOCI often selected only one or two predictor candidates in each crossvalidation run.

The resulting variable inclusion frequencies in Tab. 4 show only the variable tracheostomy to be selected in more than 50% of the crossvalidation runs. Since non-selection results in a slope of zero and the median slope is reported, only tracheostomy had a median slope bigger than zero. All other predictor candidates were **not selected** in more than 50% and thus have a median slope of zero. Despite its overall low inclusion and selection frequencies, FOCI also supports the importance of tracheostomy. Ventilation and age were both selected in about 49% of the crossvalidations and thus seem somewhat influential. In contrast, the model selection frequencies are very similar between a small model only containing age and tracheostomy, a model solely containing tracheostomy, and a model comprising solely ventilation. Overall, the results of the FOCI algorithm are of limited interpretability, since it chose from a set of variables with potentially problematic multicollinearity (VIF $\leq 3.74$) and further potentially suffers more heavily from the low EPV as a non-parametric measure.

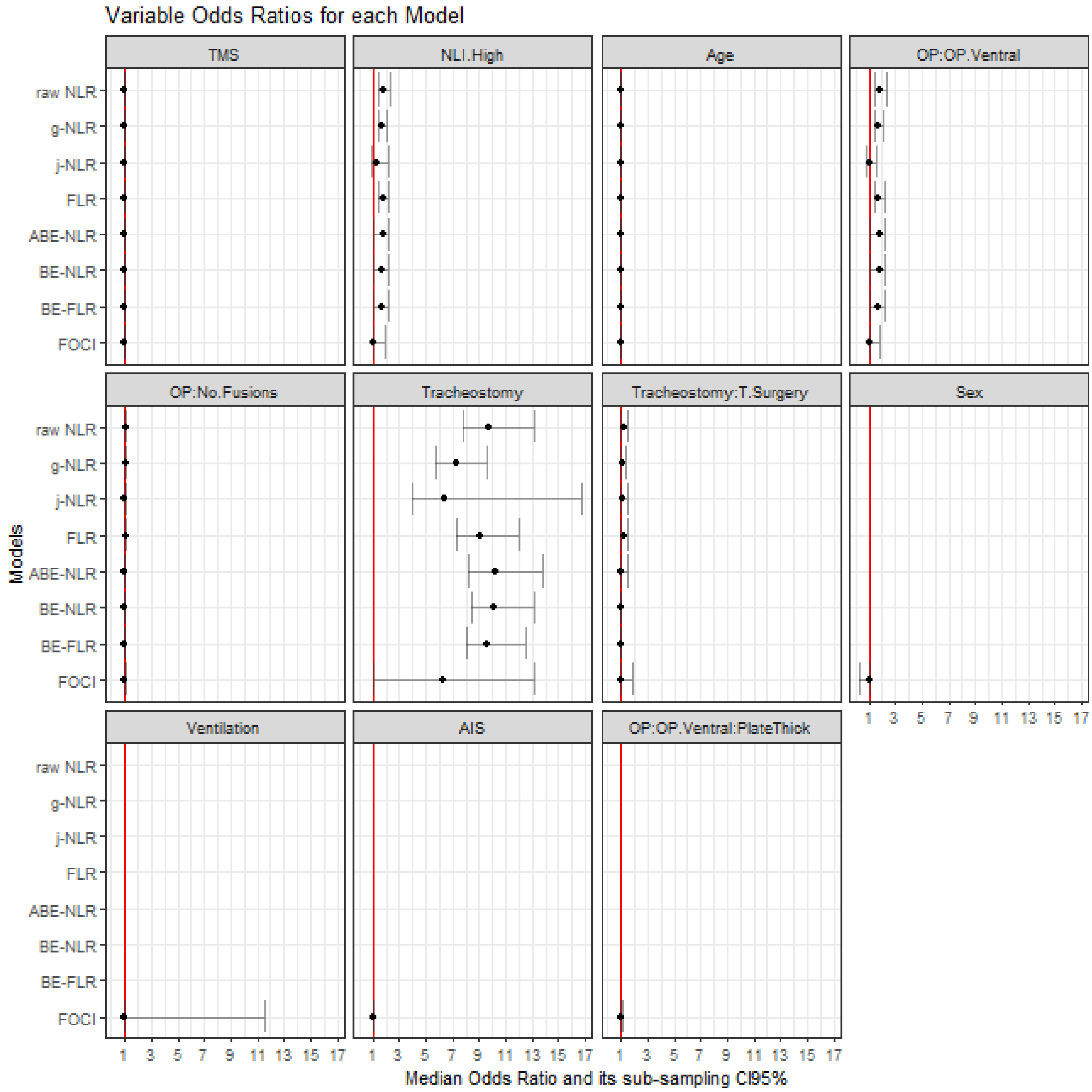

Figure 4: Forest plot of the median and subsampling $\mathrm{CI}_{\mathrm{SS}.95\%}$ of Odds Ratios per variable and model. The red vertical line depicts OR = 1.

| | **(Shrinkage) Models** | | | | **Variable Selection Models** | | | |
|---|---|---|---|---|---|---|---|---|
| Variables | Raw-NLR | g-NLR | j-NLR | FLR | ABE-NLR | BE-NLR | BE-FLR | FOCI |
| 1 | 0 [0, 0.01] | 0 [0, 0] | -0.01 [-0.02, 0.02] | 0 [0, 0.01] | 0 [0, 0] | 0 [0, 0] | 0 [0, 0] | 0 [0, 0] |
| 2 | 0.58 [0.39, 0.84] | 0.5 [0.34, 0.73] | 0.21 [-0.16, 0.78] | 0.55 [0.37, 0.8] | 0.56 [0, 0.8] | 0.55 [0, 0.78] | 0.53 [0, 0.75] | 0 [0, 0.63] |
| 3 | 0.03 [0.03, 0.04] | 0.03 [0.02, 0.03] | 0.03 [0.02, 0.03] | 0.03 [0.02, 0.03] | 0.03 [0.03, 0.04] | 0.03 [0.03, 0.04] | 0.03 [0.03, 0.04] | 0 [0, 0.03] |
| 4 | 0.59 [0.39, 0.84] | 0.51 [0.34, 0.73] | 0.1 [-0.29, 0.45] | 0.55 [0.36, 0.79] | 0.56 [0, 0.8] | 0.56 [0, 0.8] | 0.54 [0, 0.77] | 0 [0, 0.62] |
| 5 | 0.05 [-0.01, 0.09] | 0.04 [-0.01, 0.08] | 0.01 [-0.02, 0.04] | 0.04 [-0.01, 0.09] | 0 [0, 0] | 0 [0, 0] | 0 [0, 0] | 0 [0, 0.06] |
| 6 | 2.28 [2.05, 2.58] | 1.98 [1.76, 2.26] | 1.85 [1.38, 2.81] | 2.2 [1.99, 2.49] | 2.32 [2.1, 2.63] | 2.31 [2.13, 2.58] | 2.27 [2.08, 2.53] | 1.83 [0, 2.58] |
| 7 | 0.17 [-0.08, 0.4] | 0.14 [-0.07, 0.34] | 0.13 [-0.08, 0.37] | 0.17 [-0.07, 0.39] | 0 [0, 0.38] | 0 [0, 0] | 0 [0, 0] | 0 [0, 0.6] |
| 8 | / | / | / | / | / | / | / | 0 [0, 2.45] |
| 9 | / | / | / | / | / | / | / | 0 [-1.08, 0] |
| 10 | / | / | / | / | / | / | / | 0 [0, 0.13] |
| 11 | / | / | / | / | / | / | / | 0 [0, 0.09] |
| $R^2$ | 0.21 [0.19, 0.24] | / | / | 0.23 [0.21, 0.26] | 0.21 [0.19, 0.24] | 0.21 [0.19, 0.24] | 0.22 [0.19, 0.25] | 0.17 [0.03, 0.24] |
| $R^2_{\text{adj.}}$ | 0.17 [0.15, 0.2] | / | / | 0.19 [0.16, 0.22] | 0.18 [0.16, 0.21] | 0.19 [0.16, 0.21] | 0.19 [0.17, 0.22] | 0.16 [0.02, 0.21] |
| Acc. | 0.8 [0.7, 0.88] | 0.8 [0.7, 0.88] | 0.8 [0.7, 0.88] | 0.8 [0.7, 0.88] | 0.8 [0.7, 0.88] | 0.8 [0.7, 0.88] | 0.8 [0.7, 0.88] | 0.8 [0.72, 0.88] |
| Prec. | 0.5 [0, 1] | 0.5 [0, 1] | 0.5 [0, 1] | 0.5 [0, 1] | 0.5 [0, 1] | 0.5 [0, 1] | 0.5 [0, 1] | 0.56 [0, 1] |
| Sens. | 0.25 [0, 0.62] | 0.25 [0, 0.5] | 0.25 [0, 0.5] | 0.25 [0, 0.62] | 0.25 [0, 0.62] | 0.25 [0, 0.62] | 0.25 [0, 0.62] | 0 [0, 0.62] |
| Spec. | 0.94 [0.81, 1] | 0.94 [0.84, 1] | 0.94 [0.81, 1] | 0.94 [0.81, 1] | 0.91 [0.81, 1] | 0.91 [0.81, 1] | 0.94 [0.81, 1] | 1 [0.84, 1] |

