Technische Universität München

Fakultät für Mathematik
Lehrstuhl für Mathematische Modelle biologischer Systeme

Bayesian information criterion approximations for model selection in multivariate logistic regression with application to electronic medical records

Katharina Selig

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende: Prof. Dr. Silke Rolles

Prüfer der Dissertation: 1. Prof. Donna P. Ankerst, Ph.D.
2. Associate Prof. Pamela A. Shaw, Ph.D.

Die Dissertation wurde am 17.12.2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 16.03.2020 angenommen.

Summary

We prove that the Bayesian information criterion (BIC) approximates the Bayes factor for the multivariate logistic regression model to order $O_p(1)$ for arbitrary prior distributions and $O_p(n^{-\frac{1}{2}})$ for the unit information prior distribution. Thereby, we show the construction of null-orthogonal parametrizations. Further, we provide a framework in `R` for the Bayesian multivariate logistic regression with different covariates for different multivariate outcomes. Using univariate and multivariate logistic regression models we develop a risk prediction model for five adverse pathological outcomes after prostatectomy based on electronic medical records data.

Zusammenfassung

Wir beweisen für die multivariate logistische Regression, dass für beliebige A-priori-Verteilungen das Schwarz-Bayes-Informationskriterium den Bayes Faktor mit einer Ordnung von $O_p(1)$ approximiert und für die unit information A-priori-Verteilung mit einer Ordnung von $O_p(n^{-\frac{1}{2}})$. Dabei zeigen wir, wie eine Null-orthogonale Parametrisierung konstruiert werden kann. Eine Implementierung der Bayes'schen multivariaten logistischen Regression mit unterschiedlichen unabhängige Variablen für die Modellierung verschiedener abhängiger Größen wird für `R` zur Verfügung gestellt. Basierend auf elektronischen Patientendaten, erstellen wir mit univariaten und multivariaten logistischen Regressionmodellen ein Risikoprognosemodell für fünf negative pathologische Befunde.

Acknowledgements

To begin with, I want to thank my supervisor Prof. Donna Ankerst. I am deeply thankful for your constant support and valuable advice regarding my thesis as well as other statistical projects. Thank you for giving me the freedom and encouragement to pursue these projects and for making my dissertation possible.

Next, I would like to thank Prof. Pamela Shaw for taking the time to read and assess this thesis and Prof. Silke Rolles for taking the chair of the examination committee.

Additional thanks for financial support are directed towards the TUM Graduate School, the International School of Applied Mathematics, and the Global Challenges for Women in Math Science Program.

My sincere thanks go to all colleagues from the mathematics department. Thank you for your scientific and non-scientific support and joyful distractions during relaxing coffee breaks. I will preserve many enjoyable moments of the time spent in Garching.

Finally, I want to thank my friends and family for their support. In particular, my thanks go to my parents, Helena, and Jan for being a constant source of strength and encouragement. You have made this whole project possible.

Contents

1 Motivation and outline

The use and availability of electronic medical records (EMR) data for clinical research have been rising over the last years [1]. EMR data can be used to develop risk prediction models for a variety of outcomes, such as hospital readmission, mortality, and heart failure, to name just a few [2–12].

We develop a prediction model for five adverse pathological outcomes after prostatectomy based on EMR data. Given that 17,227 patients underwent a prostatectomy in Germany in 2017 this is highly relevant from a medical perspective [13]. Inspired from the data setting we take two different, but related approaches to modelling. We use separate univariate logistic regression and compare the results to a multivariate logistic regression that accounts for correlations among the outcomes. We prefer logistic regression models in this setting over other machine learning approaches as the odds ratios (OR) are easy to interpret.

Unfortunately, in the medical literature multivariable models are often misspecified as multivariate models [14]. We refer to multivariate models as those with several outcomes and univariate models as those with only one outcome. We sometimes refer to univariate models as multivariable models to indicate multiple predictors.

We are specifically concerned with model selection to develop the prediction models. Often automatic step-wise Bayesian information criterion (BIC) methods are used and one might forget that the BIC is an approximation to the Bayes factor (BF) and provides a consistent model selection procedure for the case of nested models. Kass and Vaidyanathan (1992) showed the approximation of the BF for testing the equality of two binomial distributions, stating the extension to univariate logistic regression models without proof [15]. We provide a proof for this approximation for specific and arbitrary prior distributions for univariate logistic regression and extend it to the multivariate setting.

Separate risk prediction models for adverse pathological outcomes following prostatectomy have been developed over time [16–45]. However, to our knowledge, none of them account for the correlation among all outcomes with a multivariate model. Many of the prediction models are based on retrospective data or clinical studies, and although EMR data are already used for risk models regarding diagnosis of prostate cancer or long-term outcomes, they are comparably rare for predicting adverse pathological outcomes [19, 46–49].

We present more information about the medical context and describe the EMR data provided by the Martini Klinik, Hamburg, Germany in the next chapter. After setting the mathematical foundations in Chapter 3, we get to the heart of the thesis. In Chapter 4 we prove the approximation of the BF for arbitrary prior distributions and the unit information prior distribution. Thereby, we show and prove how we can construct a null-orthogonalization of a set of parameters. In simulation studies we show the results and finally, we apply the theory to the EMR data, where we develop univariate logistic regression models for the prediction of adverse pathological outcomes. After that, we extend the theory developed for the univariate logistic regression model to the multivariate case in Chapter 5. We discuss shortly the technicalities of the implementation of the Bayesian multivariate model and then apply it to the EMR data. Finally, we compare the results of the univariate and multivariate approach for predicting outcomes in a separate validation set.

2 Medical context and description of electronic medical records data

In the following, we give some background on the medical setting and describe the electronic medical records (EMR) data set used for analysis. Further, we discuss how missing values are addressed and how the data are split for training and evaluation of the models.

2.1 Medical context and relevance

Prostate cancer is the most frequently diagnosed cancer for males in western Europe, North and South America, Australia, and large parts of Africa [50]. The detection of prostate cancer was highly influenced by the introduction of prostate-specific antigen (PSA) screening in the late 1980s [50]. PSA is a protein that is produced by the prostate gland and measured in the blood [51]. Increased levels of PSA might indicate the presence of prostate cancer, but other benign prostate conditions, such as prostatitis or benign prostatic hyperplasia (BPH), can cause a rise in PSA levels as well [51].

Most patients suspected of having prostate cancer undergo a biopsy, where several probes, so-called cores, of tissue are sampled from the prostate and then examined by a pathologist [51]. Cores that show signs of cancer are classified using Gleason grading, a grading system introduced by Donald F. Gleason in 1966 [52]. This scheme was updated over the years, aiming at a reproducible classification [52, 53]. A Gleason score consists of the sum of primary and secondary Gleason grades representing the most common classification and the highest classification among the remaining patterns [54, 55]. Figure 2.1 displays the five different Gleason patterns defined at the 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma [52]. There, pathologists agreed that a Gleason score of 4 should rarely, if ever, be diagnosed at biopsy [52]. Albertsen

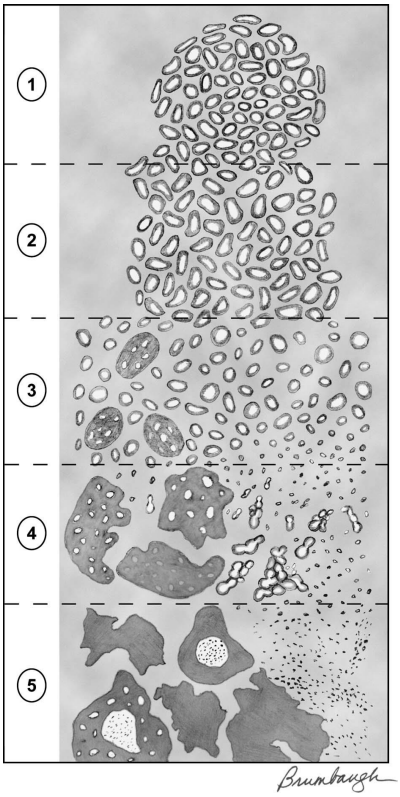

Figure 2.1 Schematic display of Gleason grading [52].

et al. (2005) compared Gleason grading over a decade and found that there has been an upward shift leading to higher classified biopsy samples [56].

Recent advances encourage the use of a different grading system, which includes five distinct grade groups based on the Gleason score [57]. However, we base the analysis separately on the primary and secondary Gleason grade, since most of the EMR data for this thesis was collected before the introduction of that grading scheme, and the results can be translated to fit into the five grading groups.

After diagnosis of prostate cancer, several treatment options depending on the severeness of the tumor are available [54]. The most common are radical prostatectomy, which is the removal of the prostate gland, radiation or radiopharmaceutical therapy, hormonal therapy, and watchful waiting or active surveillance (AS). However, the last two options are strategies that do not immediately employ a therapy [54]. Unfortunately, the distinction between watchful waiting and AS is not always clear in the literature [54]. While with watchful waiting, curative therapy is not attempted at any point and the focus lies on palliative care, AS is a strategy to avoid or delay therapy, while regularly following patients and attempting curative therapy once tumor progression is indicated [54, 58]. Protocols for surveillance and criteria for interventions vary across institutions [54, 59].

After treatment, patients are monitored to detect biochemical recurrence (BCR), which denotes an increase of PSA level in the blood [60]. After prostatectomy, the PSA level should be very low or undetectable and BCR might indicate recurrence of cancer, but not all patients experience clinical symptoms of relapse after BCR [54, 60].

The focus of this thesis is on patients that underwent prostatectomy at the Martini Klinik, Hamburg. We aim to predict five adverse pathological outcomes that are used to describe disease severeness or aggressiveness of the patient's tumor. Knowledge of these outcomes aids in the decision on further treatments after surgery, such as radiation or hormonal therapy. In the following, we describe the five pathological outcomes in detail and their prognostic or therapeutic relevance.

Extracapsular extension or extraprostatic extension (ECE) is defined by tumor growth beyond the normal boundaries of the prostate gland [61]. In the presence of ECE, patients have a higher chance of BCR and thus a poorer prognosis for survival after radical prostatectomy compared to patients without ECE [62–68]. Several models and nomograms have been developed to predict ECE before prostatectomy using patient characteristics, information available through biopsy, as well as magnetic resonance imaging (MRI) [16–26, 45].

Seminal vesicle invasion (SVI) is the infiltration of the muscular wall of the seminal vesicles by the tumor [62, 69–71]. SVI can occur as a special case of ECE, by distant metastases, or by infiltration of the ejaculatory duct followed by an invasion of the muscular wall of the seminal vesicles [72]. In our analysis, we only consider the most common route of invasion, which is a special case of ECE [72]. It has been shown that patients with SVI have a poorer prognosis after radical prostatectomy [62, 66, 73, 74]. However, Sapre et al. (2012) suggested that SVI is a surrogate marker for more aggressive tumors [70]. Seminal vesicles are commonly removed during a radical prostatectomy, which has adverse effects on urinary and erectile function [29]. Thus, there is a great interest in predicting SVI before prostatectomy, leading to several models considering pre-surgery patient characteristics, biopsy information and MRI data [26–31, 45].

Lymph node involvement (LNI) is defined as the metastasis of the tumor to the lymph nodes [75]. Diagnosis of LNI highly depends on the extent of pelvic lymph node dissection performed at radical prostatectomy [76]. Patients with LNI have an increased estimated risk for cancer-specific mortality [66, 77]. Careful pelvic lymph node dissection requires skilled surgeons and is time-consuming. Not all patients have the same risk of LNI, thus several nomograms have been developed for prediction [26, 32–38, 45, 76].

Lymphovascular invasion (LVI) is the presence of tumor cells within the lymphatic vessels [78]. It has been reported that patients with LVI have a poorer prognosis, although not all studies found LVI to be a significant predictor in a multivariable analysis [79–86]. Nevertheless, there are indications that LVI significantly increases the risk of LNI and distant metastasis [86, 87]. Risk prediction models for LVI are not as common as for the other adverse pathological outcomes.

Primary Gleason grade higher than 3 (PGG) is an indicator of whether the most prevalent pathological Gleason grade is greater than 3. There is a positive correlation between the Gleason score at biopsy and the Gleason score determined after prostatectomy. However, at biopsy, only a small sample of the prostate is examined, whereas after prostatectomy a more precise and complete grading can be performed. Therefore, cases of so-called Gleason upgrading or downgrading are not uncommon, meaning that the pathological Gleason score after surgery is higher or lower than the one determined after biopsy [39–43]. Studies have shown that patients with a primary Gleason grade higher than 3 after prostatectomy have a higher risk of BCR and worse patient survival [45, 88–93]. Thus, to improve informed decision making there have been several articles investigating Gleason score up- and down-grading and aiming to predict the pathological primary Gleason grade [39–44, 94].

Although separate risk prediction models for the five adverse pathological outcomes are available, they do not account for the correlation among them and, to our knowledge, there exists no multivariate model. We develop a multivariate risk prediction model for the five adverse pathological outcomes after prostatectomy and compare the performance to univariate models that do not account for correlation among outcomes.

2.2 Data cleaning and description

The data that we use throughout this thesis are provided by the Martini Klinik, Hamburg. The original data are three different data sets that were extracted from the EMR database and contain records for patients from 1992 to 2016. Figure 2.2 provides an overview over the different data sets and the data cleaning process. The three data sets are Pathological data, PSA data, and Biopsy data.

Pathological data contains prostatectomy information on 24,335 patients. We exclude 5 records as they either contain duplicated information or are missing the date of prostatectomy. We extract the pathological outcomes ECE, SVI, LVI, LNI, and PGG from the pathological diagnosis and additional columns. This data set only contains information available after the surgery, thus everything apart from the outcomes, the date of surgery and the patient ID is discarded.

PSA data stores PSA measurements with the corresponding dates as well as the date of surgery and the patient ID of 48,507 patients. Those without a date of surgery are excluded, as we are only interested in patients who underwent a prostatectomy. Further, records with invalid or missing PSA values are discarded as well as duplicated ones. Although additional information on specific PSA measurements, such as the percentage of free PSA, is available for some patients, overall the number of missing values is very large for these specific variables in accordance with the lack of routine use of these markers in the clinic. Thus, they are not considered further. After cleaning the data we obtain 84,177 PSA values from 24,242 patients. We join the PSA data with the Pathological data using the date of surgery and the patient ID to identify matching records.

Biopsy data contains information about biopsies for 24,335 patients and accompanying values, such as primary and secondary Gleason grades, number of cores, number of positive cores, prostate volume, patient's height and weight, PSA value at biopsy and family history of prostate cancer. Records with missing date of biopsy are excluded, as well as duplicated information. For 3,152 patients we have more than one biopsy in the data set, but we use only the most recent one. We join the Biopsy data to the other data sets using the patient ID to identify matching records.

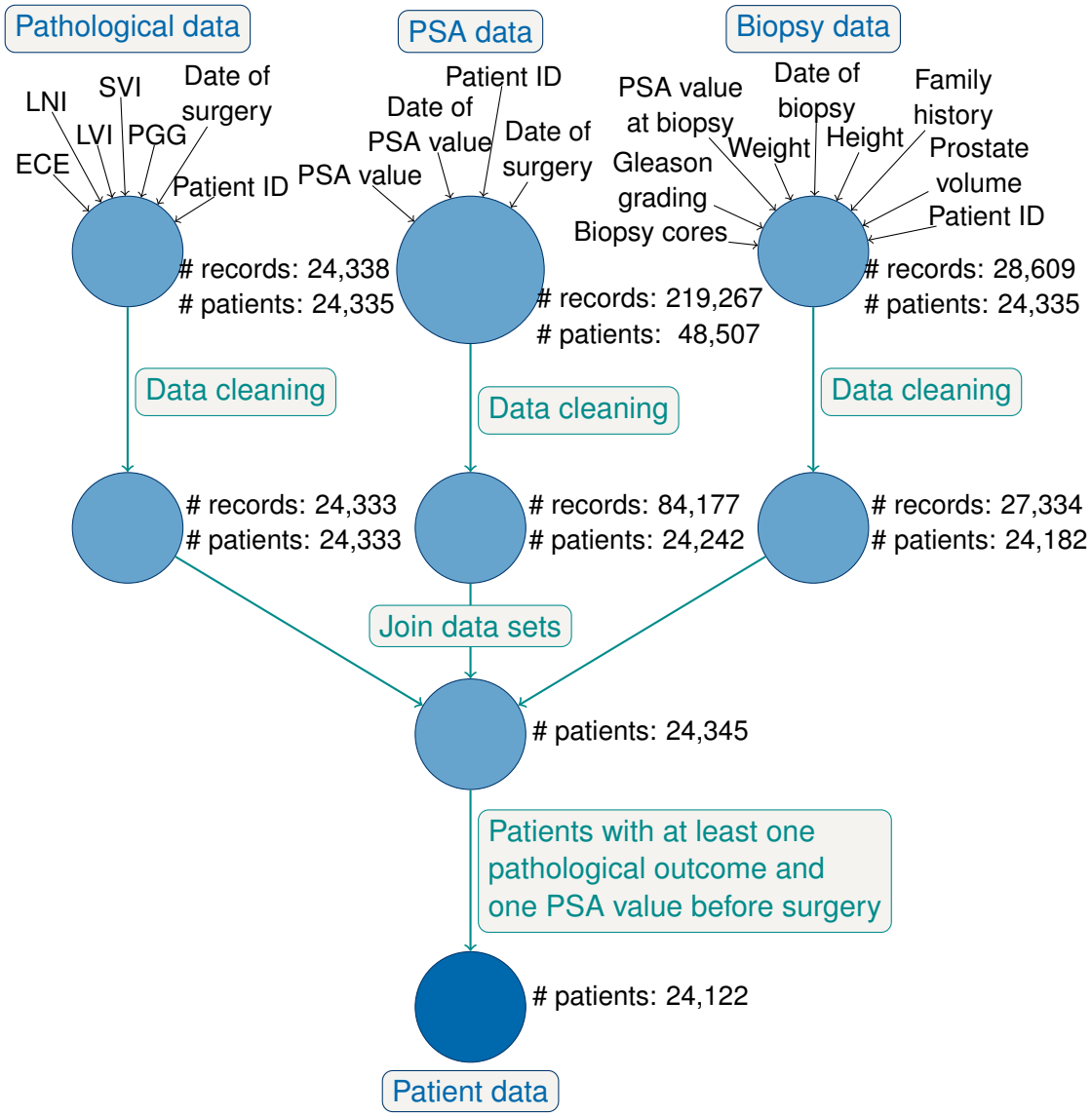

Figure 2.2 Overview of Pathology, PSA, and Biopsy data sets combined for analysis.

After we join the three data sets, we exclude 223 patients without at least one PSA value before surgery or without any pathological outcome information. In the following, we provide more detail about the final data set and additional features that are extracted from the available data.

N (%)	ECE	SVI	LVI	LNI	PGG
No	15,868 (65.8)	20,861 (86.5)	15,425 (63.9)	15,114 (62.7)	18,393 (76.2)
Yes	8,235 (34.1)	3,219 (13.3)	2,619 (10.9)	2,083 (8.6)	5,650 (23.4)
Missing	19 (0.1)	42 (0.2)	6,078 (25.2)	6,925 (28.7)	79 (0.3)

Table 2.1 Distribution of pathological outcomes with number of missing values for 24,122 patients.

ECE with an overall prevalence of 34.1% is the most common adverse pathological outcome among 24,122 prostatectomy patients that are available after initial data processing. The second most prevalent pathological outcome is PGG with 23.4%, followed by SVI with 13.3%. The lowest prevalence outcomes are LVI and LNI with 10.9% and 8.6%, respectively (Table 2.1).

Figure 2.3 shows that information on LVI is only available for patients with a date of prostatectomy after 2001. Before 2005 the available information about LVI is very sparse with 4 non-missing cases and thus the prevalence for the non-missing cases jumps from an unrealistic 0% to 100% from 1 non-missing case each, before leveling to more realistic rates. Overall we can see that the prevalence of ECE, SVI, and PGG decreases from 1992 to 2003 and then increases again. Prevalence rises more for PGG compared to ECE and SVI. This could be due to upward migration in Gleason grading discussed in Section 2.1.

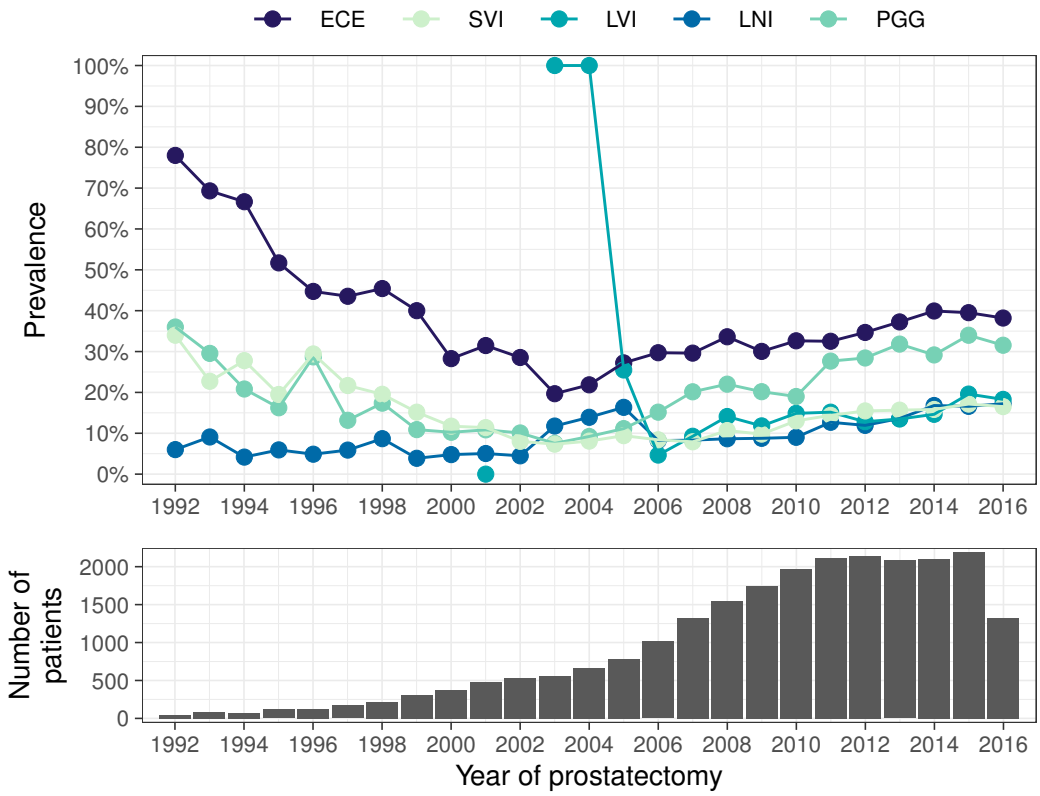

Figure 2.3 Prevalence of each pathological outcome per year of prostatectomy considering non-missing cases and total number of prostatectomies per year.

The number of prostatectomies in the data set is low from 1992 to 2000 but steadily rises to approximately 2,000 per year by 2011. The data were exported in August 2016 thus less data are available for that year.

Despite the risk of including data that are not reflective of today's practice, we use all available patient records and pool the data over years. Thus, the initial missingness of LVI information

does not have such a huge impact as well as prostatectomies with a higher prevalence of ECE. In practice, for example, implementation in a live database, a moving window approach that only considers the most recent years would be more appropriate [95, 96]. However, this is not the focus of this thesis.

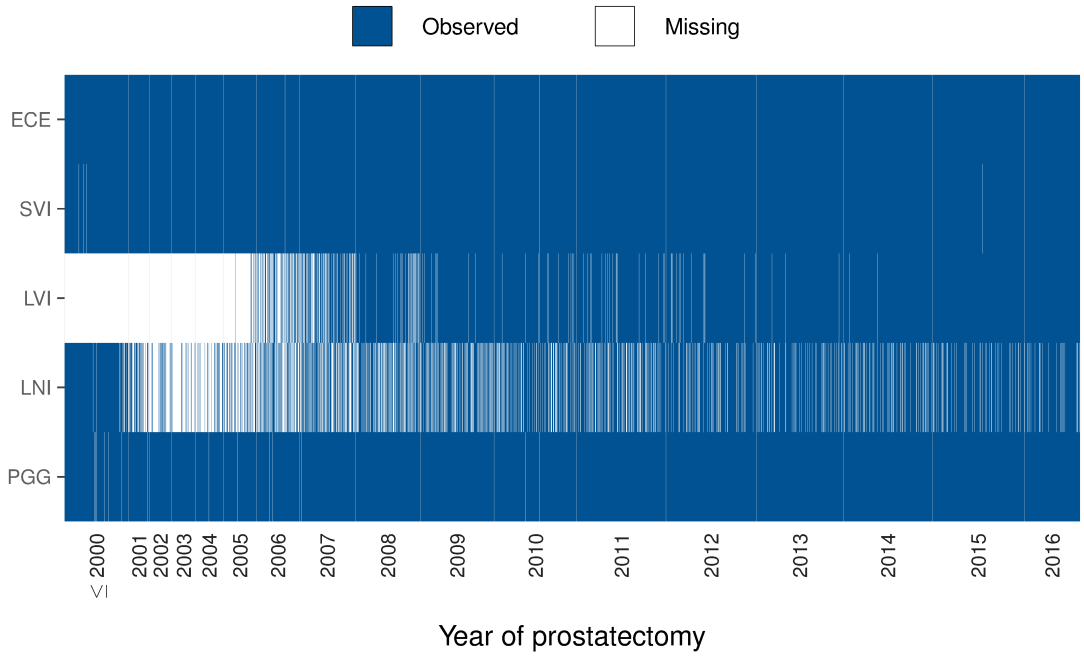

Figure 2.4 Indicator matrix for whether a pathological outcome is missing ordered by year of prostatectomy for 24,122 patients.

As the overall number of prostatectomies in the data increases, the percentage of missing data decreases over the years, which can be seen in Figure 2.4. Especially the number of missing values for LVI decreases tremendously after 2007, whereas for LNI the decrease is not as strong and the missing values are spread over the different years. For the other pathological outcomes ECE, SVI, and PGG, the number of missing values overall is less than or equal to 0.3% and thus negligible (Table 2.1).

We now take a closer look at the available patient characteristics and medical information of the patients in the data set. First, we consider the continuous variables, which we summarize in Table 2.2 and visualize in Figure 2.5.

Patients are between 32 and 81 years old at prostatectomy, with a median age of 64. We use the patient's height and weight to calculate the body mass index (BMI) that is given by $\text{BMI} = \frac{\text{weight [kg]}}{(\text{height [m]})^2}$. The BMI aids in classification of patients into categories from underweight to obese according to the World Health Organization (WHO) [97]. In the data, the BMI values range from 15.4 $^{\text{kg}}/_{\text{m}^2}$ to 51.2 $^{\text{kg}}/_{\text{m}^2}$ and the most common category is overweight. We use the classification of BMI for visualization, but include the values as a continuous covariate in the modeling process.

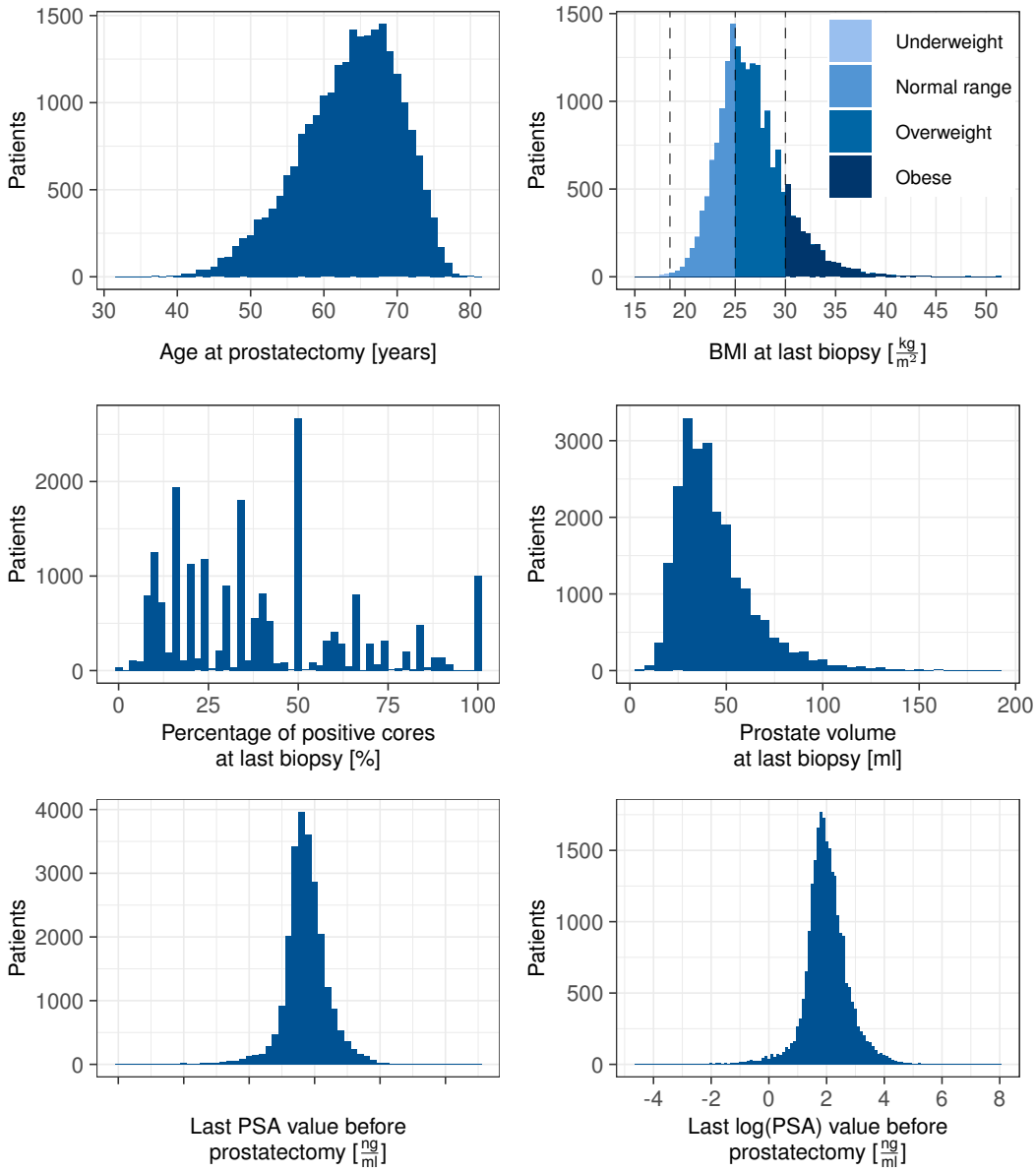

Figure 2.5 Histograms of continuous patient characteristics along with medical information on 24,122 prostatectomy patients. 18 patients (0.1%) with a prostate volume larger than 200 ml are not displayed.

From the number of cores and number of positive cores at the last biopsy, we calculate the percentage of positive cores and use this as a continuous covariate for the analysis. 33.3% is the median percentage of positive cores retrieved at biopsy and the median prostate volume is 40 ml. For 18 patients (0.1%) a prostate volume of 200ml or higher was reported and, again, we do not exclude these possible outliers. As the distribution of prostate volume is right-skewed, we additionally consider the $\log$-transformation of prostate volume for modeling, which also mitigates the impact of outliers.

The median PSA value before prostatectomy is 7.1 $^{ng}/_{ml}$ and the values range from 0.01 $^{ng}/_{ml}$

	Median	Range	Missing (%)
Age at prostatectomy [years]	64.0	32 - 81	2 (0)
BMI at last biopsy [kg/m^2]	26.2	15.4 - 51.2	5,039 (20.9)
Last PSA value before prostatectomy [ng/ml]	7.1	0.01 - 3,022	0 (0)
Percentage of positive cores at last biopsy [%]	33.3	0 - 100	3,778 (15.7)
Prostate volume at last biopsy [ml]	40.0	3.4 - 420	1,054 (4.4)

Table 2.2 Summary of continuous patient characteristics and medical information on 24,122 prostatectomy patients.

to 3,022 ng/ml. This illustrates one issue we encounter with EMR data: suspicious records that cannot be retrospectively checked and corrected. A PSA value of 0.01 ng/ml before prostatectomy is very unlikely. After prostatectomy, PSA typically drops to this level so it seems likely the incorrect PSA was entered. PSA could also be influenced by factors not available in the data, such as hormonal or radiation therapy before prostatectomy. Urologists can not provide a lower cut-off for a plausible PSA value. Overall there are 195 patients (0.8%) in the data with a PSA value less than 0.5 ng/ml. As we cannot investigate these cases further and the percentage is small, we decide not to exclude them. Similarly, we do not exclude 7 patients with a PSA value above 1000 ng/ml. For the data analysis, we transform the PSA value using the natural logarithm and obtain a Gaussian-like distribution (Figure 2.5).

As discrete and categorical covariates we mainly consider information available after biopsy, such as primary and secondary Gleason grade, and the corresponding Gleason score. For 123 patients (0.5%) we do not have information about a biopsy and for 2,601 patients more than one biopsy in the data set is available. Among patients with multiple biopsies, the maximum Gleason score differed from the Gleason score of the last biopsy before prostatectomy for only 57 patients (2.2%). The database manager from Martini Klinik, Hamburg confirmed that we can assume that data entered for the last biopsy are more complete and correct, therefore we use the latest biopsy (Dirk Pehrke, Martini Klinik, Hamburg, personal conversation, February 24, 2017). Results for the 24,122 prostatectomy patients included in the analysis are summarized in Table 2.3.

For 293 patients the result of the last biopsy is missing or negative. While this may seem impossible, it can happen, since the data contain patients that were under AS. AS patients sometimes receive several biopsies and even when they have already been diagnosed with prostate cancer, a biopsy can be negative when the biopsy needle hits an area without cancer cells [98]. It is also possible that a mistake was made during data entry, but unfortunately, we cannot retrieve information about this.

Biopsy result									
Missing	negative	positive							
256	37	23,829							
(1.1%)	(0.2%)	(98.8%)							
Number of cores taken									
Missing	≤ 5	6 - 7	8 - 9	10 - 11	12 - 13	≥ 14			
3,030	634	5,687	3,021	5,886	4,111	1,753			
(12.6%)	(2.6%)	(23.6%)	(12.5%)	(24.4%)	(17.0%)	(7.3%)			
Number of positive cores									
Missing	0	1	2	3	4	5	6	7	≥ 8
2,743	37	4,428	4,289	3,686	2,831	1,987	1,542	796	1,783
(11.4%)	(0.2%)	(18.4%)	(17.8%)	(15.3%)	(11.7%)	(8.2%)	(6.4%)	(3.3%)	(7.4%)
Primary Gleason grade					Secondary Gleason grade				
Missing	1 - 2	3	4	5	Missing	1 - 2	3	4	5
484	305	17,206	5,602	525	484	366	13,177	8,811	1,284
(2.0%)	(1.3%)	(71.3%)	(23.2%)	(2.2%)	(2.0%)	(1.5%)	(54.6%)	(36.5%)	(5.3%)
Gleason score									
Missing	≤ 4	5	6	7	8	9	10		
481	92	481	9,889	9,888	1,911	1,206	174		
(2.0%)	(0.4%)	(2.0%)	(41.0%)	(41.0%)	(7.9%)	(5.0%)	(0.7%)		

Table 2.3 Summary of discrete and categorical covariates of the last biopsy before prostatectomy for 24,122 prostatectomy patients.

The number of cores that were taken at biopsy follows an expected pattern. Six, eight, ten and twelve are the most common numbers of cores. Sometimes physicians decide to take an additional core and thus there are a few biopsies with seven, nine, eleven or thirteen cores. As the number of positive cores also depends on the number of cores that were taken, we focus on the percentage of positive cores already discussed.

The most common primary and secondary Gleason grades are three and the most common total Gleason scores are six and seven, with 3 + 4 the most prevalent combination. There are few cases with a Gleason grade lower than three, but these occur less often in the more recent years and not after 2013, which might be due to the change in Gleason grading after the 2005 ISUP consensus conference on Gleason grading of prostatic carcinoma (Section 2.1)[52].

Family history is missing for 69.3% of the patients. We see in Table 2.4 that the Martini Klinik started collecting the family history for patients in 2012, as it is missing with few exceptions before that year. We have no clear information about the degree of family history, which means the degree of relationship to the family member that was diagnosed with prostate cancer. In the following, we do not further consider family history as a covariate, due to two reasons. First, we would like to split the data into training, validation, and test set as we describe in Section 2.5 using the year of prostatectomy as a split variable. This results in training data that contain only a few patients with information about family history. Although we deal with missing data in Section 2.6, the lack of information is too large in this case. Second, in the literature, family history does not seem to be a significant factor for prostate cancer progression [99].

	Family history of prostate cancer		
Year of prostatectomy	No	Yes	Missing
≤ 2011	4 (0.0%)	2 (0.0%)	14,281 (100.0%)
2012	235 (11.0%)	68 (3.2%)	1,837 (85.8%)
2013	1,373 (66.0%)	345 (16.6%)	363 (17.4%)
2014	1,583 (75.4%)	366 (17.4%)	150 (7.1%)
2015	1,716 (78.4%)	412 (18.8%)	62 (2.8%)
2016	1,021 (77.1%)	276 (20.8%)	28 (2.1%)

Table 2.4 Summary of family history of prostate cancer for 24,122 prostatectomy patients.

Similar to Figure 2.4 we consider missing values of continuous and discrete covariates in Figure 2.6. The completeness of data increases over time and especially BMI values at last biopsy are increasingly available for patients who underwent a prostatectomy after 2004, as well as the number and percentage of positive cores for prostatectomy patients after 2007.

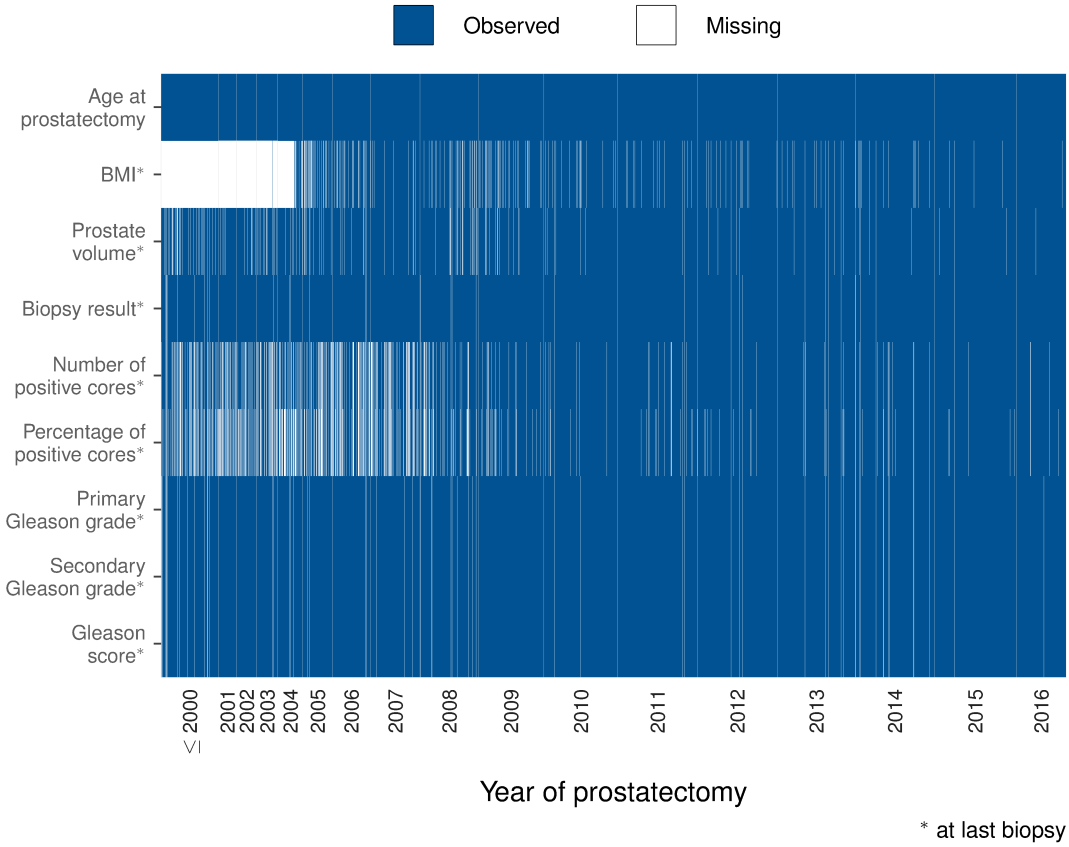

Figure 2.6 Indicator matrix for whether a covariate is missing ordered by year of prostatectomy for 24,122 prostatectomy patients.

2.3 Feature engineering

In the previous section, we discussed covariates that are readily available in the data with commonly used transformations, such as log-transformed PSA ($\log(\text{PSA})$) and BMI to summarize patient's height and weight. In the following, we extract additional covariates using the longitudinal measurements of PSA values. The extraction and transformation of covariates is called feature engineering and is an important part of machine learning. It often requires domain knowledge to create useful features that optimize learning [100]. We use transformations of the longitudinal PSA values that have been proposed in the literature and extend these further.

1	2	3	4	5	6	≥ 7
3,357	9,618	4,178	2,092	1,325	867	2,685
(13.9%)	(39.9%)	(17.3%)	(8.7%)	(5.5%)	(3.6%)	(11.1%)

Table 2.5 Distribution of the number of PSA values before prostatectomy for 24,122 prostatectomy patients.

For 20,765 patients (86.1%) more than one PSA value prior to prostatectomy is available and for 11,147 patients (46.2%) more than two PSA values are available (Table 2.5). PSA measurements are irregularly spaced and the number of PSA values differs for each patient (Table 2.5, Figure 2.7).

Figure 2.7 PSA trajectories on a log-scale for a random subset of 15 patients with at least 4 PSA measurements within 3 years before prostatectomy.

PSA velocity (PSAV) and doubling time (PSADT) have been widely used in the literature to predict prostate cancer-specific mortality, prostatectomy outcome, biochemical recurrence after prostatectomy, and disease progression in untreated patients [101–104]. However, the independent prognostic contribution of these so-called PSA dynamics to the standard static clinical risk factors is in dispute [105–110]. We use definitions of PSAV and PSADT that were reviewed by O'Brien et al. (2009), summarized in Table 2.6, and apply them to PSA measurements before prostatectomy [105].

For calculating PSAV and PSADT we use a linear regression of PSA or $\log(\text{PSA})$ on time. The model using $\log(\text{PSA})$ is

$$\log(\text{PSA}) = \beta_0 + \beta_1 \cdot \text{time}.$$

PSAV is then the slope β_1 of this model or the equivalent model for PSA, when time is measured in years. For PSADT we calculate time in months it takes for the PSA level to double given by

$$\log(2 \cdot \text{PSA}) - \log(\text{PSA}) = \beta_0 + \beta_1 \cdot t_1 - (\beta_0 + \beta_1 \cdot t_2) = \beta_1 \cdot (t_1 - t_2)$$
$$\log\left(\frac{2 \cdot \text{PSA}}{\text{PSA}}\right) = \beta_1 \cdot \text{PSADT} \quad \Rightarrow \quad \text{PSADT} = \frac{\log(2)}{\beta_1}.$$

This calculation is straightforward for the log slope method that is used by Egawa et al. (2000), the Memorial Sloan-Kettering Cancer Center (MSKCC), and Sengupta et al. (2005), but for the PSADT proposed by Stephenson et al. (2002) the doubling time depends on the PSA value [102–104, 111]. Stephenson et al. (2002) do not provide a thorough description of their definition and thus we use the untransformed slope for the Stephenson PSADT [104]. Following O'Brien et al. (2009) we truncated the doubling time at ± 240 months.

Definition	Method [a]	Restrictions on PSA measurements
Velocities		
D'Amico [112]	slope	within 1 year before prostatectomy
Thompson [107]	log slope	within 3 years before prostatectomy
MSKCC [111]	slope	all available values
Sengupta [103]	slope	within 2 years before prostatectomy, at least 90 days apart
Doubling times		
Egawa [102]	log slope	all available values, spanning at least 6 months
MSKCC [111]	log slope	all available values
Sengupta [103]	log slope	within 2 years before prostatectomy, at least 90 days apart
Stephenson [104]	slope	at least 3 measurements, with one at least 1 year before prostatectomy [b]
Discretized velocity		
D'Amico [112]	cutoff value	2.0 ng/ml/year

[a] slope - slope of a linear regression of PSA on time
log slope - slope of a linear regression of log-transformed PSA on time
[b] originally before diagnosis

Table 2.6 PSA dynamics definitions reviewed by O'Brien et al. (2009) and adapted to PSA measurements before prostatectomy [105].

In addition to the PSA dynamics in Table 2.6, we use univariate summary statistics such as the mean, minimum, and maximum, as well as regression models with several variations to summarize the longitudinal PSA measurements. We consider different subsets of the PSA values and use either all values, values within three years, two years or one year before prostatectomy for feature calculation. Furthermore, we consider the log-transformation of PSA measurements as one additional modification. Table 2.7 summarizes the considered PSA statistics.

Altogether we obtain 34 univariate summaries, 96 regression summaries, 9 PSAV definitions,

and 4 PSADT definitions as candidate features. However, not all of them apply to all patients. PSA dynamics require a minimum of two PSA values and apart from the intercept, regression coefficients cannot be calculated for patients with only one PSA measurement. Figure 2.8 visualizes the number of missing values for each predefined method, where again, the number of missing values diminish for prostatectomies performed in more recent years. Nevertheless, PSA dynamics developed by Egawa et al. (2000), Sengupta et al. (2005), and Stephenson et al. (2002) are applicable to less than half of the patients with no visible improvement in recent years [102–104].

Method	Description
Summary statistics	
Last	Last PSA value before prostatectomy
Maximum	Maximum PSA value before prostatectomy
Mean	Average PSA value before prostatectomy
Minimum	Minimum PSA value before prostatectomy
Standard deviation	Estimated standard deviation of PSA values before prostatectomy
Regression models	
Linear	PSA $= \beta_0 + \beta_1 \cdot$ time
Polynomial degree 2	PSA $= \beta_0 + \beta_1 \cdot$ time $+ \beta_2 \cdot$ time^2
Polynomial degree 3	PSA $= \beta_0 + \beta_1 \cdot$ time $+ \beta_2 \cdot$ time$^2 + \beta_3 \cdot$ time^3

Table 2.7 Engineered PSA features for consideration in the analyses.

While the simple summary PSA statistics are applicable to all patients in the data, standard deviation and simple linear regression features are not available for all as Figure 2.9 shows. Features using a higher degree polynomial regression are only available for a small subset of the patients with no clear improvement in more recent years in contrast to the other features.

An extensive list of variables used for analysis with their respective names in the data sets is given in Appendix A.1.

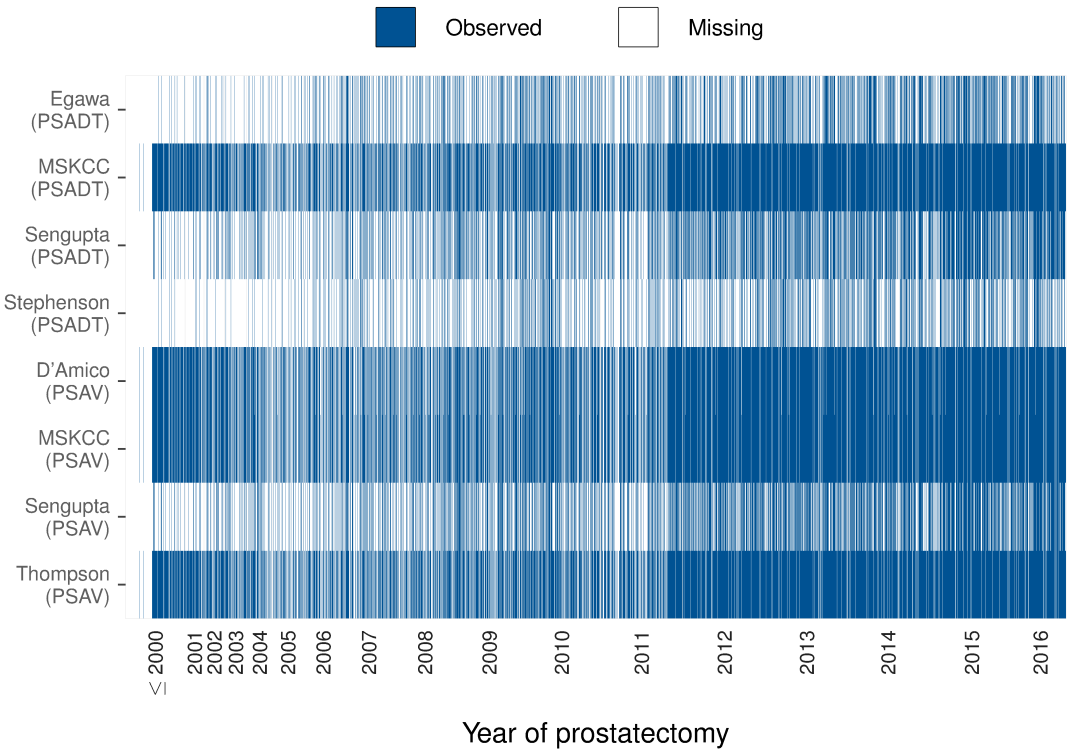

Figure 2.8 Indicator matrix for whether a predefined PSA dynamic is missing ordered by year of prostatectomy for 24,122 patients.