Table 2: This table depicts for all models the median regression coefficients and their $\text{CI}_{95\%}$ which were derived from their subsampling distribution. Such that NLR := Naive Logistic Regression (without any shrinkage), g/j-NLR := Naive Logistic Regression with either a global or a joint post-estimation shrinkage factor (PESF), and FLR = Firth's Logistic Regression (with intercept correction). In the variable selection columns, ABE := augmented Backward Elimination, BE := Backward Elimination and FOCI := Feature Ordering by Conditional Independence. With regard to the variable selection models, not selected variables were depicted with a regression coefficient of $b = 0$. The performance measures accuracy (Acc.), precision (Prec.), sensitivity (Sens.), and specificity (Spec.) were derived from the test data, while McFadden's (adjusted) pseudo $R^2$ was dervied from the training data. ***Variable Indices:*** *1 TMS, 2 NLI.High, 3 Age, 4 OP:OP.Ventral, 5 OP:No.Fusions, 6 Tracheostomy, 7 Tracheostomy:T.Surgery, 8 Ventilation, 9 Sex, 10 OP:OP.Ventral:PlateThick, & 11 AIS.*

| | *Global Model* | | *FOCI Model* | |
|---|---|---|---|---|
| | VIF | Tolerance | VIF | Tolerance |
| 1 | 1.85 [1.79 ,1.95] | 0.54 [0.51 ,0.56] | 3.74 [3.58 ,4.07] | 0.27 [0.28 ,0.25] |
| 2 | 1.15 [1.12 ,1.17] | 0.87 [0.85 ,0.89] | 1.19 [1.16 ,1.22] | 0.84 [0.86 ,0.82] |
| 3 | 1.1 [1.08 ,1.13] | 0.91 [0.88 ,0.93] | 1.15 [1.12 ,1.18] | 0.87 [0.89 ,0.85] |
| 4 | 1.04 [1.03 ,1.05] | 0.96 [0.95 ,0.97] | 1.75 [1.7 ,1.82] | 0.57 [0.59 ,0.55] |
| 5 | 1.06 [1.05 ,1.08] | 0.94 [0.93 ,0.95] | 1.11 [1.08 ,1.13] | 0.9 [0.93 ,0.88] |
| 6 | 1.75 [1.69 ,1.85] | 0.57 [0.54 ,0.59] | 3.36 [3.15 ,3.68] | 0.3 [0.32 ,0.27] |
| 7 | 1.5 [1.46 ,1.55] | 0.67 [0.65 ,0.68] | 1.51 [1.47 ,1.56] | 0.66 [0.68 ,0.64] |
| 8 | / | / | 2.82 [2.67 ,3.08] | 0.35 [0.37 ,0.32] |
| 9 | / | / | 1.03 [1.02 ,1.05] | 0.97 [0.98 ,0.95] |
| 10 | / | / | 1.53 [1.48 ,1.58] | 0.65 [0.68 ,0.63] |
| 11 | / | / | 3.02 [2.89 ,3.36] | 0.33 [0.35 ,0.3] |

Table 3: McFadden's pseudo $R^2$ was used to calculate variable inflation factors (VIF) and tolerance as multicollinearity diagnostics stratified for the *global model* and the FOCI model. ***Variable Indices:*** *1 TMS, 2 NLI.High, 3 Age, 4 OP:OP.Ventral, 5 OP:No.Fusions, 6 Tracheostomy, 7 Tracheostomy:T.Surgery, 8 Ventilation, 9 Sex, 10 OP:OP.Ventral:PlateThick, & 11 AIS.*

## 5.4 C) Model Selection and Comparison

Based on these results, two final competing models were derived: one minimal model (i.e., mini-model) with only the most robust effects of age and tracheostomy (cf. Eq. (22)). And the slightly bigger model (i.e., reduced-model) in Eq. (23), additionally comprising the ambiguous high neural level of injury (NLI.High) and ventral surgery (OP:OP.Ventral). Both models were re-fitted with a normal logistic regression. In contrast to the method pipeline in B), these two models were not parsed through shrinkage methods or variable selection algorithms, because we are now primarily interested in the unbiased and thus un-shrunken regression coefficients. Hence, model fitting was applied on the whole sample with $N = 403$ observations.

$$\text{Dysphagia} \sim \text{Age} + \text{Tracheostomy} \tag{22}$$

$$\text{Dysphagia} \sim \text{Age} + \text{Tracheostomy} + \text{NLI.High} + \text{OP:OP.Ventral} \tag{23}$$

Due to the reduction of the model sizes, the EPV was consequently higher. As an exploratory analysis, the continuous variable age is additionally applied to the multivariable fractional polynomials (MFP) algorithm by Heinze, Ambler, and Benner (2022) to determine its functional form. MFP was applied without variable selection, with the maximum complexity of a FP2 term, and a FP significance criterion of $\alpha = .157$. Con-

| **Variables** | **ABE-NLR** | **BE-NLR** | **BE-FLR** | **FOCI** |
|---|---|---|---|---|
| Age | 100% | 100% | 100% | 48% |
| Tracheostomy | 100% | 100% | 100% | 73.5% |
| NLI.High | 88.3% | 85.8% | 84.3% | 11.4% |
| OP:OP.Ventral | 73.3% | 73.9% | 68.5% | 10.9% |
| OP:No.Fusions | 0.3% | 0.1% | 0.1% | 18.7% |
| Tracheostomy:T.Surgery | 8.7% | 0% | 0% | 11.4% |
| TMS | 19% | 0% | 0% | 2.3% |
| Ventilation | / | / | / | 49.3% |
| Sex | / | / | / | 16.1% |
| OP:OP.Ventral:Platethick | / | / | / | 10% |
| AIS | / | / | / | 6% |

Table 4: Variable inclusion frequencies for stability investigations.

| **Model** | **Best 3** | **Frequency** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ABE-NLR | #1 | 44.8% |  | ■ | ■ | ■ |  | ■ |  | / | / | / | / |
|  | #2 | 16.4% |  | ■ | ■ |  |  | ■ |  | / | / | / | / |
|  | #3 | 12.6% | ■ | ■ | ■ | ■ |  | ■ |  | / | / | / | / |
| BE-NLR | #1 | 62.4% |  | ■ | ■ | ■ |  | ■ |  | / | / | / | / |
|  | #2 | 23.3% |  | ■ | ■ |  |  | ■ |  | / | / | / | / |
|  | #3 | 11.4% |  |  | ■ | ■ |  | ■ |  | / | / | / | / |
| BE-FLR | #1 | 56.5% |  | ■ | ■ | ■ |  | ■ |  | / | / | / | / |
|  | #2 | 27.7% |  | ■ | ■ |  |  | ■ |  | / | / | / | / |
|  | #3 | 11.9% |  |  | ■ | ■ |  | ■ |  | / | / | / | / |
| FOCI | #1 | 7.9% |  |  | ■ |  |  | ■ |  |  |  |  |  |
|  | #2 | 6% |  |  |  |  |  |  |  | ■ |  |  |  |
|  | #3 | 4.8% |  |  |  |  |  | ■ |  |  |  |  |  |

Table 5: Model selection frequencies for stability investigations. ***Variable Indices:*** *1 TMS, 2 NLI.High, 3 Age, 4 OP:OP.Ventral, 5 OP:No.Fusions, 6 Tracheostomy, 7 Tracheostomy:T.Surgery, 8 Ventilation, 9 Sex, 10 OP:OP.Ventral:PlateThick, & 11 AIS.*

sequently, the before mentioned mini- and reduced-model were additionally compared to their MFP counterparts (i.e., mini-MFP-model & reduced-MFP-Model).

With the remaining four predictor candidates, no multicollinearity occured (VIFs $\leq$ 1.07). In both the mini-MFP-model and the reduced-MFP-model the suggested transformation of age was of form $f(\text{age}) = (\frac{\text{age}}{100})^1$ and hence a simple linear transformation. Consequently, the MFP models were equivalent to the "normal" mini- and reduced-model, containing the same deviance, and thus showing no improvement. Ultimately, both MFP models can therefore be ignored. In comparison to an intercept model, the mini-model had a significantly lower deviance ($\text{Diff}_{\text{Deviance}} = 82.24, \text{df} = 2, p < .001, R^2_{\text{McFadden}} = .20, R^2_{\text{McFadden adj.}} = .18$), which was even significantly lower for the reduced-model ($\text{Diff}_{\text{Deviance}} = 6.46, \text{df} = 2, p = .039, R^2_{\text{McFadden}} = .21, R^2_{\text{McFadden adj.}} = .19$) compared to the mini-model. Also the AIC favored the reduced-model over the mini-model. Nevertheless, the reduced-model showed highly non-normally distributed errors (skew = 3.37, kurtosis = 13.84) which are especially high for more probable dysphagia predictions (i.e., higher standardized predictions). The residuals have a mean of M = -0.12 (SD = 2.84) and a median Mdn = -1.07 (MAD = 0.15), which both indicate the phenomenon shrinkage. This corresponds to the earlier made observations, that our (final) model has an increased specificity, but very low sensitivity.

Ultimately it can be concluded, that the reduced-model (cf. Eq. (23)) describes the available data the best, while being the most parsimonious model. As follows, the slopes are presented with their (parametric) $\text{CI}_{95\%}$. Due to the fact, that the reduced model is derived from variable selection, the $p$-values might be underestimated. But for the sake of completeness, they are nevertheless reported. Tracheostomy showed a significant positive association with dysphagia ($b = 2.30[1.69, 2.97], \text{OR} = 9.97[5.42, 19.49], p < .001$). Also age showed a significant and positive association to dysphagia with $b = 0.03[0.02, 0.05], \text{OR} = 1.03[1.02, 1.05], p < .001$. A high neurological level of injury (NLI.High) was insignificant at the 5% level ($b = 0.56[-0.04, 1.18], \text{OR} = 1.75[0.96, 3.25], p = .07$). Same applies to ventral surgeries (OP:OP.Ventral, $b = 0.56[-0.09, 1.25], \text{OR} = 1.75[0.91, 3.49], p = .103$).

When accounting for the earlier found shrinkage by estimating parameterwise PESF with jackknifing, tracheostomy was barely shrunken by a factor of $\text{PESF}_{\text{Tracheostomy}} = .98$. Also age was only shrunken by a small factor of $\text{PESF}_{\text{Age}} = .91$. In contrast, the more ambiguous effects of high neural level of injury, as well as ventral accessed surgery were shrunken more heavily with $\text{PESF}_{\text{NLI.High}} = .69$ and $\text{PESF}_{\text{OP:OP.Ventral}} = .58$.

# 6 Discussion

## 6.1 Risk-Factors for Dysphagia: Building a Descriptive Model

The aim of this thesis was to identify risk-factors for dysphagia, a clinically relevant swallowing dysfunction. For that, domain experts and myself derived a *global model* as proposed by Heinze et al. (2018). This laid the foundation for the variable selection process. Due to the low prevalence of dysphagia in the (observational) sample of $N = 403$ patients with cervical SCI, the EPV rate could be deemed too low to perform sound variable or even function selection (Heinze et al., 2018; Sauerbrei et al., 2020).

In order to counter statistical uncertainty, reduce dependencies on distributional assumptions and get robust estimates, multiple methods were applied: since Heinze et al. (2018) strongly recommended stability investigations in case of variable selection in low EPV settings, and Wallisch et al. (2020) found subsampling to be better suited when predictors might be correlated, all further analyses were wrapped in 1000 crossvalidation runs with fixed dysphagia prevalences in both training and test data. With regard to model building, a naive approach utilized normal logistic regression (”raw” NLR), but was complemented by shrinkage methods, such as post-estimation shrinkage factors (PESF; Dunkler et al., 2016) as well as likelihood penalization proposed by Firth (1993) and Puhr et al. (2017). While the ”normal” logistic regression models with either a global (g-NLR) or with joint PESF (j-NLR) were not further processed, the unshrunken ”raw” NLR model as well as Firth's corrected logistic regression model (FLR) were passed on to the (augmented) Backward Elimination variable selection algorithm (ABE-NLR, BE-NLR, BE-FLR). Ultimately, the non-parametric and relatively new FOCI algorithm was applied to a slightly increased set of predictor candidates which not only contained the ones from the *global model* but also 4 additional ones, selected by domain experts (cf. Appendix A).