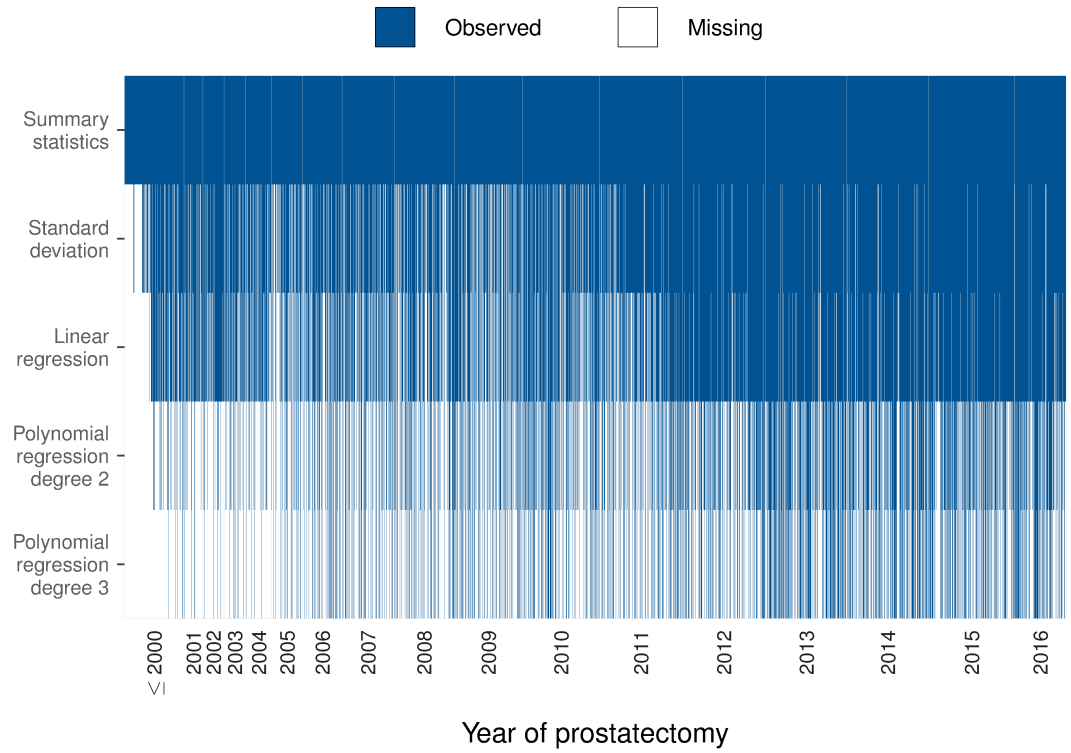

Figure 2.9 Indicator matrix for whether a PSA feature is missing ordered by year of prostatectomy for 24,122 patients.

2.4 Exploratory data analysis

In the following, we visualize and analyze the correlation between the adverse pathological outcomes, between different covariates as well as the relationship of single covariates with the pathological outcomes.

We asses the correlation between the adverse outcomes pairwise. Therefore we consider the contingency table for two binary outcomes as provided in Table 2.8, where 1 indicates that the adverse pathological outcome is present and 0 is not. We exclude missing values pairwise for testing the independence and quantifying the association between the outcomes. The number of complete cases for each pairwise combination is displayed in Table 2.9

Y_2 \ Y_1	0	1
0	n_{00}	n_{01}
1	n_{10}	n_{11}

Table 2.8 2×2 table two binary outcomes Y_1 and Y_2.

We test the pairwise independence of pathological outcomes using Pearson's χ^2 - test. The test statistic is given by

$$U = \sum_{i=0}^{1} \sum_{j=0}^{1} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \quad \text{with} \quad E_{ij} = \frac{(n_{i0} + n_{i1}) \cdot (n_{0j} + n_{1j})}{n_{00} + n_{10} + n_{01} + n_{11}},$$

and under the null hypothesis of independence $U \sim \chi_1^2$ [113]. All 10 tests for pairwise independence return a p-value smaller than 0.001 after adjusting the p-values for multiple testing using the Bonferroni method [113].

We quantify the pairwise association between the five pathological outcomes with the Φ-coefficient, Yule's Q and Yule's Y [114]. These correlations are defined for binary outcomes in the following way

$$\begin{aligned}
\Phi &= \frac{n_{11}n_{00} - n_{01}n_{10}}{\sqrt{(n_{00} + n_{01})(n_{00} + n_{10})(n_{11} + n_{01})(n_{11} + n_{10})}} \\
\text{Yule's Q} &= \frac{n_{11}n_{00} - n_{01}n_{10}}{n_{11}n_{00} + n_{01}n_{10}} \\
\text{Yule's Y} &= \frac{\sqrt{n_{11}n_{00}} - \sqrt{n_{01}n_{10}}}{\sqrt{n_{11}n_{00}} + \sqrt{n_{01}n_{10}}}
\end{aligned}$$

and range from -1 to 1, where 0 denotes no association [114]. If one of the entries in the contingency table, Table 2.8, of two binary variables is equal to 0, both Yule's Q and Yule's Y are equal to 1 or -1, whereas for $|\Phi| = 1$ both entries on one diagonal have to be 0 [114]. For the contingency table of ECE and SVI, we obtain $n_{01} = 0$, as any case of SVI is also classified as ECE. Thus, Yule's Q $= 1$ and Yule's Y $= 1$ for ECE and SVI.

Y_1	ECE				SVI			LVI		LNI
Y_2	SVI	LVI	LNI	PGG	LVI	LNI	PGG	LNI	PGG	PGG
n	24,080	18,037	17,187	24,035	18,025	17,166	24,012	14,235	18,033	17,140

Table 2.9 Number of patients with complete information on pairs of adverse pathological outcomes.

The pairwise association measures for the pathological outcomes are provided in Figure 2.10. All outcomes are pairwise positive associated as none of the coefficients is negative. The association between ECE and SVI is the strongest and ECE and LVI are least associated in terms of all association measures that we consider.

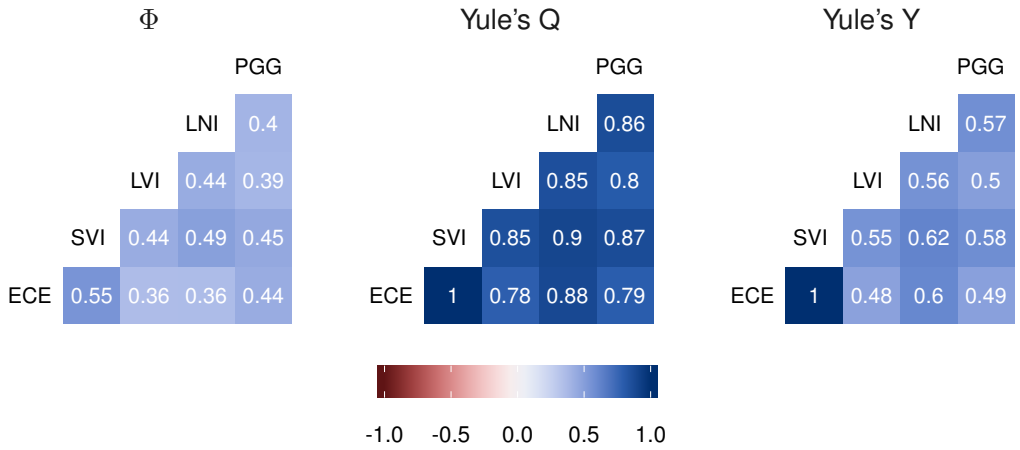

Figure 2.10 Associations between pathological outcomes measured by the Φ-coefficient, Yule's Q and Yules Y based on pairwise complete information with the sample sizes between 14,235 and 24,080 as provided in Table 2.9.

Figure 2.11 displays the distribution of different combinations of adverse pathological outcomes. Among the 14,219 patients with complete information about the pathological outcomes, 46.6% had no adverse outcome, 22.1% had one and 11.5% had two adverse pathological outcomes. The prevalence of individual combinations varies greatly, from 0.2% for LVI and LNI to 12.6% for ECE only.

We assess the correlation between covariates with the Spearman rank correlation using the approximate Student's t distribution for assessing significance. In addition, we consider the Pearson correlation coefficient to assess linear correlation and use the approximate Normal distribution to test for significance. We delete missing values in each pairwise comparison and adjust for 8,385 multiple comparisons among possible pairs of 130 features with the Bonferroni method.

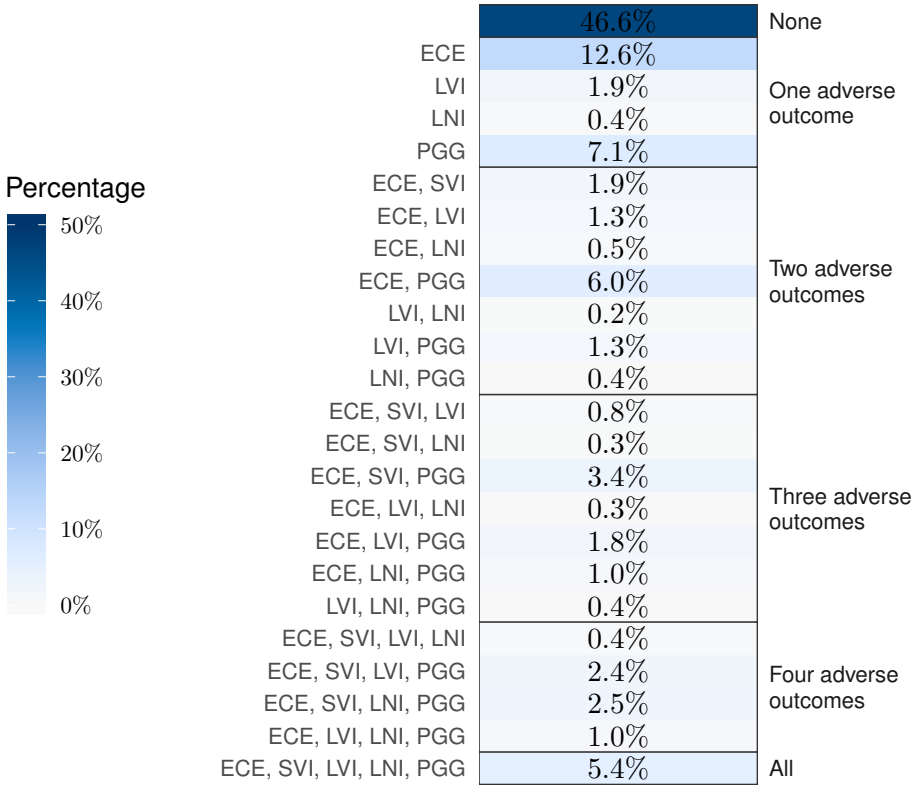

Figure 2.11 Distribution of combinations of adverse pathological outcomes for 14,219 patients with complete information on the pathological outcomes.

Figure 2.12 visualizes Spearman rank correlation for a selection of variables related to patient characteristics, biopsy, PSA statistics and dynamics defined in Section 2.3. Figure A.1 in Appendix A.2 displays Pearson correlation coefficients for the same variables with similar results, but less correlation coefficients were significant.

While correlation among age, BMI and biopsy related covariates was rather low apart from the number of positive biopsy cores and the percentage of positive cores, we detect a stronger correlation within the PSA related covariates. PSADTs defined by Egawa et al. and MSKCC were strongly positively correlated, but negatively correlated with PSADT defined by Stephenson et al. As we do not calculate the actual doubling time and take the slope for the latter PSADT definition, this negative correlation is not surprising. The PSAV definitions we consider showed strong a positive correlation with each other. For the summary statistics for PSA and $\log(\text{PSA})$, Figure 2.12 shows a strong positive correlation between the minimum, maximum and average values and a negative correlation between the minimum and the standard deviation of $\log(\text{PSA})$.

In Figure 2.13 we focus on these PSA related summary statistics, considering different time frames before prostatectomy. Thereby, we do not display minimum and maximum

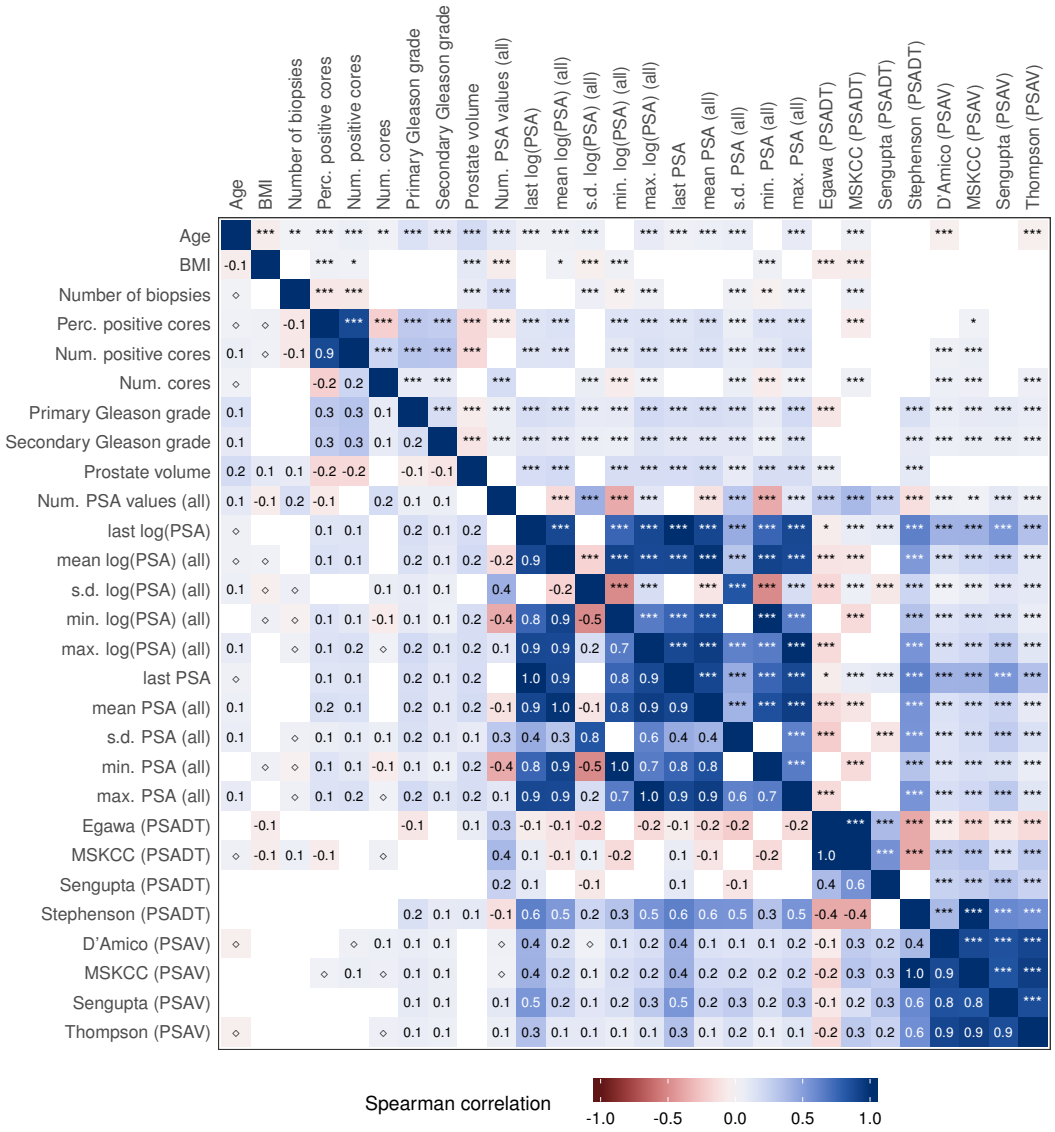

Figure 2.12 Spearman correlation coefficients for continuous covariates. (all) indicates all PSA values before prostatectomy were used, $***$ indicates a p-value < 0.001, $**$ < 0.01, and $*$ < 0.05 after adjusting for 8,385 multiple comparisons with the Bonferroni method. White color indicates no significant correlation, $\diamond$ correlation coefficients with an absolute value < 0.05.

PSA in addition to the $\log(\text{PSA})$ versions, as the Spearman correlation is rank based and thus the correlations are identical. In Figure A.2 we provide the Pearson correlation coefficients including both minimum and maximum of PSA and $\log(\text{PSA})$ since this coefficient is influenced by log-transformation.

The dark blue squares next to the diagonal in Figure 2.13 show that regardless of the specific time frame considered for the calculation of summary statistics, the correlation of the different parameters for the same summary statistics is close to 1. Besides correlations involving the standard deviation of $\log(\text{PSA})$, the minimum of $\log(\text{PSA})$, and the number of PSA

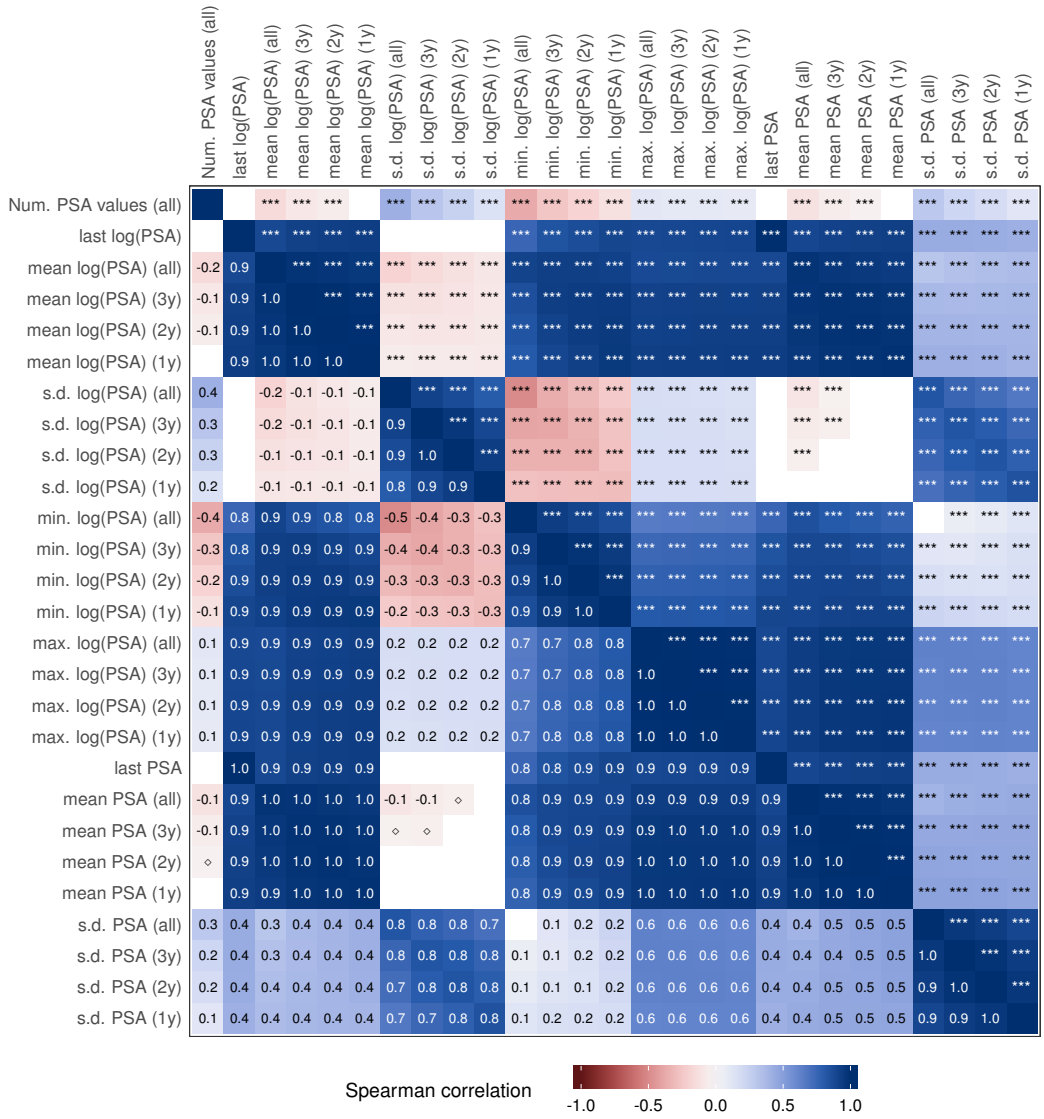

Figure 2.13 Spearman correlation coefficient of PSA related covariates for different time periods. (all), (1y), (2y), or (3y) indicate all PSA values or those within one, two or three years of prostatectomy are used. $***$ indicates a p-value < 0.001 after adjusting for 8,385 multiple comparisons with the Bonferroni method. White color indicates no significant correlation, $\diamond$ correlation coefficients with an absolute value < 0.05.

values, correlations are generally positive as would be expected. The positive correlation of standard deviation of PSA with the minimum and maximum $\log(\text{PSA})$ shows that patients with larger PSA values have larger variation. An increase of the variance with the mean is often observed in chemical experiments, where values can span orders of magnitude. The $\log$ is a well-known variance stabilizing transform that corrects this phenomenon [115]. Correlations involving the $\log(\text{PSA})$ are more likely to be negative in Figure 2.13.

We also assess the relationship between the pathological outcomes and some selected variables. Figure 2.14 shows split violin plots for age at prostatectomy, BMI, and the average

$\log(\text{PSA})$ value within one year of prostatectomy. For age at prostatectomy we can see a slight shift in the distribution for patients with the specific pathological outcome, as the median values and quartiles are higher. For BMI no clear differences are visible across all outcomes. The mean $\log(\text{PSA})$ value within one year of prostatectomy varies considerably between patients with and without each pathological outcome.

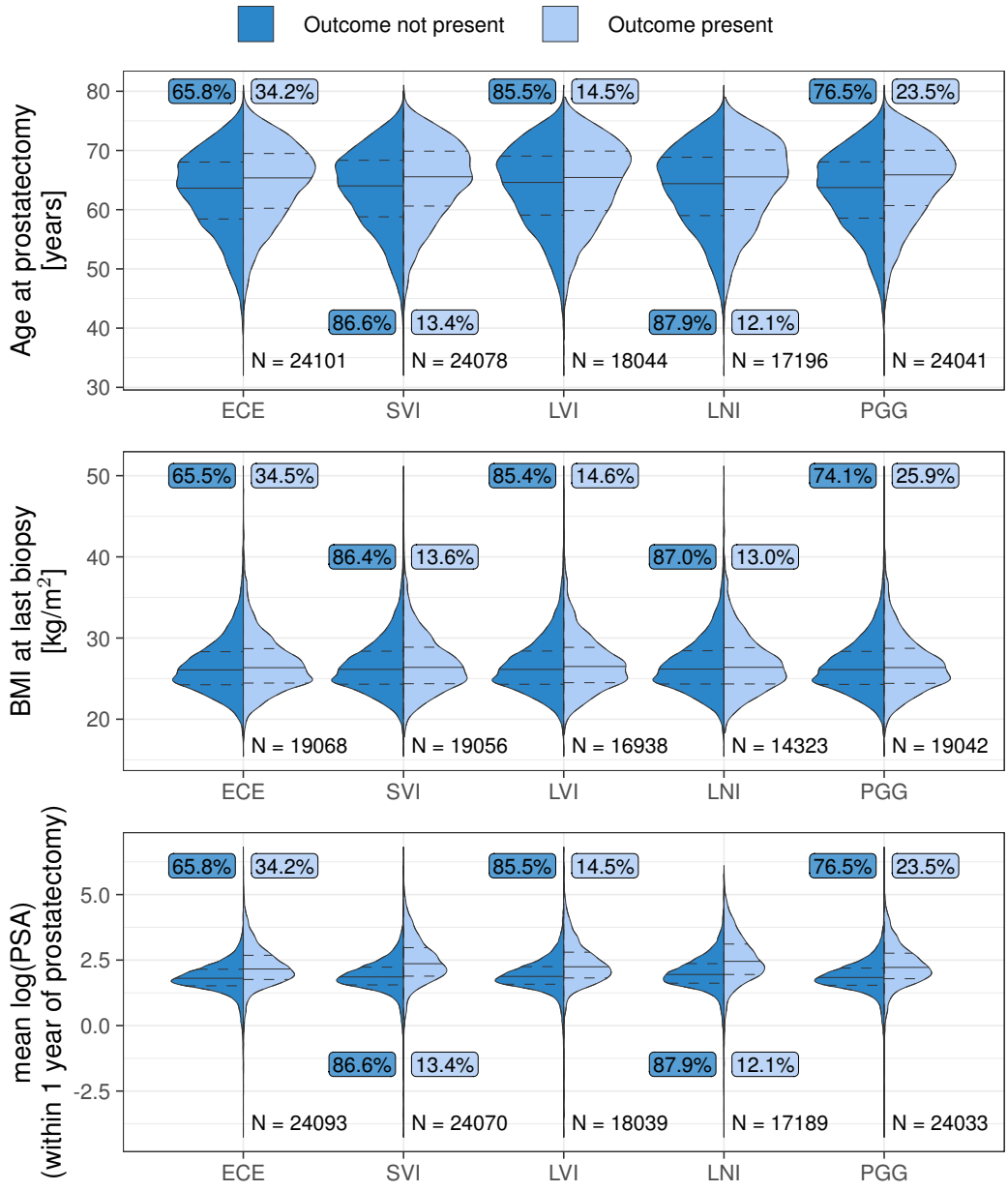

Figure 2.14 Distributions for age, BMI and average $\log(\text{PSA})$ within one year before prostatectomy split by pathological outcome with not present in dark blue versus outcome present in light blue. Median, 25% and 75% quantiles are superimposed with solid and dashed lines. Percentages indicate the distribution of pathological outcomes. Sample sizes of patients with non-missing values for both the outcome and the variable of interest are provided at the bottom.

We analyze the primary and secondary Gleason grade using a back-to-back barplot. Here,

again we detect a shift in distribution as for higher Gleason grade values the prevalence of each pathological outcome increases. Looking, for example, at the primary Gleason grade at last biopsy and ECE (top left plot), we see an increase in percentage of ECE cases from 25.0% among patients with primary Gleason grade 3 to 58.5% for those with primary Gleason grade 4.

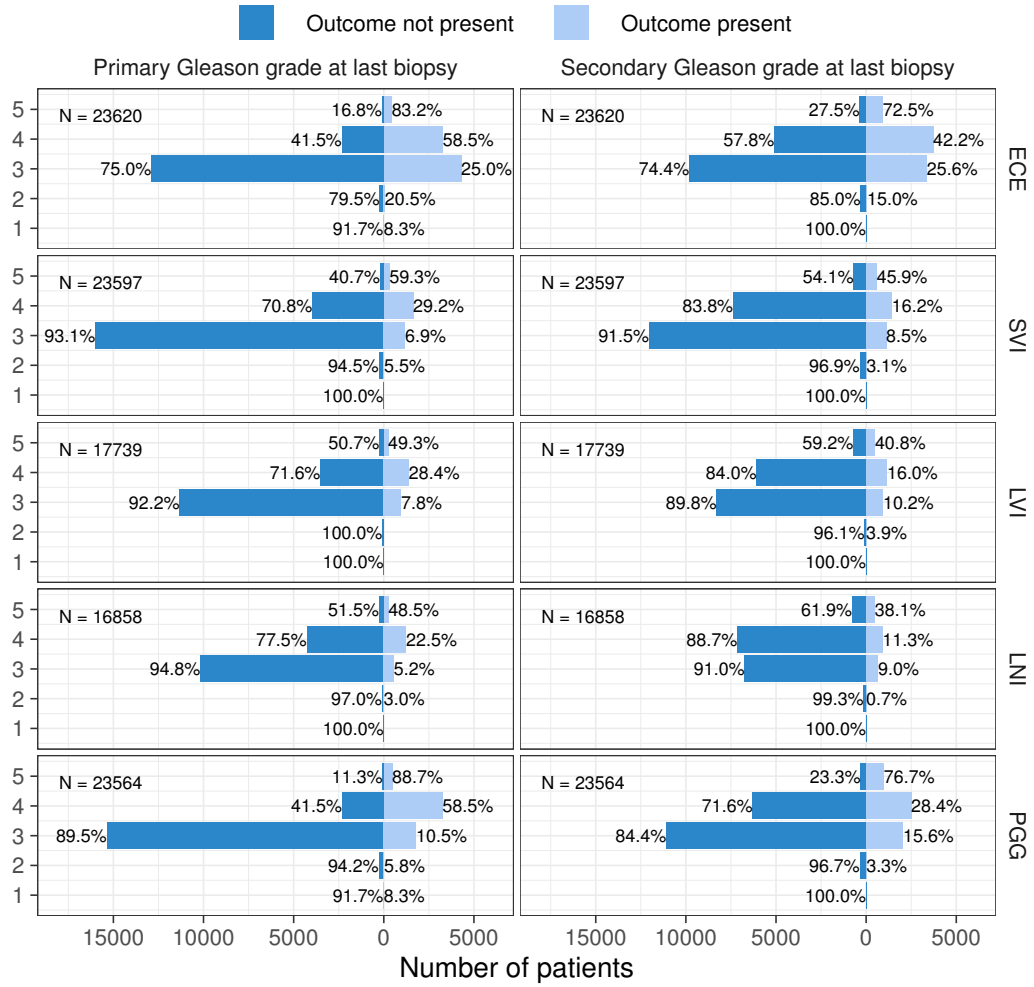

Figure 2.15 Distributions of primary and secondary Gleason grades split by pathological outcome with not present in dark blue versus outcome present in light blue. Percentages indicate the distribution for each outcome per Gleason grade value on the y-axis. Sample sizes of patients with non-missing values for both the outcome and Gleason grade is given at the top left.

2.5 Splitting data into train and test sets

In this thesis, we fit multiple models to the prostatectomy data, both univariate and multivariate. We want to select the best one and asses its performance. One risk of this procedure is overfitting, meaning that we select a model that fits the data too well and therefore models random noise of the data. We would like to generate a model that not only fits well to the data but also provides a good generalization to the underlying problem that is predicting the

prostatectomy outcomes. This is a common issue in machine learning and using a hold-out test set is one possible approach to detect overfitting [100, 116].

We use the year of prostatectomy to split the data into three parts, a training, validation, and testing set. Patient records with a date of prostatectomy before 2013 are regarded as the training set, records of patients with a prostatectomy in 2013 or 2014 comprise the validation set, and the remaining records for patients that underwent prostatectomy in 2015 or 2016 are used as a hold-out test set (Figure 2.16).

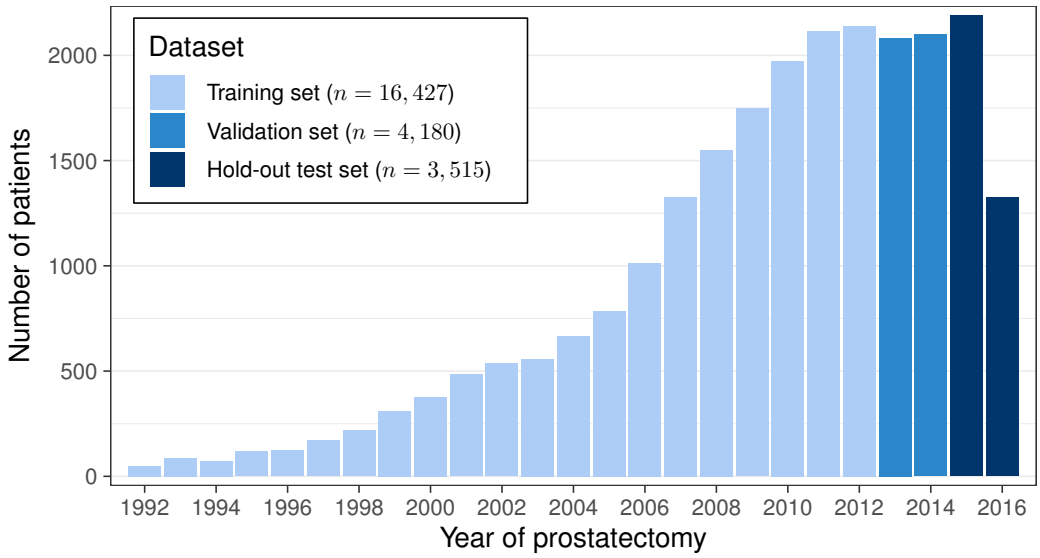

Figure 2.16 EMR data are split into training ($n = 16,427$), validation ($n = 4,180$), and hold-out test set ($n = 3,515$) using year of prostatectomy.

The hold-out test set is not used in the modeling process. With this procedure, the final model accuracy can be independently calculated and we prevent results that are too optimistic due to overfitting. Furthermore, as we use the year of prostatectomy as a splitting variable, we simulate an implementation in a real database, where predictions of future outcomes are performed based on past results. Nevertheless, there is one bias that we cannot eliminate with the data at hand. As we validate the results on a data set created in the same clinic and by the same physicians, we are unable to detect a study site-specific bias.

We use the validation set to asses model performance of intermediate models and then choose the final model. This is part of the training process and we fit the final model using the complete training data, which is the training set and validation set combined.

2.6 Imputation of missing data

In Sections 2.2 and 2.3 we have seen that various characteristics and features are missing for many patients and the outcomes of interest ECE, SVI, LVI, LNI, PGG are not observed for all patients as well. Overall only 841 patients have information on all outcomes and features and all those patients underwent a prostatectomy in or after 2012. Thus over 96.5% of the observations have at least one missing outcome or covariate and simply excluding incomplete records is not a viable solution. Moreover, premature exclusion of records can lead to biased results and loss of statistical power [117, 118].

One common approach to dealing with missing values is imputation. We assume that values are missing at random (MAR), which means that missingness of the values depends on the observed values and does not depend on the underlying value that is missing [119, 120]. Then, we use multiple imputation by chained equations (MICE) with 10 multiple imputations with 10 iterations each. We apply predictive mean matching for continuous features, logistic regression for binary features, and polytomous logistic regression for categorical features and use the `R`-function `mice` of the `mice` package for imputation [120–122]

As described in Section 2.5, we develop the models on the training data, assess the model performance on the validation set to choose the model, and test the final performance on a hold-out test set. We apply the imputation on the training and validation set together, but separately to the hold-out test set, to avoid that we accidentally use information of the hold-out test set for modeling. We include the outcome variables in the imputation for the training and validation set as it is recommended in the literature [120, 123]. Including the outcome in the imputation for the hold-out test set, might lead to overly optimistic model performance, thus we only use the covariates in the imputation and do not include the pathological outcomes of interest [124]. In the test set information about LNI is missing for 318 patients (9.0%) and information about ECE, SVI, LVI, and PGG is missing for 4, 9, 2, and 4 patients (< 0.5%), respectively. For the missing pathological outcomes, we use a second set of imputations and assume an adverse pathological outcome to be present when at least 50% of the imputed values are positive.

By default `mice` automatically removes constant and nearly collinear variables before imputation [125]. These are then not imputed at all and remain as missing variables in the imputation sets. Overwriting this default behavior leads to numerical issues and is not a feasible option. Table 2.10 lists the variables for the training data that would be removed by `mice` and the proportions of missing values. The list is dominated by covariates related to polynomial regression of PSA and $\log(\text{PSA})$ with degree 2 or 3 and restricted to different time frames before prostatectomy. Figure 2.17 shows that these coefficients are highly correlated

Variable	Number of missing values (%)
PSA poly degree 3, coef 4 (2y)	16806 (81.6%)
log(PSA) poly degree 3, coef 4 (2y)	16806 (81.6%)
PSA poly degree 3, coef 4 (3y)	16261 (78.9%)
Stephenson (PSADT)	15808 (76.7%)
PSA poly degree 3, coef 4 (all)	15633 (75.9%)
Egawa (PSADT)	14479 (70.3%)
PSA poly degree 2, coef 3 (1y)	14455 (70.1%)
log(PSA) poly degree 2, coef 3 (1y)	14455 (70.1%)
PSA poly degree 2, coef 3 (2y)	13549 (65.7%)
log(PSA) poly degree 2, coef 3 (2y)	13549 (65.7%)
PSA poly degree 3, coef 3 (2y)	13549 (65.7%)
log(PSA) poly degree 3, coef 3 (2y)	13549 (65.7%)
PSA poly degree 2, coef 3 (3y)	13232 (64.2%)
log(PSA) poly degree 2, coef 3 (3y)	13232 (64.2%)
PSA poly degree 3, coef 3 (3y)	13232 (64.2%)
log(PSA) poly degree 3, coef 3 (3y)	13232 (64.2%)
PSA reg, coef 2 (1y)	6253 (30.3%)
PSA poly degree 2, coef 2 (1y)	6253 (30.3%)
log(PSA) reg, coef 2 (2y)	6013 (29.2%)
PSA poly degree 2, coef 2 (2y)	6013 (29.2%)
PSA poly degree 3, coef 2 (2y)	6013 (29.2%)
log(PSA) reg, coef 2 (3y)	5925 (28.8%)
PSA poly degree 3, coef 2 (3y)	5925 (28.8%)
PSA reg, coef 2 (all)	5822 (28.3%)
standard deviation PSA (2y)	3439 (16.7%)
maximum PSA (1y)	9 (0%)
maximum PSA (2y)	5 (0%)
log(PSA) reg, coef 1 (2y)	5 (0%)
log(PSA) reg, coef 1 (3y)	3 (0%)
log(PSA) poly degree 2, coef 1 (3y)	3 (0%)

Table 2.10 Features nearly collinear with other variables in the training data ($n = 20,607$). reg stands for linear and poly for polynomial regression with the corresponding degree and coefficients (coef). (all), (1y), (2y), or (3y) indicate all PSA values or those within one, two or three years of prostatectomy are used.

and thus to avoid collinearity issues in the imputation we only consider the covariates related to the polynomial regression based on all PSA or $\log(\text{PSA})$ values before prostatectomy and discard variables restricted to other time frames.

The two PSADT definitions of Egawa et al. (2000) and Stephenson et al. (2002) also might

cause collinearity issues. In Figure 2.12 and Figure A.1 we see that these are perfectly correlated with the PSADT by MSKCC and the PSAV by MSKCC, respectively. When we compare the definitions of the PSA dynamics in Table 2.6, we see that although Egawa et al. (2000) and Stephenson et al. (2002) use more restrictive criteria for when the respective definition is applicable, the calculated values are then identical to those defined by MSKCC. Thus, again, to avoid collinearity issues in the imputation we remove the PSADTs of Egawa et al. (2000) and Stephenson et al. (2002) as covariates.

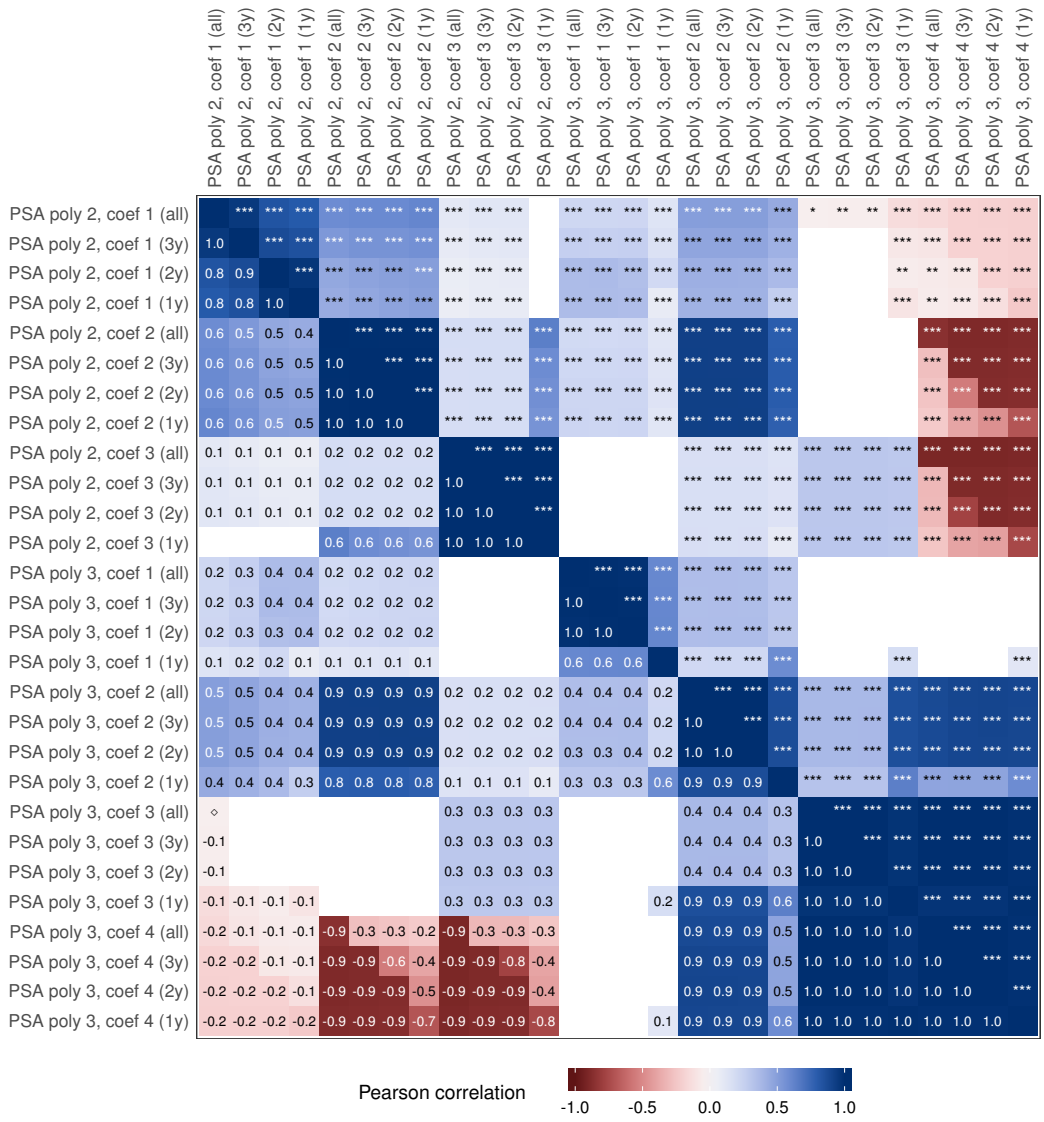

Figure 2.17 Pearson correlation coefficients for covariates related to polynomial regression of PSA for different time periods. (all), (1y), (2y), or (3y) indicate all PSA values or those within one, two or three years of prostatectomy are used. $***$ indicates a p-value < 0.001, $** < 0.01$, and $* < 0.05$ after adjusting for 8,385 multiple comparisons with the Bonferroni method. White color indicates no significant correlation, $\diamond$ correlation coefficients with an absolute value smaller than 0.05.

Finally, some direct transformations are contained in Table 2.10. The slope coefficients of linear regression models of all PSA values and values within one year of prostatectomy are

identical to the PSAVs defined by MSKCC and D'Amico et al. (2004) divided by 365.25 [111, 112]. Similarly, the slope of the linear regression model of $\log(\text{PSA})$ values within three years before prostatectomy is identical to the PSAV proposed by Thompson et al. (2006) divided by 365.25 Thompson et al. (2006). We manually exclude these slope coefficients and calculate them using the imputed PSAV values after imputation. We further calculate the D'Amico's PSAV cut-off at 2.0 ng/ml/year after imputation to ensure consistent values. The remaining nearly collinear variables that are included for the imputations do not cause numerical issues.

For the test set, we exclude the same covariates as for the training set. Further, the result of the last biopsy is missing for 12 patients (0.3%) in the test set and is positive otherwise. Thus, we impute the last biopsy value for these missing patients with "positive".

After imputation, we obtain 10 training sets with 16,427 patients, 10 validation set with 4,180 patients, and 10 test sets with 3,515 patients. Each set contains 5 pathological outcomes and 87 features.

3 Mathematical foundations

In this chapter, we introduce the mathematical background with the notation used throughout this thesis. After defining the univariate and multivariate logistic regression model, we introduce the Bayes factor (BF) and the order of stochastic approximation for random sequences with the proof for several properties. Further, we discuss metrics for evaluating prediction performance of models with binary outcomes.

3.1 Logistic regression

Logistic regression is a common method for modeling binary responses in biostatistical applications [116]. We consider the following univariate multiple logistic regression problem, where $Y_i \in \{0, 1\}$ follows the Bernoulli distribution with success probability π_i, written $Y_i \sim Ber(\pi_i)$, independently for $i = 1, \ldots, n$. Assigning 1 as the outcome to predict, this means $P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$ for $i = 1, \ldots, n$. The mean of Y_i is $\mathbb{E}[Y_i] = 1 \cdot P(Y_i = 1) + 0 \cdot P(Y_i = 0) = \pi_i$ and similar calculations show that $\mathrm{Var}(Y_i) = \pi_i(1 - \pi_i)$. With the corresponding m-dimensional vector of covariates $X_i \in \mathbb{R}^m$, $i = 1, \ldots, n$ including 1 for the intercept, the standard logistic model is given by

$$\mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = X_i'\beta,$$

where $\beta \in \mathbb{R}^m$ is the coefficient vector to be estimated [126]. We define $G : \mathbb{R} \to [0, 1]$ as $G(z) = \dfrac{\exp(z)}{1 + \exp(z)}$ and thus,

$$\pi_i = G(X_i'\beta) = \frac{\exp(X_i'\beta)}{1 + \exp(X_i'\beta)}. \tag{3.1}$$

The vector β represents log odds ratios (OR) for the respective covariates, as illustrated in Example 3.1 [126].

Example 3.1

Suppose we have n observations of $Y \in \{0, 1\}$, where 1 indicates presence of ECE and 0 no ECE. Consider one binary variable X, which represents high primary Gleason grade with

$$X = \begin{cases} 0 & \text{primary Gleason grade } \leq 3 \\ 1 & \text{primary Gleason grade } > 3. \end{cases}$$

Then the univariate logistic regression model with an intercept and a single covariate X specifies

$$\text{logit}\left(P(Y = 1|X)\right) = \begin{pmatrix} 1 & X \end{pmatrix} \beta = \beta_0 + X\beta_1,$$

where $\beta = (\beta_0, \beta_1)'$. The odds of having ECE for a patient with primary Gleason grade ≤ 3 are

$$\frac{P(Y = 1|X = 0)}{P(Y = 0|X = 0)} = \frac{\frac{\exp(\beta_0)}{1 + \exp(\beta_0)}}{\frac{1}{1 + \exp(\beta_0)}} = \exp(\beta_0),$$

while for a patient with a primary Gleason grade > 3 the odds are

$$\frac{P(Y = 1|X = 1)}{P(Y = 0|X = 1)} = \exp(\beta_0 + \beta_1).$$

Thus the OR for ECE by primary Gleason grade is

$$\frac{\frac{P(Y = 1|X = 1)}{P(Y = 0|X = 1)}}{\frac{P(Y = 1|X = 0)}{P(Y = 0|X = 0)}} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

and the log OR is β_1. With only a single binary covariate we can estimate the odds and OR with counts for each Y and X combination as shown in Table 3.1.

X \ Y	0	1
0	n_{00}	n_{01}
1	n_{10}	n_{11}

Table 3.1 2×2 table for a binary outcome Y and binary covariate X. The total sample size is $n = n_{00} + n_{01} + n_{10} + n_{11}$.

The odds are empirically estimated by

$$\frac{P(Y=1|X=0)}{P(Y=0|X=0)} = \frac{\frac{n_{01}}{n_{00}+n_{01}}}{\frac{n_{00}}{n_{00}+n_{01}}} = \frac{n_{01}}{n_{00}} \qquad \text{if } n_{00} > 0$$

$$\frac{P(Y=1|X=1)}{P(Y=0|X=1)} = \frac{\frac{n_{11}}{n_{11}+n_{10}}}{\frac{n_{10}}{n_{11}+n_{10}}} = \frac{n_{11}}{n_{10}} \qquad \text{if } n_{10} > 0,$$

and the OR with

$$\frac{\frac{P(Y=1|X=1)}{P(Y=0|X=1)}}{\frac{P(Y=1|X=0)}{P(Y=0|X=0)}} = \frac{\frac{n_{11}}{n_{10}}}{\frac{n_{01}}{n_{00}}} = \frac{n_{11}n_{00}}{n_{10}n_{01}},$$

if $n_{10}, n_{01} > 0$.

	ECE	
	No	Yes
Primary Gleason grade ≤ 3	13,142	4,359
Primary Gleason grade > 3	2,411	3,708

Table 3.2 2×2 table for ECE by primary Gleason grade among 23,620 patients. 502 patients are excluded due to missing values in either one of the variables.

Table 3.2 displays the 2×2 table of ECE by primary Gleason grade among 23,620 patients of Chapter 2 without missing information in either of the two variables. Table 3.2 yields

$$\text{OR} = \frac{\frac{3708}{2411}}{\frac{4359}{13142}} = 4.64$$

and this means that the odds of ECE were increased 4.64 fold.

For a single continuous covariate X, odds and OR correspond to a unit increase in X. For multiple covariates odds and OR correspond to holding all other covariates fixed.

Consider next a sample of n observation pairs (y_i, X_i), where $y_i \in \{0, 1\}$ are independent and X_i are fixed covariates for $i = 1, \ldots, n$. Under logistic regression the likelihood of β is given by

$$\begin{aligned}\mathcal{L}(\beta) &= \prod_{i=1}^{n} G\left(X_i'\beta\right)^{y_i} \left(1 - G\left(X_i'\beta\right)\right)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{\exp\left(X_i'\beta\right)}{1+\exp\left(X_i'\beta\right)}\right)^{y_i} \left(\frac{1}{1+\exp\left(X_i'\beta\right)}\right)^{1-y_i} = \frac{\exp\left(\sum_{i=1}^{n} y_i X_i'\beta\right)}{\prod_{i=1}^{n}\left(1+\exp\left(X_i'\beta\right)\right)}.\end{aligned}$$

The log-likelihood is

$$\ell(\beta) = \log\left(\mathcal{L}(\beta)\right) = \sum_{i=1}^{n} \left[y_i \log\left(G\left(X_i'\beta\right)\right) + (1-y_i)\log\left(1 - G\left(X_i'\beta\right)\right)\right]$$
$$= \sum_{i=1}^{n} \left[y_i X_i'\beta - \log\left(1 + \exp\left(X_i'\beta\right)\right)\right] \quad \text{[127]}.$$

With $E_0 = \{i \in \{1,\dots,n\} : y_i = 0\}$ denoting all indices that correspond to outcome 0 and $E_1 = \{i \in \{1,\dots,n\} : y_i = 1\}$ all indices that correspond to outcome 1 the log-likelihood can be compartmentalized as

$$\ell(\beta) = \sum_{i \in E_1} \log\left(G\left(X_i'\beta\right)\right) + \sum_{i \in E_0} \log\left(1 - G\left(X_i'\beta\right)\right) \quad \text{[128]}. \tag{3.2}$$

We show that for logistic regression, $0 < G(z) < 1$, $G(z)$ is strictly increasing, $-\log(G(z))$ and $-\log(1-G(z))$ are convex functions, and thus the negative log-likelihood is convex. First, $G(z) = \dfrac{\exp(z)}{1+\exp(z)} > 0$ since $\exp(z) > 0$ for all $z \in \mathbb{R}$ and $G(z) < 1$ since $\exp(z) < 1 + \exp(z)$ for all $z \in \mathbb{R}$. Further, the first derivative of $G(z)$ is given by

$$\frac{\partial}{\partial z} G(z) = \frac{(1+\exp(z))\exp(z) - \exp(z)\exp(z)}{(1+\exp(z))^2} = \frac{\exp(z)}{(1+\exp(z))^2} > 0,$$

for all $z \in \mathbb{R}$ and thus $G(z)$ is a strictly increasing function. With

$$\begin{aligned}
\frac{\partial^2}{\partial z^2}\left(-\log\left(G(z)\right)\right) &= \frac{\partial}{\partial z}\left(-\frac{1}{G(z)} \cdot \left(\frac{\partial}{\partial z} G(z)\right)\right) \\
&= \frac{\partial}{\partial z}\left(-\frac{1+\exp(z)}{\exp(z)} \cdot \frac{\exp(z)}{(1+\exp(z))^2}\right) = \frac{\partial}{\partial z}\left(-\frac{1}{(1+\exp(z))}\right) \\
&= \frac{\partial}{\partial z}\left(G(z) - 1\right) = \frac{\exp(z)}{(1+\exp(z))^2} > 0
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2}{\partial z^2}\left(-\log\left(1 - G(z)\right)\right) &= \frac{\partial}{\partial z}\left(-\frac{1}{1-G(z)} \cdot \left(\frac{\partial}{\partial z}\left(1 - G(z)\right)\right)\right) \\
&= \frac{\partial}{\partial z}\left(-(1+\exp(z)) \cdot \left(-\frac{\exp(z)}{(1+\exp(z))^2}\right)\right) \\
&= \frac{\partial}{\partial z}\frac{\exp(z)}{(1+\exp(z))} = \frac{\partial}{\partial z} G(z) = \frac{\exp(z)}{(1+\exp(z))^2} > 0
\end{aligned}$$

it follows that $-\log\left(G(z)\right)$ and $-\log(1-G(z))$ are convex functions, as the second deriva-

tives are larger than 0. Thus the negative log-likelihood

$$-\ell(\beta) = -\sum_{i \in E_1} \log\left(G\left(X_i'\beta\right)\right) - \sum_{i \in E_0} \log\left(1 - G\left(X_i'\beta\right)\right)$$

is convex as a sum of convex functions.

Silvapulle (1981) used this and showed with (3.2) that the maximum likelihood estimator (MLE) $\hat{\beta}$ exists and is uniquely defined if Assumption 3.1 holds [128].

Assumption 3.1 (Existence and uniqueness of the MLE)

1) For $X_i \in \mathbb{R}^m$, $i = 1, \ldots, n$ the design matrix $X = (X_1, \ldots, X_n)'$ has full rank m.

2) $X_{i1} = 1$ for each $i = 1, \ldots, n$ or the design matrix includes one constant term for all samples.

3) $S \cap F \neq \emptyset$, where S and F are the relative interiors of the convex cones generated by $\{X_i : i \in E_0\}$ and $\{X_i : i \in E_1\}$, respectively. Thus

$$S = \left\{ \sum_{i \in E_0} k_i X_i : k_i > 0 \right\} \quad \text{and} \quad F = \left\{ \sum_{i \in E_1} k_i X_i : k_i > 0 \right\}$$

with $E_0 = \{i \in \{1, \ldots, n\} : y_i = 0\}$ and $E_1 = \{i \in \{1, \ldots, n\} : y_i = 1\}$.

Suppose, independent observations y_i have been collected with covariate vectors X_i satisfying Assumption 3.1. Without loss of generality we can assume for Assumption 3.1 2) that $X_{i1} = 1$ for each $i = 1, \ldots, n$ as otherwise X_i can be reordered and rescaled to fulfill this assumption. Assumption 3.1 3) is defined as overlap by Albert and Anderson (1984) and implies that there exists no $\beta \in \mathbb{R}^m \backslash \{0\}$ such that

$$X_i'\beta \geq 0 \quad \forall i \in E_1 \quad \text{and} \quad X_i'\beta \leq 0 \quad \forall i \in E_0 \quad \text{[129, 130].} \tag{3.3}$$

When there exists a $\beta \in \mathbb{R}^m$ such that (3.3) holds with strict inequalities the data is completely separated. When (3.3) holds and for at least one $i = 1, \ldots, n$ holds with equality the data is quasicompletely separated [129, 130]. The MLE $\widehat{\beta}$ exists only in case of overlap [128–130].

To better understand when Assumption 3.1 fails, we illustrate it in the context of the single binary covariate example of Table 3.1. We include an intercept in the logistic regression model to fulfill Assumption 3.1 2) and consider different values for n_{00}, n_{01}, n_{10}, and n_{11}.

If $n_{00} = n_{01} = 0$ then all individuals in the data set have the same covariate value $X = 1$. The design matrix $(1\ X)$ has rank 1 and thus Assumption 3.1 1) is not fulfilled. The same

failure of Assumption 3.1 1) holds for $n_{10} = n_{11} = 0$, where all individuals have the same covariate $X = 0$. So it is not possible to estimate the OR. For $n_{00} = n_{10} = 0$ we have the problem that all individuals in the data set have the same $Y = 1$ outcome. The odds are estimated to be infinity since $P(Y = 0)$ is estimated zero. Assumption 3.1 3) fails because $S = \emptyset$ and thus $S \cap F = \emptyset$. For $n_{01} = n_{11} = 0$ all individuals have the same $Y = 0$ outcome and $S \cap F = \emptyset$ since $F = \emptyset$, violating Assumption 3.1 3).

Figure 3.1 displays the remaining scenarios for the values of n_{00}, n_{01}, n_{10}, and n_{11} with the corresponding S, F, and $S \cap F$ and we discuss in which cases Assumption 3.1 is fulfilled.

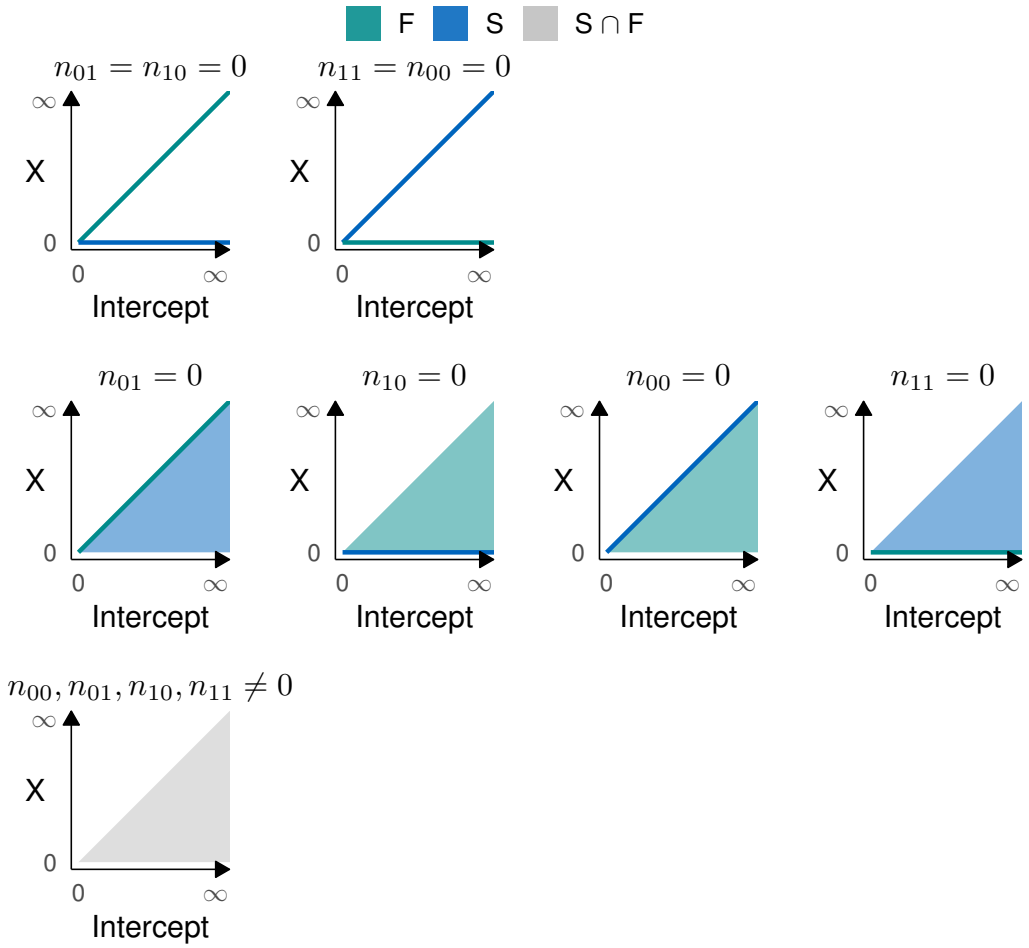

Figure 3.1 S, F, and $S \cap F$ for different values of n_{00}, n_{01}, n_{10}, and n_{11} in Table 3.1, summarizing n samples of $Y \in \{0, 1\}$ and a binary covariate X. The top row shows completely separated, the mid row quasicompletely separated and the bottom row overlapping data. Assumption 3.1 is only fulfilled for the bottom row.

The top row shows S and F for $n_{01} = n_{10} = 0$ and $n_{11} = n_{00} = 0$. When $n_{01} = n_{10} = 0$ all individuals with $X = 1$ have the same outcome $Y = 1$ and all individuals with $X = 0$ have the same outcome $Y = 0$. Thus the data is completely separated as for any $\beta = (\beta_0, \beta_1) \in \mathbb{R}^2$ with $\beta_0 < 0$ and $\beta_1 > |\beta_0|$ it holds that for all $i \in E_1$ $X_i'\beta = (1\ 1)\beta = \beta_0 + \beta_1 > 0$ and for all $i \in E_0$ $X_i'\beta = (1\ 0)\beta = \beta_0 < 0$. S and F are straight lines that do not intersect, since $(0, 0) \notin S$ and $(0, 0) \notin F$. Thus $S \cap F = \emptyset$ and Assumption 3.1 3) is not fulfilled. Similar arguments show that the data is completely separated for $n_{11} = n_{00} = 0$ and again Assumption 3.1 3) is violated.

The mid row of Figure 3.1 displays S and F when one of n_{00}, n_{01}, n_{10}, and n_{11} is equal to 0. When $n_{00} = 0$ $(n_{10} = 0)$ all individuals with $X = 0$ $(X = 1)$ have the same outcome $Y = 1$ and for $n_{01} = 0$ $(n_{11} = 0)$ all individuals with $X = 0$ $(X = 1)$ have the same outcome $Y = 0$. The data is quasicompletely separated and we illustrate this for $n_{00} = 0$. For any $\beta = (\beta_0, \beta_1) \in \mathbb{R}^2$ with $\beta_0 > 0$ and $\beta_1 = -\beta_0$ it holds that for all $i \in E_0$ $X_i'\beta = (1\ 1)\beta = \beta_0 - \beta_0 = 0$ and for all $i \in E_1$

$$X_i'\beta = \begin{cases} (1\ 0)\beta = \beta_0 > 0 & \text{for } X = 0 \\ (1\ 1)\beta = 0 & \text{for } X = 1. \end{cases}$$

For one of n_{00}, n_{01}, n_{10}, and n_{11} equal to 0 either S or F is a straight line and the other one the open set $\{(a, b) : a, b \in (0, \infty) \cap b < a\}$. Again, $S \cap F = \emptyset$ and Assumption 3.1 3) is violated.

If $n_{00}, n_{01}, n_{10}, n_{11} \neq 0$ as in the bottom row of Figure 3.1, Assumption 3.1 is fulfilled and the MLE for the logistic regression model exists and is uniquely defined.

The simple example shows that all $2 \cdot 2 = 4$ cells need to be non-empty. With a single categorical variable with k levels, $2 \cdot k$ cells must be filled and it becomes more difficult to satisfy Assumption 3.1 as the curse of dimension kicks in. The problem becomes exasperated with multiple categorical covariates and lack of convergence for logistic regression modeling is a more common experience than for normal regression.

For continuous covariates modeled by a single linear parameter or reduced smoother, Assumption 3.1 is likely to hold as long as at least some individuals of both groups have overlapping covariate values. Here, complete separation occurs if all individuals with one outcome have lower values for the continuous covariate than all individuals with the other outcome.

We now assume Assumption 3.1 holds. Although several algorithms exist, we obtain the MLE $\hat{\beta}$ for β using the Newton-Raphson algorithm [116, 131]. The score vector, or first

derivative of the log-likelihood function, is denoted by

$$\begin{aligned} u(\beta) = \nabla_\beta \ell(\beta) &= \sum_{i=1}^{n} \left[y_i X_i - \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)} \cdot X_i \right] \\ &= \sum_{i=1}^{n} (y_i - G(X_i'\beta)) X_i = X'(Y-p), \end{aligned}$$

where $Y = (y_1, \ldots, y_n)' \in \{0,1\}^n$, $p = (\pi_1, \ldots, \pi_n)' \in [0,1]^n$ and $X = (X_1, \ldots X_n)' \in \mathbb{R}^{n \times m}$. We want to solve the resulting m equations

$$u(\beta) = 0.$$

For the Newton-Raphson algorithm we obtain the Hessian matrix, which is the matrix of second derivatives of the log-likelihood. The log-likelihood is twice differentiable for the logistic regression model under Assumption 3.1.

$$\begin{aligned} D^2\ell(\beta) = \left(\frac{\partial^2 \ell(\beta)}{\partial \beta_r \partial \beta_s} \right) &= \left(\frac{\partial \ell(\beta)}{\partial \beta_r} \sum_{i=1}^{n} \left[y_i - G(X_i'\beta) \right] X_{is} \right) \\ &= \left(\sum_{i=1}^{n} -X_{ir} \frac{\exp(X_i'\beta)}{(1+\exp(X_i'\beta))^2} X_{is} \right) \\ &= \left(\sum_{i=1}^{n} -X_{ir} G(X_i'\beta)\left(1 - G(X_i'\beta)\right) X_{is} \right) \\ &= -X'WX, \end{aligned}$$

where W is a $n \times n$ diagonal matrix with elements $G(X_i'\beta)\left(1 - G(X_i'\beta)\right)$ on the diagonal [116, 127]. Then the MLE $\hat{\beta}$ can be determined iteratively. At each step

$$\begin{aligned} \beta^{(k+1)} &= \beta^{(k)} - D^2\ell(\beta^{(k)}) u(\beta^{(k)}) \\ &= \beta^{(k)} + (X'WX)^{-1} X'(Y-p) \\ &= (X'WX)^{-1} X'W \left(X\beta^{(k)} + W^{-1}(Y-p) \right) \\ &= (X'WX)^{-1} X'W z^{(k)}, \end{aligned}$$

where $z^{(k)} = X\beta^{(k)} + W^{-1}(Y-p)$ is the adjusted $n \times 1$ response vector in a weighted least squares problem [116].

3.2 Multivariate logistic regression

While logistic regression has been established as a standard approach for modeling univariate binary responses, there is no equivalent in the multivariate case for correlated binary outcomes. In the following, we provide a short overview of approaches that generalize univariate logistic regression to the multivariate setting. Then, following O'Brien and Dunson (2004), we introduce multivariate logistic regression using latent variables [132].

Adapting the notation from Bel et al. (2016), we consider $Y_i = (Y_{i1}, \ldots, Y_{iq})$ with $Y_{ij} \in \{0, 1\}$, $j = 1, \ldots, q$, the vector of $q \in \mathbb{N}$ binary outcomes for patient i with $i = 1, \ldots, n$ [133]. Further, let $X_{ij} = (x_{ij1}, \ldots, x_{ijm_j})' \in \mathbb{R}^{m_j}$ be the corresponding m_j-dimensional vector of covariates and $\beta_j \in R^{m_j}$ the coefficient vector to be estimated for outcome j. We assume that Y_i, $i = 1, \ldots, n$ are independent, but consider correlation among Y_{ij}, $j = 1, \ldots q$. We let y_{ij} denote a realization of the random variable Y_{ij}.

For a q - dimensional binary vector Y_i, we have 2^q possible observations that can be obtained. Thus, by transforming the q-dimensional vector to a categorical variable with 2^q categories, we can use a multinomial logistic regression model [133, 134]. However, with an increasing number of binary outcomes, the number of categories grows exponentially, thus requiring an ever increasing sample size n to obtain a sufficient number of observations for each category [133, 134]. For example, the 5 binary outcomes of interest result in $2^5 = 32$ categories. Although we have a large sample size, we see in Figure 2.11 that the number of patients per class varies greatly and is rather low for some classes, which might lead to identifiability problems.

Bahadur (1961) suggested a representation of the multinomial model, which was later extended by Fitzmaurice et al. (1993) [135, 136]. McCullagh and Nelder (1999), Glonek and McCullagh (1995), Ekholm et al. (1995), and Russell and Petersen (2000) proposed modifications and transformations of the parameters used for the multinominal logistic regression model to improve modeling of the dependence between outcomes [127, 133, 137–139]. Ekholm et al. (1995) described how the number of parameters can be reduced by explicitly setting higher-order interactions between outcomes to zero or assuming homogeneity, for example, that the dependence between outcomes is equal for equal-sized subsets [138].

Bel et al. (2016) proposed alternative estimation methods, including stratified importance sampling, composite conditional likelihood and a generalized method of moments, and compared these to maximum likelihood [133]. While these approaches resolved some computational issues, they only provided an approximation to the likelihood.

Dai (2012) and Dai et al. (2013) extended the Bernoulli distribution to the multivariate case and proposed multivariate Bernoulli logistic regression models [140, 141]. Their models

required estimation of $m \times (2^q - 1)$ parameters, similar to other models containing all higher-order interactions [140, 141].

Log-linear models have been commonly used for analysis of correlated categorical variables because they have convenient theoretical properties and straightforward specification and solution of the maximum likelihood [137, 142, 143]. However, Liang et al. (1992) and Glonek and McCullagh (1995) showed that these models are not reproducible in the sense that the marginal models of individual outcomes given the covariates depend on the number of outcomes, and in general are not on the logistic scale [137, 144]. Further, estimated parameters represent logits conditional on the other observed outcomes, thus providing an insufficient generalization of univariate to multivariate logistic regression [137].

Building upon Glonek and McCullagh (1995), Glonek (1996) proposed a mixture of multivariate logistic and log-linear regression models to improve computational feasibility [137, 145]. While providing a flexible class, the proposed model provided only weak reproducibility, with the log-linear component presenting the same issues pertinent to these models discussed previously [145].

Bonney (1987) discussed regressive logistic models, assuming a natural or sensible ordering of the binary outcomes [146]. This is a useful technique for multivariate binary outcomes with a natural ordering, but not applicable in this setting. Joe and Liu (1996) proposed a model for multivariate binary responses where the conditional distribution of one outcome given the other outcomes and covariates is logistic [147]. This model does not provide logistic marginals and thus we do not consider it further.

Liang et al. (1992) used generalized estimating equations (GEE) for modeling multivariate binary data, treating dependence between outcomes as a nuisance parameter and concentrating interest on the marginal means [144, 148]. Second-order generalized estimating equations (GEE2) were presented for modeling pairwise associations between outcomes [144, 148]. The GEE approach for inference concerning the mean is robust to misspecification of the ”working” correlation structure, which is in general not the true correlation [149, 150]. The approach is not likelihood-based, so not amenable to the Bayesian methodology used here [150].

Bonat and Jørgensen (2016) extended univariate generalized linear models (GLM) to the multivariate case with correlated responses and covariates, and proposed multivariate covariance generalized linear models (MCGLM) [151]. The models are implemented in the `R` package mcglm using a Newton scoring algorithm based on quasi-likelihood and Pearson estimating functions [151, 152]. Similar to the GEE approach, the MCGLM is a quasi-likelihood approach and thus not suitable for Bayesian analysis.

While direct extension from univariate to multivariate logistic regression might be desirable and most intuitive, the lack of computationally feasible and stable solutions has led to different approaches using latent variables. Hereby, we consider the univariate latent variable model with random variable Z_i, such that

$$Y_i = \mathbb{1}(Z_i > 0) = \begin{cases} 1 & \text{for } Z_i > 0 \\ 0 & \text{for } Z_i \leq 0. \end{cases}$$

Assuming different probability distributions for Z_i leads to different approaches.

For probit models we assume an underlying normal distribution with $Z_i \sim \mathcal{N}(X_i'\beta, 1)$, and thus $P(Y_i = 1) = P(Z_i > 0) = \Phi(X_i'\beta)$, where Φ denotes the cumulative density function (cdf) of the standard normal distribution [153, 154]. Ashford and Sowden (1970) proposed a comparably easy extension to the multivariate case, where the dependency structure could be described by correlation coefficients in the multivariate normal distribution [132, 154, 155]. Chen and Dey (1998) provided scale mixtures of multivariate normal distributions as a generalization to the underlying normal structure [132, 156]. Although probit models offer a parsimonious description of the correlation structure and appealing computation, interpretation of parameters for the univariate probit model is less intuitive than for the logistic model, which extends to the multivariate case [132]. We aim to extend the logistic model with the interpretation of ORs to the multivariate case. Thus, we do not further consider probit models as they lack this interpretability.

We parametrize the standard logistic regression model as given in (3.1) using a latent variable with underlying logistic distribution $Z_i \sim \mathbb{L}(X_i'\beta, \frac{\pi^2}{3})$ [132, 157]. $\mathbb{L}(\mu, \sigma^2)$ denotes the logistic distribution with density

$$l(z|\mu, \sigma) = \frac{\pi}{\sigma\sqrt{3}} \frac{\exp\left(-\frac{\pi}{\sigma\sqrt{3}}(z-\mu)\right)}{\left[1 + \exp\left(-\frac{\pi}{\sigma\sqrt{3}}(z-\mu)\right)\right]^2},$$

for $z \in \mathbb{R}$, where $\mu \in \mathbb{R}$ is the location parameter and $\sigma > 0$ the scale parameter [157].

In the following we omit the scale parameter σ and set $\mathbb{L}(\cdot) = \mathbb{L}(\cdot\,, \frac{\pi^2}{3})$, with density

$$l(z|\mu) = \frac{\exp\left(-(z-\mu)\right)}{\left[1 + \exp\left(-(z-\mu)\right)\right]^2}. \tag{3.4}$$

For univariate $Y_i = \mathbb{1}(Z_i > 0)$ with $Z_i \sim \mathbb{L}(X_i'\beta)$, we obtain

$$P(Y_i = 1) = P(Z_i > 0) = \int_0^\infty l(z|X_i'\beta)dz = \left[\frac{1}{1+\exp\left(-(z - X_i'\beta)\right)}\right]_0^\infty$$
$$= 1 - \frac{1}{1+\exp\left(X_i'\beta\right)} = \frac{\exp\left(X_i'\beta\right)}{1+\exp\left(X_i'\beta\right)},$$

and thus the probability for $Y_i = 1$ is equivalent to (3.1), the inverse logit of π_i in Section 3.1.

For the multivariate extension to $Y_i = (Y_{i1}, \ldots, Y_{iq})$, we have $Y_{ij} = \mathbb{1}(Z_{ij} > 0)$, where $Z_i = (Z_{i1}, \ldots, Z_{iq})$ are joint distributed with univariate logistic marginals, $Z_{ij} \sim \mathbb{L}(X_{ij}'\beta_j)$, with $X_{ij}, \beta_j \in R^{m_j}$ for $j = 1, \ldots, q$ [132]. We assume that Z_i, $i = 1, \ldots, n$ are independent, but consider correlation among Z_{ij}, $j = 1, \ldots q$.

Before considering the extension of the univariate to multivariate logistic distribution, we discuss the existence of the MLE for the multivariate latent variable model. Lesaffre and Kaufmann (1992) provided a proof for existence and uniqueness of the MLE for the multivariate probit model [158]. Todem and Kim (2007) extended this proof to arbitrary underlying distributions with strictly increasing cdf [159].

Analogous to Albert and Anderson (1984) for univariate logistic regression, Todem and Kim (2007) defined complete separation, quasi-complete separation, and overlap for correlated Bernoulli random variables [129, 159]. They considered

$$\mathcal{U}_j = \left\{(-1)^{y_{ij}} X_{ij} : i \in \{1, \ldots, n\}\right\}, \quad j = 1, \ldots, q,$$

and defined complete separation for outcome j if there exists a vector $\beta_j \in \mathbb{R}^{m_j} \backslash \{0\}$ such that $u'\beta_j < 0$ for all $u \in \mathcal{U}_j$. $\mathcal{U}_j$ is quasi-complete separable if there exists a vector $\beta_j \in \mathbb{R}^{m_j} \backslash \{0\}$ such that $u'\beta_j \leq 0$ for all $u \in \mathcal{U}_j$. $\mathcal{U}_j$ is overlapped if there exists a vector $\beta_j \in \mathbb{R}^{m_j} \backslash \{0\}$ such that $u'\beta_j > 0$ for some $u \in \mathcal{U}_j$ [159]. In addition to $\mathcal{U}_j$, $j = 1, \ldots, q$ they considered the polychoric correlation coefficient $\rho \in [-1, 1]^{q(q-1)/2}$. Pearson (1900) introduced the polychoric correlation coefficient for ordered categorical variables [160, 161]. In the case of two binary variables, it is also referred to as the tetrachoric correlation coefficient and defined as the solution ρ_{jk} to the integral equation

$$p_{jk} = \int_{\Phi^{-1}(1-p_j)}^\infty \int_{\Phi^{-1}(1-p_k)}^\infty \Phi_2(x_1, x_2, \rho_{jk}) dx_1 dx_2,$$

where Φ denotes the cdf of the standard normal distribution, Φ_2 the cdf of the bivariate standard normal distribution, p_j, p_k the marginal probabilities of Y_j, Y_k, respectively, and p_{jk}

the joint probability of Y_j and Y_k [162]. Todem and Kim (2007) showed that the MLE for the multivariate latent variable model exists when ρ is not on the boundary of the parameter space, that is $|\rho_{jk}| \neq 1$, and U_j is overlapped for all $j = 1, \ldots, q$ [159].

The requirements imply that the data have to fulfill the conditions for the existence of the MLE for univariate logistic regression on each outcome separately in addition to the restriction of not being perfectly correlated in terms of the polychoric correlation. We assume that the conditions for the existence of the MLE are fulfilled and continue with the extension of the univariate latent variable model to the multivariate latent variable model with logistic marginals.

Extending the logistic distribution to a multivariate logistic distribution with logistic marginals has been a challenge. Arnold (1992) and Kotz et al. (2005) provided an overview of the development of a multivariate logistic distribution starting from the initial proposal of three bivariate extensions by Gumbel (1961) [163–165]. Arnold (1992) and Kotz et al. (2005) discussed limitations of the different approaches including inflexible or restrictive modeling of the correlation and lack of closed-form representation of the joint density [163, 164]. Li and Wong (2011) proposed a generalization of the Gumbel distribution [166]. Nikoloulopoulos (2012) corrected their assumptions and pointed out that only under specific restrictive constraints on the parameters and for weakly dependent responses, the approach by Li and Wong (2011) could be used [166, 167].

Molenberghs and Lesaffre (1994) proposed the multivariate Plackett distribution for multivariate ordered responses [168]. While it provides interpretation using log odds and ORs, the complexity of parameters is high and higher-order interactions between outcomes are set to zero to ensure computational feasibility [168, 169]. Forcina and Dardanoni (2008) provided a detailed discussion and linked the distribution to the model proposed by Glonek and McCullagh (1995) [137, 169]. The multivariate Plackett distribution as well as some others discussed by Arnold (1992) and Kotz et al. (2005) can be constructed using copulas [148, 170]. Panagiotelis et al. (2012) proposed a pair copula construction for multivariate discrete responses and Stöber et al. (2015) provided an extension to multivariate mixed responses [171, 172]. We follow the approach of O'Brien and Dunson (2004) that can also be described as a t-copula [170, 173].

O'Brien and Dunson (2004) used transformations of conventional multivariate distributions, such as multivariate normal or t-distributions, to obtain logistic marginals [132]. They make use of the fact that for any continuous random variable V with cdf F, $F(V) \sim U(0, 1)$, and it follows that $\mu + \log\left(\frac{F(V)}{1 - F(V)}\right)$ is logistic distributed with location parameter μ.

To show this, let $U \sim U(0,1)$ and $Z = \mu + \log\left(\frac{U}{1-U}\right)$. Then

$$U = \frac{\exp(Z-\mu)}{1+\exp(Z-\mu)} \quad \text{with} \quad \frac{dU}{dZ} = \frac{\exp(Z-\mu)}{[1+\exp(Z-\mu)]^2},$$

and using the transformation theorem for probability distributions,

$$\begin{aligned} f_Z(z) &= \mathbb{1}\left(\frac{\exp(z-\mu)}{1+\exp(z-\mu)} \in [0,1]\right) \cdot \left|\frac{\exp(z-\mu)}{[1+\exp(z-\mu)]^2}\right| \\ &= \frac{\exp(z-\mu)}{[1+\exp(z-\mu)]^2} = \frac{\exp(z-\mu)}{[\exp(-(z-\mu))+1]^2\,[\exp(z-\mu)]^2} \\ &= \frac{\exp(-(z-\mu))}{[1+\exp(-(z-\mu))]^2} = l(z|\mu). \end{aligned}$$

O'Brien and Dunson (2004) implicitly defined the q-dimensional logistic distribution $\mathbb{L}_{q,\nu}(\mu, R)$ using a multivariate t-distribution $\mathcal{T}_{q,\nu}(0, R)$ [132]. The general density of $\mathcal{T}_{q,\nu}(\mu, \Sigma)$ is

$$t_{q,\nu}(v|\mu,\Sigma) = \frac{\Gamma\left(\frac{\nu+q}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{\frac{q}{2}}|\Sigma|^{\frac{1}{2}}} \cdot \left(1 + \frac{1}{\nu}(v-\mu)'\Sigma^{-1}(v-\mu)\right)^{-\frac{\nu+q}{2}}, \tag{3.5}$$

where $\Gamma(u)$ is the gamma function

$$\Gamma(u) = \int_0^\infty x^{u-1}\exp(-x)dx \quad \text{[174]}.$$

Let F_ν denote the cdf of the univariate t-distribution with ν degrees of freedom. Suppose $v = (v_1, \ldots, v_q)' \in \mathbb{R}^q$, $v \sim \mathcal{T}_{q,\nu}(0, R)$, then $z = (z_1, \ldots, z_q)'$ with $z_j = \mu_j + \log\left(\frac{F_\nu(v_j)}{1-F_\nu(v_j)}\right)$ follows a q-dimensional multivariate logistic distribution $\mathbb{L}_{q,\nu}(\mu, R)$. The explicit density of $\mathbb{L}_{q,\nu}(\mu, R)$ for a scale matrix R with 1's on the diagonal is

$$\begin{aligned} l_{q,\nu}(z|\mu,R) = t_{q,\nu}\left(g_\nu(z_1-\mu_1), \ldots, g_\nu(z_q-\mu_q)|0,R\right) \\ \cdot \prod_{j=1}^{q} \frac{l(z_j|\mu_j)}{t_{1,\nu}(g_\nu(z_j-\mu_j)|0,1)}, \end{aligned} \tag{3.6}$$

where $g_\nu(x) = F_\nu^{-1}\left(\frac{\exp(x)}{1+\exp(x)}\right)$ is the inverse of $\log\left(\frac{F_\nu(x)}{1-F_\nu(x)}\right)$ and $l(z_j|\mu_j)$ the univariate logistic density as defined in (3.4). Equation (3.6) reduces to the univariate logistic distribution function for $q = 1$ and the marginals are univariate logistic distributions for $q > 1$ [132].

Albert and Chib (1993) noted that the univariate t-distribution with 8 degrees of freedom

closely resembles the univariate logistic distribution [175]. O'Brien and Dunson (2004) extended this to the multivariate case and approximated $l_{q,\nu}(\cdot|\mu, R)$ with $t_{q,\nu}(\cdot|\mu, \sigma^2 R)$ using $\nu = \widetilde{\nu} = 7.3$ and $\sigma^2 = \widetilde{\sigma}^2 = \dfrac{\pi^2 \cdot (\nu - 2)}{3\nu}$ [132]. They chose these parameters to minimize the squared distance between the two univariate densities and to have equal univariate variances [132]. They then used the approximation to construct an importance sampling algorithm for estimation of the posterior distribution [132]. Following O'Brien and Dunson (2004), we drop ν from the notation, assuming it is fixed at 7.3 if not stated otherwise, and denote by $\mathbb{L}_q(\mu, R)$ the q-variate logistic distribution [132].

We consider the multivariate latent variable model

$$Y_{ij} = \mathbb{1}(Z_{ij} > 0) \quad \text{with} \quad Z_i \sim \mathbb{L}_q(\mu_i, R) \quad \text{and} \quad \mu_i = (X_{i1}'\beta_1, \ldots, X_{iq}'\beta_q), \tag{3.7}$$

where $X_{ij} = (x_{ij1}, \ldots, x_{ijm_j})' \in \mathbb{R}^{m_j}$ and $\beta_j \in R^{m_j}$ for $j = 1, \ldots, q$. Hereby, we explicitly allow for different covariates to be used for different outcomes, while O'Brien and Dunson (2004) assumed equal dimension for all coefficients and thus $m_j = m_k$ for all $j = 1, \ldots, q$ [132]. The probability of a multivariate binary outcome y_i under (3.7) is

$$\begin{aligned} P(Y_i = y_i|\mu_i, R) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} l_q(z_i|\mu_i, R) \cdot \prod_{j=1}^{q} \mathbb{1}(z_{ij} > 0)^{y_{ij}} \mathbb{1}(z_{ij} \leq 0)^{1-y_{ij}} dz_i \\ &= \int_{A_{i1}} \cdots \int_{A_{iq}} l_q(z_i|\mu_i, R) dz_i, \end{aligned}$$

where

$$A_{ij} = \begin{cases} (0, \infty) & \text{for} \quad y_{ij} = 1 \\ (-\infty, 0] & \text{for} \quad y_{ij} = 0. \end{cases}$$

Then the likelihood of (3.7) for a random sample y of size n is given by

$$\mathcal{L}(\beta, R) = \prod_{i=1}^{n} \int_{A_{i1}} \cdots \int_{A_{iq}} l_q(z_i|\mu_i, R) dz_i, \tag{3.8}$$

with $\beta = \left(\beta_{11}, \ldots, \beta_{1m_1}, \ldots, \beta_{q1}, \ldots, \beta_{qm_q}\right)' \in \mathbb{R}^m$, $m = \sum_{j=1}^{q} m_j$.

O'Brien and Dunson (2004) proposed a Bayesian approach for estimation and inference [132]. The joint posterior density of β and R given the data y is

$$p(\beta, R|y) \propto f(\beta, R)\mathcal{L}(\beta, R), \tag{3.9}$$

where $f(\beta, R)$ denotes the joint prior density of β and R [132].

Assuming for simplicity that $f(\beta, R) = f(\beta)f(R)$, O'Brien and Dunson (2004) specified a normal prior distribution $f(\beta) \sim \mathcal{N}_m(\mu_\beta, \Sigma_\beta)$ or an improper uniform prior distribution $f(\beta) \propto 1$ for β and a uniform prior distribution with support on the space of correlation matrices for R. They stated conditions required so that the improper uniform prior results in a proper posterior distribution [132].

Calculating the posterior density (3.9) is difficult as the likelihood function is complex. Therefore, O'Brien and Dunson (2004) proposed using the multivariate t-distribution $t_{q,\widetilde{\nu}}(z_i|\mu_i, \widetilde{\sigma}^2 R)$ instead of the multivariate logistic distribution in (3.8), yielding

$$\mathcal{L}^*(\beta, R) = \prod_{i=1}^{n} \int_{A_{i1}} \cdots \int_{A_{iq}} t_{q,\widetilde{\nu}}(z_i|\mu_i, \widetilde{\sigma}^2 R) dz_i, \tag{3.10}$$

with $\widetilde{\nu} = 7.3$ and $\widetilde{\sigma}^2 = \dfrac{\pi \cdot (\widetilde{\nu} - 2)}{3\widetilde{\nu}}$. The corresponding posterior distribution is

$$p^*(\beta, R|y) \propto f(\beta, R) \cdot \mathcal{L}^*(\beta, R), \tag{3.11}$$

and the exact posterior distribution $p(\beta, R|y)$ is obtained by assigning appropriate importance weights [132].

We adapt the Markov Chain Monte Carlo (MCMC) algorithm proposed by O'Brien and Dunson (2004) based on Albert and Chib (1993) to explicitly include different numbers of covariates for different outcomes [132, 175]. We specify the likelihood (3.10) alternatively with

$$\begin{aligned} y_{ij} &= \mathbb{1}(z_{ij} > 0) \\ z_i|\beta, R, \phi_i &\sim \mathcal{N}_q(\mu_i, \widetilde{\sigma}^2 \phi_i^{-1} R) \\ \phi_i|\beta, R &\sim \Gamma\left(\frac{\widetilde{\nu}}{2}, \frac{\widetilde{\nu}}{2}\right), \end{aligned}$$

where $\Gamma(a, b)$ denotes the gamma distribution with shape parameter a and rate parameter b.

Then the joint posterior distribution of the parameters and latent variables is

$$p^*(\beta, R, \phi, z|y) \propto f(\beta)\, f(R)\, f(\phi)\, f^*(z|\beta, R, \phi)\, f(y|\beta, R, \phi, z), \tag{3.12}$$

with

$$\begin{aligned} f(\beta) &= f_{\mathcal{N}_m}(\beta|\mu_\beta, \Sigma_\beta) \\ f(R) &= \text{ any distribution with support on the space of correlation matrices} \end{aligned}$$

$$f(\phi) = \prod_{j=1}^{q} f_\Gamma\left(\phi_i \Big| \frac{\widetilde{\nu}}{2}, \frac{\widetilde{\nu}}{2}\right)$$
$$f^*(z|\beta, R, \phi) = \prod_{i=1}^{n} f_{\mathcal{N}_q}(z_i|\mu_i, \widetilde{\sigma}^2\phi_i^{-1}R)$$
$$f(y|\beta, R, \phi, z) = \prod_{i=1}^{n}\prod_{j=1}^{q} \mathbb{1}(z_{ij} > 0)^{y_{ij}}\mathbb{1}(z_{ij} \leq 0)^{1-y_{ij}},$$

where $f_\mathcal{N}$ and f_Γ denote the multivariate normal and gamma distributions, respectively. An improper uniform prior distribution for β uses the normal prior, but sets $\Sigma_\beta^{-1} = 0$ [132].

The MCMC algorithm in Algorithm 3.1 alternates between sampling from the full conditional distributions of z, ϕ and β and a Metropolis step for updating R. For the algorithm, we denote the $q^* = \frac{q(q-1)}{2}$ unique elements of R by $r_1, \ldots, r_{q^*}$. Besides initial values for ϕ, β, and R, we experimentally choose Ω, the variance-covariance matrix in the normal proposal density for the elements of R, for a desired acceptance probability in the Metropolis step [132].

Algorithm 3.1 samples from the approximate full posterior distribution (3.12) and $p^*(\beta, R|y)$ defined in (3.11) is obtained as the marginal. For inference of the desired posterior $p(\beta, R|y)$ based on the multivariate logistic distribution, we need to correct the approximation by Algorithm 3.1. We use the appropriate sampling weights

$$\omega^{(t)} \propto \frac{p(\beta^{(t)}, R^{(t)}, z^{(t)}|y)}{p^*(\beta^{(t)}, R^{(t)}, z^{(t)}|y)},$$

which we compute with

$$\omega^{(t)} = \frac{f(z^{(t)}|\beta^{(t)}, R^{(t)})}{f^*(z^{(t)}|\beta^{(t)}, R^{(t)})} = \prod_{i=1}^{n} \frac{l_q\left(z_i^{(t)}|\mu_i^{(t)}, R^{(t)}\right)}{t_{q,\widetilde{\nu}}\left(z_i^{(t)}|\mu_i^{(t)}, R^{(t)}\right)}.$$

With the weights we estimate the posterior expectation of functionals $h(\beta, R)$, such as the mean or quantile of parameters, given by

$$\mathbb{E}\left[h(\beta, R)|y\right] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\beta, R)p(\beta, R|y)d\beta dR$$

with

$$\widehat{\mathbb{E}}\left[h(\beta, R)|y\right] = \frac{\sum_{t=B+1}^{T} \omega^{(t)} h(\beta^{(t)}, R^{(t)})}{\sum_{t=B+1}^{T} \omega^{(t)}},$$

where T denotes the total number of iterations and B the burn-in of the MCMC algorithm [132].

Algorithm 3.1: Markov Chain Monte Carlo algorithm [132]

Input: $y \in \mathbb{R}^{n \times q}$, $X = (X_1, \ldots, X_n)$ with $X_i \in \mathbb{R}^{q \times m}$, $\beta^{(0)} \in \mathbb{R}^m$, $R^{(0)} \in \mathbb{R}^{q \times q}$, $\phi^{(0)} \in \mathbb{R}^n$, $\mu_\beta \in \mathbb{R}^m$, $\Sigma_\beta \in \mathbb{R}^{m \times m}$, $\Omega \in \mathbb{R}^{q^* \times q^*}$, T number of iterations

for $t = 1, \ldots, T$ **do**

  for $i = 1, \ldots, n$ **do**

    Sample $z_i^{(t)} \sim \mathcal{N}_q\left(\mu_i^{(t-1)}, \frac{\widetilde{\sigma}^2}{\phi_i^{(t-1)}} R^{(t-1)}\right)$,

    with z_{ij} truncated to > 0 (< 0) for $y_{ij} = 1$ ($y_{ij} = 0$)

  end

  for $i = 1, \ldots, n$ **do**

    Sample $\phi_i^{(t)} \sim \Gamma\left(\frac{\widetilde{\nu} + q}{2}, \frac{\widetilde{\nu} + \widetilde{\sigma}^{-2}\left(z_i^{(t)} - \mu_i^{(t-1)}\right)' \left(R^{(t-1)}\right)^{-1}\left(z_i^{(t)} - \mu_i^{(t-1)}\right)}{2}\right)$

  end

  $\widetilde{\Sigma}_\beta \leftarrow \left(\Sigma_\beta^{-1} + \widetilde{\sigma}^{-2} \sum_{i=1}^n \phi_i^{(t)} X_i' \left(R^{(t-1)}\right)^{-1} X_i\right)^{-1}$

  $\widetilde{\mu}_\beta \leftarrow \widetilde{\Sigma}_\beta \left(\Sigma_\beta^{-1}\mu_\beta + \widetilde{\sigma}^{-2} \sum_{i=1}^n \phi_i^{(t)} X_i' \left(R^{(t-1)}\right)^{-1} z_i^{(t)}\right)$

  Sample $\beta^{(t)} \sim \mathcal{N}_m\left(\widetilde{\mu}_\beta, \widetilde{\Sigma}_\beta\right)$

  for $i = 1, \ldots, n$ **do**

    $\mu_i^{(t)} \leftarrow= X_i \beta^{(t)}$

  end

  Sample $\widetilde{r}_1, \ldots, \widetilde{r}_{q^*} \sim \mathcal{N}_{q^*}\left(\left(r_1^{(t-1)}, \ldots, r_{q^*}^{(t-1)}\right)', \Omega\right)$

  if $\widetilde{R}$ is positive definite **then**

    Sample $u \sim \mathcal{U}(0, 1)$

    $$\rho^{(t)} \leftarrow \frac{p(\widetilde{R}) \prod_{i=1}^n N_q\left(z_i^{(t)} \middle| \mu_i^{(t)}, \frac{\widetilde{\sigma}^2}{\phi_i^{(t)}} \widetilde{R}\right)}{p(R^{(t-1)}) \prod_{i=1}^n N_q\left(z_i^{(t)} \middle| \mu_i^{(t)}, \frac{\widetilde{\sigma}^2}{\phi_i^{(t)}} R^{(t-1)}\right)}$$

    if $u < \min(1, \rho^{(t)})$ **then**

      $R^{(t)} \leftarrow \widetilde{R}$

    else

      $R^{(t)} \leftarrow R^{(t-1)}$

    end

  else

    $R^{(t)} \leftarrow R^{(t-1)}$

  end

end

Nooraee et al. (2016) extended model (3.7) by implementing covariates in the covariance structure [173]. Using the approximation of the multivariate logistic distribution by the multivariate t-distribution, they proposed approximate maximum likelihood estimation with a quasi-Newton method, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [173]. Thereby, Nooraee et al. (2016) used an alternative representation of the t-distribution that is equivalent to (3.5) [173].

Hirk et al. (2018) applied composite likelihood methods for estimation of the MLE that is implemented in the `R` package `mvord` [176, 177]. Like GEEs, composite likelihood methods are pseudo-likelihood approaches that do not specify the full joint distribution. For a review of the method see Varin (2008) and Varin et al. (2011) [178, 179]. Gao and Song (2010) proposed a Bayesian information criterion (BIC) using the composite likelihood [180].

Caubet Fernandez et al. (2019) praised the high interpretability of model (3.7) using log odds and pointed out that dependencies are accounted for in an unrestricted fashion [181]. They implemented Algorithm 3.1 to sample from the posterior distribution, but without using importance weights for exact inference based on the multivariate logistic distribution [181]. We build on `R` and C++ code provided by Caubet Fernandez et al. (2019) to fit the multivariate logistic regression model [181].

3.3 Bayes factors

In the following we are interested in comparing two nested models M_0 and M_1 [182]. Let $\beta = (\theta, \psi) \in \Theta \times \Psi$, with $\dim(\Theta) = m_0$ and $\dim(\Psi) = m - m_0$. We compare M_0 with $\beta_0 = (\theta, \psi_0)$, where ψ_0 is fixed, to M_1 with $\beta_1 = (\theta, \psi)$. Thus we test the hypotheses $H_0 : \psi = \psi_0$ versus the alternative $H_1 : \psi \neq \psi_0$.

Following a Bayesian paradigm, we assume prior model probabilities $p(M_0)$ and $p(M_1) = 1 - p(M_0)$ and prior distributions $f(\beta_0|M_0)$ and $f(\beta_1|M_1)$ for the parameters under the respective models. Probability densities of Y under the models, denoted by $p(Y|M_0)$ and $p(Y|M_1)$, are obtained by integrating over the respective parameter spaces

$$p(Y|M_k) = \int p(Y|\beta_k, M_k) f(\beta_k|M_k) d\beta_k, \text{ for } k = 0, 1, \tag{3.13}$$

where $p(Y|\beta_k, M_k)$ is the probability density of Y given β_k under model M_k, and is also the likelihood that can be denoted by $\mathcal{L}(\beta_k)$ [182]. Posterior probabilities of each model given the data Y are $p(M_k|Y)$ for $k = 0, 1$, which by Bayes' theorem can be expressed as

$$p(M_k|Y) = \frac{p(Y|M_k)p(M_k)}{P(Y)} = \frac{p(Y|M_k)p(M_k)}{p(Y|M_0)p(M_0) + p(Y|M_1)p(M_1)}\,.$$

Thus the posterior odds comparing M_0 to M_1 are given by

$$\frac{p(M_0|Y)}{p(M_1|Y)} = \frac{\frac{p(Y|M_0)p(M_0)}{p(Y|M_0)p(M_0)+p(Y|M_1)p(M_1)}}{\frac{p(Y|M_1)p(M_1)}{p(Y|M_0)p(M_0)+p(Y|M_1)p(M_1)}} = \frac{p(Y|M_0)}{p(Y|M_1)}\frac{p(M_0)}{p(M_1)},$$

and reduce to the prior odds $\frac{p(M_0)}{p(M_1)}$ multiplied by $\frac{p(Y|M_0)}{p(Y|M_1)}$. This last fraction is called the Bayes factor (BF) and can be interpreted as the ratio of posterior odds to prior odds [182]. From (3.13)

$$\text{BF} = \frac{\int p(Y|\beta_0, M_0) f(\beta_0|M_0) d\beta_0}{\int p(Y|\beta_1, M_1) f(\beta_1|M_1) d\beta_1} = \frac{\int \mathcal{L}(\beta_0) f(\beta_0|M_0) d\beta_0}{\int \mathcal{L}(\beta_1) f(\beta_1|M_1) d\beta_1} \quad \text{[182]}.$$

We simplify the notation as ψ_0 is fixed and write $f_0(\theta) = f(\beta_0|M_0)$ and $f(\theta, \psi) = f(\beta_1|M_1)$ to yield

$$\text{BF} = \frac{\int \mathcal{L}(\theta, \psi_0) f_0(\theta) d\theta}{\iint \mathcal{L}(\theta, \psi) f(\theta, \psi) d\theta d\psi}.$$

In general $f_0(\theta)$ and $f(\theta, \psi)$ do not have to be logically related, but it might be desirable to connect them [15]. As an example $f_0(\theta)$ can be chosen to be the marginal distribution

$$f_0(\theta) = \int f(\theta, \psi) d\psi = f_\theta(\theta) \tag{3.14}$$

or the conditional distribution

$$f_0(\theta) = f(\theta|\psi_0) = \frac{f(\theta, \psi_0)}{\int f(\theta, \psi_0) d\theta} = \frac{f(\theta, \psi_0)}{f_\psi(\psi_0)}. \tag{3.15}$$

If θ and ψ are a priori independent under M_1 (3.14) and (3.15) define the same prior density

$$f_0(\theta) = \frac{f(\theta, \psi_0)}{f_\psi(\psi_0)} = \frac{f_\theta(\theta) f_\psi(\psi_0)}{f_\psi(\psi_0)} = f_\theta(\theta),$$

but they do not have to be equal in general [15].

Jeffreys introduced rules for the BF to judge the evidence against the null hypothesis [183]. Jeffreys' rule was adapted by Kass and Raftery (1995) as shown in Table 3.3 [182]. The BF is considered on a negative $\log_{10}$-scale for easier comparison. An interpretation of a BF of 10^{-1} is that the posterior probability for model M_1 is 10 times higher compared to the posterior probability of M_0 under the assumption that their prior probabilities are equal.

$-\log_{10}$(BF)	BF	Evidence against H_0
< 0	> 1	Null hypothesis supported
0 to $\frac{1}{2}$	$10^{-1/2}$ to 1	Not worth more than a bare mention
$\frac{1}{2}$ to 1	10^{-1} to $10^{-1/2}$	substantial
1 to 2	10^{-2} to 10^{-1}	strong
> 2	$< 10^{-2}$	decisive

Table 3.3 Jeffreys' rule for the BF as adapted by Kass and Raftery (1995) [182].

3.4 Order of approximation for stochastic sequences

Following Bishop et al. (2007) we introduce the o_p and O_p notation, which were unified by Wald and Mann (1943) [114, 184]. We recall the definitions of o and O for non-stochastic real sequences and introduce their extension for vector sequences as covered by Bishop et al. (2007) [114]. Matrices can be vectorized so follow the same definitions.

Definition 3.1 (O and o notation for non-stochastic sequences)
Let $\{a_n\}$ and $\{b_n\}$ be two sequences of real numbers and $\{v_n\}$ a sequence of vectors with the corresponding euclidean norm $||v_n|| = \sqrt{\sum_i v_{ni}^2}$. We define

(i) $a_n = O(b_n)$ if there exists a number K and an integer n_K such that for all $n \geq n_K$ we have $|a_n| < K|b_n|$,

(ii) $a_n = o(b_n)$ if for all $\varepsilon > 0$, there exists an integer n_ε such that for all $n \geq n_\epsilon$ we have $|a_n| < \varepsilon|b_n|$,

(iii) $v_n = O(b_n)$ if $||v_n|| = O(b_n)$, and $v_n = o(b_n)$ if $||v_n|| = o(b_n)$.