Summa summarum, the predictor candidates were estimated and assessed in 1000 randomly drawn subsamples of training and test data. Thus, each variable of the *global model* belonged to 8 different subsampling distributions, from which the median and the 2.5 and 97.5 percentiles were derived as robust estimates and their 95% subsampling Confidence Interval $\mathrm{CI}_{\mathrm{SS.95\%}}$. In addition, stability investigations comprised variable inclusion frequencies, as well as the top 3 model selection frequencies over the 1000 crossvalidation runs.

From this holistic variable selection approach, tracheostomy and age were identified as the most influential risk-factors for dysphagia. These results are supported in all models except for the FOCI model (cf. Tab. 2). The latter is discussed later separately, since its results were often less clear. Both age and tracheostomy had the highest variable selection frequencies, were both included in the top 3 selected models, and showed consistent results (i.e., barely varying effects) over our different models. Thus, their estimates can be deemed unbiased. It should be noted, that due to availability problems in the used R packages, age was introduced as a "passive" predictor (cf.; Dunkler et al., 2014) only in the ABE-NLR model, as suggested by Heinze et al. (2018). Thus it did not participate in the variable selection process of this one model. But it participated in the variable selection process of all other models. Despite this inconsistent treatment of the variable age, neither its subsampling distribution nor its derived estimates differed from each other and thus support the robustness of age's influence on dysphagia.

In contrast, a high neurological level of injury (NLI.High) and ventrally accessed surgery (OP:OP.Ventral) showed more ambiguous results over the 8 models. Some indicate an influential association and some such as the joint shrinkage model (j-NLR), and all variable selection models do not, since their $\mathrm{CI}_{\mathrm{SS.95\%}}$ include zero. Nevertheless, both predictor candidates still showed obviously higher variable inclusion frequencies, belonged to the most often selected models, and showed a significantly better model fit in the reduced-model, compared with the mini-model comprising only age and tracheostomy. Although the results are less clear, they nonetheless suggest their importance. Hence, the 4 predictors age, tracheostomy, high neurological level of injury, and a ventral surgery can be derived as the most influential predictors for dysphagia, where the former are more influential than the latter ones.

Of the total motor score (TMS), number of fusioned vertebrae (OP:No.Fusions), and the indicator for surgical tracheostomy (Tracheostomy:T.Surgery), none showed significant (linear) associations with dysphagia when simultaneously accounting for the before mentioned influential variables. TMS actually showed a Null-effect. Although one could argue for a potential non-monotonic (e.g., quadratic) association between TMS and dysphagia from a methodological point of view, there is no plausible argumentation supporting this hypothesis from a clinical point of view. Thus, TMS can indeed be deemed unrelated to dysphagia, when accounting for at least tracheostomy and age. This is actually surprising from a clinical perspective, since TMS was used as a proxy for injury severity which was expected to be associated with dysphagia. Since several significant bivariate (i.e., un-adjusted) associations could be found between TMS and other predictors in Fig. 3, one can assume that at least tracheostomy and age, if not even a high neurological level of injury and ventral surgery, account better for injury severity than TMS alone. Also the type of tracheostomy showed very weak (adj.) effects. It can be seen and argued that receiving a tracheostomy at all is so much more influential than the type of tracheostomy received in the end. Also the number of fusioned vertebrae showed weak effects over our models, since many $\mathrm{CI}_{\mathrm{SS.95\%}}$ are near zero. Although this can be deemed as not significantly linearly associated, there might be a small non-linear association with dysphagia which could not be examined in this thesis.

Overall, these conclusions and the $\mathrm{CI}_{\mathrm{SS.95\%}}$ on which they are based on are also supported by the subsampling distribution histograms in Fig. 5 in Appendix B. Again with exclusion of the FOCI results, those histograms showed mostly uni-modal, symmetric. and sometimes even somewhat visually normal looking distributions. This depicts the robustness of the median and $\mathrm{CI}_{\mathrm{SS.95\%}}$. It can be annotated, that for the more ambiguous NLI.High and OP:OP.Ventral the variable selection distributions are skewed due to the slopes of zero, if the significance (or change-in-estimate) criterion is not met. Tracheostomy shows a heavy upper tail and sometimes a bi-modal distribution which is the result of specific subsamples, where the already big effect is even more extreme. The other predictor candidates from the *global model* such as the total motor score (TMS), the number of fusioned vertebrae and a surgical tracheostomy showed clearly no (linear) association to dysphagia. Their subsampling distributions were often bi-modal with a spike-at-zero due to their frequent non-selection in the variable selection models. Their distributions in the (shrinkage) models are mostly uni-modal and somewhat symmetric around zero.

Irrespective of the 8 different model types, all performed bad with low accuracy, low precision, low sensitivity, and a low ratio of explained log-likelihood (i.e., low McFadden's Pseudo $R^2$) in the cross-validation runs. This indicates, that there are still influential risk-factors missing from the model. The compared models all perform better in identifying healthy individuals than identifying individuals with dysphagia. But nevertheless, our models do not achieve higher accuracy than guessing a patients condition based on the prevalence.

Variable selection by FOCI behaved differently. In contrast to the backward elimination algorithms which were based on the pre-defined *global model*, FOCI is a non-parametric forward selection algorithm which additionally had a larger pool of predictor candidates to choose from. Overall, FOCI selected only up to 2 predictors in each cross-validation run. This resulted in a lot of low variable inclusion frequencies. Since, the median (and not the mean) were derived from the slope distributions and most variables were selected in less than 50% of the crossvalidation runs, most median slopes are zero and all $\mathrm{CI}_{\mathrm{SS.95\%}}$ include zero. FOCI solely identified tracheostomy as an important risk-factor for dysphagia. But even there, its $\mathrm{CI}_{\mathrm{SS.95\%}}$ included zero and showed a tri-modal coefficient distribution with peaks at zero, roundabout 1.5 and roundabout 2.5.

In conclusion, the median slope and its $\mathrm{CI}_{\mathrm{SS}.95\%}$ of the FOCI model should generally not be interpreted due to their subsampling distributions. With respect to the variable inclusion frequncies, FOCI not only determines tracheostomy as very influential, but also both age and ventilation. The latter was only included in the set of predictor candidates for the FOCI model because of its clinical interdependence with tracheostomy. Thus, the FOCI model obviously suffered more from multicollinearity than the *global model*. Although the model selection frequencies of the top 3 models barely differed from each other, FOCI still identified tracheostomy and/or age, or ventilation as influential risk-factors. Since ventilation is highly associated with tracheostomy, the selected models are still reasonable from a clinical point of view.

In the end, the bad discriminatory performance of the FOCI model can be explained by the fact, that non-parametric methods use "less information" than implicitly assumed in parametric alternatives. In settings where parametric assumptions are met, non-parametric methods such as FOCI thus have less statistical power than their parametric alternatives, especially when sample sizes are small (e.g.; Prajapati et al., 2010). This actually might be the case in our analyses due to the (mostly) linearly associated predictor candidates, neglible to low multicollinearity and the low EPV rate. Nevertheless, FOCI still somewhat full-filled its task by providing an interesting and non-parametric view on the variable selection process with an increase pool of candidate variables.

Another interesting phenomenon could be observed in Tab. 2: both global and obviously also joint PESF applied more shrinkage on the estimates than Firth's penalized likelihood approach. Since Van Calster et al. (2020) found Firth's correction to best approximate the true parameters in their simulation study, it would be interesting for future simulation studies to compare the amount of shrinkage between (advanced) penalized likelihood methods and PESF.

## 6.2 Limitations

The methods applied in this thesis were chosen to meet current state-of-the-art recommendations regarding descriptive model building by Heinze et al. (2018), Sauerbrei et al. (2020), Wallisch et al. (2020), and Van Calster et al. (2020).

As such, a *global model* was pre-defined by domain experts to support descriptive modelling and thus variable selection by clinical knowledge and the current state of research. But since 7 of the 8 models are based on this *global model*, potentially missing confounders or even included unknown colliders or mediatiors might bias the results (e.g.; Heinze et al., 2018; Rohrer, 2018). The collected variables and the low prevalence of dysphagia only allowed to build a relatively small *global model*. Hence, decisions had to be made which variables to include and which not. Although the FOCI algorithm was applied in order to counter at least some potential model building errors, FOCI suffered from too low statistical power to perform adequately well. In the end, our variable selection can only be as sound as the *global model* on which it is based on and the sample used for it. Regarding the latter, the retrospective design of the observational data used has to be acknowledged.

With respect to the applied methods, it can be discussed if the joint-PESF model should have been a parametwise-PESF model instead. Originally, there was an association and thus multicollinearity assumed between the interaction terms and their main effects in the *global model*. Surprisingly, the variance inflation factors were below 2 and hence mutlicollinearity was negligible (cf. Tab. 3). Despite their semantic similarity, the application of parametwise PESF migth have led to other results.

Further, it would have been nice to test the continuous predictor candidates for non-linearity in the variable selection process. Especially, since our variable selection shows no hints that the number of fusioned vertebrae might be unrelated to dysphagia, as it is the case for TMS. But due to the noisy data and already low EPV it would not have been feasible to additionally perform MFP (Heinze, Ambler, & Benner, 2022) or alternative algorithms based on splines.

Complementary to the applied methods for model building, it would have been informative to: a) also perform Lasso regression or its advancements to compare the amount of shrinkage applied to the models and thus complement the simulation study of Van Calster et al. (2020). And b) to perform own simulations to evaluate the goodness of the selected methods in the given low EPV setting.

Last but not least, the finally derived ”reduced model” comprising the predictors age, tracheostomy, high neurological level of injury, and ventral surgery showed severely non-normal residuals and bias in form of the phenomenon shrinkage. In combination with the parameterwise PESF and the low performance measures it can be concluded that important predictors which were not included in the *global model* are still missing. Future research should address this issue.

# 7 Conclusion

This thesis addressed many issues arising from descriptive model building in low EPV settings: it combined the fixation of dysphagia prevalences in subsamples with stability analyses. It further provided both the median and $\mathrm{CI}_{\mathrm{SS.95\%}}$ as robust effect estimates and thus circumvent un-interpretable $p$-values due to multiple testing. Additionally, it fixed some researcher’s degrees of freedom by applying multiple different methods in the face of arbitrary approaches to perform variable selection and hence can be deemed a multiverse analysis (e.g.; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016).

In conclusion, descriptive model building was successful as age, tracheostomy, a high neurological level of injury, and ventrally accessed surgery were identified as influential risk-factors. The lower variable inclusion frequencies, the $\mathrm{CI}_{\mathrm{SS.95\%}}$ comprising zero, and the heavier parameterwise PESF found in the final reduced-model indicate less importance and some amount of overestimation in the latter two predictors. Also, the low performance of all models indicate that other influential predictors are still missing, and thus have to be determined in future research.

Further, future research should explicitly examine the amount of shrinkage applied by PESF in comparison to likelihood penalization approaches such as Firth's correction, (adaptive) Lasso, and Elastic Net to provide recommendations for variable selection in low EPV settings. In addition it would be helpful to examine, if variable selection approaches like (augmented) Backward Elimination should be applied on ”raw” models or shrinkage models such as Firth's logistic regression. Since the amount of data available is increasing in most research domains, the need for recommendations of reasonable variable selection methods will grow.

Nonetheless, this thesis provided some elaborate but informative approach for variable selection. Ultimately, the potpourri of applied methods provided results which are as sound as one could hope for a variable selection process in the given low EPV setting within a retrospective observational study.

# References

Alliance, U. H. D. R. (2020, July). *Trusted Research Environments (TRE): A strategy to build public trust and meet changing health data science needs.*

Alliance, U. H. D. R. (2021, August). *Building Trusted Research Environments: Principles and Best Practices; towards TRE ecosystems.*

Auguie, B., & Antonov, A. (2017, September). *gridExtra: Miscellaneous Functions for "Grid" Graphics.* Retrieved 2023-04-06, from https://CRAN.R-project.org/package=gridExtra

Azadkia, M., & Chatterjee, S. (2021, March). *A simple measure of conditional dependence.* arXiv. Retrieved 2023-03-18, from http://arxiv.org/abs/1910.12327 (arXiv:1910.12327 [cs, math, stat])

Azadkia, M., Chatterjee, S., & Matloff, N. (2021, March). *FOCI: Feature Ordering by Conditional Independence.* Retrieved 2023-04-06, from https://CRAN.R-project.org/package=FOCI

Ben-Shachar, M. S., Makowski, D., Lüdecke, D., Patil, I., Wiernik, B. M., Thériault, R., ... Karreth, J. (2023, January). *effectsize: Indices of Effect Size.* Retrieved 2023-04-06, from https://CRAN.R-project.org/package=effectsize

Binder, H., Sauerbrei, W., & Royston, P. (2013). Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Statistics in Medicine*, *32*(13), 2262–2277. Retrieved 2022-11-03, from https://onlinelibrary.wiley.com/doi/10.1002/sim.5639 doi: 10.1002/sim.5639

Blagus, R., & Babic, S. (2017, October). *abe: Augmented Backward Elimination.* Retrieved 2023-04-06, from https://CRAN.R-project.org/package=abe

Briscoe, E., & Feldman, J. (2011, January). Conceptual complexity and the bias/variance tradeoff. *Cognition*, *118*(1), 2–16. Retrieved 2023-03-06, from https://linkinghub.elsevier.com/retrieve/pii/S0010027710002295 doi: 10.1016/j.cognition.2010.10.004

Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008, December). Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, *3*(1), 17. Retrieved 2023-03-19, from https://scfbm.biomedcentral.com/articles/10.1186/1751-0473-3-17 doi: 10.1186/1751-0473-3-17

Callegaro, M. (2013, September). Paradata in Web Surveys. In F. Kreuter (Ed.), *Improving Surveys with Paradata* (pp. 259–279). Hoboken, New Jersey: John Wiley & Sons, Inc. Retrieved 2021-09-07, from https://onlinelibrary.wiley.com/doi/10.1002/9781118596869.ch11 doi: 10.1002/9781118596869.ch11

Chaw, E., Shem, K., Castillo, K., Wong, S., & Chang, J. (2012, October). Dysphagia and Associated Respiratory Considerations in Cervical Spinal Cord Injury. *Topics in Spinal Cord Injury Rehabilitation*, *18*(4), 291–299. Retrieved 2023-02-12, from https://meridian.allenpress.com/tscir/article/doi/10.1310/sci1804-291 doi: 10.1310/sci1804-291