Definition 3.2 (O_p notation for stochastic sequences)
Let $\{Y_n\}$ be a sequence of real-valued random variables and $\{b_n\}$ sequence of real numbers. We define $Y_n = O_p(b_n)$ if for every $\eta > 0$ there exists a constant K_η and an integer n_η such that for all $n \geq n_\eta$ it holds

$$P\left(\left|\frac{Y_n}{b_n}\right| \leq K_\eta\right) \geq 1 - \eta$$

or equivalently, if $\dfrac{Y_n}{b_n}$ is bounded in probability. For a sequence of random vectors $\{X_n\}$ we have $X_n = O_p(b_n)$ if $||X_n|| = O_p(b_n)$.

Definition 3.3 (o_p notation for stochastic sequences)
Let $\{Y_n\}$ be a sequence of real-valued random variables and $\{b_n\}$ a sequence of real numbers. We define $Y_n = o_p(b_n)$ if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{Y_n}{b_n}\right| \leq \varepsilon\right) = 1\,,$$

or equivalently, if $\dfrac{Y_n}{b_n}$ converges in probability to 0. For a sequence of random vectors $\{X_n\}$ we have $X_n = o_p(b_n)$ if $||X_n|| = o_p(b_n)$.

In the following we prove some properties of O_p, which we repeatedly use in proofs later on.

Proposition 3.1
Let $\{a_n\}$ and $\{b_n\}$ be two sequences of real numbers and $\{Y_n\}, \{Z_n\}$ be two sequences of real-valued random variables with $\{Y_n\} = O_p(a_n)$ and $\{Z_n\} = O_p(b_n)$. Then it follows

(i) $Y_n Z_n = O_p(a_n)O_p(b_n) = O_p(a_n b_n)$.

It further holds that if $a_n = O(b_n)$,

(ii) $Y_n = O_p(a_n) = O_p(b_n)$ and

(iii) $Y_n + Z_n = O_p(a_n) + O_p(b_n) = O_p(b_n)$.

Suppose that $b_n = O(1)$ and $a_n = O(b_n)$, then

(iv) $\dfrac{1+a_n}{1+b_n} - 1 = \dfrac{1+O(a_n)}{1+O(b_n)} - 1 = O(b_n)$ and

(v) $\dfrac{1+Y_n}{1+Z_n} - 1 = \dfrac{1+O_p(a_n)}{1+O_p(b_n)} - 1 = O_p(b_n)$.

Proof.

(i) With $\{Y_n\} = O_p(a_n)$ and $\{Z_n\} = O_p(b_n)$ it follows that for every $\eta_a, \eta_b > 0$ there exist constants K_{η_a}, K_{η_b} and integers n_{η_a}, n_{η_b} such that for all $n_a \geq n_{\eta_a}, n_b \geq n_{\eta_b}$ it holds

$$P\left(\left|\frac{Y_{n_a}}{a_{n_a}}\right| \leq K_{\eta_a}\right) \geq 1-\eta_a \quad \text{and} \quad P\left(\left|\frac{Z_{n_b}}{b_{n_b}}\right| \leq K_{\eta_b}\right) \geq 1-\eta_b.$$

Let $\eta > 0$ be arbitrary. For $n \geq \max(n_{\eta_a}, n_{\eta_b})$ and $K_\eta = K_{\eta_a} K_{\eta_b}$ we have

$$\begin{aligned}
P\left(\left|\frac{Y_n Z_n}{a_n b_n}\right| \leq K_\eta\right) &= P\left(\left|\frac{Y_n}{a_n}\right| \left|\frac{Z_n}{b_n}\right| \leq K_{\eta_a} K_{\eta_b}\right) \\
&\geq P\left(\left|\frac{Y_n}{a_n}\right| \left|\frac{Z_n}{b_n}\right| \leq K_{\eta_a} K_{\eta_b} \cap \left|\frac{Z_n}{b_n}\right| \leq K_{\eta_b}\right) \\
&= P\left(\left|\frac{Y_n}{a_n}\right| \left|\frac{Z_n}{b_n}\right| \leq K_{\eta_a} K_{\eta_b} \;\middle|\; \left|\frac{Z_n}{b_n}\right| \leq K_{\eta_b}\right) \cdot P\left(\left|\frac{Z_n}{b_n}\right| \leq K_{\eta_b}\right) \\
&\geq P\left(\left|\frac{Y_n}{a_n}\right| \leq K_{\eta_a}\right) P\left(\left|\frac{Z_n}{b_n}\right| \leq K_{\eta_b}\right) \\
&\geq (1-\eta_a)(1-\eta_b) = 1 - \eta_a - \eta_b + \eta_a \eta_b \\
&\geq 1 - \eta,
\end{aligned}$$

where we choose $\eta_a, \eta_b > 0$ such that $\eta_a + \eta_b - \eta_a \eta_b \leq \eta$. Thus $Y_n Z_n = O_p(a_n) O_p(b_n) = O_p(a_n b_n)$.

(ii) With $a_n = O(b_n)$ it follows that there exists a constant K_a and an integer n_{K_a} such that for all $n \geq n_{K_a}$ we have

$$\frac{|a_n|}{|b_n|} < K_a.$$

Further assume $\eta_a, K_{\eta_a}, n_{\eta_a}$ as defined in (i). Let $\eta > 0$ be arbitrary. For $n \geq \max(n_{\eta_a}, n_{K_a})$ and $K_\eta = K_a K_{\eta_a}$ we have

$$\begin{aligned}
P\left(\left|\frac{Y_n}{b_n}\right| \leq K_\eta\right) = P\left(\left|\frac{Y_n}{a_n}\right| \left|\frac{a_n}{b_n}\right| \leq K_a K_{\eta_a}\right) &\geq P\left(\left|\frac{Y_n}{a_n}\right| K_a \leq K_a K_{\eta_a}\right) \\
= P\left(\left|\frac{Y_n}{a_n}\right| \leq K_{\eta_a}\right) &\geq 1 - \eta_a = 1 - \eta,
\end{aligned}$$

where we choose $\eta_a = \eta$.

(iii) We have $Z_n = O_p(b_n)$ and from (ii) it follows that $Y_n = O_p(b_n)$. Thus for every $\eta_b > 0$ there exist constants K_{η_Y}, K_{η_Z} and integers n_{η_Y}, n_{η_Z} such that for all $n \geq \max(n_{\eta_Y}, n_{\eta_Z})$ it holds

$$P\left(\left|\frac{Y_n}{b_n}\right| \leq K_{\eta_Y}\right) \geq 1 - \eta_b \quad \text{and} \quad P\left(\left|\frac{Z_n}{b_n}\right| \leq K_{\eta_Z}\right) \geq 1 - \eta_b.$$

Let $\eta > 0$ be arbitrary. For $n \geq \max(n_{\eta_Y}, n_{\eta_Z})$ and $K_\eta = K_{\eta_Y} + K_{\eta_Z}$ we have

$$\begin{aligned}
P\left(\left|\frac{Y_n + Z_n}{b_n}\right| \leq K_\eta\right) &= P\left(\left|\frac{Y_n}{b_n} + \frac{Z_n}{b_n}\right| \leq K_{\eta_Y} + K_{\eta_Z}\right) \\
&\geq P\left(\left|\frac{Y_n}{b_n}\right| + \left|\frac{Z_n}{b_n}\right| \leq K_{\eta_Y} + K_{\eta_Z}\right) \\
&\geq P\left(\left|\frac{Y_n}{b_n}\right| + \left|\frac{Z_n}{b_n}\right| \leq K_{\eta_Y} + K_{\eta_Z} \cap \left|\frac{Z_n}{b_n}\right| \leq K_{\eta_Z}\right) \\
&= P\left(\left|\frac{Y_n}{b_n}\right| + \left|\frac{Z_n}{b_n}\right| \leq K_{\eta_Y} + K_{\eta_Z} \;\middle|\; \left|\frac{Z_n}{b_n}\right| \leq K_{\eta_Z}\right) P\left(\left|\frac{Z_n}{b_n}\right| \leq K_{\eta_Z}\right) \\
&\geq P\left(\left|\frac{Y_n}{b_n}\right| \leq K_{\eta_Y}\right) P\left(\left|\frac{Z_n}{b_n}\right| \leq K_{\eta_Z}\right) \\
&\geq (1-\eta_b)(1-\eta_b) = 1 - 2\eta_b + \eta_b^2 \geq 1 - \eta,
\end{aligned}$$

where we choose $\eta_b > 0$ such that $\eta_b^2 - 2\eta_b \leq \eta$.

(iv) With $b_n = O(1)$ it follows that there exists a constant K_b and an integer n_{K_b} such that for all $n \geq n_{K_b}$ we have

$$\frac{|b_n|}{|1|} = |b_n| < K_b.$$

Further we consider K_a, n_{K_a} as defined in (ii). For $n \geq \max(n_{K_a}, n_{K_b})$ and $K_{ab} = \frac{1+K_a}{|1-K_b|}$ we have

$$\begin{aligned}
\left|\frac{\frac{1+a_n}{1+b_n} - 1}{b_n}\right| = \left|\frac{a_n - b_n}{(1+b_n)b_n}\right| &\leq \frac{1}{|1+b_n|}\left(\left|\frac{a_n}{b_n}\right| + \left|\frac{b_n}{b_n}\right|\right) < \frac{1}{|1+b_n|}(K_a + 1) \\
&\leq \frac{1+K_a}{|1-|b_n||} \leq \frac{1+K_a}{|1-K_b|} = K_{ab}.
\end{aligned}$$

Here we used that $|1+x| \geq |1-|x||$ for $x \in \mathbb{R}$, since

$$|1+x| = \begin{cases} |1+|x|| > |1-|x|| & \text{for } x \geq 0 \\ |1-|x|| & \text{for } x < 0 \end{cases}.$$

It follows $\dfrac{1}{|1+x|} \leq \dfrac{1}{|1-|x||}$.

(v) We consider again $\eta_b, K_{\eta_Y}, K_{\eta_Z}, n_{\eta_Y}, n_{\eta_Z}$ as defined in (iii) and K_b, n_{K_b} as defined in (iv). Let $\eta > 0$ be arbitrary. For $n \geq \max(n_{\eta_Y}, n_{\eta_Z}, n_{K_b})$ and

$$K_\eta = \frac{K_{\eta_Y} + K_{\eta_Z}}{|1 - K_{\eta_Z} K_b|}$$

we have again with $\frac{1}{|1+x|} \leq \frac{1}{|1-|x||}$ for $x \in \mathbb{R}$

$$
\begin{aligned}
P\left(\left|\frac{\frac{1+Y_n}{1+Z_n}-1}{b_n}\right| \leq K_\eta\right) &= P\left(\left|\frac{Y_n - Z_n}{b_n(1+Z_n)}\right| \leq K_\eta\right) \\
&\geq P\left(\frac{1}{|1+Z_n|}\left(\left|\frac{Y_n}{b_n}\right| + \left|\frac{Z_n}{b_n}\right|\right) \leq K_\eta\right) \\
&\geq P\left(\frac{1}{|1-|Z_n||}\left(\left|\frac{Y_n}{b_n}\right| + \left|\frac{Z_n}{b_n}\right|\right) \leq K_\eta\right) \\
&= P\left(\frac{1}{\left|1-\frac{|Z_n|}{|b_n|}|b_n|\right|}\left(\left|\frac{Y_n}{b_n}\right| + \left|\frac{Z_n}{b_n}\right|\right) \leq K_\eta\right) \\
&\geq P\left(\frac{1}{\left|1-\left|\frac{Z_n}{b_n}\right| K_b\right|}\left(\left|\frac{Y_n}{b_n}\right| + \left|\frac{Z_n}{b_n}\right|\right) \leq \frac{K_{\eta_Y}+K_{\eta_Z}}{|1-K_{\eta_Z}K_b|}\right) \\
&\geq P\left(\frac{1}{\left|1-\left|\frac{Z_n}{b_n}\right| K_b\right|}\left(\left|\frac{Y_n}{b_n}\right| + \left|\frac{Z_n}{b_n}\right|\right) \leq \frac{K_{\eta_Y}+K_{\eta_Z}}{|1-K_{\eta_Z}K_b|} \cap \left|\frac{Z_n}{b_n}\right| \leq K_{\eta_Z}\right) \\
&= P\left(\frac{1}{\left|1-\left|\frac{Z_n}{b_n}\right| K_b\right|}\left(\left|\frac{Y_n}{b_n}\right| + \left|\frac{Z_n}{b_n}\right|\right) \leq \frac{K_{\eta_Y}+K_{\eta_Z}}{|1-K_{\eta_Z}K_b|} \;\middle|\; \left|\frac{Z_n}{b_n}\right| \leq K_{\eta_Z}\right) \\
&\quad \cdot P\left(\left|\frac{Z_n}{b_n}\right| \leq K_{\eta_Z}\right) \\
&\geq P\left(\frac{1}{|1-K_{\eta_Z}K_b|}\left(\left|\frac{Y_n}{b_n}\right| + K_{\eta_Z}\right) \leq \frac{K_{\eta_Y}+K_{\eta_Z}}{|1-K_{\eta_Z}K_b|}\right) \cdot P\left(\left|\frac{Z_n}{b_n}\right| \leq K_{\eta_Z}\right) \\
&\geq P\left(\left|\frac{Y_n}{b_n}\right| + K_{\eta_Z} \leq K_{\eta_Y} + K_{\eta_Z}\right) \cdot P\left(\left|\frac{Z_n}{b_n}\right| \leq K_{\eta_Z}\right) \\
&\geq P\left(\left|\frac{Y_n}{a_n}\right| \leq K_{\eta_Y}\right) \cdot P\left(\left|\frac{Z_n}{b_n}\right| \leq K_{\eta_Z}\right) \\
&= (1-\eta_b)(1-\eta_b) = 1 - 2\eta_b + \eta_b^2 \geq 1 - \eta,
\end{aligned}
$$

where we choose $\eta_b > 0$ such that $\eta_b^2 - 2\eta_b \leq \eta$.

□

Remark
Similar to the proofs in Proposition 3.1 it can be shown that

- $O(a_n)O(b_n) = O(a_n b_n)$,
- $O_p(a_n)O(b_n) = O_p(a_n b_n)$,
- $O(a_n) + O(b_n) = \begin{cases} O(a_n) & \text{if } b_n = O(a_n) \\ O(b_n) & \text{if } a_n = O(b_n), \end{cases}$
- $O_p(a_n) + O(b_n) = \begin{cases} O_p(a_n) & \text{if } b_n = O(a_n) \\ O_p(b_n) & \text{if } a_n = O(b_n), \end{cases}$
- $\frac{1 + O(a_n)}{1 + O(b_n)} = \begin{cases} 1 + O(a_n) & \text{if } a_n = O(1) \text{ and } b_n = O(a_n) \\ 1 + O(b_n) & \text{if } b_n = O(1) \text{ and } a_n = O(b_n), \text{ and} \end{cases}$
- $\frac{1 + O_p(a_n)}{1 + O_p(b_n)} = \begin{cases} 1 + O_p(a_n) & \text{if } a_n = O(1) \text{ and } b_n = O(a_n) \\ 1 + O_p(b_n) & \text{if } b_n = O(1) \text{ and } a_n = O(b_n). \end{cases}$

Proposition 3.2
Let $\{a_n\}$ be a sequence of real numbers and $\{Y_n\}$ be a sequence of real-valued random variables with $\{Y_n\} = O_p(a_n)$. Then it follows

(i) $\log(1 + Y_n) = \log(1 + O_p(a_n)) = O_p(a_n)$, if $1 + Y_n > 0$.

(ii) $(1 + Y_n)^{\frac{1}{2}} - 1 = (1 + O_p(a_n))^{\frac{1}{2}} - 1 = O_p(a_n)$, if $1 + Y_n > 0$.

Suppose that $a_n = O(1)$, then

(iii) $\exp(Y_n) = \exp\left(O_p(a_n)\right) = 1 + O_p(a_n)$.

Proof.

(i) With $\{Y_n\} = O_p(a_n)$ it follows that for every $\eta_a > 0$ there exist a constant K_{η_a} and an integer n_{η_a} such that for all $n \geq n_{\eta_a}$ it holds $P\left(\left|\frac{Y_n}{a_n}\right| \leq K_{\eta_a}\right) \geq 1 - \eta_a$. Let $\eta > 0$ be arbitrary. For $n \geq n_{\eta_a}$ and $K_\eta = K_{\eta_a}$ we have

$$P\left(\left|\frac{\log(1 + Y_n)}{a_n}\right| \leq K_\eta\right) \geq P\left(\left|\frac{Y_n}{a_n}\right| \leq K_{\eta_a}\right) \geq 1 - \eta_a = 1 - \eta,$$

where we used that $\log(x) \leq x - 1$ for $x \in (0, \infty)$ and chose $\eta_a = \eta$.

(ii) Let K_{η_a} and n_{η_a} be defined as in (i) and suppose $\eta > 0$ is arbitrary. For $n \geq n_{\eta_a}$ and $K_\eta = K_{\eta_a}$ we have

$$\begin{aligned} P\left(\left|\frac{(1+Y_n)^{\frac{1}{2}}-1}{a_n}\right| \leq K_\eta\right) &\geq P\left(\left|\frac{1+|Y_n|-1}{a_n}\right| \leq K_\eta\right) \\ &= P\left(\left|\frac{Y_n}{a_n}\right| \leq K_\eta\right) \\ &\geq 1-\eta_a = 1-\eta, \end{aligned}$$

where we used that $\sqrt{1+x} \leq 1+|x|$ for $x \in [-1,\infty)$ and chose $\eta_a = \eta$.

(iii) Let K_{η_a} and n_{η_a} be defined as in (i) and with $a_n = O(1)$ it follows that there exists a constant K_a and an integer n_{K_a} such that for all $n \geq n_{K_a}$ we have

$$\frac{|a_n|}{|1|} = |a_n| < K_a.$$

Let $\eta > 0$ be arbitrary. For $n \geq \max(n_{\eta_a}, n_{K_a})$ and

$$K_\eta = \frac{\exp(K_{\eta_a}K_a)-1}{K_a}$$

we obtain with $\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}$

$$\begin{aligned} P\left(\left|\frac{\exp(Y_n)-1}{a_n}\right| \leq K_\eta\right) &\geq P\left(\left|\frac{\exp(|Y_n|)-1}{a_n}\right| \leq \frac{\exp(K_{\eta_a}K_a)-1}{K_a}\right) \\ &= P\left(\frac{1}{|a_n|}\left|\sum_{k=0}^{\infty}\frac{|Y_n|^k}{k!}-1\right| \leq \frac{\exp(K_{\eta_a}K_a)-1}{K_a}\right) \\ &= P\left(\frac{1}{|a_n|}\left|1+\sum_{k=1}^{\infty}\frac{|Y_n|^k}{k!}-1\right| \leq \frac{\exp(K_{\eta_a}K_a)-1}{K_a}\right) \\ &= P\left(\frac{1}{|a_n|}\sum_{k=1}^{\infty}\frac{|Y_n|^k}{k!} \leq \frac{\exp(K_{\eta_a}K_a)-1}{K_a}\right) \\ &= P\left(\sum_{k=1}^{\infty}\frac{|Y_n|^k}{|a_n|^k}\frac{|a_n|^k}{|a_n|}\frac{1}{k!} \leq \frac{\exp(K_{\eta_a}K_a)-1}{K_a}\right) \\ &= P\left(\sum_{k=1}^{\infty}\frac{|Y_n|^k}{|a_n|^k}|a_n|^{k-1}\frac{1}{k!} \leq \frac{\exp(K_{\eta_a}K_a)-1}{K_a}\right) \\ &\geq P\left(\sum_{k=1}^{\infty}\left(\frac{|Y_n|}{|a_n|}\right)^k K_a^{k-1}\frac{1}{k!} \leq \frac{\exp(K_{\eta_a}K_a)-1}{K_a}\right), \end{aligned}$$

using $|a_n|^{k-1} \leq K_a^{k-1}$ for $k \geq 1$. It follows

$$\begin{aligned} P\left(\left|\frac{\exp(Y_n)-1}{a_n}\right| \leq K_\eta\right) &\geq P\left(\sum_{k=0}^{\infty}\left(\frac{|Y_n|}{|a_n|}\right)^k \frac{K_a^k}{k!} - 1 \leq \exp(K_{\eta_a}K_a) - 1\right) \\ &= P\left(\exp\left(\frac{|Y_n|}{|a_n|}K_a\right) \leq \exp(K_{\eta_a}K_a)\right) \\ &= P\left(\frac{|Y_n|}{|a_n|}K_a \leq K_{\eta_a}K_a\right) \\ &= P\left(\left|\frac{Y_n}{a_n}\right| \leq K_{\eta_a}\right) \geq 1-\eta_a = 1-\eta, \end{aligned}$$

where we chose $\eta_a = \eta$.

□

3.5 Evaluation metrics for binary classification

We have two main goals in the modeling process. First, we are interested in statistical inference that provides insight into potential factors correlated with a higher risk of an adverse pathological outcome. The interpretation of ORs as described in Section 3.1 aids in understanding these associations. Second, we want to predict the risk of an adverse pathological outcome for a new patient. Thus, we are interested in the predictive performance of the model on new data not used during the modeling process. In the following, we describe several evaluation metrics for this purpose.

We assume that a general binary classification model has been fit to a training data set and is to be evaluated on an independent test set. Each patient in the test set receives a model based probability of the outcome of interest, obtained by substituting his individual characteristics into the model probability function based on coefficients estimated from the training set. A patient is classified as positive if the probability exceeds a specified cut-off value c. The potential errors are summarized by the confusion matrix, also commonly referred to as a contingency table in Table 3.4.

		True class: P	True class: N
Predicted class	P	True positives (tp)	False positives (fp)
	N	False negatives (fn)	True negatives (tn)

Table 3.4 Confusion matrix for true versus predicted class; P: positive, N: negative.

Based on Table 3.4 we define several evaluation metrics [185–187].

Accuracy is the number of correctly predicted patients divided by the total number of patients

$$Accuracy = \frac{tp + tn}{tp + fn + fp + tn}.$$

Precision or positive predictive value (PPV) is the number of correctly predicted positive patients divided by the number of positive predicted patients

$$Precision = \frac{tp}{tp + fp}.$$

Sensitivity, recall, or true positive rate is the number of correctly predicted positive patients divided by the number of positive patients

$$Sensitivity = \frac{tp}{tp + fn}.$$

Specificity is the number of correctly predicted negative patients divided by the number of negative patients

$$Specificity = \frac{tn}{fp + tn}.$$

We use these evaluation metrics to assess the performance of each model during the modeling process. Although it might be tempting to only use accuracy for assessment, we also consider the other measures as they distinguish specific types of errors [185, 188].

As noted earlier, the metrics depend on choice of cut-off for which patients with probability above are classified as positive. It is common practice, therefore, to evaluate models independent of the cut-off value c. The receiver operating characteristic (ROC) curve plots the sensitivity against $1-$ specificity for all possible cut-off values c [189]. The area underneath the receiver operating characteristic curve (AUC) defined by the integral of the ROC curve is equivalent to the probability that the model will rank a randomly chosen true positive case higher than a randomly chosen true negative case, in the sense that the former will have a higher probability of the outcome [189]. The AUC falls between 0.5, corresponding to random prediction and 1.0, corresponding to perfect prediction [189].

We extend the evaluation metrics defined above to summarize the classification performance on all pathological outcomes simultaneously by defining a multi-label classification problem. We set

$$Y_{ij} = \begin{cases} 1 & \text{if patient } i \text{ has the corresponding adverse pathological outcome } j \\ 0 & \text{otherwise,} \end{cases}$$

for $j = 1, \ldots, q$, $i = 1, \ldots, n$ and let $\hat{Y}_{ij}$ denote the predicted value of Y_{ij}. We define evaluation metrics adapting the approach of Yang (1999) and Kazawa et al. (2004), which Tsoumakas et al. (2010) discussed in more detail [190–192].

Exact match ratio is the proportion of correctly classified patients

$$Exact\ match\ ratio = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{q} \mathbb{1}(\hat{Y}_{ij} = Y_{ij}),$$

where $\mathbb{1}$ denotes the indicator function. The product within the sum evaluates to 1 if and only if all five outcomes are correctly classified for the patient.

The exact match ratio does not distinguish between partially correct and incorrect classification and thus is a strict evaluation measure [192]. We therefore additionally consider macro- and micro-averaged approaches that can be applied to any binary evaluation measure B, such as precision, sensitivity, and specificity, and take partially correct classification of multiple outcomes into account [192]. We consider the binary evaluation metric $B(tp, tn, fp, fn)$ and let tp_j, tn_j, fp_j, fn_j denote the corresponding number of true positives, true negatives, false positives and false negatives for outcome j [192]. Then the macro- and micro-averaged versions of B are given by

$$B_{macro} = \frac{1}{q} \sum_{j=1}^{q} B(tp_j, tn_j, fp_j, fn_j),$$

$$B_{micro} = B\left(\sum_{j=1}^{q} tp_j, \sum_{j=1}^{q} tn_j, \sum_{j=1}^{q} fp_j, \sum_{j=1}^{q} fn_j \right).$$

While the macro- and micro-averaged accuracies are equal, the averages differ for other measures [192]. Summary measures, such as the ROC and AUC, can be calculated using the B_{macro} or the B_{micro} values for each cut-off.

4 Approximation of the Bayes Factor for nested logistic regression models

4.1 Mathematical derivation

Kass and Vaidyanathan (1992) showed that under certain regularity conditions, parametrizations and prior distributions, the Schwarz criterion approximates the BF up to order $O_p(n^{-\frac{1}{2}})$ for the problem of testing the equality of two binomial proportions [15]. They mentioned the extension to multiple logistic regression without proof. Raftery (1995) outlined a proof for general statistical models for independent and identically distributed (iid) observations and Kass and Wasserman (1995) applied the results to a larger variety of models and prior distributions without proof [193, 194]. Raftery (1996) outlined a proof for generalized linear models based on the Laplace approximation without providing a proof for the order of approximation [195]. Cavanaugh and Neath (1999) provided a more general derivation of the BIC based on regularity conditions similar to the ones we use in the following [196]. However, they did not provide an order of approximation. In this section we give the theory for approximation of the Schwarz criterion to the BF for multiple logistic regression, providing approximations of greater order and for arbitrary prior distributions along the way. Finally, we show the extension to the BIC using a base null model for comparison.

In the following, we first consider the univariate logistic regression model introduced in Section 3.1, $\text{logit}(\pi_i) = X_i'\beta$, with independent observations $Y_i \sim Ber(\pi_i)$, $i = 1, \ldots, n$, corresponding covariates $X_i \in \mathbb{R}^m$ including 1 as the intercept, and $\beta \in \mathbb{R}^m$ as the coefficient vector. Let $\beta = (\theta, \psi) \in \Theta \times \Psi$, with $\dim(\Theta) = m_0$, $\dim(\Psi) = m - m_0$, and suppose we are interested in testing the hypotheses

$$H_0 : \psi = \psi_0 \quad \text{versus} \quad H_A : \psi \neq \psi_0.$$

We specify prior densities $f_0(\theta)$ and $f(\theta, \psi)$ under H_0 and H_A, respectively. With the likelihood function $\mathcal{L}(\beta)$ based on the observed data, the BF is given by

$$\text{BF} = \frac{\int \mathcal{L}(\theta, \psi_0) f_0(\theta) d\theta}{\iint \mathcal{L}(\theta, \psi) f(\theta, \psi) d\theta d\psi}. \tag{4.1}$$

In order to provide an approximation to the BF, following regularity conditions are required to apply Laplace's method to both the numerator and denominator of (4.1) [15, 197].

Definition 4.1 (Laplace regularity)
Assume that $\Omega \subseteq \mathbb{R}^m$ is an open subset and $\{\ell_n = \log(\mathcal{L}_n) : \ n = 1, 2, \ldots\}$ a sequence of log-likelihood functions on Ω. Let $-nh_n = \ell_n$ define a sequence of real functions having local minima $\{\widehat{\beta}_n : \ n = 1, 2, \ldots\}$ and assume $h_n \in \mathcal{C}^6(\Omega)$, which means h_n is six times continuously differentiable on Ω. Denote by $\mathcal{B}_\delta(\beta)$ the open ball of radius δ centered at β and the Hessian matrix of h_n at β by $D^2 h_n(\beta)$. The sequence of log-likelihood functions $\ell_n = -nh_n$ is Laplace regular if there exist $\varepsilon, M, \eta > 0$ and $n_0 \in \mathbb{N}$ such that $n \geq n_0$ implies

(i) $\forall \beta \in \mathcal{B}_\varepsilon(\widehat{\beta}_n), \forall d \in \{0, \ldots, 6\}, \forall j_1, \ldots j_d \in \{1, \ldots m\}$: $|\partial_{j_1 \cdots j_d} h_n(\beta)| < M$;

(ii) $\det\left(D^2 h_n(\widehat{\beta}_n)\right) > \eta$;

(iii) $\forall \delta \in (0, \varepsilon) : \ \mathcal{B}_\delta(\widehat{\beta}_n) \subseteq \Omega$ and $\limsup_{n\to\infty} \sup_{\beta} \left\{ h_n(\widehat{\beta}_n) - h_n(\beta) : \beta \in \Omega \backslash \mathcal{B}_\delta(\widehat{\beta}_n) \right\} < 0.$

Part (i) of Definition 4.1 implies that the log-likelihood is six times partially differentiable with derivatives bounded for β in the area of the MLE $\widehat{\beta}_n$. We assume the log-likelihood function is smooth and bounded near the MLE, and thus in our case of logistic regression, that the log ORs β are bounded as well. In medicine, ORs are bounded except for extreme imbalanced cases where the MLE does not exist, as it was discussed in Section 3.1. We assume a sufficiently balanced data set without complete separation so Definition 4.1 (i) holds.

Part (ii) ensures that the determinant of the negative Hessian matrix of ℓ_n is bounded away from zero and thus that the matrix is non-singular. In Section 3.1 we show that the negative log-likelihood is convex and thus the Hessian matrix is positive semidefinite. With Part (ii) of the definition we assume that at the MLE the Hessian matrix is positive definite.

With (iii) we ensure that $\widehat{\beta}_n$ does not lie on the boundary of Ω and approximates the global maximum of ℓ_n asymptotically.

Thus, assuming Laplace regularity for log-likelihoods overlaps with the restrictions on MLE existence and uniqueness required for estimation in Section 3.1.

Assumption 4.1 (Validity of Laplace approximation)
Assume that $\Omega \subseteq \mathbb{R}^m$ is an open subset.

1) The observations $Y = (Y_1 \cdots Y_n)'$ are embedded in an infinite sequence of data vectors with the resulting sequences of log-likelihoods $\{\ell_n\}$ on Ω Laplace regular.

2) $b \in \mathcal{C}^4(\Omega)$, which implies b is a four times continuously differentiable function on Ω.

Under the Laplace regularity conditions and Assumption 4.1, the Laplace's approximation with $h(\beta) = -n^{-1}\ell_n(\beta)$, $\Omega \subseteq \mathbb{R}^m$ is given by

$$\int b(\beta)\exp(-nh(\beta))d\beta = \left(\frac{2\pi}{n}\right)^{\frac{m}{2}} \det(\Sigma)^{\frac{1}{2}} \exp(-nh(\widehat{\beta}))b(\widehat{\beta})(1+O(n^{-1})), \tag{4.2}$$

for $n \to \infty$, where $\widehat{\beta}$ maximizes $-h$ and $\Sigma = (-D^2h(\widehat{\beta}))^{-1}$ defines the negative inverse of the Hessian matrix $D^2h(\beta)$ evaluated at $\widehat{\beta}$ [197].

Let $b_0(\theta)$ and $b(\theta,\psi)$ be positive functions on Θ and $\Theta \times \Psi$, respectively, and set

$$\widetilde{\ell}_0(\theta) = \log\left(\mathcal{L}(\theta,\psi_0)\cdot\frac{f_0(\theta)}{b_0(\theta)}\right) \quad \text{and}$$
$$\widetilde{\ell}(\theta,\psi) = \log\left(\mathcal{L}(\theta,\psi)\cdot\frac{f(\theta,\psi)}{b(\theta,\psi)}\right).$$

The BF (4.1) can be rewritten as

$$\mathsf{BF} = \frac{\int \exp(\widetilde{\ell}_0(\theta))b_0(\theta)d\theta}{\iint \exp(\widetilde{\ell}(\theta,\psi))b(\theta,\psi)d\theta d\psi}, \tag{4.3}$$

which leads to the following proposition.

Proposition 4.1
Suppose Assumption 4.1 holds for the sequences $\{\widetilde{\ell}_{0,n}\}$, $\{\widetilde{\ell}_n\}$, and the positive functions $b_0(\theta)$, and $b(\theta,\psi)$ on Θ and $\Theta\times\Psi$, respectively. Further assume that $\widetilde{\theta}_0$ and $(\widetilde{\theta},\widetilde{\psi})$ are the maxima of $\widetilde{\ell}_0$ and $\widetilde{\ell}$, respectively, and the integrals in (4.3) exist and are finite. Then

$$\mathsf{BF} = (2\pi)^{\frac{m_0-m}{2}} \frac{\det(\widetilde{\Sigma}_0)^{\frac{1}{2}}\exp(\widetilde{\ell}_0(\widetilde{\theta}_0))b_0(\widetilde{\theta}_0)}{\det(\widetilde{\Sigma})^{\frac{1}{2}}\exp(\widetilde{\ell}(\widetilde{\theta},\widetilde{\psi}))b(\widetilde{\theta},\widetilde{\psi})}(1+O(n^{-1}))\,,$$

where $\widetilde{\Sigma}_0 = (-D^2\widetilde{\ell}_0(\widetilde{\theta}_0))^{-1}$ and $\widetilde{\Sigma} = (-D^2\widetilde{\ell}(\widetilde{\theta},\widetilde{\psi}))^{-1}$.

Proof. Under Assumption 4.1 we can apply the Laplace approximation (4.2) to the numerator and denominator of (4.3) [15, 197]. Recall that $m_0 = \dim(\Theta)$ and $m = \dim(\Theta\times\Psi)$. With

$$\det\left(\left(-\frac{1}{n}D^2\widetilde{\ell}_0(\widetilde{\theta}_0)\right)^{-1}\right)^{\frac{1}{2}} = \det\left(n\cdot\left(-D^2\widetilde{\ell}_0(\widetilde{\theta}_0)\right)^{-1}\right)^{\frac{1}{2}}$$
$$= n^{\frac{m_0}{2}}\det\left(\left(-D^2\widetilde{\ell}_0(\widetilde{\theta}_0)\right)^{-1}\right)^{\frac{1}{2}}$$

and

$$\det\left(\left(-\frac{1}{n}D^2\widetilde{\ell}(\widetilde{\theta},\widetilde{\psi})\right)^{-1}\right)^{\frac{1}{2}} = n^{\frac{m}{2}}\det\left(\left(-D^2\widetilde{\ell}(\widetilde{\theta},\widetilde{\psi})\right)^{-1}\right)^{\frac{1}{2}}$$

we obtain the approximation

$$\begin{aligned}\mathsf{BF} &= \frac{\left(\frac{2\pi}{n}\right)^{\frac{m_0}{2}}\det\left(\left(-\frac{1}{n}D^2\widetilde{\ell}_0(\widetilde{\theta}_0)\right)^{-1}\right)^{\frac{1}{2}}\exp\left(\widetilde{\ell}_0(\widetilde{\theta}_0)\right)b_0(\widetilde{\theta}_0)\,(1+O(n^{-1}))}{\left(\frac{2\pi}{n}\right)^{\frac{m}{2}}\det\left(\left(-\frac{1}{n}D^2\widetilde{\ell}(\widetilde{\theta},\widetilde{\psi})\right)^{-1}\right)^{\frac{1}{2}}\exp\left(\widetilde{\ell}(\widetilde{\theta},\widetilde{\psi})\right)b(\widetilde{\theta},\widetilde{\psi})\,(1+O(n^{-1}))}\\ &= \frac{(2\pi)^{\frac{m_0}{2}}\det\left(\left(-D^2\widetilde{\ell}_0(\widetilde{\theta}_0)\right)^{-1}\right)^{\frac{1}{2}}\exp\left(\widetilde{\ell}_0(\widetilde{\theta}_0)\right)b_0(\widetilde{\theta}_0)\,(1+O(n^{-1}))}{(2\pi)^{\frac{m}{2}}\det\left(\left(-D^2\widetilde{\ell}(\widetilde{\theta},\widetilde{\psi})\right)^{-1}\right)^{\frac{1}{2}}\exp\left(\widetilde{\ell}(\widetilde{\theta},\widetilde{\psi})\right)b(\widetilde{\theta},\widetilde{\psi})\,(1+O(n^{-1}))}\,.\end{aligned}$$

Applying $\frac{1+O(n^{-1})}{1+O(n^{-1})} = 1+O(n^{-1})$, shown in Proposition 3.1, we obtain

$$\mathsf{BF} = (2\pi)^{\frac{m_0-m}{2}}\,\frac{\det\left(\left(-D^2\widetilde{\ell}_0(\widetilde{\theta}_0)\right)^{-1}\right)^{\frac{1}{2}}\exp\left(\widetilde{\ell}_0(\widetilde{\theta}_0)\right)b_0(\widetilde{\theta}_0)}{\det\left(\left(-D^2\widetilde{\ell}(\widetilde{\theta},\widetilde{\psi})\right)^{-1}\right)^{\frac{1}{2}}\exp\left(\widetilde{\ell}(\widetilde{\theta},\widetilde{\psi})\right)b(\widetilde{\theta},\widetilde{\psi})}\left(1+O(n^{-1})\right)\,.$$

Thus a first approximation to the BF is

$$\widehat{\mathsf{BF}} = (2\pi)^{\frac{m_0-m}{2}}\,\frac{\det(\widetilde{\Sigma}_0)^{\frac{1}{2}}\exp(\widetilde{\ell}_0(\widetilde{\theta}_0))b_0(\widetilde{\theta}_0)}{\det(\widetilde{\Sigma})^{\frac{1}{2}}\exp(\widetilde{\ell}(\widetilde{\theta},\widetilde{\psi}))b(\widetilde{\theta},\widetilde{\psi})}\,, \tag{4.4}$$

where $\widetilde{\Sigma}_0 = \left(-D^2\widetilde{\ell}_0(\widetilde{\theta}_0)\right)^{-1}$ and $\widetilde{\Sigma} = \left(-D^2\widetilde{\ell}(\widetilde{\theta},\widetilde{\psi})\right)^{-1}$ [15]. □

For prior densities $b_0(\theta) = f_0(\theta)$ and $b(\theta,\psi) = f(\theta,\psi)$ the approximation converts to the expression to be used here

$$\widehat{\mathsf{BF}} = (2\pi)^{\frac{m_0-m}{2}}\,\frac{\det(\widehat{\Sigma}_0)^{\frac{1}{2}}\exp(\ell_0(\widehat{\theta}_0))f_0(\widehat{\theta}_0)}{\det(\widehat{\Sigma})^{\frac{1}{2}}\exp(\ell(\widehat{\theta},\widehat{\psi}))f(\widehat{\theta},\widehat{\psi})}\,, \tag{4.5}$$

where $\widehat{\theta}_0$ and $(\widehat{\theta},\widehat{\psi})$ are the maxima of $\ell_0(\cdot) = \log(\mathcal{L}(\cdot,\psi_0))$ and $\ell = \log(\mathcal{L})$, respectively, and $\widehat{\Sigma}_0 = \left(-D^2\ell_0(\widehat{\theta}_0)\right)^{-1}$ and $\widehat{\Sigma} = \left(-D^2\ell(\widehat{\theta},\widehat{\psi})\right)^{-1}$.

To move towards the Schwarz approximation, we define null orthogonality and add further structural and regularity assumptions following Kass and Vaidyanathan (1992), Kass and Wasserman (1995), and Pauler (1998) [15, 194, 198].

Consider the expected value of the negative Hessian matrix in a sample of size one that is denoted as the expected Fisher information matrix

$$I(\theta, \psi) = \mathbb{E}\left[-D^2\ell(\theta, \psi)\right] = \mathbb{E}\left[X'WX\right],$$

where $X \in \mathbb{R}^{1\times m}$ and $W = G(X\beta)\left(1 - G(X\beta)\right)$ is the one-dimensional diagonal matrix analogously defined to Section 3.1. We partition the Fisher information matrix

$$I(\theta, \psi) = \begin{pmatrix} I_{\theta\theta}(\theta, \psi) & I_{\psi\theta}(\theta, \psi) \\ I_{\theta\psi}(\theta, \psi) & I_{\psi\psi}(\theta, \psi) \end{pmatrix} = \begin{pmatrix} \mathbb{E}\left[-\dfrac{\ell(\theta, \psi)}{\partial\theta\partial\theta'}\right] & \mathbb{E}\left[-\dfrac{\ell(\theta, \psi)}{\partial\theta\partial\psi'}\right] \\ \mathbb{E}\left[-\dfrac{\ell(\theta, \psi)}{\partial\psi\partial\theta'}\right] & \mathbb{E}\left[-\dfrac{\ell(\theta, \psi)}{\partial\psi\partial\psi'}\right] \end{pmatrix}, \tag{4.6}$$

where $I_{\psi\psi}(\theta, \psi)$, $I_{\theta\theta}(\theta, \psi)$ denote the blocks for ψ and θ, respectively. With the symmetry of the Hessian matrix it follows that $I(\theta, \psi)$ is symmetric and thus $I_{\psi\theta}(\theta, \psi) = I_{\theta\psi}(\theta, \psi)$. Null orthogonality was defined by Kass and Vaidyanathan (1992) as orthogonality of the expected Fisher information under the null hypothesis [15].

Definition 4.2 (Null orthogonality)
Parameters θ and ψ are said to be null orthogonal if $I_{\theta\psi}(\theta, \psi_0) = 0$ for all $\theta \in \Theta$.

Proposition 4.2
Given a non-null orthogonal parametrization (ξ, ψ), we can construct a transformation $(\xi, \psi) \to (\theta, \psi)$ such that θ and ψ are null orthogonal with

$$\theta = \Phi(\xi, \psi) = \xi + I_{\xi\xi}(\xi, \psi_0)^{-1} I_{\xi\psi}(\xi, \psi_0)(\psi - \psi_0), \tag{4.7}$$

where $I_{\xi\xi}$ and $I_{\xi\psi}$ correspond to respective blocks of the expected Fisher information matrix [15].

Proof. We follow and extend the argumentation by Cox and Reid (1987) and Huzurbazar and Jeffreys (1950) [199, 200].

With (4.7) we can rewrite the log-likelihood as follows

$$\ell(\xi, \psi) = \ell^*(\Phi(\xi, \psi), \psi) = \ell^*(\theta, \psi). \tag{4.8}$$

For the first and second order partial derivatives with respect to ξ and ψ, we consider their single entries. Then,

$$\underbrace{\frac{\partial \ell(\xi,\psi)}{\partial \psi_j}}_{1\times 1} = \underbrace{\frac{\partial \ell^*(\Phi(\xi,\psi),\psi)}{\partial \psi_j}}_{1\times 1} + \underbrace{\frac{\partial \Phi(\xi,\psi)}{\partial \psi_j}}_{1\times m_0} \cdot \underbrace{\frac{\partial \ell^*(\Phi(\xi,\psi),\psi)}{\partial \Phi}}_{m_0\times 1},$$

using partial derivatives of functions and the chain rule. Thus, first derivative with respect to ψ is a $m - m_0$ dimensional vector.

The mixed second order partial derivatives are

$$\begin{aligned}
\underbrace{\frac{\partial^2 \ell(\xi,\psi)}{\partial \xi_k \partial \psi_j}}_{1\times 1} &= \underbrace{\frac{\partial \Phi(\xi,\psi)}{\partial \xi_k}}_{1\times m_0} \cdot \underbrace{\frac{\partial^2 \ell^*(\Phi(\xi,\psi),\psi)}{\partial \Phi \partial \psi_j}}_{m_0\times 1} + \frac{\partial}{\partial \xi_k}\left[\underbrace{\frac{\partial \Phi(\xi,\psi)}{\partial \psi_j}}_{1\times m_0} \cdot \underbrace{\frac{\partial \ell^*(\Phi(\xi,\psi),\psi)}{\partial \Phi}}_{m_0\times 1}\right] \\
&= \frac{\partial \Phi(\xi,\psi)}{\partial \xi_k} \cdot \frac{\partial^2 \ell^*(\Phi(\xi,\psi),\psi)}{\partial \Phi \partial \psi_j} + \left[\frac{\partial}{\partial \xi_k}\frac{\partial \ell^*(\Phi(\xi,\psi),\psi)}{\partial \Phi}\right] \cdot \left(\frac{\partial \Phi(\xi,\psi)}{\partial \psi_j}\right)' \\
&\quad + \left[\frac{\partial}{\partial \xi_k}\left(\frac{\partial \Phi(\xi,\psi)}{\partial \psi_j}\right)'\right] \cdot \frac{\partial \ell^*(\Phi(\xi,\psi),\psi)}{\partial \Phi} \\
&= \frac{\partial \Phi(\xi,\psi)}{\partial \xi_k} \cdot \frac{\partial^2 \ell^*(\Phi(\xi,\psi),\psi)}{\partial \Phi \partial \psi_j} \\
&\quad + \underbrace{\frac{\partial \Phi(\xi,\psi)}{\partial \xi_k}}_{1\times m_0} \cdot \underbrace{\frac{\partial^2 \ell^*(\Phi(\xi,\psi),\psi)}{\partial \Phi \partial \Phi'}}_{m_0\times m_0} \cdot \underbrace{\left(\frac{\partial \Phi(\xi,\psi)}{\partial \psi_j}\right)'}_{m_0\times 1} \\
&\quad + \underbrace{\frac{\partial^2 \Phi(\xi,\psi)'}{\partial \xi_k \partial \psi_j}}_{1\times m_0} \cdot \underbrace{\frac{\partial \ell^*(\Phi(\xi,\psi),\psi)}{\partial \Phi}}_{m_0\times 1},
\end{aligned} \tag{4.9}$$

using partial derivatives of functions, the chain rule, and product derivative rules. $\dfrac{\partial^2 \ell(\xi,\psi)}{\partial \xi \partial \psi'}$ is a $m_0 \times (m - m_0)$ - matrix with second order partial derivatives.

For the expectation of the last part of the sum in (4.9) it holds that

$$\mathbb{E}\left[\frac{\partial^2 \Phi(\xi,\psi)'}{\partial \xi_k \partial \psi_j} \cdot \frac{\partial \ell^*(\Phi(\xi,\psi),\psi)}{\partial \Phi}\right] = \frac{\partial^2 \Phi(\xi,\psi)'}{\partial \xi_k \partial \psi_j} \cdot \mathbb{E}\left[\frac{\partial \ell^*(\Phi(\xi,\psi),\psi)}{\partial \Phi}\right]$$

as $\Phi(\xi,\psi)$ is constant with respect to X.

Further,

$$\begin{aligned}\mathbb{E}\left[\frac{\partial \ell^*(\Phi(\xi,\psi),\psi)}{\partial \Phi}\right] &= \int \frac{\partial \log\left(P(x|\Phi(\xi,\psi),\psi)\right)}{\partial \Phi} P(x|\Phi(\xi,\psi),\psi)dx \\ &= \int \frac{1}{P(x|\Phi(\xi,\psi),\psi)} P(x|\Phi(\xi,\psi),\psi)\frac{\partial P(x|\Phi(\xi,\psi),\psi)}{\partial \Phi}dx \\ &= \frac{\partial}{\partial \Phi}\int P(x|\Phi(\xi,\psi),\psi)dx = \frac{\partial}{\partial \Phi}1 = 0,\end{aligned}$$

where we could exchange the order of differentiation and integration since for each $(\Phi(\xi,\psi),\psi)$ $P(x|\Phi(\xi,\psi),\psi)$ is integrable of x. Thus, for $I_{\xi\psi}(\xi,\psi) = \mathbb{E}\left[-\frac{\partial^2 \ell(\xi,\psi)}{\partial \xi \partial \psi'}\right]$, similar to (4.6), we have with (4.9)

$$\begin{aligned}I_{\xi\psi}(\xi,\psi) = &\frac{\partial \Phi(\xi,\psi)}{\partial \xi}\cdot I^*_{\Phi\psi}(\Phi(\xi,\psi),\psi) \\ &+ \frac{\partial \Phi(\xi,\psi)}{\partial \xi}\cdot I^*_{\Phi\Phi}(\Phi(\xi,\psi),\psi)\cdot\left(\frac{\partial \Phi(\xi,\psi)}{\partial \psi}\right)',\end{aligned} \tag{4.10}$$

where we denote by I^* the expected Fisher information matrix for ℓ^*.