Chen, E. E., & Wojcik, S. P. (2016, December). A practical guide to big data research in psychology. *Psychological Methods*, *21*(4), 458–474. Retrieved 2023-02-14, from http://doi.apa.org/getdoi.cfm?doi=10.1037/met0000111 doi: 10.1037/met0000111

Courvoisier, D. S., Combescure, C., Agoritsas, T., Gayet-Ageron, A., & Perneger, T. V. (2011, September). Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *Journal of Clinical Epidemiology*, *64*(9), 993–1000. Retrieved 2023-02-15, from https://linkinghub.elsevier.com/retrieve/pii/S0895435610004245 doi: 10.1016/j.jclinepi.2010.11.012

Cox, L. A. (2018, September). Modernizing the Bradford Hill criteria for assessing causal relationships in observational data. *Critical Reviews in Toxicology*, *48*(8), 682–712. Retrieved 2023-06-06, from https://www.tandfonline.com/doi/full/10.1080/10408444.2018.1518404 doi: 10.1080/10408444.2018.1518404

Diedenhofen, B., & Musch, J. (2017, August). PageFocus: Using paradata to detect and prevent cheating on online achievement tests. *Behavior Research Methods*, *49*(4), 1444–1459. Retrieved 2021-09-07, from http://link.springer.com/10.3758/s13428-016-0800-7 doi: 10.3758/s13428-016-0800-7

Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., … Schwen, B. (2023, February). *data.table: Extension of 'data.frame'.* Retrieved 2023-04-06, from https://CRAN.R-project.org/package=data.table

Dunkler, D., & Heinze, G. (2022, February). *shrink: Global, Parameterwise and Joint Shrinkage Factor Estimation.* Retrieved 2023-04-06, from https://CRAN.R-project.org/package=shrink

Dunkler, D., Plischke, M., Leffondré, K., & Heinze, G. (2014, November). Augmented Backward Elimination: A Pragmatic and Purposeful Way to Develop Statistical Models. *PLoS ONE*, *9*(11), e113677. Retrieved 2022-11-05, from https://dx.plos.org/10.1371/journal.pone.0113677 doi: 10.1371/journal.pone.0113677

Dunkler, D., Sauerbrei, W., & Heinze, G. (2016). Global, Parameterwise and Joint Shrinkage Factor Estimation. *Journal of Statistical Software*, *69*(8). Retrieved 2022-11-03, from http://www.jstatsoft.org/v69/i08/ doi: 10.18637/jss.v069.i08

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38.

Gauthier, J., Wu, Q. V., & Gooley, T. A. (2020, April). Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians. *Bone Marrow Transplantation*, *55*(4), 675–680. Retrieved 2023-03-01, from https://www.nature.com/articles/s41409-019-0679-x doi: 10.1038/s41409-019-0679-x

Harrell, F. E. (2015a). Describing, Resampling, Validating, and Simplifying the Model. In *Regression Modeling Strategies* (pp. 103–126). Cham: Springer International Publishing. Retrieved 2023-02-15, from https://link.springer.com/10.1007/978-3-319-19425-7_5 (Series Title: Springer Series in Statistics) doi: 10.1007/978-3-319-19425-7_5

Harrell, F. E. (2015b). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* Cham: Springer International Publishing. Retrieved 2023-02-15, from https://link.springer.com/10.1007/978-3-319-19425-7 doi: 10.1007/978-3-319-19425-7

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models.* Boca Raton, Fla: Chapman & Hall/CRC.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction* (2nd ed.).

Hayashi, T., Fujiwara, Y., Ariji, Y., Sakai, H., Kubota, K., Kawano, O., … Maeda, T. (2020, November). Mechanism of Dysphagia after Acute Traumatic Cervical Spinal Cord Injury. *Journal of Neurotrauma*, *37*(21), 2315–2319. Retrieved 2023-03-20, from https://www.liebertpub.com/doi/10.1089/neu.2020.6983 doi: 10.1089/neu.2020.6983

Hayashi, T., Fujiwara, Y., Sakai, H., Maeda, T., Ueta, T., & Shiba, K. (2017, October). Risk factors for severe dysphagia in acute cervical spinal cord injury. *Spinal Cord*, *55*(10), 940–943. Retrieved 2023-02-12, from http://www.nature.com/articles/sc201763 doi: 10.1038/sc.2017.63

Heinze, G., Ambler, G., & Benner, A. (2022, January). *mfp: Multivariable Fractional Polynomials.* Retrieved 2023-04-28, from https://cran.r-project.org/web/packages/mfp/index.html

Heinze, G., & Dunkler, D. (2017, January). Five myths about variable selection. *Transplant International*, *30*(1), 6–10. Retrieved 2023-03-14, from https://onlinelibrary.wiley.com/doi/10.1111/tri.12895 doi: 10.1111/tri.12895

Heinze, G., Ploner, M., Dunkler, D., Southworth, H., & Jiricka, L. (2022, January). *logistf: Firth's Bias-Reduced Logistic Regression.* Retrieved 2023-04-06, from https://CRAN.R-project.org/package=logistf

Heinze, G., Wallisch, C., & Dunkler, D. (2018, May). Variable selection - A review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431–449. Retrieved 2022-11-03, from https://onlinelibrary.wiley.com/doi/10.1002/bimj.201700067 doi: 10.1002/bimj.201700067

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Third edition ed.) (No. 398). Hoboken, New Jersey: Wiley.

Hulsen, T., Jamuar, S. S., Moody, A. R., Karnes, J. H., Varga, O., Hedensted, S., . . . McKinney, E. F. (2019). From Big Data to Precision Medicine. *Frontiers in Medicine*, *6*. Retrieved 2023-02-14, from https://www.frontiersin.org/articles/10.3389/fmed.2019.00034

Hüller, G. (2022). *Paradata: An economic source for conscientiousness indicators?* (Unpublished master's thesis). Paris-Lodron University, Salzburg.

Iruthayarajah, J., McIntyre, A., Mirkowski, M., Welch-West, P., Loh, E., & Teasell, R. (2018, December). Risk factors for dysphagia after a spinal cord injury: a systematic review and meta-analysis. *Spinal Cord*, *56*(12), 1116–1123. Retrieved 2023-03-20, from http://www.nature.com/articles/s41393-018-0170-3 doi: 10.1038/s41393-018-0170-3

Jeppson, H., Hofmann, H., Cook, D., & Wickham, H. (2021, February). *ggmosaic: Mosaic Plots in the 'ggplot2' Framework.* Retrieved 2023-04-06, from https://CRAN.R-project.org/package=ggmosaic

Junker, R. R., Griessenberger, F., & Trutschnig, W. (2019, February). *A copula-based measure for quantifying asymmetry in dependence and associations.* arXiv. Retrieved 2023-04-07, from http://arxiv.org/abs/1902.00203 (arXiv:1902.00203 [stat]) doi: 10.48550/arXiv.1902.00203

Kasper, T., Griessenberger, F., Junker, R. R., Petzel, V., & Trutschnig, W. (2022, December). *qad: Quantification of Asymmetric Dependence.* Retrieved 2023-04-06, from https://CRAN.R-project.org/package=qad

Le Cessie, S., & Houwelingen, J. C. V. (1992). Ridge Estimators in Logistic Regression. *Applied Statistics*, *41*(1), 191. Retrieved 2023-03-08, from https://www.jstor.org/stable/10.2307/2347628?origin=crossref doi: 10.2307/2347628

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, London, New Delhi: SAGE Publications Ltd.