We are interested in whether $I^*_{\Phi\psi}(\Phi(\xi,\psi_0),\psi_0) = 0$ for all ξ and thus we evaluate (4.10) at $\psi = \psi_0$. First note by taking the partial derivative with respect to ξ of (4.7) we obtain

$$\begin{aligned}\frac{\partial \Phi(\xi,\psi)}{\partial \xi} = &\mathrm{Id}_{m_0} + \left(\frac{\partial}{\partial \xi} I_{\xi\xi}(\xi,\psi_0)^{-1}\right) I_{\xi\psi}(\xi,\psi_0)(\psi-\psi_0) \\ &+ I_{\xi\xi}(\xi,\psi_0)^{-1}\left(\frac{\partial}{\partial \xi} I_{\xi\psi}(\xi,\psi_0)\right)(\psi-\psi_0),\end{aligned}$$

where Id_{m_0} is the $m_0 \times m_0$ identity matrix. This implies

$$\left.\frac{\partial \Phi(\xi,\psi)}{\partial \xi}\right|_{\psi=\psi_0} = \mathrm{Id}_{m_0}. \tag{4.11}$$

Similarly, by taking the partial derivative with respect to ψ of (4.7)

$$\frac{\partial \Phi(\xi,\psi)}{\partial \psi} = \left(I_{\xi\xi}(\xi,\psi_0)^{-1} I_{\xi\psi}(\xi,\psi_0)\right)'$$

so that (4.10) evaluated at $\psi = \psi_0$ yields

$$\begin{aligned}I_{\xi\psi}(\xi,\psi_0) = &\mathrm{Id}_{m_0}\cdot I^*_{\Phi\psi}(\Phi(\xi,\psi_0),\psi_0) \\ &+ \mathrm{Id}_{m_0}\cdot I^*_{\Phi\Phi}(\Phi(\xi,\psi_0),\psi_0) I_{\xi\xi}(\xi,\psi_0)^{-1} I_{\xi\psi}(\xi,\psi_0).\end{aligned} \tag{4.12}$$

Next we note that

$$\begin{aligned}&\frac{\partial^2\Phi(\xi,\psi)}{\partial\xi_k\partial\xi_j}\\&\quad=\frac{\partial}{\partial\xi_k}\left[\left(\frac{\partial}{\partial\xi_j}I_{\xi\xi}(\xi,\psi_0)^{-1}\right)I_{\xi\psi}(\xi,\psi_0)+I_{\xi\xi}(\xi,\psi_0)^{-1}\left(\frac{\partial}{\partial\xi_j}I_{\xi\psi}(\xi,\psi_0)\right)\right](\psi-\psi_0)\end{aligned}$$

and this implies

$$\left.\frac{\partial^2\Phi(\xi,\psi)}{\partial\xi_k\partial\xi_j}\right|_{\psi=\psi_0}=0. \tag{4.13}$$

With (4.8) it follows that

$$\begin{aligned}\frac{\partial^2\ell(\xi,\psi)}{\partial\xi_k\partial\xi_j}&=\frac{\partial^2\ell^*(\Phi(\xi,\psi),\psi)}{\partial\xi_k\partial\xi_j}=\frac{\partial}{\partial\xi_k}\left(\frac{\partial\Phi(\xi,\psi)}{\partial\xi_j}\cdot\frac{\partial\ell^*(\Phi(\xi,\psi),\psi)}{\partial\Phi}\right)\\&=\frac{\partial\Phi(\xi,\psi)}{\partial\xi_k}\cdot\frac{\partial^2\ell^*(\Phi(\xi,\psi),\psi)}{\partial\Phi\partial\Phi'}\cdot\left(\frac{\partial\Phi(\xi,\psi)}{\partial\xi_j}\right)'\\&\quad+\frac{\partial^2\Phi(\xi,\psi)'}{\partial\xi_k\partial\xi_j}\cdot\frac{\partial\ell^*(\Phi(\xi,\psi),\psi)}{\partial\Phi},\end{aligned}$$

using partial derivatives of functions, the chain rule and product derivatives rules. With (4.11) and (4.13) we obtain

$$\left.\frac{\partial^2\ell(\xi,\psi)}{\partial\xi_k\partial\xi_j}\right|_{\psi=\psi_0}=e_k\cdot\frac{\partial^2\ell^*(\Phi(\xi,\psi_0),\psi_0)}{\partial\Phi\partial\Phi'}\cdot e_j'=\frac{\partial^2\ell^*(\Phi(\xi,\psi_0),\psi_0)}{\partial\Phi_k\partial\Phi_j},$$

where e_k, e_j denote the unit vectors with one at entry k, j respectively, and zero otherwise.

It follows

$$I_{\xi\xi}(\xi,\psi_0)=\mathbb{E}\left[-\frac{\partial^2\ell(\xi,\psi_0)}{\partial\xi\partial\xi'}\right]=\mathbb{E}\left[-\frac{\partial^2\ell^*(\Phi(\xi,\psi_0),\psi_0)}{\partial\Phi\partial\Phi'}\right]=I^*_{\Phi\Phi}(\Phi(\xi,\psi_0),\psi_0).$$

Solving (4.12) for $I^*_{\Phi\psi}(\Phi(\xi,\psi_0),\psi_0)$ yields

$$\begin{aligned}I^*_{\Phi\psi}(\Phi(\xi,\psi_0),\psi_0)&=I_{\xi\psi}(\xi,\psi_0)-I^*_{\Phi\Phi}(\Phi(\xi,\psi_0),\psi_0)I_{\xi\xi}(\xi,\psi_0)^{-1}I_{\xi\psi}(\xi,\psi_0)\\&=I_{\xi\psi}(\xi,\psi_0)-I_{\xi\xi}(\xi,\psi_0)I_{\xi\xi}(\xi,\psi_0)^{-1}I_{\xi\psi}(\xi,\psi_0)\\&=I_{\xi\psi}(\xi,\psi_0)-I_{\xi\psi}(\xi,\psi_0)=0.\end{aligned}$$

Thus, θ and ψ are null orthogonal. □

The following structural and regularity assumptions are necessary in addition to Assumption 4.1 for further approximation of the BF [15, 194, 198].

Assumption 4.2
Let $\widehat{\theta}_0$ and $(\widehat{\theta}, \widehat{\psi})$ be the MLEs of $\ell_0(\theta) = \ell(\theta, \psi_0)$ and $\ell(\theta, \psi)$, respectively.

1) θ and ψ are null orthogonal.

2) The marginal prior distribution for θ is the same under both hypotheses

$$f_0(\theta) = f_\theta(\theta) = \int f(\theta, \psi) d\psi.$$

3) The matrix of second derivatives converges asymptotically to the Fisher information matrix

$$-\frac{1}{n} D^2 \ell(\widehat{\theta}, \widehat{\psi}) - I(\theta, \psi) = O_p(n^{-1/2}).$$

4) The MLE $\widehat{\psi}$ under the alternative hypothesis satisfies

$$\widehat{\psi} - \psi_0 = O_p\left(n^{-\frac{1}{2}}\right).$$

This implies that the true value of ψ is ψ_0 or a neighboring alternative ψ_n with $\psi_n - \psi_0 = O(n^{-\frac{1}{2}})$.

Assumption 4.2 1) implies that the expected Fisher information matrix $I(\theta, \psi)$ is block diagonal under the null hypothesis. We have shown in Proposition 4.2 that for a given non-null orthogonal parametrization (ξ, ψ) we can construct a transformation (θ, ψ) that is null orthogonal. Under invariance of the likelihood we have that

$$\max_{\xi, \psi} \ell(\xi, \psi) = \max_{\theta, \psi} \ell^*(\theta, \psi)$$

and thus $(\hat{\xi}, \hat{\psi}) \to (\hat{\theta}, \hat{\psi})$ [201]. This implies, that reparametrization of ξ does not influence interpretation of the MLEs. Therefore, we can without loss of generality assume that θ and ψ are null orthogonal.

Assumption 4.2 2) implies that the conditional prior distribution for ψ under the alternative hypothesis is

$$f_{\psi|\theta}(\psi|\theta) = \frac{f(\theta, \psi)}{f_\theta(\theta)} = \frac{f(\theta, \psi)}{f_0(\theta)},$$

and under Assumption 4.2 1) of null orthogonality, does not impose a substantial additional restriction on the model.

Assumption 4.2 3) of asymptotic convergence to the Fisher information matrix is intuitive for iid outcomes Y_i's as it can be shown using the law of large numbers [113]. We do not assume identical distributions, but independence of the Y_i's. Thus we assume that asymptotically and for a finite parameter space of X, the expected Fisher information is approximated by the sample average of the negative second derivative of the log-likelihood. For the covariates, this does not impose additional restrictions on model assumptions to those required for existence of the MLE.

The BF is exponentially small for large samples when Assumption 4.2 4) does not hold and thus an approximation is no longer meaningful [15, 182, 198].

Proposition 4.3
Suppose Assumption 4.1 holds for $\{\ell_{0,n}\}$, $\{\ell_n\}$, and the prior densities f_0, and f on Θ and $\Theta \times \Psi$, respectively. Let $\dim(\Theta) = m_0$ and $\dim(\Psi) = m - m_0$ and $\widehat{\theta}_0$ and $(\widehat{\theta}, \widehat{\psi})$ be the maxima of ℓ_0 and ℓ, respectively. Assume the prior density $f_0(\theta)$ and its first derivative are bounded for $\theta \in \mathcal{B}_\delta(\widehat{\theta}_0)$, $\delta > 0$. Suppose that the integrals in (4.1) are finite and Assumption 4.2 holds. Then

$$\mathsf{BF} = \left(\frac{2\pi}{n}\right)^{\frac{m_0-m}{2}} \det\left(I_{\psi\psi}(\widehat{\theta}, \psi_0)\right)^{\frac{1}{2}} \frac{\exp(\ell_0(\widehat{\theta}_0))}{\exp(\ell(\widehat{\theta}, \widehat{\psi}))} \frac{1}{f_{\psi|\theta}(\widehat{\psi}|\widehat{\theta})} \left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right).$$

Proof. Substituting (4.5) into Proposition 4.1 yields

$$\begin{aligned}
\mathsf{BF} &= \widehat{\mathsf{BF}}\left(1 + O(n^{-1})\right) \\
&= (2\pi)^{\frac{m_0-m}{2}} \frac{\det((-D^2\ell_0(\widehat{\theta}_0))^{-1})^{\frac{1}{2}} \exp(\ell_0(\widehat{\theta}_0)) f_0(\widehat{\theta}_0)}{\det((-D^2\ell(\widehat{\theta}, \widehat{\psi}))^{-1})^{\frac{1}{2}} \exp(\ell(\widehat{\theta}, \widehat{\psi})) f(\widehat{\theta}, \widehat{\psi})} \left(1 + O(n^{-1})\right) \\
&= \left(\frac{2\pi}{n}\right)^{\frac{m_0-m}{2}} \frac{\det(-\frac{1}{n}D^2\ell_0(\widehat{\theta}_0))^{-\frac{1}{2}}}{\det(-\frac{1}{n}D^2\ell(\widehat{\theta}, \widehat{\psi}))^{-\frac{1}{2}}} \frac{\exp(\ell_0(\widehat{\theta}_0))}{\exp(\ell(\widehat{\theta}, \widehat{\psi}))} \frac{f_0(\widehat{\theta}_0)}{f(\widehat{\theta}, \widehat{\psi})} \left(1 + O(n^{-1})\right). \qquad (4.14)
\end{aligned}$$

First, we show that $\widehat{\theta}_0 - \widehat{\theta} = O_p\left(n^{-\frac{1}{2}}\right)$. Since $\widehat{\theta}_0$ is the MLE of $\ell_0(\theta)$ we have

$$\frac{\partial \ell_0(\widehat{\theta}_0)}{\partial \theta} = \frac{\partial \ell(\widehat{\theta}_0, \psi_0)}{\partial \theta} = \begin{pmatrix} \frac{\partial \ell(\widehat{\theta}_0, \psi_0)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell(\widehat{\theta}_0, \psi_0)}{\partial \theta_{m_0}} \end{pmatrix} = 0.$$

We expand each component of $\frac{\partial \ell_0(\widehat{\theta}_0)}{\partial \theta}$ around $(\widehat{\theta}, \widehat{\psi})$ using a Taylor approximation and obtain for $k = 1, \ldots, m_0$ [202],

$$\begin{aligned} 0 &= \frac{\partial \ell(\widehat{\theta}_0, \psi_0)}{\partial \theta_k} \\ &= \frac{\partial \ell(\widehat{\theta}, \widehat{\psi})}{\partial \theta_k} + \sum_{j=1}^{m_0} (\widehat{\theta}_{0j} - \widehat{\theta}_j) \frac{\partial^2 \ell(\widehat{\theta}, \widehat{\psi})}{\partial \theta_k \partial \theta_j} + \sum_{j=1}^{m-m_0} (\widehat{\psi}_{0j} - \psi_j) \frac{\partial^2 \ell(\widehat{\theta}, \widehat{\psi})}{\partial \theta_k \partial \psi_j} \\ &\quad + o_p\left(\left\| \begin{pmatrix} \widehat{\theta}_0 \\ \psi_0 \end{pmatrix} - \begin{pmatrix} \widehat{\theta} \\ \widehat{\psi} \end{pmatrix} \right\| \right). \end{aligned} \tag{4.15}$$

With Definition 3.3 and Assumption 4.2 4) we note that

$$\begin{aligned} o_p\left(\left\| \begin{pmatrix} \widehat{\theta}_0 \\ \psi_0 \end{pmatrix} - \begin{pmatrix} \widehat{\theta} \\ \widehat{\psi} \end{pmatrix} \right\| \right) &= o_p\left(\begin{pmatrix} \widehat{\theta}_0 \\ \psi_0 \end{pmatrix} - \begin{pmatrix} \widehat{\theta} \\ \widehat{\psi} \end{pmatrix} \right) \\ &= O_p\left(\begin{pmatrix} \widehat{\theta}_0 \\ 0 \end{pmatrix} - \begin{pmatrix} \widehat{\theta} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \psi_0 \end{pmatrix} - \begin{pmatrix} 0 \\ \widehat{\psi} \end{pmatrix} \right) \\ &= O_p\left(\widehat{\theta}_0 - \widehat{\theta} \right) + O_p(n^{-\frac{1}{2}}). \end{aligned}$$

We divide (4.15) by n and with $\frac{\partial \ell(\widehat{\theta}, \widehat{\psi})}{\partial \theta_k} = 0$, since $(\widehat{\theta}, \widehat{\psi})$ is the MLE of $\ell(\theta, \psi)$, we obtain

$$\begin{aligned} 0 &= \frac{1}{n} \sum_{j=1}^{m_0} (\widehat{\theta}_{0j} - \widehat{\theta}_j) \frac{\partial^2 \ell(\widehat{\theta}, \widehat{\psi})}{\partial \theta_k \partial \theta_j} + \frac{1}{n} \sum_{j=1}^{m-m_0} (\widehat{\psi}_{0j} - \psi_j) \frac{\partial^2 \ell(\widehat{\theta}, \widehat{\psi})}{\partial \theta_k \partial \psi_j} \\ &\quad + \frac{1}{n} O_p\left(\widehat{\theta}_0 - \widehat{\theta} \right) + \frac{1}{n} O_p(n^{-\frac{1}{2}}) \\ &= \frac{1}{n} \sum_{j=1}^{m_0} (\widehat{\theta}_{0j} - \widehat{\theta}_j) \frac{\partial^2 \ell(\widehat{\theta}, \widehat{\psi})}{\partial \theta_k \partial \theta_j} + \frac{1}{n} \sum_{j=1}^{m-m_0} O_p\left(n^{-\frac{1}{2}} \right) \frac{\partial^2 \ell(\widehat{\theta}, \widehat{\psi})}{\partial \theta_k \partial \psi_j} \\ &\quad + \frac{1}{n} O_p\left(\widehat{\theta}_0 - \widehat{\theta} \right) + O_p\left(n^{-\frac{3}{2}} \right), \end{aligned} \tag{4.16}$$

where we use Assumption 4.2 4) in the last step. Recall from Section 3.1 that

$$D^2 \ell(\theta, \psi) = -X' W(\theta, \psi) X,$$

where $W(\theta, \psi)$ is a diagonal matrix with elements

$$0 \leq W_{ii}(\theta, \psi) = G\left(X_i' \begin{pmatrix} \theta \\ \psi \end{pmatrix} \right) \left(1 - G\left(X_i' \begin{pmatrix} \theta \\ \psi \end{pmatrix} \right) \right) \leq 1$$

with $G(z) = \frac{\exp(z)}{1+\exp(z)}$. Therefore, for any entry of $d_{kj} = (D^2\ell(\theta,\psi))_{kj}$ we obtain

$$|d_{kj}| = \left|\sum_{i=1}^{n} X_{ik}W_{ii}(\theta,\psi)X_{ij}\right| \leq n\left|\max_{i=1,\dots,n}(X_{ik}W_{ii}X_{ij})\right| \leq n\left|\max_{i=1,\dots,n}(X_{ik}X_{ij})\right|$$

and it follows $d_{kj} = O_p(n)$. We apply this result to the entries $\frac{\partial^2\ell(\widehat{\theta},\widehat{\psi})}{\partial\theta_k\partial\theta_j}$ and $\frac{\partial^2\ell(\widehat{\theta},\widehat{\psi})}{\partial\theta_k\partial\psi_j}$ in (4.16) and use the multiplicative and additive properties of O_p

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{j=1}^{m_0}(\widehat{\theta}_{0j}-\widehat{\theta}_j)O_p(n) + \frac{1}{n}\sum_{j=1}^{m-m_0} O_p\left(n^{-\frac{1}{2}}\right)O_p(n) + \frac{1}{n}O_p\left(\widehat{\theta}_0-\widehat{\theta}\right) + O_p\left(n^{-\frac{3}{2}}\right)\\
&= \sum_{j=1}^{m_0}(\widehat{\theta}_{0j}-\widehat{\theta}_j)O_p(1) + (m-m_0)O_p\left(n^{-\frac{1}{2}}\right) + \frac{1}{n}O_p\left(\widehat{\theta}_0-\widehat{\theta}\right) + O_p\left(n^{-\frac{3}{2}}\right)\\
&= \sum_{j=1}^{m_0}(\widehat{\theta}_{0j}-\widehat{\theta}_j)O_p(1) + \frac{1}{n}O_p\left(\widehat{\theta}_0-\widehat{\theta}\right) + O_p\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

Thus, we obtain

$$\sum_{j=1}^{m_0}(\widehat{\theta}_{0j}-\widehat{\theta}_j)O_p(1) + \frac{1}{n}O_p\left(\widehat{\theta}_0-\widehat{\theta}\right) = O_p\left(n^{-\frac{1}{2}}\right)$$

and it follows that

$$\widehat{\theta}-\widehat{\theta}_0 = O_p\left(n^{-\frac{1}{2}}\right). \tag{4.17}$$

Next, we show that $D^2\ell_0(\widehat{\theta}_0) = D^2_{\theta\theta}\ell(\widehat{\theta},\psi_0)\left(1+O_p\left(n^{-\frac{1}{2}}\right)\right)$, where $D^2_{\theta\theta}$ denotes the part of the Hessian matrix of ℓ corresponding to θ. With (4.17), Assumption 4.2 4), and the properties of O_p we obtain

$$\begin{aligned}
\exp\left(X_i'\begin{pmatrix}\widehat{\theta}_0\\ \psi_0\end{pmatrix}\right) &= \exp\left(\sum_{j=1}^{m_0}X_{ij}\widehat{\theta}_{0j} + \sum_{j=1}^{m-m_0}X_{i\,m_0+j}\psi_{0j}\right)\\
&= \exp\left(\sum_{j=1}^{m_0}X_{ij}\left(\widehat{\theta}_j + O_p\left(n^{-\frac{1}{2}}\right)\right) + \sum_{j=1}^{m-m_0}X_{i\,m_0+j}\psi_{0j}\right)\\
&= \exp\left(\sum_{j=1}^{m_0}X_{ij}\widehat{\theta}_j + \sum_{j=1}^{m-m_0}X_{i\,m_0+j}\psi_{0j} + \sum_{j=1}^{m_0}X_{ij}O_p\left(n^{-\frac{1}{2}}\right)\right)\\
&= \exp\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix} + m_0\cdot O(1)O_p\left(n^{-\frac{1}{2}}\right)\right)\\
&= \exp\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix} + O_p\left(n^{-\frac{1}{2}}\right)\right)\\
&= \exp\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\exp\left(O_p\left(n^{-\frac{1}{2}}\right)\right).
\end{aligned}$$

With $\exp\left(O_p\left(n^{-\frac{1}{2}}\right)\right) = 1 + O_p\left(n^{-\frac{1}{2}}\right)$ as shown in Proposition 3.2 this yields

$$\exp\left(X_i'\begin{pmatrix}\widehat{\theta}_0\\ \psi_0\end{pmatrix}\right) = \exp\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right).$$

We note that with the additive properties of O_p it follows that

$$\begin{aligned}
1+\exp\left(X_i'\begin{pmatrix}\widehat{\theta}_0\\ \psi_0\end{pmatrix}\right) &= 1 + \exp\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right)\\
&= 1 + \exp\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right) + \exp\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right) O_p\left(n^{-\frac{1}{2}}\right)\\
&= 1 + \exp\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right) + \exp\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right) O_p\left(n^{-\frac{1}{2}}\right) + O_p\left(n^{-\frac{1}{2}}\right)\\
&= \left[1 + \exp\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\right]\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right).
\end{aligned}$$

Thus, with Proposition 3.1 we obtain

$$\begin{aligned}
G\left(X_i'\begin{pmatrix}\widehat{\theta}_0\\ \psi_0\end{pmatrix}\right) &= \frac{\exp\left(X_i'\begin{pmatrix}\widehat{\theta}_0\\ \psi_0\end{pmatrix}\right)}{1 + \exp\left(X_i'\begin{pmatrix}\widehat{\theta}_0\\ \psi_0\end{pmatrix}\right)} = \frac{\exp\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right)}{\left[1 + \exp\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\right]\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right)}\\
&= G\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right). \qquad (4.18)
\end{aligned}$$

Using

$$\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right)^2 = 1 + 2O_p\left(n^{-\frac{1}{2}}\right) + O_p\left(n^{-1}\right) = 1 + O_p\left(n^{-\frac{1}{2}}\right)$$

and (4.18), we approximate the diagonal elements of $W\left(\widehat{\theta}_0, \psi_0\right)$ by

$$\begin{aligned}
W_{ii}\left(\widehat{\theta}_0, \psi_0\right) &= G\left(X_i'\begin{pmatrix}\widehat{\theta}_0\\ \psi_0\end{pmatrix}\right)\left(1 - G\left(X_i'\begin{pmatrix}\widehat{\theta}_0\\ \psi_0\end{pmatrix}\right)\right)\\
&= G\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right)\left(1 - G\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right)\right)\\
&= G\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\left(\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right) - G\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right)^2\right)\\
&= G\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\left(\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right) - G\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right)\right)\\
&= G\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\left(1 - G\left(X_i'\begin{pmatrix}\widehat{\theta}\\ \psi_0\end{pmatrix}\right)\right)\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right)\\
&= W_{ii}\left(\widehat{\theta}, \psi_0\right)\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right).
\end{aligned}$$

We obtain for any entry $d_{kj}^{(0)} = \left(D^2\ell_0(\widehat{\theta}_0)\right)_{kj}$

$$\begin{aligned} d_{kj}^{(0)} &= -\sum_{i=1}^{n} X_{ik}W_{ii}\left(\widehat{\theta}_0, \psi_0\right) X_{ij} \\ &= -\sum_{i=1}^{n} X_{ik}W_{ii}\left(\widehat{\theta}, \psi_0\right)\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right) X_{ij} \\ &= -\left[\sum_{i=1}^{n} X_{ik}W_{ii}\left(\widehat{\theta}, \psi_0\right) X_{ij}\right]\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right) \\ &= d_{kj}\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right) = \left(D^2_{\theta\theta}\ell(\widehat{\theta}, \psi_0)\right)_{kj}\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right). \end{aligned}$$

We note that the indices k and j only index the part of $D^2_{\theta\theta}\ell(\widehat{\theta}, \psi_0)$ corresponding to θ. We obtain with Assumption 4.2 3)

$$\begin{aligned} -\frac{1}{n}D^2\ell_0(\widehat{\theta}_0) &= -\frac{1}{n}D^2_{\theta\theta}\ell(\widehat{\theta}, \psi_0)\left(\mathrm{Id}_{m_0} + O_p\left(n^{-\frac{1}{2}}\right)\right) \\ &= \left(I_{\theta\theta}(\widehat{\theta}, \psi_0) + O_p\left(n^{-\frac{1}{2}}\right)\right)\left(\mathrm{Id}_{m_0} + O_p\left(n^{-\frac{1}{2}}\right)\right). \end{aligned}$$

$I(\theta, \psi)$ is the expected Fisher information matrix for a single observation X and thus

$$I(\theta, \psi) = \mathbb{E}\left[X' \cdot G(X'\beta)(1 - G(X'\beta)) \cdot X\right] \leq \mathbb{E}\left[X'X\right] = O_p(1), \tag{4.19}$$

where we assume finite second moments for the distribution of data X. We obtain with $I(\theta, \psi) = O_p(1)$

$$I(\theta, \psi) + O_p\left(n^{-\frac{1}{2}}\right) = I(\theta, \psi)\left(\mathrm{Id}_m + O_p\left(n^{-\frac{1}{2}}\right)\right) \tag{4.20}$$

and it follows

$$\begin{aligned} -\frac{1}{n}D^2\ell_0(\widehat{\theta}_0) &= I_{\theta\theta}(\widehat{\theta}, \psi_0)\left(\mathrm{Id}_{m_0} + O_p\left(n^{-\frac{1}{2}}\right)\right)\left(\mathrm{Id}_{m_0} + O_p\left(n^{-\frac{1}{2}}\right)\right) \\ &= I_{\theta\theta}(\widehat{\theta}, \psi_0)\left(\mathrm{Id}_{m_0} + 2O_p\left(n^{-\frac{1}{2}}\right) + O_p\left(n^{-1}\right)\right) \\ &= I_{\theta\theta}(\widehat{\theta}, \psi_0)\left(\mathrm{Id}_{m_0} + O_p\left(n^{-\frac{1}{2}}\right)\right) \end{aligned} \tag{4.21}$$

Analogously to above, we can show

$$\exp\left(X_i'\begin{pmatrix}\widehat{\theta} \\ \widehat{\psi}\end{pmatrix}\right) = \exp\left(X_i'\begin{pmatrix}\widehat{\theta} \\ \psi_0\end{pmatrix}\right)\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right)$$

and

$$G\left(X_i'\begin{pmatrix}\widehat{\theta} \\ \widehat{\psi}\end{pmatrix}\right) = G\left(X_i'\begin{pmatrix}\widehat{\theta} \\ \psi_0\end{pmatrix}\right)\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right)$$

which yields

$$W_{ii}\left(\widehat{\theta},\widehat{\psi}\right) = W_{ii}\left(\widehat{\theta},\psi_0\right)\left(1+O_p\left(n^{-\frac{1}{2}}\right)\right)$$

and thus

$$D^2\ell(\widehat{\theta},\widehat{\psi}) = D^2\ell(\widehat{\theta},\psi_0)\left(\mathrm{Id}_m + O_p\left(n^{-\frac{1}{2}}\right)\right).$$

With Assumption 4.2 3) and (4.20) we obtain

$$\begin{aligned}-\frac{1}{n}D^2\ell(\widehat{\theta},\widehat{\psi}) &= \left(I(\widehat{\theta},\psi_0)+O_p\left(n^{-\frac{1}{2}}\right)\right)\left(\mathrm{Id}_m+O_p\left(n^{-\frac{1}{2}}\right)\right)\\ &= I(\widehat{\theta},\psi_0)\left(\mathrm{Id}_m+O_p\left(n^{-\frac{1}{2}}\right)\right)\left(\mathrm{Id}_m+O_p\left(n^{-\frac{1}{2}}\right)\right)\\ &= I(\widehat{\theta},\psi_0)\left(\mathrm{Id}_m+O_p\left(n^{-\frac{1}{2}}\right)\right). \end{aligned} \tag{4.22}$$

Under Assumption 4.2 1) and with the determinant product rule it follows

$$\det\left(I(\widehat{\theta},\psi_0)\right) = \det\left(I_{\theta\theta}(\widehat{\theta},\psi_0)\right)\det\left(I_{\psi\psi}(\widehat{\theta},\psi_0)\right).$$

With the determinant product rule, (4.21), and (4.22) this yields

$$\begin{aligned}\frac{\det\left(-\frac{1}{n}D^2\ell_0(\widehat{\theta}_0)\right)}{\det\left(-\frac{1}{n}D^2\ell(\widehat{\theta},\widehat{\psi})\right)} &= \frac{\det\left(I_{\theta\theta}(\widehat{\theta},\psi_0)\left(\mathrm{Id}_{m_0}+O_p\left(n^{-\frac{1}{2}}\right)\right)\right)}{\det\left(I(\widehat{\theta},\psi_0)\left(\mathrm{Id}_m+O_p\left(n^{-\frac{1}{2}}\right)\right)\right)}\\ &= \frac{\det\left(I_{\theta\theta}(\widehat{\theta},\psi_0)\right)\det\left(\mathrm{Id}_{m_0}+O_p\left(n^{-\frac{1}{2}}\right)\right)}{\det\left(I_{\theta\theta}(\widehat{\theta},\psi_0)\right)\det\left(I_{\psi\psi}(\widehat{\theta},\psi_0)\right)\det\left(\mathrm{Id}_m+O_p\left(n^{-\frac{1}{2}}\right)\right)}\\ &= \frac{\det\left(\mathrm{Id}_{m_0}+O_p\left(n^{-\frac{1}{2}}\right)\right)}{\det\left(I_{\psi\psi}(\widehat{\theta},\psi_0)\right)\det\left(\mathrm{Id}_m+O_p\left(n^{-\frac{1}{2}}\right)\right)}. \end{aligned} \tag{4.23}$$

We obtain with Hadamard's inequality and the multiplicative and additive properties of O_p [203]

$$\begin{aligned}\det\left(\mathrm{Id}_{m_0}+O_p\left(n^{-\frac{1}{2}}\right)\right) &\leq \left(1+O_p\left(n^{-\frac{1}{2}}\right)\right)^{m_0} = 1+O_p\left(n^{-\frac{1}{2}}\right),\\ \det\left(\mathrm{Id}_m+O_p\left(n^{-\frac{1}{2}}\right)\right) &\leq \left(1+O_p\left(n^{-\frac{1}{2}}\right)\right)^{m} = 1+O_p\left(n^{-\frac{1}{2}}\right). \end{aligned} \tag{4.24}$$

$-D^2\ell(\widehat{\theta},\widehat{\psi})$ is positive semidefinite since ℓ is convex as shown in Section 3.1. Thus $I_{\theta\theta}(\widehat{\theta},\psi_0) = \mathbb{E}\left[-D^2_{\theta\theta}\ell(\widehat{\theta},\psi_0)\right]$ and $I(\widehat{\theta},\psi_0) = \mathbb{E}\left[-D^2\ell(\widehat{\theta},\psi_0)\right]$ are positive semidefinite and it follows $\mathrm{Id}_m + O_p\left(n^{-\frac{1}{2}}\right)$ and $\mathrm{Id}_{m_0} + O_p\left(n^{-\frac{1}{2}}\right)$ are positive semidefinite. Let $\mu_i \geq 0$ denote the eigenvalues of the random matrix $Y_n = O_p(n^{-\frac{1}{2}})$.

Then

$$\det\left(\mathrm{Id}_m + Y_n\right) = \prod_{i=1}^{m}(1+\mu_i) \geq 1 + \prod_{i=1}^{m}\mu_i = 1 + \det\left(Y_n\right) \geq 1 + O_p\left(n^{-\frac{1}{2}}\right)$$

and analogously $\det\left(\mathrm{Id}_{m_0} + Y_n\right) \geq 1 + O_p\left(n^{-\frac{1}{2}}\right)$. It follows with (4.24)

$$\det\left(\mathrm{Id}_m + O_p\left(n^{-\frac{1}{2}}\right)\right) = 1 + O_p\left(n^{-\frac{1}{2}}\right).$$

Substituting this into (4.23) yields with Proposition 3.1

$$\begin{aligned}\frac{\det\left(-\frac{1}{n}D^2\ell_0(\widehat{\theta}_0)\right)}{\det\left(-\frac{1}{n}D^2\ell(\widehat{\theta},\widehat{\psi})\right)} &= \frac{1 + O_p\left(n^{-\frac{1}{2}}\right)}{\det\left(I_{\psi\psi}(\widehat{\theta},\psi_0)\right)\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right)} \\ &= \det\left(I_{\psi\psi}(\widehat{\theta},\psi_0)\right)^{-1}\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right).\end{aligned}$$

and using $\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right)^{\frac{1}{2}} = 1 + O_p\left(n^{-\frac{1}{2}}\right)$, shown in Proposition 3.2, we obtain

$$\begin{aligned}\frac{\det\left(-\frac{1}{n}D^2\ell_0(\widehat{\theta}_0)\right)^{-\frac{1}{2}}}{\det\left(-\frac{1}{n}D^2\ell(\widehat{\theta},\widehat{\psi})\right)^{-\frac{1}{2}}} &= \left(\frac{\det\left(-\frac{1}{n}D^2\ell_0(\widehat{\theta}_0)\right)}{\det\left(-\frac{1}{n}D^2\ell(\widehat{\theta},\widehat{\psi})\right)}\right)^{-\frac{1}{2}} \\ &= \det\left(I_{\psi\psi}(\widehat{\theta},\psi_0)\right)^{\frac{1}{2}}\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right)^{\frac{1}{2}} \\ &= \det(I_{\psi\psi}(\widehat{\theta},\psi_0))^{\frac{1}{2}}\left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right). \qquad (4.25)\end{aligned}$$

It is left to show that $\frac{f_0(\widehat{\theta}_0)}{f(\widehat{\theta},\widehat{\psi})}$ is approximated by $\frac{1}{f_{\psi|\theta}(\widehat{\psi}|\widehat{\theta})}$. Using a Taylor expansion of f_0 around $\widehat{\theta}$ we show with (4.17), Definition 3.3, and under the assumption that f_0 and its derivative is bounded in an area around $\widehat{\theta}$ that

$$\begin{aligned}f_0(\widehat{\theta}_0) &= f_0(\widehat{\theta}) + \sum_{j=1}^{m_0}\left(\widehat{\theta}_j - \widehat{\theta}_{0j}\right)\partial_{\theta_j}f_0(\widehat{\theta}) + o\left(||\widehat{\theta} - \widehat{\theta}_0||\right) \\ &= f_0(\widehat{\theta}) + \sum_{j=1}^{m_0}O_p\left(n^{-\frac{1}{2}}\right)\partial_{\theta_j}f_0(\widehat{\theta}) + o\left(O_p\left(n^{-\frac{1}{2}}\right)\right) \\ &= f_0(\widehat{\theta}) + O_p\left(n^{-\frac{1}{2}}\right) = f_0(\widehat{\theta})(1 + O_p\left(n^{-1}\right)).\end{aligned}$$

Thus, with Assumption 4.2 2) it follows

$$\frac{f_0(\widehat{\theta}_0)}{f(\widehat{\theta},\widehat{\psi})} = \frac{f_0(\widehat{\theta}_0)}{f_{\psi|\theta}(\widehat{\psi}|\widehat{\theta})f_0(\widehat{\theta})} = \frac{f_0(\widehat{\theta})(1+O_p\left(n^{-1}\right))}{f_{\psi|\theta}(\widehat{\psi}|\widehat{\theta})f_0(\widehat{\theta})} = \frac{1+O_p\left(n^{-1}\right)}{f_{\psi|\theta}(\widehat{\psi}|\widehat{\theta})}. \tag{4.26}$$

Substituting (4.25) and (4.26) into (4.14) yields with the additive and multiplicative properties of O_p following approximation for the BF

$$\begin{aligned}\mathsf{BF} &= \left(\frac{2\pi}{n}\right)^{\frac{m_0-m}{2}} \det(I_{\psi\psi}(\widehat{\theta},\psi_0))^{\frac{1}{2}} \left(1+O_p\left(n^{-\frac{1}{2}}\right)\right) \frac{\exp(\ell_0(\widehat{\theta}_0))}{\exp(\ell(\widehat{\theta},\widehat{\psi}))} \frac{1+O_p\left(n^{-1}\right)}{f_{\psi|\theta}(\widehat{\psi}|\widehat{\theta})} \\ &= \left(\frac{2\pi}{n}\right)^{\frac{m_0-m}{2}} \det(I_{\psi\psi}(\widehat{\theta},\psi_0))^{\frac{1}{2}} \frac{\exp(\ell_0(\widehat{\theta}_0))}{\exp(\ell(\widehat{\theta},\widehat{\psi}))} \frac{1}{f_{\psi|\theta}(\widehat{\psi}|\widehat{\theta})} \left(1+O_p\left(n^{-\frac{1}{2}}\right)\right).\end{aligned}$$

□

We define the Schwarz criterion, which was first introduced by Schwarz (1978) and relate this to the result obtained in Proposition 4.3 [15, 204].

Definition 4.3 (Schwarz criterion)
The Schwarz criterion for the multiple logistic regression model is given by

$$S = \ell_0(\widehat{\theta}) - \ell(\widehat{\theta},\widehat{\psi}) + \frac{m-m_0}{2}\log(n). \tag{4.27}$$

Corollary 4.1
Under the assumptions of Proposition 4.3 and if the conditional prior density $f_{\psi|\theta}(\psi|\theta)$ is bounded for $(\theta,\psi) \in \mathcal{B}_\delta(\widehat{\theta},\widehat{\psi})$, $\delta > 0$ it holds that

$$\log(\mathsf{BF}) = S + O_p(1).$$

Proof. We take the logarithm on both sides of the result from Proposition 4.3 and obtain

$$\begin{aligned}\log(\mathsf{BF}) &= \frac{m_0-m}{2}\log\left(\frac{2\pi}{n}\right) + \frac{1}{2}\log\left(\det\left(I_{\psi\psi}(\widehat{\theta},\psi_0)\right)\right) + \ell_0(\widehat{\theta}_0) - \ell(\widehat{\theta},\widehat{\psi}) \\ &\quad - \log\left(f_{\psi|\theta}(\widehat{\psi}|\widehat{\theta})\right) + \log\left(1+O_p\left(n^{-\frac{1}{2}}\right)\right) \\ &= \ell_0(\widehat{\theta}_0) - \ell(\widehat{\theta},\widehat{\psi}) + \frac{m-m_0}{2}\log(n) + \frac{m_0-m}{2}\log(2\pi) \\ &\quad + \frac{1}{2}\log\left(\det\left(I_{\psi\psi}(\widehat{\theta},\psi_0)\right)\right) - \log\left(f_{\psi|\theta}(\widehat{\psi}|\widehat{\theta})\right) + O_p\left(n^{-\frac{1}{2}}\right),\end{aligned}$$

where we use $\log\left(1+O_p\left(n^{-\frac{1}{2}}\right)\right)=O_p\left(n^{-\frac{1}{2}}\right)$ shown in Proposition 3.2. Thus we have

$$\begin{aligned}\log(\mathsf{BF}) = S + \frac{m_0-m}{2}\log(2\pi) + \frac{1}{2}\log\left(\det\left(I_{\psi\psi}(\widehat{\theta},\psi_0)\right)\right) \\ - \log\left(f_{\psi|\theta}(\widehat{\psi}|\widehat{\theta})\right) + O_p\left(n^{-\frac{1}{2}}\right).\end{aligned} \tag{4.28}$$

Using (4.19) we obtain

$$\log\left(\det\left(I_{\psi\psi}(\widehat{\theta},\psi_0)\right)\right) = \log\left(\det\left(O_p(1)\right)\right) = O_p(1).$$

With $\frac{m_0-m}{2}\log(2\pi)=O(1)$ and $\log\left(f_{\psi|\theta}(\widehat{\psi}|\widehat{\theta})\right)=O_p(1)$ this yields

$$\log(\mathsf{BF}) = S + O_p(1).$$

□

Theorem 4.1

Suppose the assumptions of Proposition 4.3 hold and let the prior density $f_{\psi|\theta}(\psi|\theta)$ for ψ under the alternative hypothesis have the form

$$\psi|\theta \sim \mathcal{N}_{m-m_0}(\psi_0, I_{\psi\psi}(\theta,\psi_0)^{-1}).$$

Then we have

$$\log(\mathsf{BF}) = S + O_p\left(n^{-\frac{1}{2}}\right).$$

Proof. The multivariate normal density of $\psi|\theta$ at $\widehat{\psi}|\widehat{\theta}$ is given by

$$\begin{aligned}f_{\psi|\theta}(\widehat{\psi}|\widehat{\theta}) &= (2\pi)^{\frac{m_0-m}{2}}\det(I_{\psi\psi}(\widehat{\theta},\psi_0))^{\frac{1}{2}}\exp\left(-\frac{1}{2}(\widehat{\psi}-\psi_0)'I_{\psi\psi}(\widehat{\theta},\psi_0)(\widehat{\psi}-\psi_0)\right) \\ &= (2\pi)^{\frac{m_0-m}{2}}\det(I_{\psi\psi}(\widehat{\theta},\psi_0))^{\frac{1}{2}}\exp\left(O_p\left(n^{-\frac{1}{2}}\right)O_p(1)O_p\left(n^{-\frac{1}{2}}\right)\right) \\ &= (2\pi)^{\frac{m_0-m}{2}}\det(I_{\psi\psi}(\widehat{\theta},\psi_0))^{\frac{1}{2}} + O_p\left(n^{-1}\right),\end{aligned}$$

where we use (4.19) and $\exp(O_p\left(n^{-1}\right))=1+O_p\left(n^{-1}\right)$ shown in Proposition 3.2. Next, note that

$$\begin{aligned}\log\left(f_{\psi|\theta}(\widehat{\psi}|\widehat{\theta})\right) &= \log\left((2\pi)^{\frac{m-m_0}{2}}\det(I_{\psi\psi}(\widehat{\theta},\psi_0))^{\frac{1}{2}}\left(1+O_p\left(n^{-1}\right)\right)\right) \\ &= \frac{m-m_0}{2}\log(2\pi) + \frac{1}{2}\log\left(\det(I_{\psi\psi}(\widehat{\theta},\psi_0))\right) + \log\left(1+O_p\left(n^{-1}\right)\right) \\ &= \frac{m-m_0}{2}\log(2\pi) + \frac{1}{2}\log\left(\det(I_{\psi\psi}(\widehat{\theta},\psi_0))\right) + O_p\left(n^{-1}\right),\end{aligned}$$

since $\log\left(1 + O_p\left(n^{-1}\right)\right) = O_p\left(n^{-1}\right)$ as shown in Proposition 3.2. Substitution into (4.28) yields

$$\log\left(\mathsf{BF}\right) = S + O_p\left(n^{-\frac{1}{2}}\right) - O_p\left(n^{-1}\right) = S + O_p\left(n^{-\frac{1}{2}}\right).$$

□

The prior distribution used in Theorem 4.1 is called a unit information prior distribution as it contains the amount of information of one observation [194, 205, 206]. It sometimes is classified as a reference prior in the sense that the prior distribution is specified by a formal rule [205].

Jeffreys used a unit information prior distribution in terms of a Cauchy distribution with the Fisher information as scale parameter for normal location testing problems [183, 194]. Kass and Wasserman (1995) used a normal unit information prior for testing iid variables and Raftery (1996) for stating an approximation to the BIC [194, 195]. Pauler (1998) approximated the Schwarz criterion for normal linear models with a normal unit information prior distribution [198].

The unit information prior contains the information of one observation and thus we expect the variance $I_{\psi\psi}(\theta, \psi_0)^{-1}$ to be large. For illustration, we consider random samples of size $n = 10, 100$, and 1000 from the data described in Chapter 2. We fit a logistic regression model for the pathological outcome ECE with an intercept and $\log(\text{PSA})$ as the only covariate. We are interested in testing whether the coefficient ψ corresponding to the log OR for $\log(\text{PSA})$ is equal to zero.

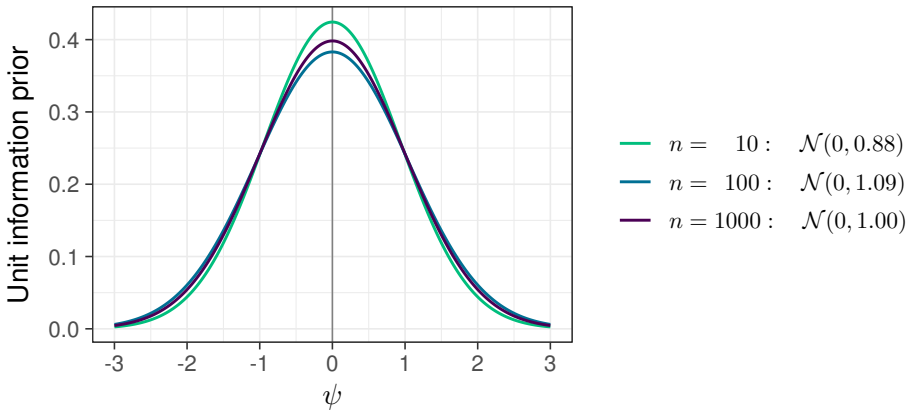

Figure 4.1 Unit information prior distribution for testing whether the coefficient ψ of $\log(\text{PSA})$ is equal to zero in a logistic regression model for ECE on random samples of the EMR data using different sample sizes n.

Figure 4.1 displays the observed normal unit information prior distribution from random samples of the EMR data corresponding to the different sample sizes. The variance of the

prior distribution stays relatively constant with increasing sample size and thus the observed unit information prior contains a similar amount of information.

In Figure 4.2 we see the likelihood functions for ψ, assuming the coefficient for the intercept is fixed. In contrast to the unit information prior distribution the variance of the likelihood function decreases with increasing sample size.

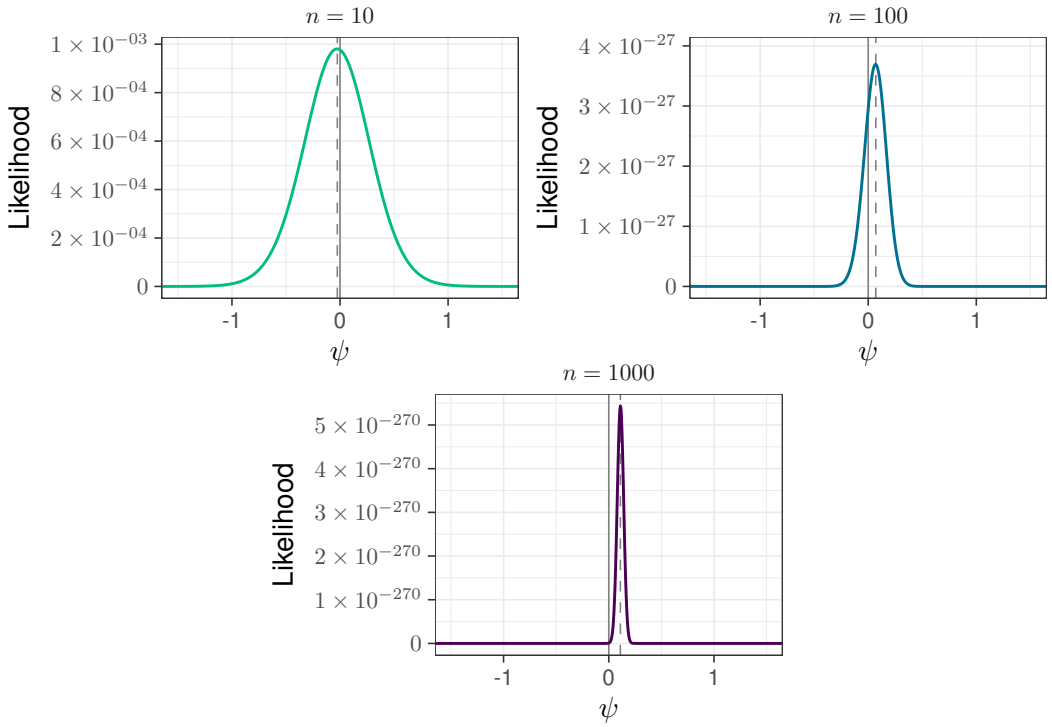

Figure 4.2 Likelihood function for the coefficient ψ of $\log(\text{PSA})$ in a logistic regression model for ECE containing an intercept and $\log(\text{PSA})$ as covariate fitted to random samples of the EMR data using different sample sizes.

The Schwarz criterion provides a good approximation to the Bayes factor for a specific parametrization, certain assumptions, and a choice of unit information prior. Its approximation is restricted to the comparison of two nested models, whereas it is often desirable to compare multiple models. A possible approach is to compare each model with a reference model, for example the null or intercept model M_0 with no independent variables, or the saturated model M_S with an exact fit for each data point [193].

Let ℓ_i and ℓ_j denote the log-likelihood and m_i and m_j the number of parameters for M_i and M_j, respectively. The approximate $\log(\text{BF}_{ij})$ for comparing models M_i and M_j is given by

$$\log\left(\text{BF}_{ij}\right) \approx \ell_i - \ell_j + \frac{m_j - m_i}{2}\log(n).$$

For two models M_1 and M_2 each nested within the saturated model M_S, we have

$$\begin{aligned}\log\left(\mathsf{BF}_{1S}\right) - \log\left(\mathsf{BF}_{2S}\right) &\approx \ell_1 - \ell_S + \frac{m_S - m_1}{2}\log(n) - \ell_2 + \ell_S - \frac{m_S - m_2}{2}\log(n) \\ &= \ell_1 - \ell_2 + \frac{m_2 - m_1}{2}\log(n) \approx \log\left(\mathsf{BF}_{12}\right).\end{aligned}$$

Thus, by comparing each model to the same saturated model we can construct an approximation for comparing the two models directly [193]. Analogously, using the null model M_0 that is nested within all models we have that

$$\begin{aligned}\log\left(\mathsf{BF}_{02}\right) - \log\left(\mathsf{BF}_{01}\right) &\approx \ell_0 - \ell_2 + \frac{m_2 - m_0}{2}\log(n) - \ell_0 + \ell_1 - \frac{m_1 - m_0}{2}\log(n) \\ &= \ell_1 - \ell_2 + \frac{m_2 - m_1}{2}\log(n) \approx \log\left(\mathsf{BF}_{12}\right).\end{aligned}$$

Although the choice of the comparison model M_0 or M_S results in a numerical difference in absolute values, both can be used to reconstruct $\log\left(\mathsf{BF}_{12}\right)$. We use the null model M_0 and follow convention to define the Bayesian information criterion (BIC) [193, 204, 207].

Definition 4.4 (Bayesian information criterion (BIC))
The BIC for a multiple logistic regression model M_m with m the dimension of β is given by

$$\mathsf{BIC} = -2\ell(\widehat{\beta}) + m\log(n).$$

Let BIC_1 and BIC_2 be the BIC for M_1 and M_2, respectively. Then,

$$\begin{aligned}\frac{1}{2}\mathsf{BIC}_2 - \frac{1}{2}\mathsf{BIC}_1 &= -\ell_2 + \frac{m_2}{2}\log(n) - \left(-\ell_1 + \frac{m_1}{2}\log(n)\right) \\ &= \ell_1 - \ell_2 + \frac{m_2 - m_1}{2}\log(n) \\ &\approx \log\left(\mathsf{BF}_{12}\right).\end{aligned}$$

Thus, we can use the BIC to reconstruct the approximation to $\log\left(\mathsf{BF}_{12}\right)$.

Here we did not explicitly specify that the two models M_1 and M_2 need to be nested, although we used that assumption throughout Section 4.1. While the approximation holds specifically for nested models, using a single comparison model M_0 or M_S provides a possibility to compare more general specified models [193].

We have shown how the Schwarz criterion approximates the BF for a specific unit information prior distribution for logistic regression models. Similar results have been shown for other

regression models such as normal linear models, linear mixed models, and censored survival models [15, 195, 198, 208, 209].