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models* (2nd ed.). London; New York: Chapman and Hall.

Meissner, I., Dietmann, S., Hüller, G., Mach, O., Vogel, M., Ehret, M., . . . Leister, I. (2023). *Identification of risk factors for dysphagia after traumatic cervical spinal cord injury: A retrospective study* [Manuscripft Draft].

Menard, S. (2000, February). Coefficients of Determination for Multiple Logistic Regression Analysis. *The American Statistician*, *54*(1), 17. Retrieved 2023-

02-14, from https://www.jstor.org/stable/2685605?origin=crossref doi: 10.2307/2685605

Myers, R. H., & Montgomery, D. C. (1997, July). A Tutorial on Generalized Linear Models. *Journal of Quality Technology*, *29*(3), 274–291. Retrieved 2023-02-17, from https://www.tandfonline.com/doi/full/10.1080/00224065.1997.11979769 doi: 10.1080/00224065.1997.11979769

Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S., & Mitliagkas, I. (2019, December). *A Modern Take on the Bias-Variance Trade-off in Neural Networks.* arXiv. Retrieved 2023-03-06, from http://arxiv.org/abs/1810.08591 (arXiv:1810.08591 [cs, stat])

Prajapati, B., Dunne, M., & Armstrong, R. (2010). Sample size estimation and statistical power analyses.

Puhr, R., Heinze, G., Nold, M., Lusa, L., & Geroldinger, A. (2017). Firth's logistic regression with rare events: accurate effect estimates and predictions?: R. PUHR *ET AL.* *Statistics in Medicine*. Retrieved 2023-02-14, from https://onlinelibrary.wiley.com/doi/10.1002/sim.7273 doi: 10.1002/sim.7273

Revelle, W. (2023, March). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Retrieved 2023-04-06, from https://CRAN.R-project.org/package=psych

Rinker, T., Kurkiewicz, D., Hughitt, K., Wang, A., Aden-Buie, G., Wang, A., & Burk, L. (2019, March). *pacman: Package Management Tool.* Retrieved 2023-04-06, from https://CRAN.R-project.org/package=pacman

Rohrer, J. M. (2018, March). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. Retrieved 2022-11-03, from http://journals.sagepub.com/doi/10.1177/2515245917745629 doi: 10.1177/2515245917745629

Royston, P., & Sauerbrei, W. (2007, February). Multivariable Modeling with Cubic Regression Splines: A Principled Approach. *The Stata Journal: Promoting communications on statistics and Stata*, *7*(1), 45–70. Retrieved 2023-03-19, from http://journals.sagepub.com/doi/10.1177/1536867X0700700103 doi: 10.1177/1536867X0700700103

Royston, P., & Sauerbrei, W. (2008). *Multivariable Model-Building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables* (1st ed.). Wiley. Retrieved 2023-02-12, from https://onlinelibrary.wiley.com/doi/book/10.1002/9780470770771 doi: 10.1002/9780470770771

Sauerbrei, W., Perperoglou, A., Schmid, M., Abrahamowicz, M., Becher, H., Binder, H., ... Heinze, G. (2020, December). State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. *Diagnostic and Prognostic Research*, *4*(1), 3, s41512–020–00074–3. Retrieved 2023-02-12, from https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-020-00074-3 doi: 10.1186/s41512-020-00074-3

Sauerbrei, W., & Royston, P. (1999, January). Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*(1), 71–94. Retrieved 2023-02-14, from https://onlinelibrary.wiley.com/doi/10.1111/1467-985X.00122 doi: 10.1111/1467-985X.00122

Sauerbrei, W., Royston, P., & Binder, H. (2007, December). Selection of important variables and determination of functional form for continuous predictors in multivariable model building: SELECTION OF VARIABLES AND FUNCTIONAL FORMS. *Statistics in Medicine*, *26*(30), 5512–5528. Retrieved 2023-02-14, from https://onlinelibrary.wiley.com/doi/10.1002/sim.3148 doi: 10.1002/sim.3148

Schaefer, R., Roi, L., & Wolfe, R. (1984, January). A ridge logistic estimator. *Communications in Statistics - Theory and Methods*, *13*(1), 99–113. Retrieved 2023-03-08, from http://www.tandfonline.com/doi/abs/10.1080/03610928408828664 doi: 10.1080/03610928408828664

Shem, K., Wong, J., Dirlikov, B., & Castillo, K. (2019, September). Pharyngeal Dysphagia in Individuals With Cervical Spinal Cord Injury: A Prospective Observational Cohort Study. *Topics in Spinal Cord Injury Rehabilitation*, *25*(4), 322–330. Retrieved 2023-02-12, from https://meridian.allenpress.com/tscir/article/doi/10.1310/sci2504-322 doi: 10.1310/sci2504-322

Shmueli, G. (2010, August). To Explain or to Predict? *Statistical Science*, *25*(3). Retrieved 2023-02-14, from https://projecteuclid.org/journals/statistical-science/volume-25/issue-3/To-Explain-or-to-Predict/10.1214/10-STS330.full doi: 10.1214/10-STS330

Simonsohn, U. (2018, December). Two Lines: A Valid Alternative to the Invalid Testing of U-Shaped Relationships With Quadratic Regressions. *Advances in Methods and Practices in Psychological Science*, *1*(4), 538–555. Retrieved 2021-09-07, from http://journals.sagepub.com/doi/10.1177/2515245918805755 doi: 10.1177/2515245918805755

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016, September). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *11*(5), 702–712. doi:

10.1177/1745691616658637

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., ... Robinson, G. E. (2015, July). Big Data: Astronomical or Genomical? *PLOS Biology*, *13*(7), e1002195. Retrieved 2023-02-14, from https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195 (Publisher: Public Library of Science) doi: 10.1371/journal.pbio.1002195

Van Calster, B., van Smeden, M., De Cock, B., & Steyerberg, E. W. (2020, November). Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Statistical Methods in Medical Research*, *29*(11), 3166–3178. Retrieved 2023-02-14, from http://journals.sagepub.com/doi/10.1177/0962280220921415 doi: 10.1177/0962280220921415

Verweij, P. J. M., & Van Houwelingen, H. C. (1993, December). Cross-validation in survival analysis. *Statistics in Medicine*, *12*(24), 2305–2314. Retrieved 2023-03-07, from https://onlinelibrary.wiley.com/doi/10.1002/sim.4780122407 doi: 10.1002/sim.4780122407

Wallisch, C., Dunkler, D., Rauch, G., Bin, R., & Heinze, G. (2020). Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling. *Statistics in Medicine*, *40*(2), 369–381. Retrieved 2023-02-11, from https://onlinelibrary.wiley.com/doi/10.1002/sim.8779 doi: 10.1002/sim.8779

Wickham, H., & RStudio. (2023, February). *tidyverse: Easily Install and Load the 'Tidyverse'*. Retrieved 2023-04-06, from https://CRAN.R-project.org/package=tidyverse

Wolf, C., & Meiners, T. H. (2003, June). Dysphagia in patients with acute cervical spinal cord injury. *Spinal Cord*, *41*(6), 347–353. Retrieved 2023-02-12, from http://www.nature.com/articles/3101440 doi: 10.1038/sj.sc.3101440

Wolpert, D., & Macready, W. (1997, April). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. Retrieved 2023-03-16, from http://ieeexplore.ieee.org/document/585893/ doi: 10.1109/4235.585893

Wood, S. N. (2017). *Generalized Additive Models* (2nd ed.). Chapman & Hall/CRC.