The BIC is now commonly used for model selection, as well as in Bayesian model averaging (BMA) to account for model uncertainty. Here the BIC approximation to the BF provides computationally simple and conservative weights for different models under consideration [210–213].

It has been extended to numerous complex modeling situations, including to account for large model spaces and small-n-large-p model selection problems, for sparse GLM or for singular model selection problems [214–218]. Further extensions have been proposed including Gaussian graphical models and high-dimensional Ising models, non-linear models estimated by penalized likelihood methods, to account for longitudinal and clustered data, for mixture models, quantile regression, for selection of tuning parameters in bridge regression models, or for the Cox model with a high-dimensional feature space [219–226]. Recently, Bayarri et al. (2019) proposed a modification to incorporate prior information [227].

Besides modifications for multiple change-point models, to locate multiple interacting quantitative trait loci, or for genome-wide association studies (GWAS), the BIC is adapted for peptide identification or source enumeration in array processing [228–235]. In applications for speech recognition, the BIC has been modified to identify the number of clusters for clustering, for decision tree state tying or for speaker diarization [236–241]. Furthermore, the BIC has many applications in medical use cases. A query for the exact term ”Bayesian information criterion” yielded 773 results in PubMed, a free search engine containing more than 30 million citations and abstracts primarily from biomedicine and health-related fields (September 13, 2019 [242]).

The BIC is implemented in R via the BIC function as well as a special case of the extractAIC-function provided by the MASS package [243]. This package includes also the stepAIC-function for implementation of a step-wise model selection procedure with the Akaike's information criterion (AIC) and BIC [243]. We describe this procedure in more detail in Section 4.3.

4.2 Simulation studies

We validate the results derived in Section 4.1 with simulation studies. We consider two nested models with a unit information prior distribution and compare the approximate BF (BF_{approx}) with the calculated BF (BF_{calc}) obtained with numerical integration using the integrate-function in R [244]. Table 4.1 summarizes the simulation specifications for two scenarios that use different covariates and are similar otherwise.

Covariates	Scenario 1:	Intercept $X_0 = 1$ and one binary covariate $X_1 \sim Ber(0.5)$, re-scaled to mean 0
	Scenario 2:	Intercept $X_0 = 1$ and one continuous covariate X_1 sampled from $\log(PSA)$ values in the EMR data, re-scaled to mean 0
β-values	$\beta_0 = 1$; $\beta_1 = 0, 0.01, 0.1, 0.5, 1, 3$	
Hypotheses	H_0: $\beta_1 = 0$ vs. H_A: $\beta_1 \neq 0$	
Sample sizes	$n = 10, 100, 1000$	
Samples	$y_i \sim Ber(p_i)$ with $p_i = G(X'\beta) = \frac{\exp(X'\beta)}{1+\exp(X'\beta)}$	
Evaluation	- Calculate the 'exact' BF_{calc} using numeric integration with prior densities $f_0(\beta_0) \sim 1$ and $f_{\beta_1\|\beta_0}(\beta_1\|\beta_0) \sim N(0, I_{\beta_1\beta_1}(\beta_0, 0)^{-1})$ - Compare BF_{calc} with $BF_{approx} = \exp(S)$ - Compare the BF decision according to Jeffreys' scale with p - values obtained from the Wald test	
Number of repetitions	$k = 1000$	
Total number of simulations	18,000 for each scenario	

Table 4.1 Specifications of the simulation study for the BF approximation for nested logistic regression models.

In the following, we discuss the results for Scenario 1. Results for Scenario 2 are provided in Appendix B.1. Overall the two scenarios returned similar results and we discuss differences where they occurred.

Figure 4.3 shows that for an increasing sample size the difference between BF_{approx} and BF_{calc} decreases. For a sample size of $n = 100$ the absolute difference is less than 0.31 and for $n = 1000$ less than 0.11. With a small sample size of $n = 10$, we detect some outlying values for $BF_{calc} - BF_{approx}$, which we now discuss.

In 6,000 simulation runs for Scenario 1 with $n = 10$ and different β_1-values the absolute difference between BF_{approx} and BF_{calc} is greater than 2 for 228 cases (3.8%). These outliers occur in those cases where either $Y_i = 1$ for all $i \in \{1, \ldots, 10\}$, $X_i = 1$ for all $i \in \{1, \ldots, 10\}$, or $X_i = 0$ for all $i \in \{1, \ldots, 10\}$. In the latter two cases BF_{calc} is infinite and in all these cases the ORs for either β_0 or β_1 are unbounded, since $n_{00} = 0$ or $n_{11} = 0$, where $n_{00}, n_{01}, n_{10}, n_{11}$ are the counts of different combinations for X_1 and Y as defined in Table 3.1. Therefore, the sequence of log-likelihoods is not Laplace regular and Assumption 4.1 is not fulfilled. The

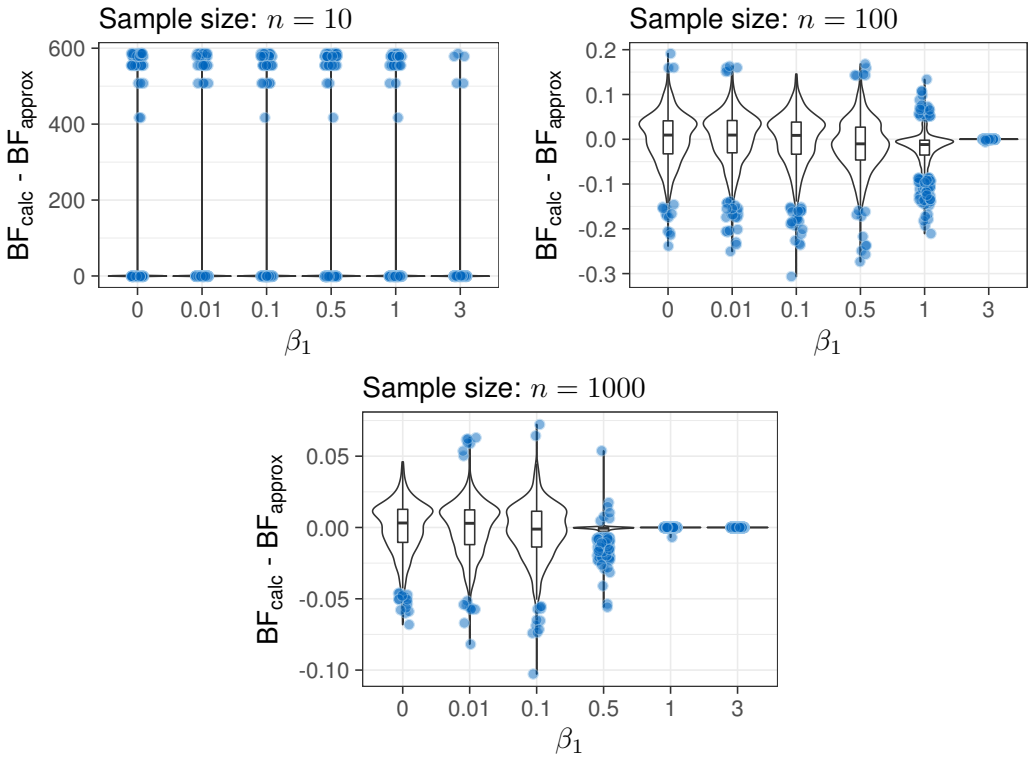

Figure 4.3 Difference between BF_{calc} and BF_{approx} for the first scenario of the simulation study for logistic regression models with sample sizes $n = 10$, $n = 100$, and $n = 1000$ based on 6,000 simulations for each sample size.

228 outliers are excluded in Figure 4.4 as well as in the following analysis. In Figure 4.4 we see that in general the absolute difference between BF_{approx} and BF_{calc} is smaller than 0.5, but there are outlying values up to 1.5 for all cases of β_1.

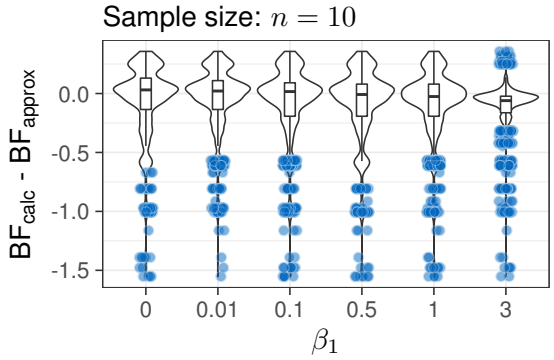

Figure 4.4 Difference between BF_{calc} and BF_{approx} for the first scenario of the simulation study for logistic regression models with a sample size $n = 10$ excluding 228 outlying cases out of 6,000.

In simulations with a continuous covariate (Scenario 2) we detect 202 outliers for $n = 10$ (3.4%) with an absolute difference between BF_{approx} and BF_{calc} larger than 2 and all those occur if $Y_i = 1$ for all $i \in \{1, \ldots, 10\}$ (Fig. B.1). Thus, again the definition of Laplace regularity is not fulfilled for the sequence of log-likelihoods and these cases are excluded from additional analyses in Appendix B.1. After removing the outliers for $n = 10$ the absolute

difference between BF_{approx} and BF_{calc} is smaller than 0.5 in general and the distribution is more symmetric around 0 than for the first scenario (Fig. 4.4, Fig. B.2). For $n = 100$ and $n = 1000$ the maximum absolute differences are 1.32 and 0.61 and thus larger compared to Scenario 1 (Fig. 4.3, Fig. B.1).

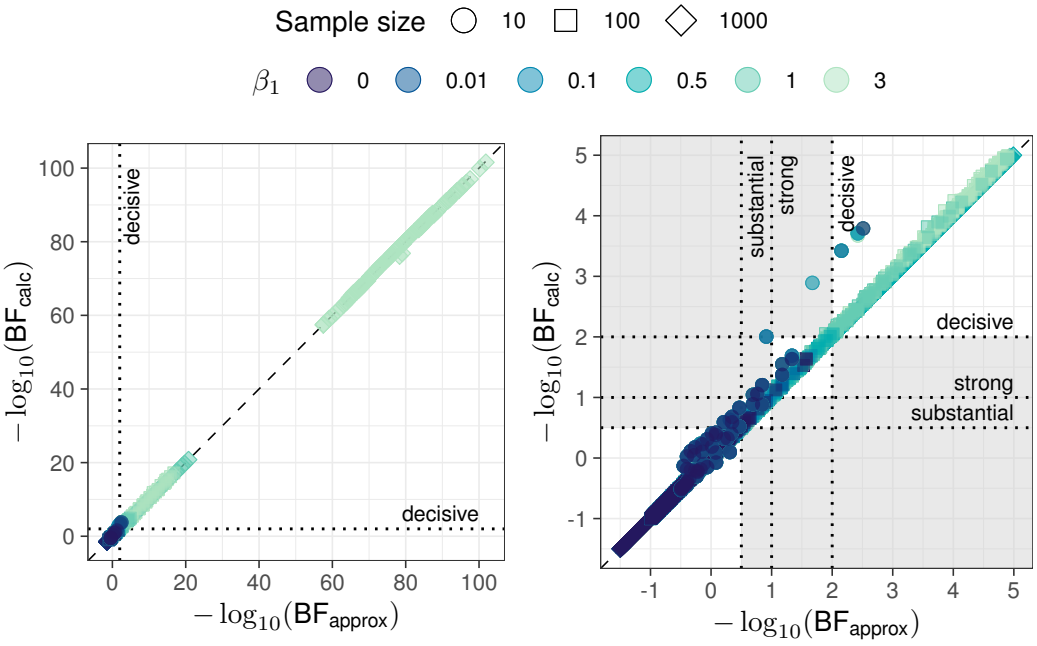

Figure 4.5 Comparison of BF_{approx} to BF_{calc} on a negative $\log_{10}$ - scale using Jeffreys' rule for the first scenario. On the left side all simulation runs are shown, whereas on the right side the focus is on the range between -1.5 and 5 on a negative $\log_{10}$-scale.

For a better judgment of the difference between BF_{approx} and BF_{calc}, we plot them on a negative $\log_{10}$-scale and classify the evidence against the null hypothesis according to Jeffreys' rule as introduced in Section 3.3 into substantial, strong, and decisive. Figure 4.5 shows that in general $-\log_{10}(BF_{approx})$ and $-\log_{10}(BF_{calc})$ agree well, but the classification differs in 525 of 17,772 cases (3.0%) located in the gray areas. Stratified by sample size the misclassification rates are 8.6% (499/5,772) for $n = 10$, 0.4% (23/6,000) for $n = 100$, and 0.1% (3/6,000) for $n = 1000$.

When investigating the 499 misclassified cases for $n = 10$ in more detail, we find that at least one of n_{00}, n_{01}, n_{10}, or n_{11} is equal to 0 and thus the underlying data is completely or quasicompletely separated as defined in Section 3.1. Therefore, the assumptions for Laplace regularity are not fulfilled and the approximation fails. With an increasing sample size no such outliers occur in the simulations. In medical applications, statistical analysis would not proceed in the presence of small samples with unstable effects, justifying exclusion of the 499 outlying cases.

On closer inspection misclassification of 26 cases for $n = 100$ and $n = 1000$ occurs on

the boundaries of Jeffreys' rule, where $-\log_{10}(\text{BF}_{\text{approx}})$ and $-\log_{10}(\text{BF}_{\text{calc}})$ do not differ to a large extent. The median, minimum, and maximum difference between $-\log_{10}(\text{BF}_{\text{approx}})$ and $-\log_{10}(\text{BF}_{\text{calc}})$ for these cases are -0.03, -0.17, and 0.02, respectively. Thus, $\text{BF}_{\text{approx}}$ rather underestimates BF_{calc} and in all but 2 misclassified cases the $\text{BF}_{\text{approx}}$ gave a more conservative classification according to the Jeffreys' rule. We do not exclude these 26 misclassified cases.

In Scenario 2 simulating with a continuous covariate we detect 633 misclassified cases out of 17,798 (3.6%) (Fig. B.3). Stratified by sample size the misclassification rate are 8.3% (484/5,798) for $n = 10$, 2.4% (146/6,000) for $n = 100$, and 0.1% (3/6,000) for $n = 1000$. For $n = 10$ misclassification in 200 cases occurs when the minimum $\log(\text{PSA})$ value of one group is larger than the maximum $\log(\text{PSA})$ value of the other group and the data is completely separated. For 192 misclassified cases for $n = 10$ we observe a quasicomplete separation in the data in the sense that the two groups barely overlap and only one observation of one group has a $\log(\text{PSA})$-value within the range of the other group. The assumptions for Laplace regularity are not fulfilled and the approximation fails in these cases and similar to the Scenario 1, we exclude these 392 misclassified cases for the further analysis in Appendix B.1. With an increasing sample size no such separation occurs in the simulations. For the remaining cases and the misclassified cases for $n = 100$ and $n = 1000$ misclassification occurs on the boundaries of the Jeffreys' rule, similar to Scenario 1. The difference between $\text{BF}_{\text{approx}}$ and BF_{calc} is slightly larger with the continuous covariate in Scenario 2 than for the binary covariate in Scenario 1 and this leads to more misclassifications on the boundaries of Jeffreys' rule.

In addition to the difference between $\text{BF}_{\text{approx}}$ and BF_{calc}, we evaluate $\text{BF}_{\text{approx}}$ in comparison to the p-value obtained from the Wald test for the coefficient β_1. Again, we use a negative $\log_{10}$-scale and classify $-\log_{10}(\text{BF}_{\text{approx}})$ according to Jeffreys' rule. We consider 3 different significance levels for the p-value, 0.05, 0.01, and 0.001, which correspond to $\approx$ 1.3, 2, and 3 on the negative $\log_{10}$-scale.

On the left side in Figure 4.6, we see for a large sample size and large β_1-value both the p-value and $\text{BF}_{\text{approx}}$ are highly significant ($p < 0.001$) and decisive, respectively. However, for $n = 100$, we detect 33 cases (0.6%), located in the lower gray area, where the p-value is large and low on the $\log_{10}$-scale, but Jeffreys' rule classifies these as decisive. These cases occur for $\beta_1 = 3$ and we detect that $n_{10} = 0$. Thus, the calculation of the p - value fails and results in values close to 1, whereas the $\text{BF}_{\text{approx}}$ yields a correct decision as we expect decisive evidence against the null hypothesis $\beta_1 = 0$ for a true value $\beta_1 = 3$.

On the right side in Figure 4.6, we see similar cases for $n = 10$, where the p-value is large, but the $\text{BF}_{\text{approx}}$ provides substantial, strong or decisive evidence against the null-

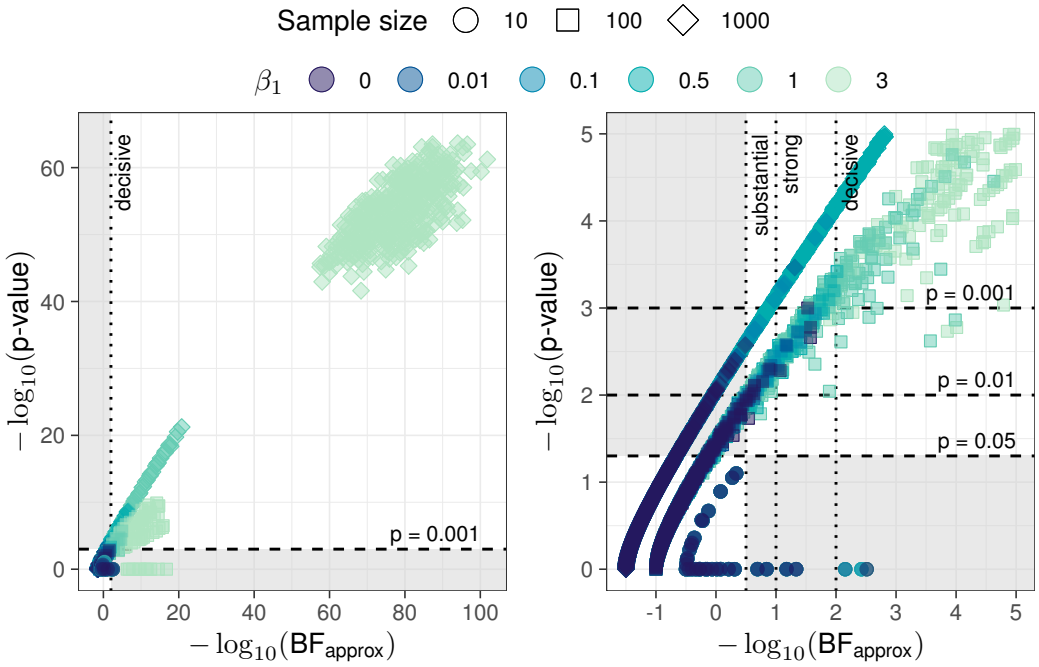

Figure 4.6 Comparison of $\text{BF}_{\text{approx}}$ and the p-value on a negative $\log_{10}$ - scale for the first scenario. On the left side all simulation runs are shown, whereas on the right side the focus is on the range between -1.5 and 5 on a negative $\log_{10}$-scale.

hypothesis. Those 485 cases out of 5,273 remaining simulations for $n = 10$ (9.2 %) occur when least one of n_{00}, n_{01}, n_{10}, or n_{11} is equal to 0. We note that only a comparably few combinations of $-\log_{10}(\text{BF}_{\text{approx}})$ and p-values are obtained for $n = 10$ and thus in Figure 4.6 many points overlap and are not visible. As already mentioned, in medical applications, statistical analysis would not proceed in the presence of small samples with unstable effects. However, the $\text{BF}_{\text{approx}}$ seems to yield more stable results compared to the p-value, although the assumptions for Laplace regularity are not fully satisfied.

For $n = 100$ and $n = 1000$ we split the right side in Figure 4.6 and display the results separated by samples size and β_1-value in Figure 4.7 for better visibility.

For $n = 1000$ and $\beta_1 = 1$ only very few p-values and $\text{BF}_{\text{approx}}$-values are smaller than 5 on a negative $\log_{10}$-scale and for $\beta_1 = 3$ no values fall within the displayed range (Fig.4.7, bottom row). For $n = 100$ and $\beta_1 = 3$ all p-values are less than 0.01 and $\text{BF}_{\text{approx}}$-values indicate at least strong evidence against the null hypothesis. These results are not surprising as we expect evidence against the null hypothesis for these parameter settings.

For $\beta_1 \leq 0.5$ we can see a shift of the p-values for a sample size of $n = 1000$ compared to $n = 100$, where for a similar $-\log_{10}(\text{BF}_{\text{approx}})$-value the p-value tends to be smaller, and thus larger on a negative $\log_{10}$-scale for the larger sample size. Therefore, for comparable evidence against the null hypothesis based on $\text{BF}_{\text{approx}}$, the p-value yields more significant results for a larger sample size. Moreover, we detect several cases, where $\text{BF}_{\text{approx}}$ does not

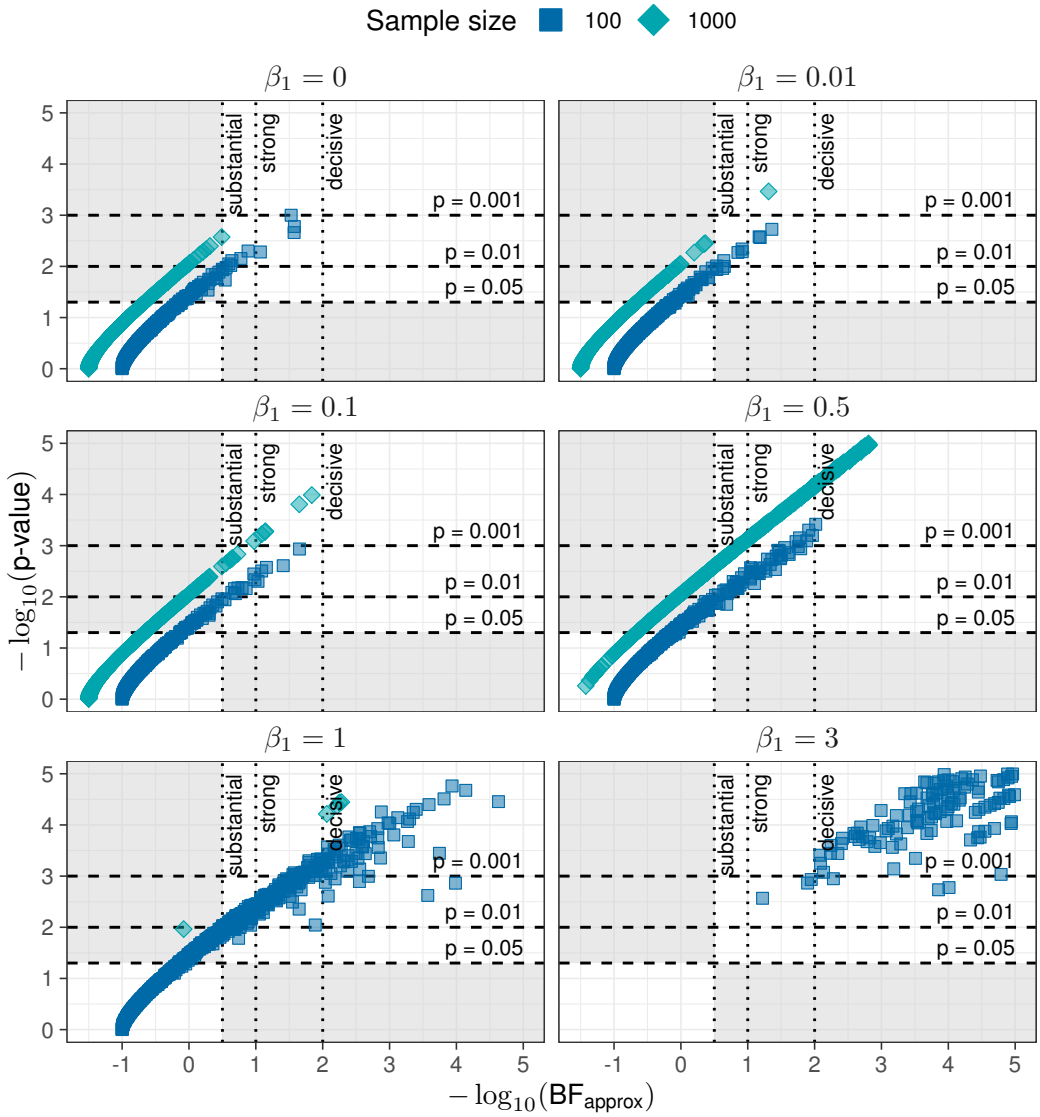

Figure 4.7 Comparison of $\text{BF}_{\text{approx}}$ and the p-value on a negative $\log_{10}$ - scale for the first scenario for $n = 100$ and $n = 1000$ with focus is on the range between -1.5 and 5 on a negative $\log_{10}$-scale.

provide evidence against the null hypothesis with $-\log_{10}(\text{BF}_{\text{approx}}) < 0.5$, but the Wald test yields p-values at least smaller than 0.05. These disagreeing cases are located in the upper gray areas. Stratified by sample size the disagreement proportions are 7.7% (460/6,000) for $n = 100$ and 8.1% (485/6,000) for $n = 1000$. There are no cases, where $\text{BF}_{\text{approx}}$ indicates evidence against the null hypothesis and the p-value is larger than 0.05.

For Scenario 2 we observe a similar shift in p-values and the disagreement proportions are 7.3% (441/6,000) for $n = 100$ and 4.7% (281/6,000) for $n = 1000$. In 415 cases out of 4,525 (9.2%) for $n = 10$, $\text{BF}_{\text{approx}}$ indicates evidence against the null hypothesis, but the p-value is larger than 0.05. However, for larger sample sizes we do not detect such cases. Overall we can consider the $\text{BF}_{\text{approx}}$ to be more conservative.

4.3 Separate logistic regression models for pathological outcomes after prostatectomy

In the following, we consider data from Chapter 2 and fit logistic regression models to each of the 5 pathological outcomes after prostatectomy separately. We use the BIC derived in Section 4.1 as a selection criteria in a step-wise variable selection procedure described in Algorithm 4.1 starting with a model containing only an intercept [245]. The aim of the algorithm is to find an optimal model in terms of minimizing the BIC. Thus, at each step we either include or exclude a variable from the set of all available variables without interactions and repeat until the BIC cannot be further minimized.

Algorithm 4.1: Stepwise variable selection procedure

Input: Initial model M, e.g. model containing only the intercept, available variables

$M_{new} \leftarrow M$

`vars_in` $\leftarrow$ model variables contained in M

`vars_out` $\leftarrow$ model variables available for selection and not contained in M

repeat

  $M^* \leftarrow M_{new}$

  if `vars_out` is **not** empty **then**

    $\mathtt{v}^{(1)} \leftarrow \underset{\mathtt{v} \in \mathtt{vars\_out}}{\operatorname{argmin}}$ BIC(M^* with $\mathtt{v}$)

    $M^{(1)} \leftarrow M^*$ with $\mathtt{v}^{(1)}$

  end

  if `vars_in` is **not** empty **then**

    $\mathtt{v}^{(2)} \leftarrow \underset{\mathtt{v} \in \mathtt{vars\_in}}{\operatorname{argmin}}$ BIC(M^* without $\mathtt{v}$)

    $M^{(2)} \leftarrow M^*$ without $\mathtt{v}^{(2)}$

  end

  if *BIC*$(M^{(1)}) <$ *BIC*$(M^{(2)})$ **then**

    $\mathtt{vars\_in} \leftarrow \mathtt{vars\_in} \cup \mathtt{v}^{(1)}$

    $\mathtt{vars\_out} \leftarrow \mathtt{vars\_out} \setminus \mathtt{v}^{(1)}$

    $M_{new} \leftarrow M^{(1)}$

  else

    $\mathtt{vars\_in} \leftarrow \mathtt{vars\_in} \setminus \mathtt{v}^{(2)}$

    $\mathtt{vars\_out} \leftarrow \mathtt{vars\_out} \cup \mathtt{v}^{(2)}$

    $M_{new} \leftarrow M^{(2)}$

  end

until *BIC*$(M_{new}) \geq$ *BIC*(M^*)

return M^*

There are many other model selection procedures available and we choose two most commonly used for comparison [246]. First, we consider the AIC defined by

$$\text{AIC} = -2\ell(\widehat{\beta}) + 2\dim(\widehat{\beta}),$$

as a selection criteria in the step-wise approach described in Algorithm 4.1 [247]. Second, we use the least absolute shrinkage and selection operator (LASSO) estimate defined by

$$\widehat{\beta}_{\text{LASSO}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i \log\left(G(X_i'\beta)\right) + (1-y_i) \log\left(1 - G(X_i'\beta)\right) + \lambda ||\beta||_1 \right) \right\},$$

with sample size n, $G(X_i'\beta) = \dfrac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}$, the $\mathcal{L}_1$-norm $||\cdot||_1$ and λ as penalization parameter [248, 249]. We determine λ using a 10-fold cross validation on the training set with 16,427 observations resulting in folds with 1,643 observations. Thereby we consider those models that are selected using λ at maximum AUC (LASSO$_{max}$) or λ within one standard error of the maximum AUC (LASSO$_{1se}$), where the AUC defined in Section 3.5 is evaluated on the remaining fold of the cross validation in each step. Using the selected variables we refit the models without shrinkage parameter to avoid bias for the coefficients towards 0 [116].

As described in Section 2.6, we use 10 imputation sets to account for the missing data. We apply the different model selection algorithms to each imputation of the training set with 16,427 observations and evaluate those models on the corresponding imputation of the validation set with 4,180 observations using the AUC. Figure 4.8 visualizes the AUC with 95%-confidence interval (CI) as proposed by DeLong et al. (1988) for each imputation set and each model selection procedure [250]. Across the different selection algorithms the AUC does not vary substantially for each outcome suggesting choice of which procedure to use is not relevant. The ranking of the outcomes in terms of highest discriminative power of the selected models was LNI, PGG, and SVI (AUC = 0.85) followed by ECE (AUC = 0.80) and LVI (AUC = 0.77).

Figure 4.9 provides an overview of the variables that were selected by the BIC and LASSO$_{1se}$. AIC and LASSO$_{max}$ selected a greater variety and more variables without visibly improvement regarding the AUC (Fig. 4.8). Therefore, we focus on BIC and LASSO$_{1se}$, but for reference Figure B.7 in Appendix B.2 displays covariates included in at least 4 out of 10 imputation sets using AIC and LASSO$_{max}$.

Overall, BIC and LASSO$_{1se}$ selected similar variables. The methods agreed in selecting primary and secondary biopsy Gleason grade for all outcomes in all imputation sets and per-

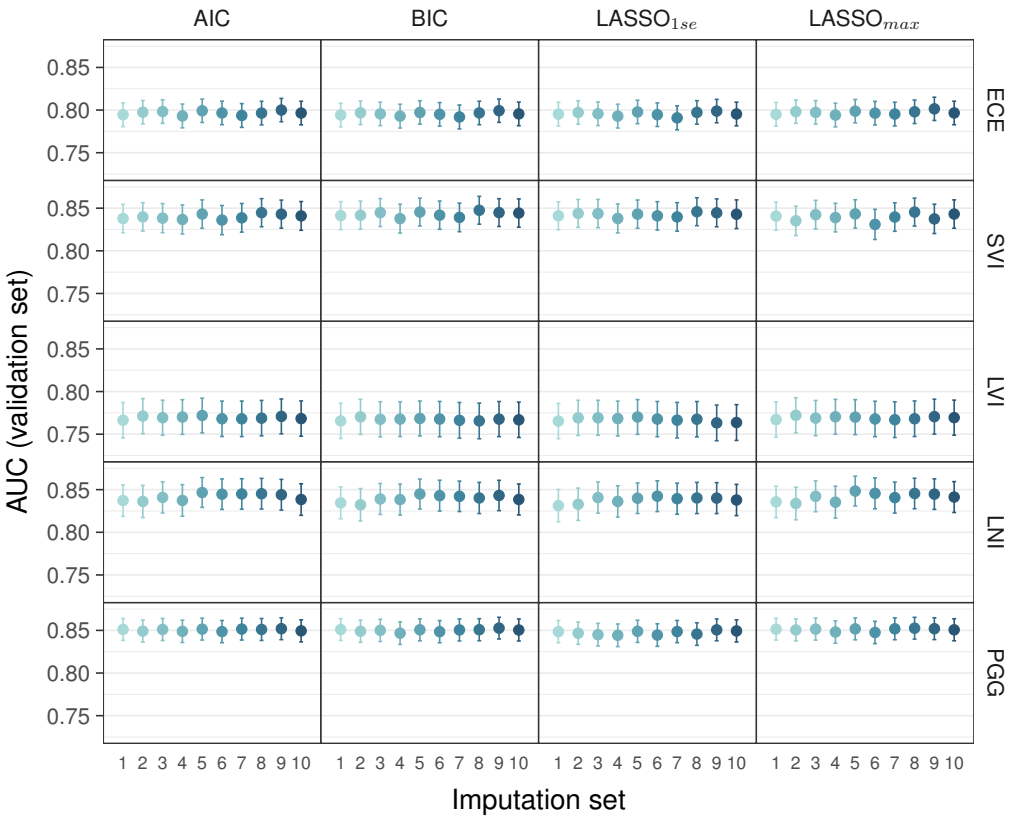

Figure 4.8 AUCs and 95%-CIs for selected models by each model selection algorithm for each imputation set ($n = 16,427$) validated on the corresponding imputation set of the validation data ($n = 4,180$).

centage of positive biopsy cores in the majority of all imputation sets for all outcomes. Further, both selection algorithms chose age, BMI, prostate volume, minimum log-transformed PSA ($\log(\text{PSA})$) within on year of prostatectomy, and the D'Amico PSAV cut-off in the majority of the imputation sets for the same outcomes. BIC and LASSO$_{1se}$ included mean $\log(\text{PSA})$ within the last year before prostatectomy and the number of positive cores for almost all pathological outcomes, but BIC did not select the first one for LVI and the latter one not for PGG in the majority of the imputation sets.

Both methods included the number biopsies for SVI in the majority of the imputation sets, and BIC selected it also for PGG, whereas LASSO$_{1se}$ chose it for ECE. BIC and LASSO$_{1se}$ selected the number of psa values within three years of prostatectomy for ECE and the last $\log(\text{PSA})$ value for LVI. LASSO$_{1se}$ included these covariates also for SVI and PGG, respectively.

BIC selected the standard deviation of all $\log(\text{PSA})$ values for SVI, standard deviation of all $\log(\text{PSA})$ values within three years of prostatectomy for PGG, and minimum $\log(\text{PSA})$ value within three years of prostatectomy for LNI. LASSO$_{1se}$ included the maximum $\log(\text{PSA})$ of

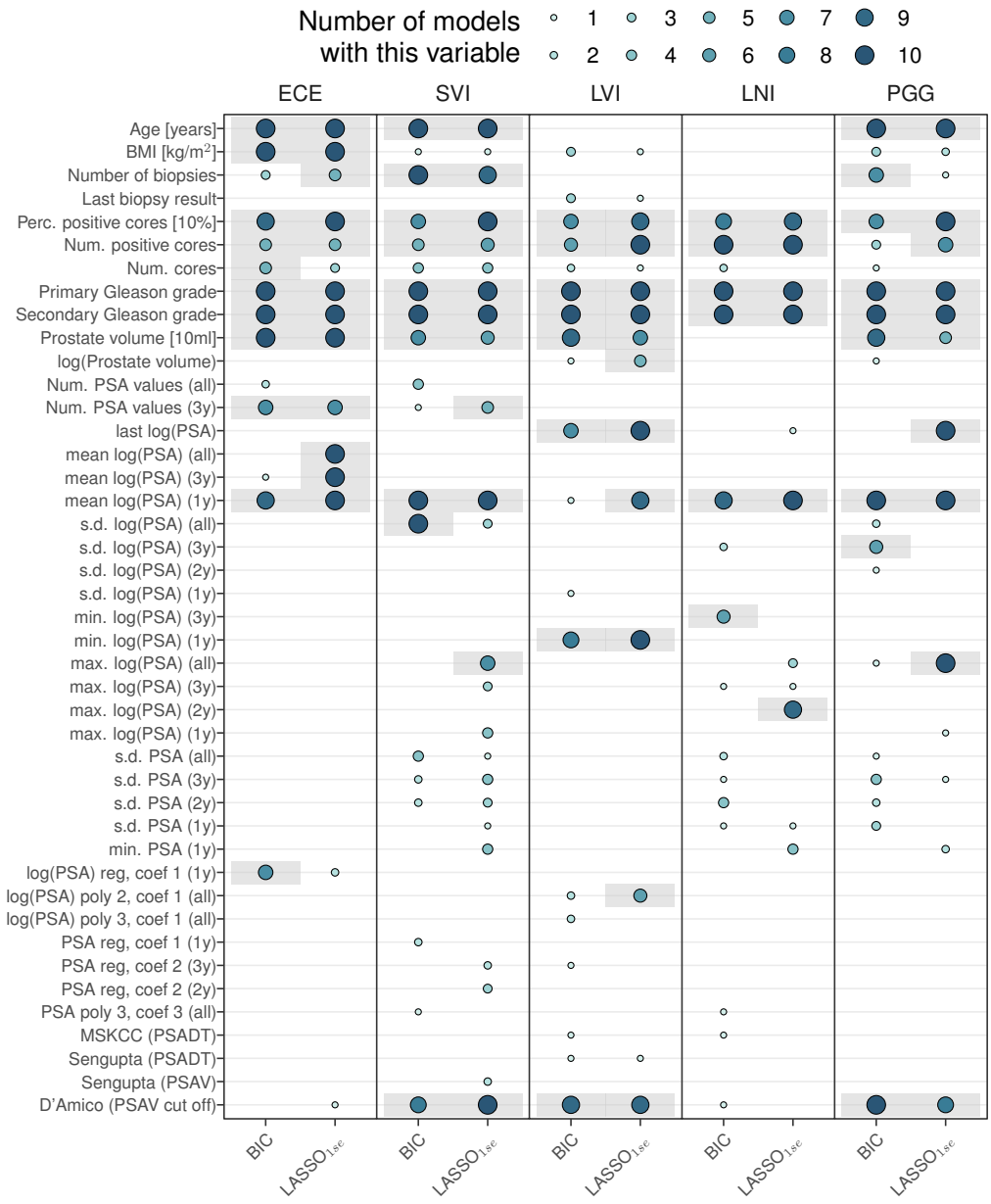

Figure 4.9 Variables selected in at least one of the 10 imputation sets ($n = 16,427$) using BIC or the LASSO$_{1se}$ method. Size and color indicate the number of imputation sets on which the specific variable was selected. Grey backgrounds indicate variables selected in at least 50% of the imputed sets. poly stands for polynomial and reg for linear regression with the corresponding degree and coefficients (coef), (1y), (2y), (3y), and (all) for the time frame in years before prostatectomy used for the calculation of the PSA related features.

all PSA values in the models for SVI and PGG in the majority of the imputation sets and the maximum $\log(\text{PSA})$ within two years before prostatectomy for LNI.

BIC included the number of biopsy cores and the intercept of a linear regression model on all $\log(\text{PSA})$ values within one year of prostatectomy for ECE in the models on the majority of the imputation sets, and LASSO$_{1se}$ included the mean of all $\log(\text{PSA})$ values

and of those within three years before prostatectomy for this outcome. LASSO$_{1se}$ chose the log-transformed prostate volume and the intercept of a polynomial regression on all $\log(\text{PSA})$ values before prostatectomy for LVI as covariates.

We use the majority rule as described by Wood et al. (2008) to select those variables that should be included in the final models for each selection procedure [251]. This implies that we select those variables that were included in at least 50% of the models on the imputation sets. Additionally we perform the likelihood ratio test for multiple imputed data sets introduced by Meng and Rubin (1992) to test whether variables selected in 4 of the 10 imputations should be included as well [251, 252]. This procedure yields one set of variables per selection algorithm for each pathological outcome and with those variables we fit models on each imputation set and pool the coefficients according to Rubin's rules to give

$$\bar{\beta} = \frac{1}{10} \sum_{k=1}^{10} \widehat{\beta}^{(k)}, \tag{4.29}$$

with $\widehat{\beta}^{(k)}$ the estimated coefficient vector of imputation set k [253, 254].

We evaluate the model performance for each model selection algorithm and each outcome on the 10 imputations of the validation set with 4,180 observations [251]. We calculate the AUC as well as the accuracy, precision, sensitivity, and specificity for a cut-off value of 0.5 as defined in Section 3.5 [255]. Then we use the average values to compare the different models in Figure 4.10 [124]. This pooled performance approach as described by Wood et al. (2015) assumes an ideal clinical setting in the future, where no variables are missing [124]. At this step, we evaluate the model selection algorithm and not the final model performance and thus the assumption of non-missing values is not very restrictive. For the final model evaluation we use a less optimistic model performance evaluation based on the hold-out test set [124].

We see no large differences between the model selection algorithms for each outcome, only that model performance varies between outcomes. While the AUC and accuracy at a cut-off of 0.5 range between 0.75 and 0.85, the precision, sensitivity and specificity have a larger range across outcomes. For SVI, LVI, and LNI the models have a high specificity above 0.95, lower sensitivity with less than 0.30, and a precision between 0.70 and 0.77. One explanation for the poorer performance might be the lower prevalence of SVI, LVI, and LNI as indicated by the dotted black lines in Figure 4.10. For ECE and PGG the specificity is lower, but this is accompanied by a higher precision and higher sensitivity. Thus these models can better predict true positive values, while the trade off is that more false positive values occur at the cut-off of 0.5.

In the above analysis, we used 0.5 as a cut-off value for accuracy, precision, sensitivity and specificity, whereas there might be more optimal and different values for each pathological outcome. Comparison between model selection algorithms is the main purpose of choosing a cut-off for these evaluation metrics in Figure 4.10. However, as we cannot detect a large difference between the algorithms and also for the cut-off independent AUC-value, we do not explore different values of for the cut-off.

We have seen that all models perform similarly and thus we proceed with the simpler models selected by the BIC and LASSO$_{1se}$ as the additional variables in models selected by the AIC and LASSO$_{max}$ did not provide a benefit in predictions (Fig. 4.10). For the simple models Figure 4.11 visualizes ORs and their significance at a 0.05-level for multiple imputed data as introduced by Rubin (1987) [253]. Figure B.8 similarly visualizes the AIC and LASSO$_{max}$ models and the non-significance of many covariates support the choice to proceed with the simpler models.

Overall, ORs for variables included by both selection algorithms are similar, but we detect some issues and variation that might be due to multicollinearity. In the following we remove those features and discuss the final model selection for each pathological outcome.

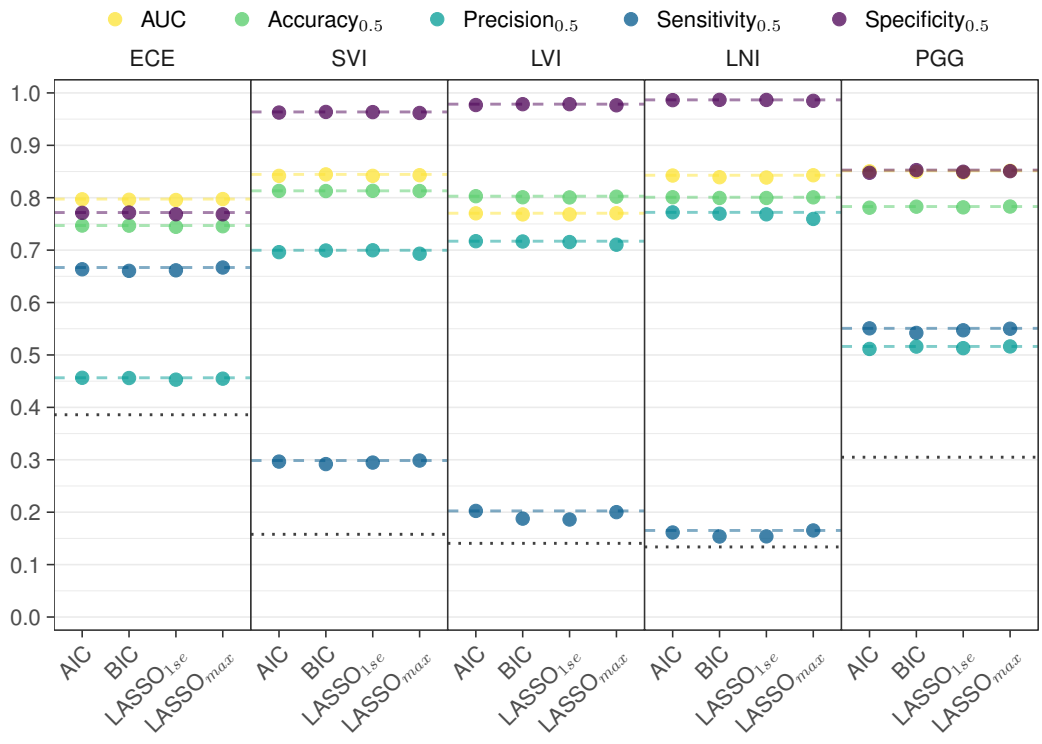

Figure 4.10 AUC and accuracy, precision, sensitivity, and specificity at a cut-off of 0.5 for pooled logistic regression models evaluated on the imputed validation sets ($n = 4,180$). The dashed lines indicate the maximum values for each evaluation metric and the dotted black lines the prevalances of each outcome in the validation set.

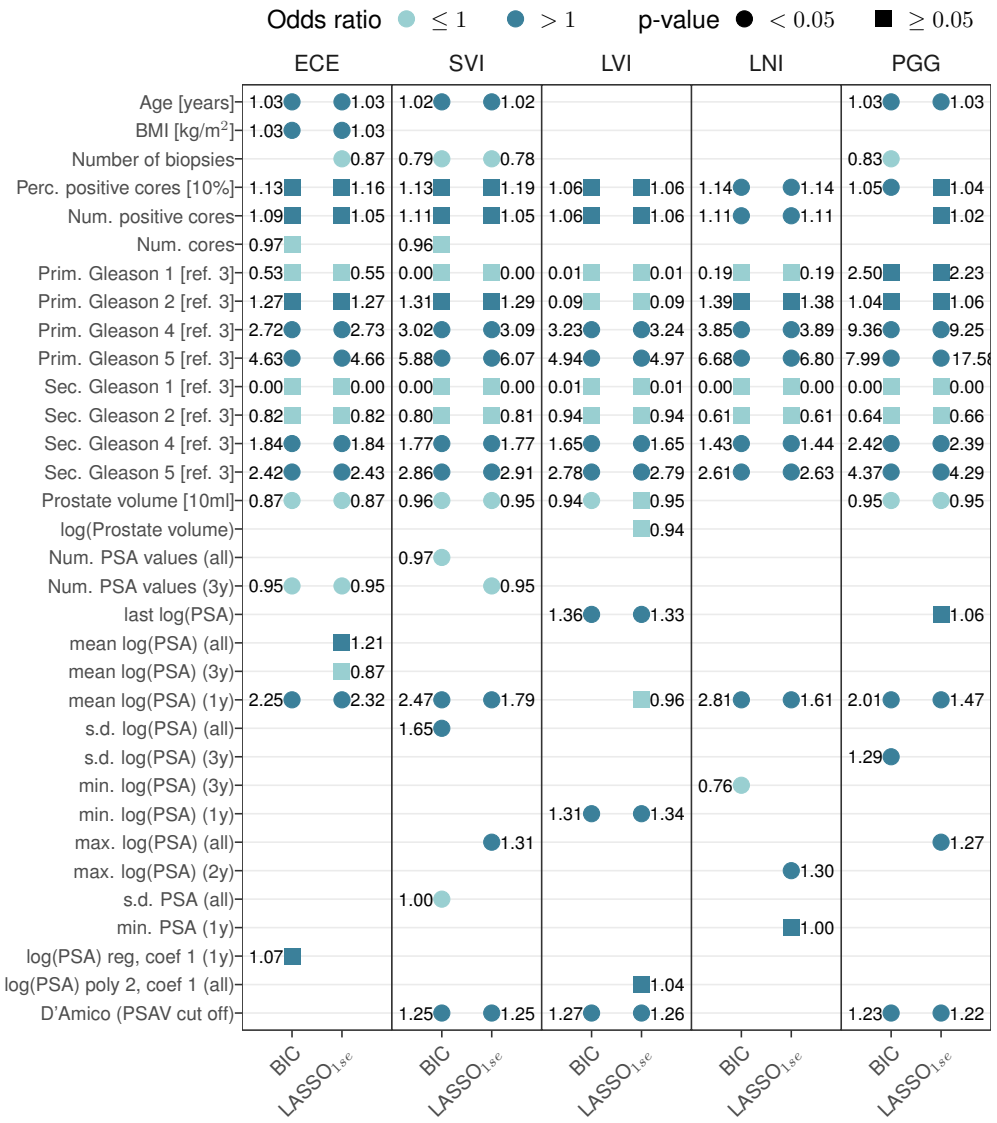

Figure 4.11 Odds ratios for models selected by the BIC and LASSO$_{1se}$, with magnitude greater than 1 indicated by dark blue and significance at the 0.05 level, by circles. poly stands for polynomial and reg for linear regression with the corresponding degree and coefficients (coef), (1y), (2y), (3y), and (all) for the time frame in years before prostatectomy used for the calculation of the PSA related features.

For all outcomes, the ORs for the primary and secondary Gleason grade are taken with respect to the reverence level 3. The coefficients for grades 1 and 2 are not significant. As we discussed in Section 2.2, these grades are not present in the data set after 2013, which is probably due to the change in the Gleason grading scheme in 2005 [52, 256]. In the following, we exclude patients with a primary or secondary Gleason grade less than 3 in the modeling to better reflect modern practice. Primary and secondary Gleason grade are included in the model for each pathological outcome.

For ECE we additionally include age, BMI, prostate volume, the number of PSA values within

three years, and the mean $\log(\text{PSA})$ within on year before prostatectomy as covariates in the model as these were selected by both methods. In addition, we include the number of biopsies, selected by LASSO$_{1se}$. LASSO$_{1se}$ included also the average $\log(\text{PSA})$ value for all values and those within three years before prostatectomy, however, the coefficients are not significant at a 0.05 level. For patients with one PSA measurement, these values are equal to the average $\log(\text{PSA})$ value within one year of prostatectomy and for patients with more than one value these variables are highly correlated as shown in Figures 2.12 and A.1. When included in a model these might cause multicollinearity issues as OR for one variable is smaller than one, whereas ORs for the other two are greater than one. We include only the mean $\log(\text{PSA})$ within one year of prostatectomy in the model for ECE. BIC selected the intercept of a linear regression model on all $\log(\text{PSA})$ values within one year of prostatectomy. Since this coefficient is not significant, not selected by the LASSO$_{1se}$ and does not provide a sensible correlation with the outcome, we do not choose it as a covariate. The covariates related to the number of positive biopsy cores are not significant. This is probably due to the high correlation among them as it is shown in Figures 2.12 and A.1. The percentage of positive cores summarizes the number of positive cores divided by the number of cores taken, thus we chose only this covariate and do not include the other in the model.

In the models for SVI, we detect similar issues as described for ECE regarding the number and percentage of positive cores. Again, we include the percentage of positive cores. Further, we include age, the number of biopsies, prostate volume, the mean $\log(\text{PSA})$ value within one year of prostatectomy, and PSAV cut off proposed by D'Amico et al. (2004) as covariates. The overall number of PSA values and those within three years before prostatectomy do not differ by a large extent, thus we chose the latter one as covariate. We include the standard deviation of all $\log(\text{PSA})$ values, but not the standard deviation of all PSA values as these two covariates are positively correlated and one OR is greater than one whereas the other one is smaller. This, again, hints at multicollinearity issues, which we aim to avoid. The maximum $\log(\text{PSA})$ and the mean $\log(\text{PSA})$ are highly correlated and therefore, we only include the first one as covariate in the model for SVI.

For LVI we include the percentage of positive cores in the model, for the same reasons as described above. For the other covariates, we choose those selected by the BIC, as Figure 4.10 shows that this is comparable to the model selected by the LASSO$_{1se}$, but the latter one includes more correlated variables, such as the log-transformed prostate volume in addition to the prostate volume and the mean $\log(\text{PSA})$ within one year of prostatectomy in addition to the last $\log(\text{PSA})$ value. Thus, the covariates included in the model for the LVI are the last $\log(\text{PSA})$ value before prostatectomy, the minimum PSA value within one year of

prostatectomy and the D'Amico PSAV cut-off in addition to the percentage of positive cores and the primary and secondary Gleason grade.

For LNI and PGG we select the BIC models as all variables are significant, apart from the primary and secondary Gleason grade 1 and 2. Models selected by the LASSO_{1se} have similar issues with correlated variables that are not significant or yield different directions for the OR values as we already described for ECE, SVI, and LVI.

We fit the selected models to the combined imputation sets of the training and validation data but exclude patients with a primary or secondary Gleason grade less than 3. Thus, we can use a total of 19,967 to 19,989 samples. The variation is due to the different imputations for some patients. We pool the regression coefficients according to Rubin's rules (4.29) and calculate 95% CI as well as the significance level for multiple imputed data sets [253]. Table 4.2 summarizes the results for each pathological outcome.

Across all outcomes, a higher percentage of positive cores and a higher primary and higher secondary Gleason grade at biopsy were associated with an increased risk of an adverse pathological outcome after prostatectomy. A higher number of positive cores at biopsy was correlated with a higher risk of LNI.

Increased risk of ECE, SVI, and PGG was correlated with higher age at prostatectomy and for ECE with higher BMI. For ECE, SVI, LNI, and PGG a higher mean $\log(\text{PSA})$ within the last year before prostatectomy was correlated with a higher risk of these outcomes, whereas for LVI the last $\log(\text{PSA})$ before prostatectomy was selected for the model and also associated with a higher risk. It is important to note that this mean $\log(\text{PSA})$ and the last $\log(\text{PSA})$ are identical for patients with only one PSA measurement within the last year before prostatectomy. However, for patients with more than one PSA-value, the mean within the last year of prostatectomy might be a good summary statistic. Increased risk of SVI and PGG was associated with a higher standard deviation of $\log(\text{PSA})$, where all measurements before prostatectomy were used for SVI and only the measurements within the last 3 years before prostatectomy for PGG. Further, for these pathological outcomes and LVI a PSAV above the discrete cut-off of 2 ng/ml/yr as proposed by D'Amico et al. (2004) was associated with a higher risk.

Increased risk of LVI was associated with a higher minimum $\log(\text{PSA})$ within 1 year of prostatectomy. For patients with only one PSA measurement this value is identical to the last $\log(\text{PSA})$ and thus these two features result in a total OR of 1.80 for a unit increase in the single $\log(\text{PSA})$ value. However, for patients with multiple PSA measurements, a decline for $\log(\text{PSA})$ within the last year before prostatectomy was associated with a higher risk of LVI compared to patients with an increase in $\log(\text{PSA})$, given that they had the same last

	ECE	SVI	LVI	LNI	PGG
Age [years]	1.04*** 1.03 - 1.04	1.02*** 1.02 - 1.03			1.03*** 1.02 - 1.03
BMI [kg/m^2]	1.02*** 1.01 - 1.04				
Number of biopsies	0.85*** 0.77 - 0.93	0.78** 0.67 - 0.91			0.82*** 0.74 - 0.92
Perc. positive cores [10%]	1.19** 1.08 - 1.31	1.24*** 1.12 - 1.36	1.10*** 1.05 - 1.16	1.14* 1.03 - 1.27	1.06*** 1.03 - 1.09
Number of positive cores				1.11* 1.02 - 1.21	
Prim. Gleason grade 4 [ref. 3]	2.76*** 2.36 - 3.24	3.13*** 2.58 - 3.80	3.19*** 2.78 - 3.66	3.79*** 3.31 - 4.35	8.94*** 8.10 - 9.86
Prim. Gleason grade 5 [ref. 3]	4.89*** 3.54 - 6.75	6.48*** 4.97 - 8.44	5.05*** 4.00 - 6.37	7.52*** 5.80 - 9.73	19.40*** 13.70 - 27.48
Sec. Gleason grade 4 [ref. 3]	1.82*** 1.55 - 2.14	1.72*** 1.43 - 2.06	1.62*** 1.40 - 1.86	1.48*** 1.28 - 1.71	2.33*** 2.12 - 2.57
Sec. Gleason grade 5 [ref. 3]	2.41*** 1.92 - 3.03	2.50*** 2.02 - 3.10	2.33*** 1.93 - 2.81	2.42*** 2.00 - 2.92	4.24*** 3.50 - 5.14
Prostate volume [10ml]	0.87*** 0.85 - 0.90	0.96* 0.92 - 1.00	0.93*** 0.90 - 0.96		0.94*** 0.91 - 0.96
Number of PSA values (3y)	0.96** 0.93 - 0.99	0.95** 0.91 - 0.99			
last log(PSA)			1.46*** 1.26 - 1.70		
mean log(PSA) (1y)	2.47*** 2.29 - 2.66	2.31*** 2.10 - 2.55		3.04*** 2.60 - 3.56	2.07*** 1.94 - 2.20
s.d. log(PSA) (all)		1.57*** 1.38 - 1.78			
s.d. log(PSA) (3y)					1.52*** 1.35 - 1.72
min. log(PSA) (3y)				0.70*** 0.62 - 0.79	
min. log(PSA) (1y)			1.23** 1.07 - 1.42		
D'Amico (PSAV cut off)		1.25** 1.10 - 1.42	1.24*** 1.11 - 1.39		1.36*** 1.24 - 1.49

Table 4.2 Odds ratios with 95% confidence intervals of selected final multivariable models fitted on complete training data, excluding patients with primary or secondary Gleason grade < 3 ($n = 19,967$ to $19,989$). $***$ indicates a p-value < 0.001, $** < 0.01$, and $* < 0.05$. (1y), (3y), and (all) denote the time frame in years before prostatectomy for calculating PSA related features.

$\log(\text{PSA})$ before prostatectomy. This association might be due to patients who underwent a treatment such as chemotherapy or radiation before prostatectomy that can lead to a decrease in PSA. Such treatments are probably more likely described for patients with higher aggressive prostate cancer and thus might be an important factor that should be included in the models. Unfortunately, information about treatments before prostatectomy were not available.

Variables associated with a decreased risk of an adverse pathological outcome were an increased prostate volume at biopsy for ECE, SVI, LVI and PGG and higher minimum $\log(\text{PSA})$ within 3 years of prostatectomy for LNI. Similar to LVI, including the minimum $\log(\text{PSA})$ within 3 years before prostatectomy in the model for LNI is relevant for patients with more than one PSA measurement.

A higher total number of biopsies for ECE, SVI, and PGG, or a higher number of PSA measurements within 3 years before prostatectomy for ECE and SVI were associated with a decreased risk for the outcomes. These two variables might serve as an indication for patients on AS or patients who undergo more prevention procedures. Here we assume that the design of the models would highly benefit from the additional information of when the patients were diagnosed with prostate cancer and whether they were under AS. Unfortunately, this information was not available in the EMR data.

In the PGG model, the primary Gleason grade dominates the potential risk for those patients with a grade higher than 3. This is not particularly surprising, as one would expect that in general when a primary Gleason score higher than 3 is detected during biopsy it is very likely that this is confirmed after prostatectomy. Although there are some cases of downgrading for Gleason grades, the risk for PGG is expected to be high in such cases. Apart from these trivial insights, the model for PGG might provide interesting associations for patients with a primary Gleason grade of 3.

The final model performance is assessed using the hold-out test set. For evaluation, we use the pooled performance following Wood et al. (2015), where we calculate the predictions for each pathological outcome on each imputation set and then average the probabilities to obtain one prediction for each outcome [124]. In Figure 4.12 we evaluate the predictions using ROC-curves with AUC values and 95% CIs. On the hold-out test set, predictions of PGG and SVI yield the highest AUC with 0.86 and thus a higher value than on the validation set (AUC = 0.85). For LNI we obtain an AUC of 0.83 and slightly lower compared to the AUC on the validation set (AUC = 0.85). Predictions for ECE yield the same AUC value as on the validation set with 0.80 and for LVI a slightly higher value with 0.78 (AUC = 0.77).

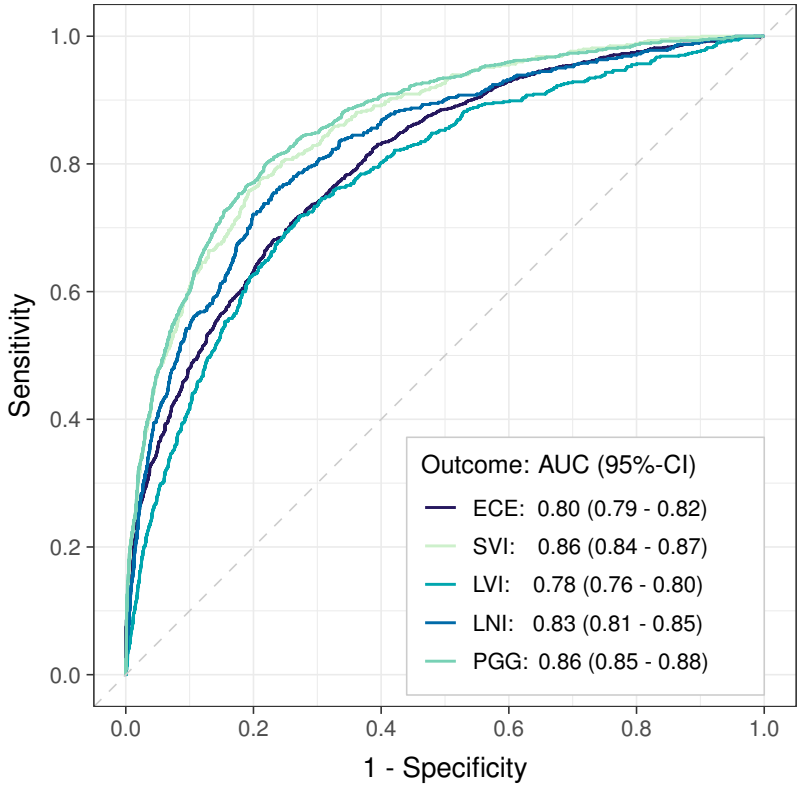

Figure 4.12 Receiver operating characteristic curves with AUC values and 95% confidence intervals for predictions on the hold-out test set $n = 3,515$ for each pathological outcome.

In addition to the ROC curves and AUC values, we use calibration plots to more precisely identify areas where the models do not fit well. For each pathological outcome, we divide the predicted risk into groups by decile and compare the mean predicted risk with the percentage of observed outcomes [257]. For further illustration, we add the distribution of predicted risk to the calibration plots.

Figure 4.13 shows the models for the pathological outcomes are well-calibrated overall as the slope of the calibration graphs is close to one. For ECE the model slightly underestimates the risk for patients at higher risk with a predicted risk between 60% and 80%. The models for LVI and LNI could benefit from recalibration as they overestimate the risk of an adverse pathological outcome for patients around 20% predicted risk and they underestimate risk for high-risk patients around 70% predicted risk. However, for both pathological outcomes comparably few patients have a high predicted risk and overall the prevalence of these outcomes is low.

We see in Table 4.2 that the number of positive cores and the percentage of positive cores are borderline significant for LNI. To test whether we should only include the percentage of positive cores in the model we use the theory developed in Section 4.1 to compare two

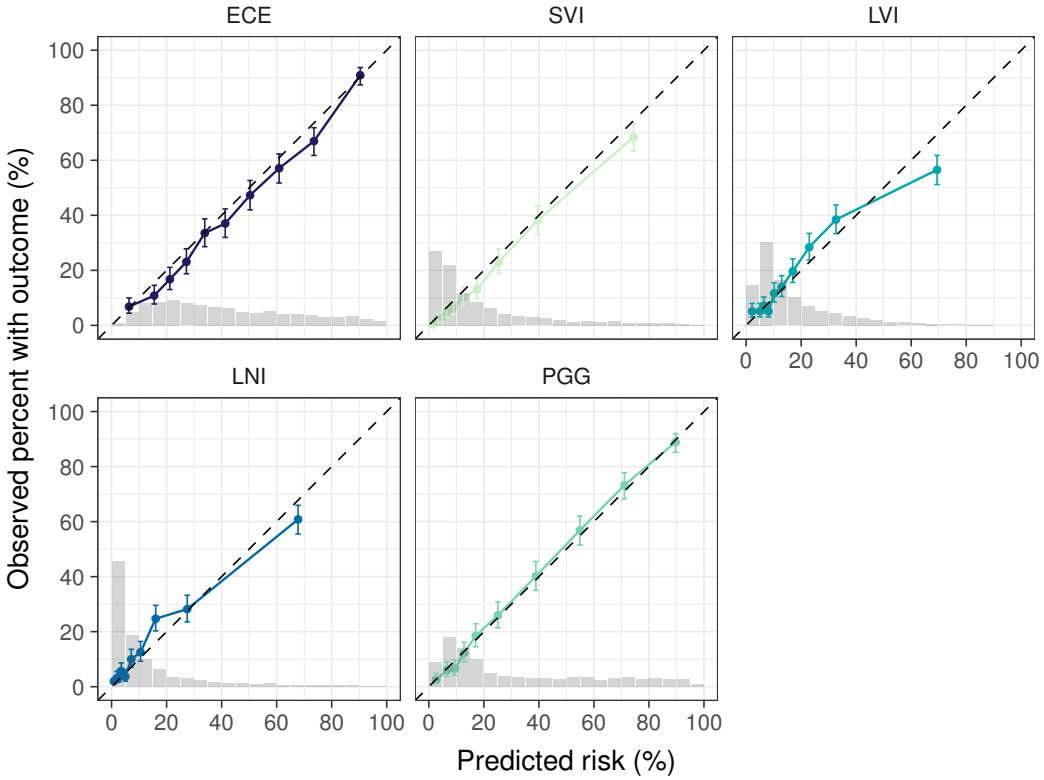

Figure 4.13 Calibration curves for predictions on the hold-out test set $n = 3,515$ for each pathological outcome. The gray bars visualize the distribution of patients across the range of predicted risk.

nested models for LNI, one including the number of positive cores as a variable (M_1) and one without it (M_0). For illustration, we only consider patients with no missing value for LNI or in any of the covariates included in M_1 from the complete training data, which is train and validation set combined. We can use 11,955 out of 20,607 patients (58.0%).

A summary of M_0 and M_1 is given in Table 4.3. The Schwarz criterion is given by

$$S = \ell_{M_0} - \ell_{M_1} + \frac{1}{2}\log(11,955) = -8.86$$

and this yields an approximation to BF of

$$\text{BF}_{\text{approx}} = \exp(S) = 1.42 \times 10^{-4},$$

which evaluates to 3.85 on a negative $\log 10$-scale. According to Jeffreys' rule this provides decisive evidence against M_0 for M_1.

An 'exact' numeric integration of the likelihood for M_0 and M_1 is infeasible, as the dimensions 8 and 9 of the respective parameter spaces are too large for standard methods. Therefore,

	M_0		M_1	
	OR	p-value	OR	p-value
(Intercept)	0.002	< 0.001	0.002	< 0.001
Perc. positive cores [10%]	1.274	< 0.001	1.186	< 0.001
Prim. Gleason grade 4 [ref. 3]	3.924	< 0.001	3.833	< 0.001
Prim. Gleason grade 5 [ref. 3]	8.616	< 0.001	8.282	< 0.001
Sec. Gleason grade 4 [ref. 3]	1.501	< 0.001	1.472	< 0.001
Sec. Gleason grade 5 [ref. 3]	2.337	< 0.001	2.262	< 0.001
mean log(PSA) (1y)	2.816	< 0.001	2.795	< 0.001
min. log(PSA) (3y)	0.706	< 0.001	0.713	< 0.001
Num. positive cores	-	-	1.090	< 0.001

Table 4.3 M_0 and M_1 for LNI based on 11,955 patients of the training and validation set with no missing values for LNI or any of the listed variables.

we use Stan for MCMC sampling to obtain samples from the posterior distributions [258]. Stan provides Hamiltonian Monte Carlo (HMC) sampling, an efficient adaptive MCMC method introduced by Duane et al. (1987) and Neal (1994) and the No-U-Turn sampler (NUTS) developed by Hoffman and Gelman (2014) [258–262]. With bridge sampling we obtain the marginal likelihood needed for the BF calculation [262, 263]. For implementation in `R` we use the `rstan` package and the `bridgesampling` package [264, 265].

As the sampling procedure is computationally intense, we use RStudio Server with the `R` version 3.6.0 running on a Linux distribution. We can use 8 cores in parallel and therefore choose to estimate 8 chains for each model with 250, 500, 1000, and 2000 iterations. A smaller number of iterations resulted in convergence issues. For the warm-up, we use half of the number of iterations and thus retrieve 1000, 2000, 4000, and 8000 samples respectively.

The Stan model specification for M_0 and M_1 are given in Appendix B.3. We use independent flat normal prior distributions $\mathcal{N}(0, \sigma^2)$, with $\sigma = 100$ for the parameters shared by M_0 and M_1 and the unit information prior $\psi_{\text{Num pos cores}}|\theta \sim \mathcal{N}(0, I_{\psi\psi}(\theta, 0)^{-1})$ on the coefficient for the number of positive cores.

The results for the BF estimation with different number of iterations are summarized in Table 4.4. 250 iterations seem to be sufficient in this case as we do not detect changes in the estimation with a larger number of iterations. The resulting BF is 2.03×10^{-5} and lower compared to the $\text{BF}_{\text{approx}}$ with 1.42×10^{-4}, agreeing with the observation in the simulation study in Section 4.2 that the approximation is more conservative. On a negative $\log 10$-scale

both BFs yield the same conclusion according to Jeffreys' rule with decisive evidence against M_0.

The results are comparable, but while the calculation of BF_{approx} is immediately possible using standard output of `R` the HMC takes significantly longer.

Chains	Iterations	Samples	Time	BF	$-\log 10(\text{BF})$	Jeffrey's rule
8	250	1000	4 min 35 sec	2.03×10^{-5}	4.69	decisive
8	500	2000	8 min 21 sec	2.00×10^{-5}	4.70	decisive
8	1000	4000	10 min 55 sec	2.03×10^{-5}	4.69	decisive
8	2000	8000	21 min 17 sec	2.03×10^{-5}	4.69	decisive

Table 4.4 BF results for M_0 versus M_1 using HMC and bridge sampling on 8 parallel cores on RStudio Server with different number of iterations.

5 Extension to multivariate logistic regression models

In the previous chapter, we modeled the adverse pathological outcomes separately with independent univariate logistic regression models. Thereby, we ignored the correlation among the outcomes that we detected in Section 2.2. In this chapter, we extend the approximation to the BF to the multivariate case.

5.1 Mathematical derivation

We consider n independent observations $Y_i \in \mathbb{R}^q$, $i = 1, \ldots, n$ of q potentially correlated binary outcomes. The multivariate logistic regression model introduced in Section 3.2 is given by $Y_{ij} = \mathbb{1}(Z_{ij} > 0)$ and $Z_i \sim \mathbb{L}_q(\mu_i, R)$ with the scale matrix R having 1's on the diagonal, mean vectors $\mu_i = (X_{i1}'\beta_1, \ldots, X_{iq}'\beta_q)$, corresponding covariates $X_{ij} \in \mathbb{R}^{m_j}$, and coefficient vectors $\beta_j \in R^{m_j}$ for $j = 1, \ldots, q$. The likelihood for $\beta = \left(\beta_{11}, \ldots, \beta_{1m_1}, \ldots, \beta_{q1}, \ldots, \beta_{qm_q}\right)' \in \mathbb{R}^m$, $m = \sum_{j=1}^q m_j$, is given by

$$\mathcal{L}(\beta, R) = \prod_{i=1}^n \int_{A_{i1}} \cdots \int_{A_{iq}} l_q(z_i|\mu_i, R) dz_i,$$

where l_q denotes the multivariate logistic density as defined in (3.6) and

$$A_{ij} = \begin{cases} (0, \infty) & \text{for} \quad y_{ij} = 1 \\ (-\infty, 0] & \text{for} \quad y_{ij} = 0. \end{cases}$$

Then the log-likelihood is

$$\ell(\beta, R) = \sum_{i=1}^n \log \left(\int_{A_{i1}} \cdots \int_{A_{iq}} l_q(z_i|\mu_i, R) dz_i \right).$$

Analogous to Section 4.1 we consider $\beta = (\theta, \psi) \in \Theta \times \Psi$ with $\dim(\Theta) = m_0$ and $\dim(\Psi) = m - m_0$, and are interested in testing the hypotheses

$$H_0 : \psi = \psi_0 \quad \text{versus} \quad H_A : \psi \neq \psi_0.$$

We might need to reorder vector β such that it matches the desired hypothesis in practice. Note that we are able to test several coefficients for different outcomes simultaneously with the appropriate reordering of β. Further, we consider prior distributions $f_0(\theta)$ and $f(\theta, \psi)$ under H_0 and H_A, respectively. Analogous to the univariate logistic regression, the BF is given by

$$\text{BF} = \frac{\int \mathcal{L}(\theta, \psi_0) f_0(\theta) d\theta}{\iint \mathcal{L}(\theta, \psi) f(\theta, \psi) d\theta d\psi} \,. \tag{5.1}$$

Definition 4.1 for the Laplace regularity of a sequence of log-likelihood functions can be applied to the multivariate logistic regression model as well. Part (i) is fulfilled since the log-likelihood is a sum of the composition of smooth functions and thus sufficiently smooth, and analogous to the univariate case, we assume that the ORs are bounded in case the MLE exists.

Further, we assume that the Hessian matrix of the negative log-likelihood at the MLE is positive definite and thus the determinant bounded away from zero. With the existence of the MLE we can assume that the Hessian matrix is positive semidefinite, but we ensure the positive definiteness with Part (ii) of Definition 4.1.

Again, we would like to ensure that the MLE does not lie on the boundary of the parameter space and that it approximates the global maximum of the log-likelihood, which is fulfilled with part (iii) of Definition 4.1.

Let $\ell_0(\theta) = \log(\mathcal{L}(\theta, \psi_0))$ and $\ell(\theta, \psi) = \log(\mathcal{L}(\theta, \psi))$ denote the log-likelihoods for the null and alternative hypotheses. Suppose that Assumption 4.1 holds for ℓ_0 and ℓ as well as the prior distributions $f_0(\theta)$ and $f(\theta, \psi)$ under H_0 and H_A, respectively. Assume that the integrals in (5.1) are finite. Then, Proposition 4.1 can be applied in the multivariate case and we can approximate the BF with

$$\text{BF} = (2\pi)^{\frac{m_0 - m}{2}} \frac{\det((-D^2 \ell_0(\widehat{\theta}_0))^{-1})^{\frac{1}{2}} \exp(\ell_0(\widehat{\theta}_0)) f_0(\widehat{\theta}_0)}{\det((-D^2 \ell(\widehat{\theta}, \widehat{\psi}))^{-1})^{\frac{1}{2}} \exp(\ell(\widehat{\theta}, \widehat{\psi})) f(\widehat{\theta}, \widehat{\psi})} \left(1 + O(n^{-1})\right) .$$

We apply the same concept of null orthogonality as defined in Definition 4.2 to the expected Fisher information matrix of the multivariate log-likelihood. In Proposition 4.2 we proved that we can construct null orthogonal parameters from a non-null orthogonal parametrization. The proof does not depend on the explicit form of the Fisher information matrix and holds for the multivariate log-likelihood as well.

We suppose that Assumption 4.2 holds for θ, ψ, ψ_0, the prior densities $f_0(\theta)$ and $f(\theta, \psi)$, and the log-likelihood. Thus, we assume that θ and ψ are null orthogonal, the marginal

prior for θ is the same under both hypotheses, the matrix of second derivatives converges asymptotically to the Fisher information matrix, and the MLE $\widehat{\psi}$ approximates ψ_0 with an order of $O_p(n^{-\frac{1}{2}})$. Analogous to the univariate logistic regression models, Assumption 4.2 does not imply an extensive restriction for the likelihood and parameters.

Under the assumptions of Proposition 4.3 the approximation to the BF is

$$\text{BF} = \left(\frac{2\pi}{n}\right)^{\frac{m_0-m}{2}} \det\left(I_{\psi\psi}(\widehat{\theta},\psi_0)\right)^{\frac{1}{2}} \frac{\exp(\ell_0(\widehat{\theta}_0))}{\exp(\ell(\widehat{\theta},\widehat{\psi}))} \frac{1}{f_{\psi|\theta}(\widehat{\psi}|\widehat{\theta})} \left(1 + O_p\left(n^{-\frac{1}{2}}\right)\right).$$

The proof extends from the univariate case, but we need to show that $D^2\ell(\widehat{\theta},\widehat{\psi}) = O_p(n)$, $D^2\ell_0(\widehat{\theta}_0) = O_p(n)$, $I(\theta,\psi) = O_p(1)$, $\frac{1}{n}D^2\ell_0(\widehat{\theta}_0) = \frac{1}{n}D^2_{\theta\theta}\ell(\widehat{\theta},\psi_0) + O_p\left(n^{-\frac{1}{2}}\right)$, and $\frac{1}{n}D^2\ell(\widehat{\theta},\widehat{\psi}) = \frac{1}{n}D^2\ell(\widehat{\theta},\psi_0) + O_p\left(n^{-\frac{1}{2}}\right)$, without using the explicit form of the Hessian matrix as in Section 4.1.

The first two expressions follow from the assumptions of Laplace regularity for ℓ and ℓ_0. This implies that any partial derivative up to the 6th derivative of $-\frac{1}{n}\ell$ and $-\frac{1}{n}\ell_0$ in an open ball around the MLE is bounded and thus these second derivatives are $O_p(1)$, which shows that $D^2\ell(\widehat{\theta},\widehat{\psi}) = O_p(n)$ and $D^2\ell_0(\widehat{\theta}_0) = O_p(n)$. Then, we can show $\widehat{\theta} - \widehat{\theta}_0 = O_p\left(n^{-\frac{1}{2}}\right)$ analogous to the univariate case. The Fisher information matrix $I(\theta,\psi)$ is the expected value of the negative Hessian matrix for a single observation. Thus, $I(\theta,\psi) = O_p(1)$.

Using a Taylor expansion for $d^{(0)}_{kj} = \left(D^2\ell_0(\widehat{\theta}_0)\right)_{kj}$ around $\widehat{\theta}$ we obtain with $\widehat{\theta}-\widehat{\theta}_0 = O_p\left(n^{-\frac{1}{2}}\right)$ and the assumptions of Laplace regularity for the partial derivatives of ℓ

$$\begin{aligned} d^{(0)}_{kj} &= \frac{\partial^2\ell(\widehat{\theta}_0,\psi_0)}{\partial\theta_k\partial\theta_j} = \frac{\partial^2\ell(\widehat{\theta},\psi_0)}{\partial\theta_k\partial\theta_j} + \sum_{i=1}^{m_0}(\widehat{\theta}_{0i}-\widehat{\theta}_i)\frac{\partial^3\ell(\widehat{\theta},\psi_0)}{\partial\theta_k\partial\theta_j\partial\theta_i} + o_p\left(||\widehat{\theta}_0-\widehat{\theta}||\right) \\ &= \frac{\partial^2\ell(\widehat{\theta},\psi_0)}{\partial\theta_k\partial\theta_j} + m_0 O_p\left(n^{-\frac{1}{2}}\right) O_p(n) + O_p\left(n^{-\frac{1}{2}}\right) = \frac{\partial^2\ell(\widehat{\theta},\psi_0)}{\partial\theta_k\partial\theta_j} + O_p\left(n^{\frac{1}{2}}\right). \end{aligned}$$

Thus, $\frac{1}{n}D^2\ell_0(\widehat{\theta}_0) = \frac{1}{n}D^2_{\theta\theta}\ell(\widehat{\theta},\psi_0) + O_p\left(n^{-\frac{1}{2}}\right)$. Arguing similarly with a Taylor expansion for $d_{kj} = \left(D^2\ell(\widehat{\theta},\widehat{\psi})\right)_{kj}$ around ψ_0 and using $\widehat{\psi} - \psi_0 = O_p\left(n^{-\frac{1}{2}}\right)$, we have $\frac{1}{n}D^2\ell(\widehat{\theta},\widehat{\psi}) = \frac{1}{n}D^2\ell(\widehat{\theta},\psi_0) + O_p\left(n^{-\frac{1}{2}}\right)$. With this, the remaining proof follows analogous to the one given for Proposition 4.3 in Section 4.1.

Without additional restrictions we can use the Schwarz criterion defined in Definition 4.3 as approximation to the $\log$(BF) with

$$\log(\text{BF}) = S + O_p(1),$$

where again the proof extends from the univariate to the multivariate case. Analogous to Theorem 4.1 we obtain with a unit information prior distribution

$$\psi|\theta \sim \mathcal{N}_{m-m_0}(\psi_0, I_{\psi\psi}(\theta, \psi_0)^{-1})$$

the approximation

$$\log(\mathsf{BF}) = S + O_p\left(n^{-\frac{1}{2}}\right).$$

We note that for the Schwarz criterion we need the difference in the number of parameters between the model for the null hypothesis versus the model for the alternative hypothesis. This is again the dimension of ψ, however, thereby we assume that we model the same number of outcomes, thus the dimension of the scale matrix is $q \times q$ for both models.

5.2 Implementation in R and C++

In the following, we discuss details of the implementation of the MCMC Algorithm 3.1 for the Bayesian multivariate logistic regression model in `R`. We implement main parts of the algorithm in C++ using the Armadillo library for linear algebra as this tremendously decreases the time of computation [266]. Thereby, we rely on the `Rcpp` and `RcppArmadillo` packages to make functions available in `R` [267, 268].

For the implementation of Algorithm 3.1, we need to sample from the truncated multivariate normal distribution denoted by $\mathcal{TN}_q(\mu, \Sigma, D, a, b)$ with mean vector $\mu \in \mathbb{R}^q$, covariance matrix $\Sigma \in \mathbb{R}^{q\times q}$, linear transformation matrix $D \in \mathbb{R}^{k\times q}$ with rank $k \leq q$, lower bound vector $a \in \mathbb{R}^k$, and upper bound vector $b \in \mathbb{R}^k$ [269]. A q-dimensional random vector $X \sim \mathcal{TN}_q(\mu, \Sigma, D, a, b)$ has the probability density function (pdf)

$$f(x) = \frac{\exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)}{\int_{a\leq Dx\leq b}\exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)dx} \cdot \mathbb{1}\left(a \leq Dx \leq b\right),$$

where the inequality $a \leq Dx \leq b$ holds element-wise, $a_i \leq (Dx)_i \leq b_i$, $i = 1, \ldots, q$ [269].

In the following, we compare four algorithms for sampling from the truncated multivariate normal distribution. Thereby, we restrict the comparison to the relevant cases for Algorithm 3.1, where $D = \mathrm{Id}_q$ and $a, b \in \{-\infty, 0, \infty\}$.

The first sampling algorithm uses the multivariate normal distribution $\mathcal{N}_q(\mu, \Sigma)$ as proposal density in an acceptance-rejection algorithm. While, the acceptance-rejection algorithm performs well for $D\mu \in (a, b)$ or $D\mu$ close to (a, b), it has a low acceptance rate otherwise. This algorithm is easy to implement in `R` or C++ as sampling from the multivariate normal

distribution is available in the R package mvtnorm and the C++ library Armadillo [270]. The tmvnorm package provides the implementation in R and we translate it to C++ [271].

Geweke (1991) proposed a Gibbs sampler using the property that the conditional distribution of $X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_q$ is truncated univariate normal distributed [272]. For the sampling from a truncated univariate normal distribution, he used a mixed rejection algorithm based on normal, half-normal, uniform and exponential rejection sampling where the choice of algorithm depends on the values for the lower and upper bound [272]. Li and Ghosh (2015) criticized the poor mixing properties of the Gibbs sampler for different sampling parameters and regions [269]. They proposed an improvement of the mixed rejection sampling algorithm for the truncated univariate distribution that incorporates an approach by Robert (1995), while using the same rejection samplers as Geweke (1991) [269, 273]. Further, with an alternative transformation for the truncated multivariate normal distribution, the algorithm presented better mixing properties even for difficult sampling regions [269]. The tmvmixnorm package provides the R implementation of the Gibbs sampler proposed by Li and Ghosh (2015) and the tmvn package available on GitHub offers an implementation in C++ using the Rcpp and RcppArmadillo packages for the connection to R [274, 275]. We adapt the latter approach for our implementation.

Botev (2017) proposed a modified acceptance-rejection algorithm using minimax tilting [276]. A central step of this algorithm solves a system of non-linear equations to find the optimal tilting parameter. The sampler does not rely on a MCMC approach and thus does not need to converge to a stationary distribution and therefore no burn-in samples as for the Gibbs sampler are required. While the algorithm is available in R through the TruncatedNormal package that uses the non-linear equation solver provided in nleqslv, there exists no equivalent implementation in C++ [277, 278]. Parts of the TruncatedNormal package already rely on C++ and Rcpp, but the implementation of the solver for the system of non-linear equations is non-trivial in C++. The package also provides an estimation of the acceptance probability for the basic acceptance-rejection sampler.

Koch and Bopp (2019) proposed in a preprint available on ArXiv a direct sampling algorithm for sampling from the truncated multivariate normal distribution under box constraints [279]. Unlike the other approaches, this algorithm requires $D = \mathrm{Id}_q$, which is fulfilled in Algorithm 3.1. The algorithm exploits the structure of the Cholesky decomposition $\Sigma = LL'$, where L is a lower triangular matrix. A C++ implementation with Rcpp is available on GitHub [280].

Before we compare the time each algorithm takes for sampling from the truncated normal

distribution, we assess whether their results are similar. We sample 10,000 times from a

$$\mathcal{TN}_3\left(\begin{pmatrix}1\\-5\\2\end{pmatrix}, \begin{pmatrix}1 & 0.3 & -0.6\\0.3 & 1 & -0.1\\0.6 & -0.1 & 1\end{pmatrix}, \mathrm{Id}_q, \begin{pmatrix}0\\0\\-\infty\end{pmatrix}, \begin{pmatrix}\infty\\\infty\\0\end{pmatrix}\right) - \text{distribution}$$

using the described algorithms. Unfortunately, the basic acceptance-rejection algorithm using $\mathcal{N}_q(\mu, \Sigma)$ as proposal density has a very low acceptance probability with approximately 2×10^{-8} for this example. Thus, we compare the results of the three remaining algorithms, the Gibbs sampler proposed by Li and Ghosh (2015) with 100 burn-in samples, the acceptance-rejection algorithm with minimax tilting by Botev (2017) and the direct sampling method by Koch and Bopp (2019). Figure 5.1 displays the combination of different dimensions with a two dimensional density estimate superimposed.

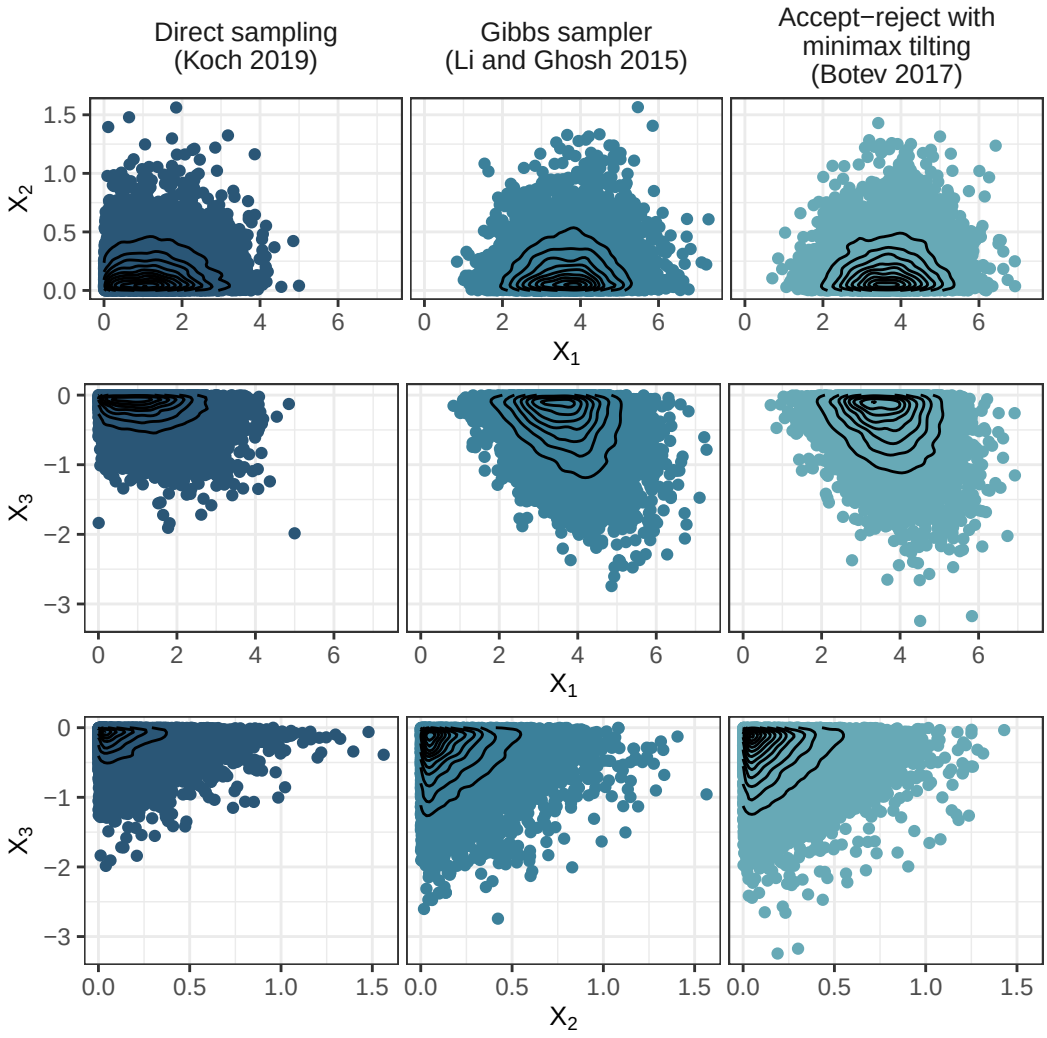

Figure 5.1 Pairwise results for 10,000 samples from a three dimensional truncated multivariate normal distribution for three different sampling algorithms.

With the plots, it becomes apparent that the direct sampling algorithm by Koch and Bopp

(2019) yields different results than the other two sampling methods. We also compare the results with the Gibbs sampler proposed by Geweke (1991) and the estimated densities are similar to those of the samplers by Li and Ghosh (2015) and Botev (2017). Thus, we have reasonable doubt that the direct sampling algorithm generates samples from the desired distribution and we do not consider this approach further.

We compare the remaining three different sampling algorithms on various scenarios that are outlined in Table 5.1 to assess their speed. For each scenario, we test the algorithms in two ways. First, we generate 10,000 samples and second we generate 1 sample 1,000 times, which implies that we call the sampling algorithm 1,000 times. In the MCMC Algorithm 3.1 we need one sample per observation in each iteration and thus the second timings are more interesting.

	μ	Σ	a	b
Scenario 1	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}\infty\\\infty\end{pmatrix}$
Scenario 2	$\begin{pmatrix}0.25\\1\end{pmatrix}$	$\begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}$	$\begin{pmatrix}-\infty\\0\end{pmatrix}$	$\begin{pmatrix}0\\\infty\end{pmatrix}$
Scenario 3	$\begin{pmatrix}-5\\5\end{pmatrix}$	$\begin{pmatrix}1 & -0.5\\-0.5 & 1\end{pmatrix}$	$\begin{pmatrix}-\infty\\0\end{pmatrix}$	$\begin{pmatrix}0\\\infty\end{pmatrix}$
Scenario 4	$\begin{pmatrix}-5\\-5\end{pmatrix}$	$\begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}\infty\\\infty\end{pmatrix}$

Table 5.1 Scenarios for comparing sampling algorithms for the truncate multivariate normal distribution.

To compare the algorithms properly, we use the `microbenchmark` package and repeat the sampling 100 times. For the Gibbs sampling algorithm, we use 100 burn-in samples. Figure 5.2 displays the minimum, mean, and maximum times for each scenario and sampling algorithm. For the basic acceptance-rejection algorithm the acceptance probability for scenario 4 is low with approximately 8×10^{-10}, thus we exclude this algorithm for this scenario.

The basic acceptance-rejection sampler is the fasted algorithm overall when the mean of the normal distribution lies in or close to the truncated area as in Scenario 1 to 3. Although the Gibbs sampler is slower than the basic acceptance-rejection algorithm in these cases, the difference is relatively small and the times for Scenario 4 are comparable to the other Scenarios. The acceptance-rejection algorithm with minimax tilting is considerably slower, especially in generating one sample 1,000 times. While the algorithm, in general, might be fast, repeatedly calling a R-function is slow in comparison to the C++ implementation of the other two samplers. Thus, for the implementation of the MCMC Algorithm 3.1, we use the Gibbs sampler proposed by Li and Ghosh (2015).

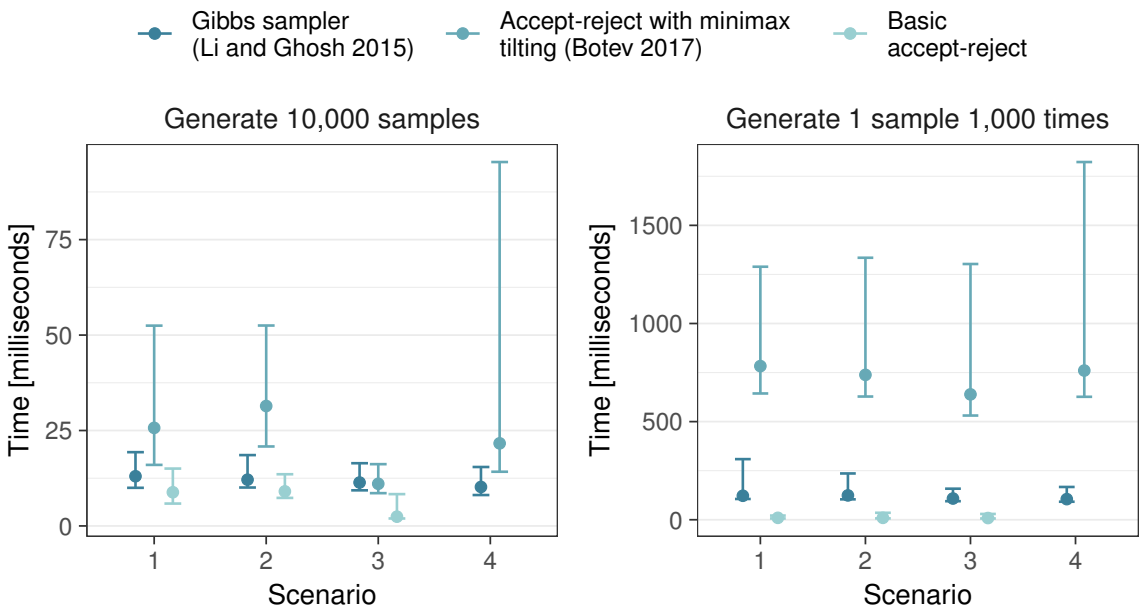

Figure 5.2 Sampling times for different scenarios and sampling algorithms. We use 100 burn-in samples for the Gibbs sampler and compare the times with the `microbenchmark` package and 100 repetitions. The dots indicate the average time for the scenario and the error bars the minimum and maximum values. Calculations are performed on a Dell Latitude E7440 with Intel(R) Core(TM) i7-4600U CPU @ 2.10GHz.

Caubet Fernandez et al. (2019) used this algorithm for sampling from the truncated multivariate normal distribution as well, however without any burn-in samples and with a different transformation for the variables [181]. We correct their approach to properly match the algorithm by Li and Ghosh (2015) and add the number of burn-in samples for the sampler as an input parameter for Algorithm 3.1 with a default value of 10.

With the sampling algorithm for the truncated multivariate normal distribution, we covered the first step of Algorithm 3.1, updating the latent variables z. For updating ϕ, β, and the scale matrix R, `Rcpp` and Armadillo provide samplers for the Gamma and multivariate normal distribution. To evaluate the acceptance probability of R in the Metropolis step, we require the calculation of the multivariate normal density. Further, for calculating the weights in each step, we need the density for the multivariate t-distribution and the univariate logistic distribution. The last one is, again, available in `Rcpp` and we discuss the calculation of the multivariate densities in the following.

The density of the q-dimensional multivariate normal distribution $\mathcal{N}_q(\mu, \Sigma)$ is given by

$$f(x) = (2\pi)^{-\frac{q}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \text{ [164].}$$

Σ is symmetric and positive definite with the Cholesky decomposition $\Sigma = Q'Q$, where Q is

an upper triangular matrix. With

$$\Sigma^{-1} = (Q'Q)^{-1} = Q^{-1}(Q')^{-1} = Q^{-1}\left(Q^{-1}\right)',$$
$$|\Sigma| = |Q'Q| = |Q'||Q| = |Q|^2 = \left(\prod_{i=1}^{q} Q_{ii}\right)^2,$$

the log-transformed density of $\mathcal{N}_k(\mu, \Sigma)$ is

$$\log(f(x)) = -\frac{1}{2}(x-\mu)'Q^{-1}\left(Q^{-1}\right)'(x-\mu) - \frac{q}{2}\log(2\pi) + \sum_{i=1}^{q}\log\left(Q_{ii}^{-1}\right).$$

Similarly, for the multivariate t-distribution given by

$$t_{q,\nu}(x|\mu,\Sigma) = \frac{\Gamma\left(\frac{\nu+q}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{\frac{q}{2}}|\Sigma|^{\frac{1}{2}}} \cdot \left(1 + \frac{1}{\nu}(x-\mu)'\Sigma^{-1}(x-\mu)\right)^{-\frac{\nu+q}{2}},$$

we obtain the log-transformed density

$$\begin{aligned}\log\left(t_{q,\nu}(x|\mu,\Sigma)\right) = \log\left(\Gamma\left(\frac{\nu+q}{2}\right)\right) - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) \\ - \frac{q}{2}\log((\nu\pi)) + \sum_{i=1}^{q}\log\left(Q_{ii}^{-1}\right) \\ - \frac{\nu+q}{2}\log\left(1 + \frac{1}{\nu}(x-\mu)'Q^{-1}\left(Q^{-1}\right)'(x-\mu)\right).\end{aligned}$$

By exploiting the structure of the Cholesky decomposition, we can decrease the computation times compared to a direct calculation of the determinant and inverse of the covariance matrix Σ [281]. Using the log-transformed density instead of the untransformed normal density, allows us to calculate the quotient of the acceptance probability as the exponential of the difference of the numerator and denominator.

Based on the multivariate t-distribution, we can calculate the multivariate logistic distribution defined in (3.6).

5.3 Multivariate logistic regression models for pathological outcome after prostatectomy

In the following, we fit the multivariate logistic regression model to the EMR data set introduced in Chapter 2. As the method is computationally intense, we first use a subsample of the training data to identify promising initial values for the parameters. Further, we use this subset to compare the algorithm to other methods. In general, we only consider the first imputation set and exclude patients with a primary or secondary Gleason grade smaller than three, but we compare the results to the univariate models obtained in Section 4.3 based on all 10 imputations sets.

As we discussed in Section 3.2, the MLE for the multivariate latent variable model exists, when there is overlap for all outcomes and when the absolute value of the polychoric correlation between the different outcomes is not equal to one. Figure 5.3 displays the polychoric correlation for the five adverse pathological outcomes, based on the complete data set, where missing values are deleted pairwise. The correlation coefficient is equal to one for ECE and SVI, thus when we include both outcomes in the multivariate logistic regression model the MLE might not exist. In the following, we, therefore, exclude SVI from modelling.

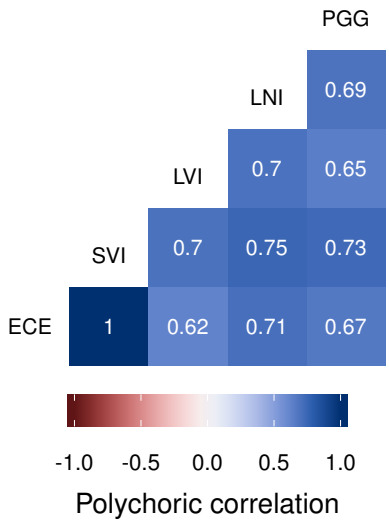

Figure 5.3 Pairwise polychoric correlation for the adverse pathological outcomes, based on the full data set ($n = 24,122$) with missing values pairwise deleted.

We generate a subsample of the first imputation set for the training data using stratified subsampling to ensure that the proportions of adverse outcome combinations in the subset match the data. We stratify the data by all combinations of ECE, LVI, LNI, and PGG and sample 20% of each group. This yields a data set with 3,994 observations.