# A Larger Pool of Predictor Candidates for FOCI

The following list describes variables and their coding, which are at least temporally associated with the (non-)manifestation of dysphagia and are based on domain expertise. This pool also contains the predictor candidates assigned to the *global model* in Eq. 21 among other variables, where no literature or experience directly suggests their influence on the response.

### Variable Description

- **Dysphagia**: a swallowing disorder; the binary response variable
- **TMS**: the (continuous) Total Motor Score (TMS) of the International Standards for Classification of Spinal Cord Injury Motor Score (ISNCSCI); used as an inverted proxy for injury severity
- **NLI.High**: dichotomized Neurological Level of Injury (NLI); NLI.High = 1 indicates a high neurological level of injury on the vertebrae C1-C4, NLI.High = 0 a lower neurological level of injury (i.e., C5 and lower)
- **Age**: the patient's age at the time of the accident
- **OP**: binary indicator if patient had a surgery (i.e., OP = 1: ”yes”)
- **OP.Ventral**: dichotomized surgery access variable; OP.Ventral = 1 indicates either solely ventrally accessed surgery or a combined ventral and dorsal access, OP.Ventral = 0 indicates solely dorsal access
- **No.Fusions**: the (continous) number of fusioned vertebrae in a patient
- **Tracheostomy**: binary indicator if a patient was tracheostomized (i.e., Tracheostomy = 1: ”yes”)
- **T.Surgery**: binary indicator how the tracheostomy was applied; T.Surgery = 1 indicates a surgical tracheostomy, while T.Surgery = 0 indicates a dilatative tracheostomy
- **Sex**: (binary) indicator of a patient's gender; Sex = 0 indicates male patients, Sex = 1 females, and Sex = 2 divers patients, although there were no diverse cases in our sample
- **Ventilation**: binary indicator if a patient received ventilation (i.e., Ventilation = 1: ”yes”)

- **AIS**: the ordinal score of the *Asia Impairment Scale*; with AIS = 1: grade A, i.e., complete impairment - no motor or sensory function is left below the level of injury;...; AIS = 5: grade E, i.e., no impairment - all motor and sensory functions are unhindered; AIS = 6: despite earlier neurological impairments (i.e., lower AIS), a remission was observed and currently no impairments are left; AIS = 7: missing data

- **OP:OP.Ventral:PlateThick**: the continuous thickness of plates, which were used to fuse vertebrae of the spinal cord, in case of a ventral surgery. The plate-thickness were observed in groups of 1.45mm indicating plates with thickness 1.4-1.5mm; 2.0mm; and 2.45mm indicating plates of thickness 2.4-2.5mm

# B Histograms of Subsampling Distributions

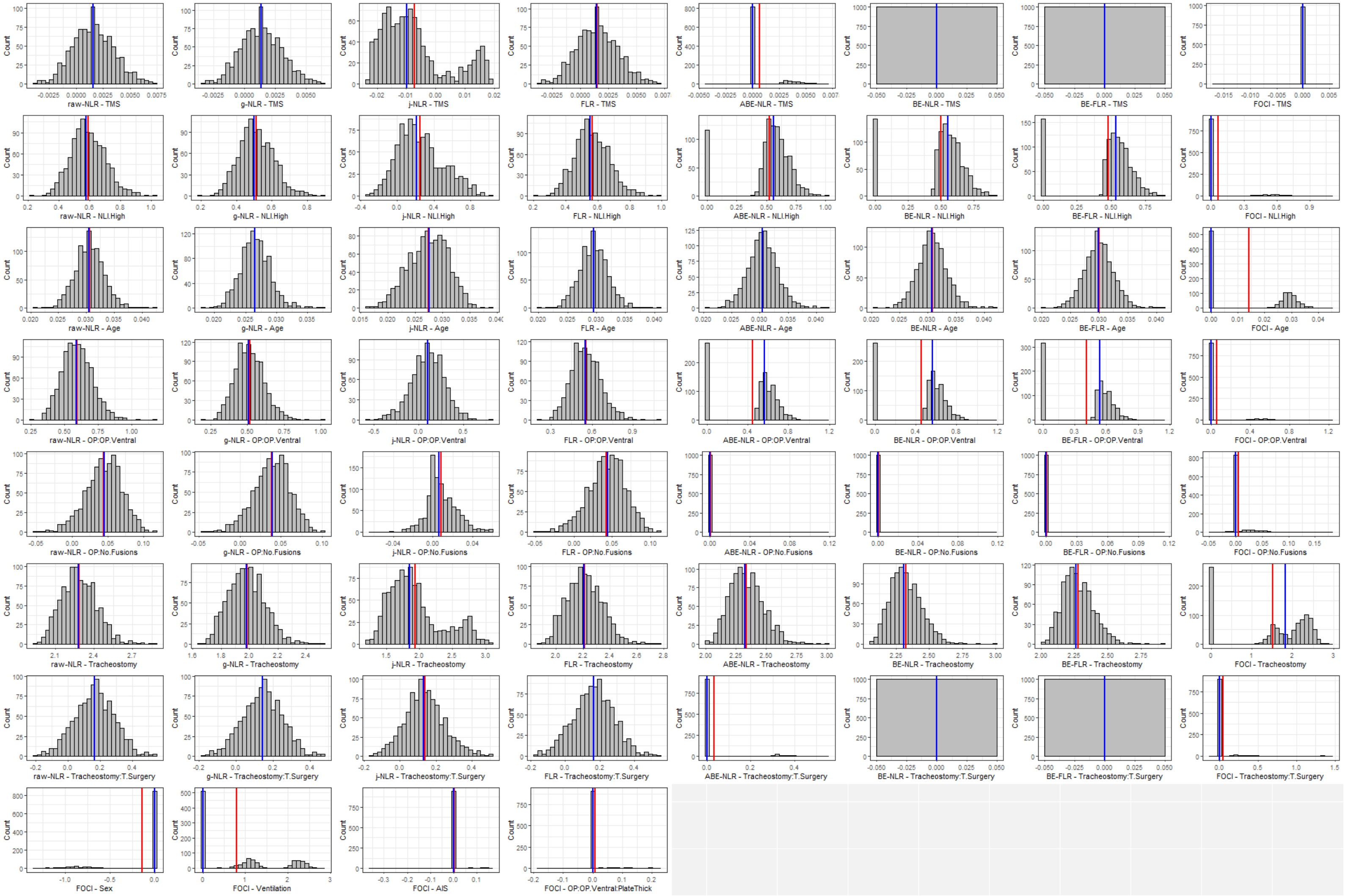

Figure 5: Histogram for each coefficient subsampling distribution stratified for variables and models. The red vertical line represents the arithmetic mean and the blue vertical line the median of the distribution.