With this data, we try different values for Ω, the covariance of the proposal density for the unique values of the scale matrix R in the random Metropolis step. For the proposal density in the random Metropolis step we need to find a balance between the number of accepted values, which should not be too small, and the movement in the Markov Chain, which should

also not be too small. For the other input parameters and initial values we choose

$$\beta^{(0)} = 0 \in \mathbb{R}^m \quad \phi^{(0)} = 1 \in \mathbb{R}^n \quad \mu_\beta = 0 \in \mathbb{R}^m$$

$$R^{(0)} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{pmatrix} \quad \Sigma_\beta = (\sigma_{ij}) \in \mathbb{R}^m, \text{ with } \sigma_{ij} = \begin{cases} 1000 & i = j = 1 \\ 4 & i = j \neq 1 \\ 0 & i \neq j \end{cases},$$

where $n = 3,994$ and m is the total number of coefficients. For sampling from the truncated normal distribution, we use 10 burn-in samples in each iteration and for the overall sampling, we use 1,000 burn-in samples.

We compare the chains for the unique values of R using following different values for Ω,

$$\Omega_1 = 5 \cdot 10^{-4} \cdot \mathrm{Id}_6, \qquad \Omega_2 = 10^{-5} \cdot \mathrm{Id}_6,$$

$$\Omega_3 = 10^{-4} \cdot \begin{pmatrix} 10 & 5 & 5 & 5 & 5 & 5 \\ 5 & 10 & 5 & 5 & 5 & 5 \\ 5 & 5 & 10 & 5 & 5 & 5 \\ 5 & 5 & 5 & 10 & 5 & 5 \\ 5 & 5 & 5 & 5 & 10 & 5 \\ 5 & 5 & 5 & 5 & 5 & 10 \end{pmatrix}, \qquad \Omega_4 = 10^{-4} \cdot \begin{pmatrix} 5 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2.5 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2.5 & 1 & 1 & 1 \\ 1 & 1 & 1 & 5 & 1 & 1 \\ 1 & 1 & 1 & 1 & 2.5 & 1 \\ 1 & 1 & 1 & 1 & 1 & 2.5 \end{pmatrix},$$

$$\Omega_5 = 2.5 \cdot 10^{-4} \cdot \mathrm{Id}_6.$$

For Ω_1 to Ω_4 we use 25,000 iterations, where we use a thinning of 10 on the chain yielding 2,500 samples for each coefficient. For Ω_5 we use a thinning of 20 instead of 10 and in total 50,000 iterations, which also yields 2,500 samples. In Figure 5.4 we compare the chains of the unique values of R after thinning. The acceptance rate and movement of the chains are low for Ω_2 and Ω_3 in comparison to the others. The extra thinning of the MCMC samples seems to aid the mixing, especially for r_{31}, r_{42}, and r_{43}. Thus, we use Ω_5 with a thinning of 20.

Figure 5.5 displays the 5,000 sampling values for the coefficients of ECE using the initial parameters describe above, $\Omega = 2.5 \cdot 10^{-4} \cdot \mathrm{Id}_6$, 1,000 burn-in samples, and 100,000 iterations with a thinning of 20. The sampling values for the coefficients related to the other pathological outcomes LVI, LNI, and PGG are given in Appendix C.1. The chains mixed well for the coefficients.

In addition to the Bayesian multivariate logistic regression model, we fit separate univariate logistic regression models to the subsample of the data. Further, we fit a GEE model, but we have to use the full training set for this as otherwise, the sample size is to small for this

Figure 5.4 MCMC samples for the unique values of the scale matrix R based on 2,500 samples using a 1,000 burn-in samples. For Ω_1 to Ω_4 a thinning of 10 was used and for Ω_5 a thinning of 20.

model. Thus, we use the first imputation set of the complete training data containing 19,977 observations for the GEE model. Finally, we also compare the results with the univariate logistic regression models fit in Section 4.3 on all 10 imputation sets where the coefficients are pooled using Rubin's rules (4.29).

Table 5.2 summarizes the times it took to fit the four different models. The differences between the univariate models and the GEE model are negligible, but fitting of the multivariate logistic regression model takes tremendously longer, as expected.

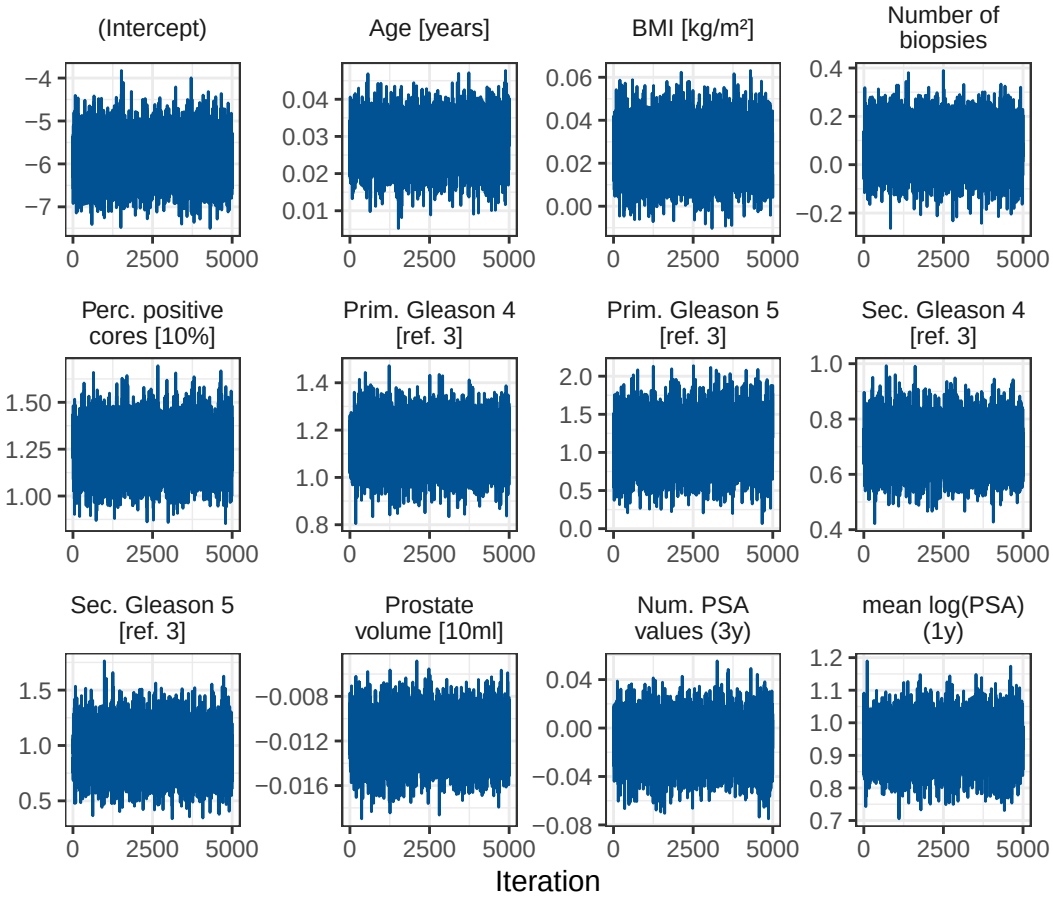

Figure 5.5 MCMC samples for the log OR for ECE based on 5,000 samples after thinning of 20 and 1000 bun-in samples.

Model	Data set size	Times	Comments
Univariate logistic regressions	3,994	0.08 sec	
Univariate logistic regressions	10 imputation sets 19,967 to 19,989	5.81 sec	includes pooling of the coefficients
GEE	19,977	10.22 sec	
Multivariate logistic regression	3,994	5 h 11 min	100,000 iterations, 1000 burn-in, thinning of 20

Table 5.2 Computation times for different models with the corresponding data set sizes. Parameter estimation was performed on an Dell Latitude E7440 with Intel(R) Core(TM) i7-4600U CPU @ 2.10GHz for all models besides the multivariate logistic regression, which was fit using RStudio Server.

Figure 5.6 displays the ORs with 95%-credibility intervals for the multivariate logistic regression in comparison to the OR with 95%-confidence intervals for the other models. The ORs are in general very similar, but the length of the intervals differ. In general, the credibility intervals for the multivariate logistic model are slightly smaller compared to the confidence intervals univariate models on the same data set, and the confidence interval for the GEE

Figure 5.6 ORs with 95%-confidence or credibility intervals for different models fit to the stratified subsample of the training data (subsample, $n = 3,994$), the first imputation of the complete training data (first training set, $n = 19,977$), or all 10 imputation sets of the complete training data (all training sets). Gleason grade coefficients are shown on a different scale on the right for better visibility.

models are slightly smaller as well. However, we should note that apart from the univariate models fit to all imputation sets, we did not adjust the confidence interval to account for the imputed data. Fitting the multivariate models to several or all of the imputed data sets would better reflect the uncertainty in the data due to the missing values and we can expect more sensible estimates for the coefficients.

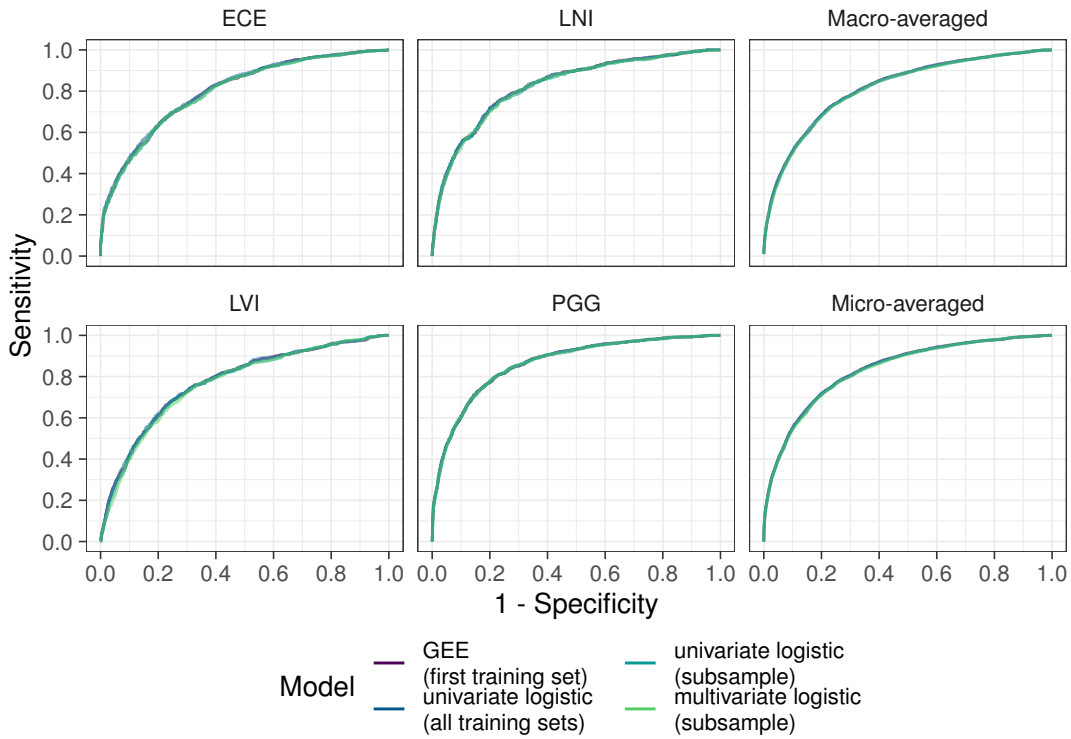

Figure 5.7 Individual, micro- and macro-averaged ROC curves for the prediction of pathological outcomes based on the first imputation set of the test data ($n = 3,515$). Different models are fit to the stratified subsample of the training data (subsample, $n = 3,994$), the first imputation of the complete training data (first training set, $n = 19,977$), or all 10 imputation sets of the complete training data (all training sets).

We assess the predictive value of the models on the first imputation set of the test data with 3,515 observations. Figure 5.7 displays the individual ROC curves for the pathological outcomes as well as the macro- and micro-averaged ROC curves. The ROC curves for the models overlay and are very similar. Thus, regardless of the small differences in the coefficients, there is a negligible difference in the predictive value for the pathological outcomes.

There is little to no hope that fitting the multivariate model to the complete training set results in a model with a higher predictive value than the once we already explored. Given that the fit on the subsample required over 5 hours, we estimate the training time for the complete data set at approximately 25 to 26 hours. While this training time is definitely manageable, fitting this model would most probably be a great waste of resources. Therefore, we do not fit the multivariate model to the complete training data.

6 Discussion

This thesis provides the proof that the BIC approximates the BF for the multivariate logistic regression model to order $O_p(1)$ for arbitrary prior distributions and $O_p(n^{-\frac{1}{2}})$ for the unit information prior distribution. On the way, we show the construction of null-orthogonal parametrizations for arbitrary models. Further, we provide a framework in `R` for the Bayesian multivariate logistic regression with different covariates for different multivariate outcomes. The code for the `R` package `multlogreg` is available on GitHub [282].

Only in the application to simulations studies and especially real-world problems, statistical theory can show its potential and usefulness. However, the results in Section 5.3 are not as promising and optimistic as we would have envisioned. Despite the higher complexity of the model and tremendously longer training time, the predictive performance for the Bayesian multivariate model is no much better than the separate univariate ones. We have a slight improvement for the model inference as the credibility intervals are smaller compared to univariate models on the same data set, but there we did not account for the fact that we used imputed data, which in general leads to larger confidence/credibility intervals.

We did not include one pathological outcome for the multivariate modelling, due to a high polychoric correlation with another outcome, but this could be addressed by using a multivariate multinomial model. The univariate models are scalable to an arbitrary number of outcomes and given the speed of implementation are therefore preferred in this case. Nevertheless, failing to improve the predictive performance in this particular case, does not imply that in general separate univariate regression models are better. Simulation studies could explore the type of models, where accounting for the correlation among outcomes is beneficial. Models dealing with time-dependent data of the same or similar measurements might be such a model type.

Working with EMR data is blessing and curse at the same time. The sheer amount of data that is readily available without conducting an expensive study offers the opportunity to explore different scientific questions. However, it is attributed to the lack of thorough design of experiment that essential variables, such as an indicator for patients under AS, are missing, and extensive data pre-processing is necessary. Further, as Wang et al. (2016) show, the evaluation of the model performance might be biased when events are classified as non-events [283]. Depending on whether the marker is positively or negatively associated with the

event, this biases the performance measures in the respective way [283]. For the application to the prostatectomy data, this can, for example, occur when lymph node involvement was not assessed but then falsely entered as no involvement instead of missing. Again, the lack of a pre-defined protocol leaves more room for such mistakes and misclassification.

With this thesis, we laid the foundation to further explore the EMR data and develop prediction models for prostatectomy patients. The model performance can be used as a benchmark for future machine learning models, but what inevitable needs to be done is to improve the data quality and add more relevant predictors, such as image or genetic data.

The framework for Bayesian multivariate logistic regression model reduces the implementation time for future applications so that exploring the possible benefits of this model approach has been made easier and is available to a broader community.

List of Figures

List of Tables

References

[1] Nordo, A. H., Levaux, H. P., Becnel, L. B., et al. (2019). Use of EHRs data for clinical research: Historical progress and current applications. *Learning Health Systems* 3 (1). DOI: 10.1002/lrh2.10076.

[2] Amarasingham, R., Moore, B. J., Tabak, Y. P., et al. (2010). An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Medical Care* 48 (11), pp. 981–988. DOI: 10.1097/MLR.0b013e3181ef60d9.

[3] Amarasingham, R., Velasco, F., Xie, B., et al. (2015). Electronic medical record-based multicondition models to predict the risk of 30 day readmission or death among adult medicine patients: Validation and comparison to existing models. *BMC Medical Informatics and Decision Making* 15, pp. 15–39. DOI: 10.1186/s12911-015-0162-6.

[4] Escobar, G. J., Gardner, M. N., Greene, J. D., Draper, D., and Kipnis, P. (2013). Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system. *Medical Care* 51 (5), pp. 446–453. DOI: 10.1097/MLR.0b013e3182881c8e.

[5] Goldstein, B. A., Navar, A. M., Pencina, M. J., and Ioannidis, J. P. A. (2017a). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *Journal of the American Medical Informatics Association* 24 (1), pp. 198–208. DOI: 10.1093/jamia/ocw042.

[6] Goldstein, B. A., Pomann, G. M., Winkelmayer, W. C., and Pencina, M. J. (2017b). A comparison of risk prediction methods using repeated observations: An application to electronic health records for hemodialysis. *Statistics in Medicine* 36 (17), pp. 2750–2763. DOI: 10.1002/sim.7308.

[7] Kennedy, E. H., Wiitala, W. L., Hayward, R. A., and Sussman, J. B. (2013). Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. *Medical Care* 51 (3), pp. 251–258. DOI: 10.1097/MLR.0b013e31827da594.

[8] Luo, L., Small, D., Stewart, W. F., and Roy, J. A. (2013). Methods for estimating kidney disease stage transition probabilities using electronic medical records. *eGEMs* 1 (3). DOI: 10.13063/2327-9214.1040.

[9] Singh, A., Nadkarni, G., Gottesman, O., et al. (2015). Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. *Journal of Biomedical Informatics* 53, pp. 220–228. DOI: 10.1016/j.jbi.2014.11.005.

[10] Wu, J., Roy, J. A., and Stewart, W. F. (2010). Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches. *Medical Care* 48 (6), pp. S106–S113. DOI: 10.1097/MLR.0b013e3181de9e17.

[11] Wolfson, J., Bandyopadhyay, S., Elidrisi, M., et al. (2015). A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data. *Statistics in Medicine* 34 (21), pp. 2941–2957. DOI: 10.1002/sim.6526.

[12] Rajkomar, A., Oren, E., Chen, K., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine* 1, p. 18. DOI: 10.1038/s41746-018-0029-1.

[13] Burchardt, M., Fichtner, J., Wesselmann, S., et al. (2019). Kennzahlenauswertung 2019: Jahresbericht der zertifizierten Prostatakrebszentren. Auditjahr 2018 / Kennzahlenjahr 2017. Deutsche Krebsgesellschaft e.V. and OnkoZert.

[14] Hidalgo, B. and Goodman, M. (2013). Multivariate or multivariable regression? *American Journal of Public Health* 103 (1), pp. 39–40. DOI: 10.2105/AJPH.2012.300897.

[15] Kass, R. E. and Vaidyanathan, S. K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society. Series B. (Methodological)* 54 (1), pp. 129–144. DOI: 10.2307/2345950.

[16] Martini, A., Gupta, A., Lewis, S. C., et al. (2018). Development and internal validation of a side-specific, multiparametric magnetic resonance imaging-based nomogram for the prediction of extracapsular extension of prostate cancer. *BJU International* 122 (6), pp. 1025–1033. DOI: 10.1111/bju.14353.

[17] Gaunay, G. S., Patel, V., Shah, P., et al. (2017). Multi-parametric MRI of the prostate: Factors predicting extracapsular extension at the time of radical prostatectomy. *Asian Journal of Urology* 4 (1), pp. 31–36. DOI: 10.1016/j.ajur.2016.07.002.

[18] Feng, T. S., Sharif-Afshar, A. R., Smith, S. C., et al. (2015). Multiparametric magnetic resonance imaging localizes established extracapsular extension of prostate cancer. *Urologic Oncology: Seminars and Original Investigations* 33 (3), pp. 109.e15–109.e22. DOI: 10.1016/j.urolonc.2014.11.007.

[19] Woo, S., Cho, J. Y., Kim, S. Y., and Kim, S. H. (2015). Extracapsular extension in prostate cancer: Added value of diffusion-weighted MRI in patients with equivocal findings on T2-weighted imaging. *American Journal of Roentgenology* 204 (2), pp. W168–W175. DOI: 10.2214/AJR.14.12939.

[20] Eifler, J. B., Feng, Z., Lin, B. M., et al. (2013). An updated prostate cancer staging nomogram (Partin tables) based on cases from 2006 to 2011. *BJU International* 111 (1), pp. 22–29. DOI: 10.1111/j.1464-410X.2012.11324.x.

[21] Chung, J. S., Choi, H. Y., Song, H.-R., et al. (2010). Preoperative nomograms for predicting extracapsular extension in Korean men with localized prostate cancer: A multi-institutional clinicopathologic study. *Journal of Korean Medical Science* 25 (10), pp. 1443–1448. DOI: 10.3346/jkms.2010.25.10.1443.

[22] Tarjan, M. and Tot, T. (2006). Prediction of extracapsular extension of prostate cancer based on systematic core biopsies. *Scandinavian Journal of Urology and Nephrology* 40 (6), pp. 459–464. DOI: 10.1080/00365590600795446.

[23] Ohori, M., Kattan, M. W., Koh, H., et al. (2004). Predicting the presence and side of extracapsular extension: A nomogram for staging prostate cancer. *The Journal of Urology* 171 (5), pp. 1844–1849. DOI: 10.1097/01.ju.0000121693.05077.3d.

[24] Ou, Y.-C. (2002). Preoperative prediction of extracapsular tumor extension at radical retropubic prostatectomy in Taiwanese patients with T1c prostate cancer. *Japanese Journal of Clinical Oncology* 32 (5), pp. 172–176. DOI: 10.1093/jjco/hyf036.

[25] Egawa, S., Suyama, K., Matsumoto, K., et al. (1998). Improved predictability of extracapsular extension and seminal vesicle involvement based on clinical and biopsy findings in prostate cancer in Japanese men. *Urology* 52 (3), pp. 433–440. DOI: 10.1016/S0090-4295(98)00207-6.

[26] Partin, A. W. (1997). Combination of prostate-specific antigen, clinical stage, and Gleason score to predict pathological stage of localized prostate cancer. *JAMA* 277 (18), p. 1445. DOI: 10.1001/jama.1997.03540420041027.

[27] Ohori, M., Kattan, M. W., Yu, C., et al. (2010). Nomogram to predict seminal vesicle invasion using the status of cancer at the base of the prostate on systematic biopsy. *International Journal of Urology* 17 (6), pp. 534–540. DOI: 10.1111/j.1442-2042.2010.02513.x.

[28] Jung, D. C., Lee, H. J., Kim, S. H., Choe, G. Y., and Lee, S. E. (2008). Preoperative MR imaging in the evaluation of seminal vesicle invasion in prostate cancer: Pattern analysis of seminal vesicle lesions. *Journal of Magnetic Resonance Imaging* 28 (1), pp. 144–150. DOI: 10.1002/jmri.21422.

[29] Gallina, A., Chun, F. K.-H., Briganti, A., et al. (2007). Development and split-sample validation of a nomogram predicting the probability of seminal vesicle invasion at radical prostatectomy. *European Urology* 52 (1), pp. 98–105. DOI: 10.1016/j.eururo.2007.01.060.

[30] Baccala, A., Reuther, A. M., Bianco, F. J., et al. (2007). Complete resection of seminal vesicles at radical prostatectomy results in substantial long-term disease-

free survival: Multi-institutional study of 6740 patients. *Urology* 69 (3), pp. 536–540. DOI: 10.1016/j.urology.2006.12.013.

[31] Koh, H., Kattan, M. W., Scardino, P. T., et al. (2003). A nomogram to predict seminal vesicle invasion by the extent and location of cancer in systematic biopsy results. *The Journal of Urology* 170 (4 Pt 1), pp. 1203–1208. DOI: 10.1097/01.ju.0000085074.62960.7b.

[32] Briganti, A., Chun, F. K.-H., Salonia, A., et al. (2007). A nomogram for staging of exclusive nonobturator lymph node metastases in men with localized prostate cancer. *European Urology* 51 (1), pp. 112–119, discussion 119–120. DOI: 10.1016/j.eururo.2006.05.045.

[33] Briganti, A., Chun, F. K.-H., Salonia, A., et al. (2006). Validation of a nomogram predicting the probability of lymph node invasion among patients undergoing radical prostatectomy and an extended pelvic lymphadenectomy. *European Urology* 49 (6), pp. 1019–1026, discussion 1026–1027. DOI: 10.1016/j.eururo.2006.01.043.

[34] Wang, L., Hricak, H., Kattan, M. W., et al. (2006). Combined endorectal and phased-array MRI in the prediction of pelvic lymph node metastasis in prostate cancer. *American Journal of Roentgenology* 186 (3), pp. 743–748. DOI: 10.2214/AJR.04.1682.

[35] Cagiannos, I., Karakiewicz, P., Eastham, J. A., et al. (2003). A preoperative nomogram identifying decreased risk of positive pelvic lymph nodes in patients with prostate cancer. *The Journal of Urology* 170 (5), pp. 1798–1803. DOI: 10.1097/01.ju.0000091805.98960.13.

[36] Crawford, E. D., Batuello, J. T., Snow, P., et al. (2000). The use of artificial intelligence technology to predict lymph node spread in men with clinically localized prostate carcinoma. *Cancer* 88 (9), pp. 2105–2109. DOI: 10.1002/(SICI)1097-0142(20000501)88:9<2105::AID-CNCR16>3.0.CO;2-3.

[37] Han, M., Snow, P. B., Brandt, J. M., and Partin, A. W. (2001). Evaluation of artificial neural networks for the prediction of pathologic stage in prostate carcinoma. *Cancer* 91 (8 Suppl), pp. 1661–1666.

[38] Narayan, P., Fournier, G., Gajendran, V., et al. (1994). Utility of preoperative serum prostate-specific antigen concentration and biopsy Gleason score in predicting risk of pelvic lymph node metastases in prostate cancer. *Urology* 44 (4), pp. 519–524. DOI: 10.1016/S0090-4295(94)80050-2.

[39] Epstein, J. I., Feng, Z., Trock, B. J., and Pierorazio, P. M. (2012). Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: Incidence and predictive factors using the modified Gleason grading system and factoring in tertiary grades. *European Urology* 61 (5), pp. 1019–1024. DOI: 10.1016/j.eururo.2012.01.050.

[40] Nayyar, R., Singh, P., Gupta, N. P., et al. (2010). Upgrading of Gleason score on radical prostatectomy specimen compared to the pre-operative needle core biopsy: An Indian experience. *Indian Journal of Urology* 26 (1), pp. 56–59. DOI: 10.4103/0970-1591.60445.

[41] Alchin, D. R., Murphy, D., and Lawrentschuk, N. (2016). What are the predictive factors for Gleason score upgrade following RP? *Urologia Internationalis* 96 (1), pp. 1–4. DOI: 10.1159/000439139.

[42] Gondo, T., Poon, B. Y., Matsumoto, K., et al. (2015). Clinical role of pathological downgrading after radical prostatectomy in patients with biopsy confirmed Gleason score 3 + 4 prostate cancer. *BJU International* 115 (1), pp. 81–86. DOI: 10.1111/bju.12769.

[43] Sarici, H., Telli, O., Yigitbasi, O., et al. (2014). Predictors of Gleason score upgrading in patients with prostate biopsy Gleason score $\leq$ 6. *Canadian Urological Association Journal* 8 (5-6), pp. E342–E346. DOI: 10.5489/cuaj.1499.

[44] Coley, R. Y., Zeger, S. L., Mamawala, M., Pienta, K. J., and Carter, H. B. (2017). Prediction of the pathologic Gleason score to inform a personalized management program for prostate cancer. *European Urology* 72 (1), pp. 135–141. DOI: 10.1016/j.eururo.2016.08.005.

[45] *Prostate cancer nomograms* (2018). Memorial Sloan Kettering Cancer Center. URL: https://www.mskcc.org/nomograms/prostate (visited on 12/07/2019).

[46] Erickstad, L., Reed, G., Bhat, D., Roehrborn, C. G., and Lotan, Y. (2011). Use of electronic medical records to identify patients at risk for prostate cancer in an academic institution. *Prostate Cancer and Prostatic Diseases* 14 (1), pp. 85–89. DOI: 10.1038/pcan.2010.50.

[47] Kishan, A. U., Cook, R. R., Ciezki, J. P., et al. (2018). Radical prostatectomy, external beam radiotherapy, or external beam radiotherapy with Brachytherapy boost and disease progression and mortality in patients with Gleason score 9-10 prostate cancer. *JAMA* 319 (9), pp. 896–905. DOI: 10.1001/jama.2018.0587.

[48] Raheem, O. A., Cohen, S. A., Parsons, J. K., Palazzi, K. L., and Kane, C. J. (2015). A family history of lethal prostate cancer and risk of aggressive prostate cancer in patients undergoing radical prostatectomy. *Scientific Reports* 5. DOI: 10.1038/srep10544.

[49] Seneviratne, M. G., Banda, J. M., Brooks, J. D., Shah, N. H., and Hernandez-Boussard, T. M. (2018). Identifying cases of metastatic prostate cancer using machine learning on electronic health records. *AMIA Annual Symposium Proceedings* 2018, pp. 1498–1504.

[50] Torre, L. A., Siegel, R. L., Ward, E. M., and Jemal, A. (2016). Global cancer incidence and mortality rates and trends: An update. *Cancer Epidemiology, Biomarkers & Prevention* 25 (1), pp. 16–27. DOI: 10.1158/1055-9965.EPI-15-0578.

[51] National Cancer Institute (2017). *Prostate-Specific Antigen (PSA) Test*. National Cancer Institute. URL: https://www.cancer.gov/types/prostate/psa-fact-sheet (visited on 12/09/2019).

[52] Epstein, J. I., Allsbrook, W. C., Amin, M. B., and Egevad, L. L. (2005). The 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *The American Journal of Surgical Pathology* 29 (9), pp. 1228–1242. DOI: 10.1097/01.pas.0000173646.99337.b1.

[53] Epstein, J. I., Egevad, L., Amin, M. B., et al. (2016). The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system. *The American Journal of Surgical Pathology* 40 (2), pp. 244–252. DOI: 10.1097/PAS.0000000000000530.

[54] National Cancer Institute (2019). *Prostate Cancer Treatment (PDQ®): Health Professional Version*. National Cancer Institute. URL: https://www.cancer.gov/types/prostate/hp/prostate-treatment-pdq (visited on 12/09/2019).

[55] Cosma, G., Acampora, G., Brown, D., et al. (2016). Prediction of pathological stage in patients with prostate cancer: A Neuro-Fuzzy model. *PLOS ONE* 11 (6). DOI: 10.1371/journal.pone.0155856.

[56] Albertsen, P. C., Hanley, J. A., Barrows, G. H., et al. (2005). Prostate cancer and the Will Rogers phenomenon. *Journal of the National Cancer Institute* 97 (17), pp. 1248–1253. DOI: 10.1093/jnci/dji248.

[57] Gordetsky, J. and Epstein, J. (2016). Grading of prostatic adenocarcinoma: Current state and prognostic implications. *Diagnostic Pathology* 11, p. 25. DOI: 10.1186/s13000-016-0478-2.

[58] *NCI dictionary of cancer terms* (2018). *Active surveillance*. National Cancer Institute. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/active-surveillance (visited on 05/14/2018).

[59] Dall'Era, M. A., Albertsen, P. C., Bangma, C., et al. (2012). Active surveillance for prostate cancer: A systematic review of the literature. *European Urology* 62 (6), pp. 976–983. DOI: 10.1016/j.eururo.2012.05.072.

[60] *NCI dictionary of cancer terms* (2018). *Biochemical recurrence*. National Cancer Institute. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biochemical-recurrence (visited on 05/14/2018).

[61] Amin, A. (2016). Pitfalls of diagnosis of extraprostatic extension in prostate adenocarcinoma. *Annals of Clinical Pathology* 4 (6).

[62] Fischer, S., Lin, D., Simon, R. M., et al. (2016). Do all men with pathological Gleason score 8-10 prostate cancer have poor outcomes? Results from the SEARCH database. *BJU International* 118 (2), pp. 250–257. DOI: 10.1111/bju.13319.

[63] Mikel Hubanks, J., Boorjian, S. A., Frank, I., et al. (2014). The presence of extracapsular extension is associated with an increased risk of death from prostate cancer after radical prostatectomy for patients with seminal vesicle invasion and negative lymph nodes. *Urologic Oncology: Seminars and Original Investigations* 32 (1), pp. 26.e1–26.e7. DOI: 10.1016/j.urolonc.2012.09.002.

[64] Watkins, J. M., Watkins, P. L., Laszewski, M., et al. (2014). A closer look at extraprostatic extension (EPE): Evaluation of PSA relapse rates following prostatectomy (RP) for pT3aN0 prostate cancer with or without margin involvement. *Journal of Clinical Oncology* 32 (4_suppl), p. 215. DOI: 10.1200/jco.2014.32.4_suppl.215.

[65] Billis, A., Meirelles, L. L., Freitas, L. L. L., et al. (2012). Influence of focal and diffuse extraprostatic extension and positive surgical margins on biochemical progression following radical prostatectomy. *International Brazilian Journal of Urology* 38 (2), pp. 175–184. DOI: 10.1590/s1677-55382012000200005.

[66] Eggener, S. E., Scardino, P. T., Walsh, P. C., et al. (2011). Predicting 15-year prostate cancer specific mortality after radical prostatectomy. *The Journal of Urology* 185 (3), pp. 869–875. DOI: 10.1016/j.juro.2010.10.057.

[67] D'Amico, A. V., Whittington, R., Malkowicz, S. B., et al. (2000). Prostate specific antigen outcome based on the extent of extracapsular extension and margin status in patients with seminal vesicle negative prostate carcinoma of Gleason score ≤ 7. *Cancer* 88 (9), pp. 2110–2115. DOI: 10.1002/(SICI)1097-0142(20000501)88:9<2110::AID-CNCR17>3.0.CO;2-E.

[68] Wheeler, T. M., Dillioglugil, Ö., Kattan, M. W., et al. (1998). Clinical and pathological significance of the level and extent of capsular invasion in clinical stage T1–2 prostate cancer. *Human Pathology* 29 (8), pp. 856–862. DOI: 10.1016/S0046-8177(98)90457-9.

[69] Adamis, S. and Varkarakis, I. (2016). High risk prognostic factors after radical prostatectomy. *Hellenic Urology* 28 (1), pp. 34–43.

[70] Sapre, N., Pedersen, J., Hong, M. K., et al. (2012). Re-evaluating the biological significance of seminal vesicle invasion (SVI) in locally advanced prostate cancer. *BJU International* 110 (S4), pp. 58–63. DOI: 10.1111/j.1464-410X.2012.11477.x.

[71] Potter, S. R., Epstein, J. I., and Partin, A. W. (2000). Seminal vesicle invasion by prostate cancer: Prognostic significance and therapeutic implications. *Reviews in Urology* 2 (3), pp. 190–195.

[72] Billis, A., Teixeira, D. A., Stelini, R. F., et al. (2007). Seminal vesicle invasion in radical prostatectomies: Which is the most common route of invasion? *International Urology and Nephrology* 39 (4), pp. 1097–1102. DOI: 10.1007/s11255-007-9189-7.

[73] Pierorazio, P. M., Ross, A. E., Schaeffer, E. M., et al. (2011). A contemporary analysis of outcomes of adenocarcinoma of the prostate with seminal vesicle invasion (pT3b) after radical prostatectomy. *The Journal of Urology* 185 (5), pp. 1691–1697. DOI: 10.1016/j.juro.2010.12.059.

[74] Epstein, J. I., Partin, A. W., Sauvageot, J., and Walsh, P. C. (1996). Prediction of progression following radical prostatectomy. A multivariate analysis of 721 men with long-term follow-up. *The American Journal of Surgical Pathology* 20 (3), pp. 286–292.

[75] Datta, K., Muders, M., Zhang, H., and Tindall, D. J. (2010). Mechanism of lymph node metastasis in prostate cancer. *Future Oncology* 6 (5), pp. 823–836. DOI: 10.2217/fon.10.33.

[76] Briganti, A., Blute, M. L., Eastham, J. H., et al. (2009). Pelvic lymph node dissection in prostate cancer. *European Urology* 55 (6), pp. 1251–1265. DOI: 10.1016/j.eururo.2009.03.012.

[77] Abdollah, F., Karnes, R. J., Suardi, N., et al. (2014). Predicting survival of patients with node-positive prostate cancer following multimodal treatment. *European Urology* 65 (3), pp. 554–562. DOI: 10.1016/j.eururo.2013.09.025.

[78] Mydlo, J. H. and Godec, C. J., eds. (2016). *Prostate cancer: Science and clinical practice*. Second edition. London: Elsevier. ISBN: 9780128005927.

[79] Park, Y. H., Kim, Y., Yu, H., et al. (2016). Is lymphovascular invasion a powerful predictor for biochemical recurrence in pT3 N0 prostate cancer? Results from the K-CaP database. *Scientific Reports* 6, p. 25419. DOI: 10.1038/srep25419.

[80] Yee, D. S., Shariat, S. F., Lowrance, W. T., et al. (2011). Prognostic significance of lymphovascular invasion in radical prostatectomy specimens. *BJU International* 108 (4), pp. 502–507. DOI: 10.1111/j.1464-410X.2010.09848.x.

[81] Cheng, L., Jones, T. D., Lin, H., et al. (2005). Lymphovascular invasion is an independent prognostic factor in prostatic adenocarcinoma. *The Journal of Urology* 174 (6), pp. 2181–2185. DOI: 10.1097/01.ju.0000181215.41607.c3.

[82] Shariat, S. F., Khoddami, S. M., Saboorian, H., et al. (2004). Lymphovascular invasion is a pathological feature of biologically aggressive disease in patients treated with radical prostatectomy. *The Journal of Urology* 171 (3), pp. 1122–1127. DOI: 10.1097/01.ju.0000113249.82533.28.

[83] Herman, C. M., Wilcox, G. E., Kattan, M. W., Scardino, P. T., and Wheeler, T. M. (2000). Lymphovascular invasion as a predictor of disease progression in prostate cancer. *The American Journal of Surgical Pathology* 24 (6), pp. 859–863. DOI: 10.1097/00000478-200006000-00012.

[84] May, M., Kaufmann, O., Hammermann, F., Loy, V., and Siegsmund, M. (2007). Prognostic impact of lymphovascular invasion in radical prostatectomy specimens. *BJU International* 99 (3), pp. 539–544. DOI: 10.1111/j.1464-410X.2006.06650.x.

[85] Yamamoto, S., Kawakami, S., Yonese, J., et al. (2008). Lymphovascular invasion is an independent predictor of prostate-specific antigen failure after radical prostatectomy in patients with pT3aN0 prostate cancer. *International Journal of Urology* 15 (10), pp. 895–899. DOI: 10.1111/j.1442-2042.2008.02140.x.

[86] Jiang, W., Zhang, L., Wu, B., et al. (2018). The impact of lymphovascular invasion in patients with prostate cancer following radical prostatectomy and its association with their clinicopathological features: An updated PRISMA-compliant systematic review and meta-analysis. *Medicine* 97 (49). DOI: 10.1097/MD.0000000000013537.

[87] Das, S. and Skobe, M. (2008). Lymphatic vessel activation in cancer. *Annals of the New York Academy of Sciences* 1131, pp. 235–241. DOI: 10.1196/annals.1413.021.

[88] Chan, T. Y., Partin, A. W., Walsh, P. C., and Epstein, J. I. (2000). Prognostic significance of Gleason score 3+4 versus Gleason score 4+3 tumor at radical prostatectomy. *Urology* 56 (5), pp. 823–827. DOI: 10.1016/s0090-4295(00)00753-6.

[89] Tollefson, M. K., Leibovich, B. C., Slezak, J. M., Zincke, H., and Blute, M. L. (2006). Long-Term prognostic significance of primary Gleason pattern in patients with Gleason score 7 prostate cancer: Impact on prostate cancer specific survival. *The Journal of Urology* 175 (2), pp. 547–551. DOI: 10.1016/S0022-5347(05)00152-7.

[90] Gandaglia, G., Karnes, R. J., Sivaraman, A., et al. (2017). Are all grade group 4 prostate cancers created equal? Implications for the applicability of the novel grade grouping. *Urologic Oncology* 35 (7), pp. 461.e7–461.e14. DOI: 10.1016/j.urolonc.2017.02.012.

[91] Zhu, X., Gou, X., and Zhou, M. (2019). Nomograms predict survival advantages of Gleason score 3+4 over 4+3 for prostate cancer: A SEER-Based Study. *Frontiers in Oncology* 9, p. 646. DOI: 10.3389/fonc.2019.00646.

[92] Kang, D. E., Fitzsimons, N. J., Presti, J. C., et al. (2007). Risk stratification of men with Gleason score 7 to 10 tumors by primary and secondary Gleason score: Results from the SEARCH database. *Urology* 70 (2), pp. 277–282. DOI: 10.1016/j.urology.2007.03.059.

[93] Park, J., Yoo, S., Cho, M. C., et al. (2018). The impact of pathologic upgrading of Gleason score 7 prostate cancer on the risk of the biochemical recurrence after radical prostatectomy. *BioMed Research International* 2018. DOI: 10.1155/2018/4510149.

[94] Wang, J.-Y., Zhu, Y., Wang, C.-F., et al. (2014). A nomogram to predict Gleason sum upgrading of clinically diagnosed localized prostate cancer among Chinese patients. *Chinese Journal of Cancer* 33 (5), pp. 241–248. DOI: 10.5732/cjc.013.10137.

[95] Strobl, A. N., Thompson, I. M., Vickers, A. J., and Ankerst, D. P. (2015a). The next generation of clinical decision making tools: Development of a real-time prediction tool for outcome of prostate biopsy in response to a continuously evolving prostate cancer landscape. *The Journal of Urology* 194 (1), pp. 58–64. DOI: 10.1016/j.juro.2015.01.092.

[96] Strobl, A. N., Vickers, A. J., van Calster, B., et al. (2015b). Improving patient prostate cancer risk assessment: Moving from static, globally-applied to dynamic, practice-specific risk calculators. *Journal of Biomedical Informatics* 56, pp. 87–93. DOI: 10.1016/j.jbi.2015.05.001.

[97] *Global database on body mass index* (2018). *BMI classification*. World Health Organization. URL: http://apps.who.int/bmi/index.jsp?introPage=intro_3.html (visited on 05/16/2018).

[98] The American Cancer Society medical and editorial content team (2019). *Tests to diagnose and stage prostate cancer*. American Cancer Society, Inc. URL: https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/how-diagnosed.html (visited on 09/11/2019).

[99] Telang, J. M., Lane, B. R., Cher, M. L., Miller, D. C., and Dupree, J. M. (2017). Prostate cancer family history and eligibility for active surveillance: A systematic review of the literature. *BJU International* 120 (4), pp. 464–467. DOI: 10.1111/bju.13862.

[100] Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM* 55 (10), p. 78. DOI: 10.1145/2347736.2347755.

[101] Loeb, S., Sutherland, D. E., D'Amico, A. V., Roehl, K. A., and Catalona, W. J. (2008). PSA velocity is associated with gleason score in radical prostatectomy specimen: Marker for prostate cancer aggressiveness. *Urology* 72 (5), 1116–20, discussion 1120. DOI: 10.1016/j.urology.2008.01.082.

[102] Egawa, S., Arai, Y., Tobisu, K., et al. (2000). Use of pretreatment prostate specific antigen doubling time to predict outcome after radical prostatectomy. *Prostate Cancer and Prostatic Diseases* 3 (S11), pp. 269–274. DOI: 10.1038/sj.pcan.4500435.

[103] Sengupta, S., Myers, R. P., Slezak, J. M., et al. (2005). Preoperative prostate specific antigen doubling time and velocity are strong and independent predictors of outcomes following radical prostatectomy. *The Journal of Urology* 174 (6), pp. 2191–2196. DOI: 10.1097/01.ju.0000181209.37013.99.

[104] Stephenson, A. J., Aprikian, A. G., Souhami, L., et al. (2002). Utility of PSA doubling time in follow-up of untreated patients with localized prostate cancer. *Urology* 59 (5), pp. 652–656. DOI: 10.1016/S0090-4295(02)01526-1.

[105] O'Brien, M. F., Cronin, A. M., Fearn, P. A., et al. (2009). Pretreatment prostate-specific antigen (PSA) velocity and doubling time are associated with outcome but neither improves prediction of outcome beyond pretreatment PSA alone in patients treated with radical prostatectomy. *Journal of Clinical Oncology* 27 (22), pp. 3591–3597. DOI: 10.1200/JCO.2008.19.9794.

[106] O'Brien, M. F., Cronin, A. M., Fearn, P. A., et al. (2011). Evaluation of prediagnostic prostate-specific antigen dynamics as predictors of death from prostate cancer in patients treated conservatively. *International Journal of Cancer* 128 (10), pp. 2373–2381. DOI: 10.1002/ijc.25570.

[107] Thompson, I. M., Ankerst, D. P., Chi, C., et al. (2006). Assessing prostate cancer risk: Results from the Prostate Cancer Prevention Trial. *Journal of the National Cancer Institute* 98 (8), pp. 529–534. DOI: 10.1093/jnci/djj131.

[108] Etzioni, R. D., Ankerst, D. P., Weiss, N. S., Inoue, L. Y. T., and Thompson, I. M. (2007). Is prostate-specific antigen velocity useful in early detection of prostate cancer? A critical appraisal of the evidence. *Journal of the National Cancer Institute* 99 (20), pp. 1510–1515. DOI: 10.1093/jnci/djm171.

[109] Vickers, A. J., Savage, C., O'Brien, M. F., and Lilja, H. (2009). Systematic review of pretreatment prostate-specific antigen velocity and doubling time as predictors for prostate cancer. *Journal of Clinical Oncology* 27 (3), pp. 398–403. DOI: 10.1200/JCO.2008.18.1685.

[110] Vickers, A. J. and Brewster, S. F. (2012). PSA velocity and doubling time in diagnosis and prognosis of prostate cancer. *British Journal of Medical & Surgical Urology* 5 (4), pp. 162–168. DOI: 10.1016/j.bjmsu.2011.08.006.

[111] Memorial Sloan-Kettering Cancer Center (2014). *Prostate cancer nomograms: PSA doubling time*. MSKCC. URL: http://nomograms.mskcc.org/Prostate/PsaDoublingTime.aspx (visited on 05/12/2018).

[112] D'Amico, A. V., Chen, M.-H., Roehl, K. A., and Catalona, W. J. (2004). Preoperative PSA velocity and the risk of death from prostate cancer after radical prostatectomy. *New England Journal of Medicine* 351 (2), pp. 125–135. DOI: 10.1056/NEJMoa032975.

[113] Wasserman, L. (2005). *All of Statistics: A concise course in statistical inference*. Corrected 2nd printing. New York, NY: Springer. 442 pp. DOI: 10.1007/978-0-387-21736-9.

[114] Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (2007). *Discrete multivariate analysis: Theory and applications*. New York, NY: Springer Science+Business Media LLC. 558 pp. DOI: 10.1007/978-0-387-72806-3.

[115] Sokal, R. R. and Rohlf, F. J. (2010). *Biometry: The principles and practice of statistics in biological research*. 3rd ed. New York, NY: Freeman. 887 pp. ISBN: 0716724111.

[116] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. 2nd ed. New York, NY: Springer. 745 pp. DOI: 10.1007/978-0-387-84858-7.

[117] Janssen, K. J. M., Donders, A. R. T., Harrell, F. E., et al. (2010). Missing covariate data in medical research: To impute is better than to ignore. *Journal of Clinical Epidemiology* 63 (7), pp. 721–727. DOI: 10.1016/j.jclinepi.2009.12.008.

[118] Harrell, J. F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. 2nd ed. Cham: Springer. 598 pp. DOI: 10.1007/978-3-319-19425-7.

[119] Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. 2nd ed. Wiley series in probability and statistics. Hoboken: Wiley. 409 pp. DOI: 10.1002/9781119013563.

[120] White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* 30 (4), pp. 377–399. DOI: 10.1002/sim.4067.

[121] van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* 45 (3), pp. 1–67. DOI: 10.18637/jss.v045.i03.

[122] Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American Journal of Epidemiology* 179 (6), pp. 764–774. DOI: 10.1093/aje/kwt312.

[123] Kontopantelis, E., White, I. R., Sperrin, M., and Buchan, I. (2017). Outcome-sensitive multiple imputation: A simulation study. *BMC Medical Research Methodology* 17 (1). DOI: 10.1186/s12874-016-0281-5.

[124] Wood, A. M., Royston, P., and White, I. R. (2015). The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biometrical Journal* 57 (4), pp. 614–632. DOI: 10.1002/bimj.201400004.

[125] van Buuren, S. (2018). *Flexible imputation of missing data*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC. 1433 pp. ISBN: 9781138588318.

[126] Collett, D. (2003). *Modelling binary data*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC. 387 pp. ISBN: 1-58488-324-3.

[127] McCullagh, P. and Nelder, J. A. (1999). *Generalized linear models*. 2nd ed. Vol. 37. Monographs on statistics and applied probability. London: Chapman & Hall. 511 pp. ISBN: 978-0-412-31760-6.

[128] Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B. (Methodological)* 43 (3), pp. 310–313.

[129] Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 71 (1), pp. 1–10. DOI: 10.2307/2336390.

[130] Santner, T. J. and Duffy, D. E. (1986). A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 73 (3), pp. 755–758. DOI: 10.1093/biomet/73.3.755.

[131] Jennrich, R. I. and Sampson, P. F. (1976). Newton-Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics* 18 (1), p. 11. DOI: 10.2307/1267911.

[132] O'Brien, S. M. and Dunson, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics* 60 (3), pp. 739–746. DOI: 10.1111/j.0006-341X.2004.00224.x.

[133] Bel, K., Fok, D., and Paap, R. (2016). Parameter estimation in multivariate logit models with many binary choices. *Econometric Reviews* 37 (5), pp. 534–550. DOI: 10.1080/07474938.2015.1093780.

[134] Cox, D. R. (1972). The analysis of multivariate binary data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 21 (2), pp. 113–120.

[135] Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In: *Studies in Item Analysis and Prediction*. Ed. by H. Solomon. Stanford, California: Stanford University Press, pp. 158–168.

[136] Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses. *Statistical Science* 8 (3), pp. 284–299. DOI: 10.1214/ss/1177010899.

[137] Glonek, G. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society. Series B. (Methodological)* 57 (3), pp. 533–546. URL: http://www.jstor.org/stable/2346155.

[138] Ekholm, A., Smith, P. W. F., and McDonald, J. W. (1995). Marginal regression analysis of a multivariate binary response. *Biometrika* 82 (4), pp. 847–854. DOI: 10.1093/biomet/82.4.847.

[139] Russell, G. J. and Petersen, A. (2000). Analysis of cross category dependence in market basket selection. *Journal of Retailing* 76 (3), pp. 367–392. DOI: 10.1016/S0022-4359(00)00030-0.

[140] Dai, B. (2012). Multivariate Bernoulli distribution models. Department of Statistics. Dissertation. Madison, Wisconsin: University of Wisconsin. 110 pp.

[141] Dai, B., Ding, S., and Wahba, G. (2013). Multivariate Bernoulli distribution. *Bernoulli* 19 (4), pp. 1465–1483. DOI: 10.3150/12-BEJSP10.

[142] Bergsma, W. P. (1997). Marginal models for categorical data. Dissertation. Tilburg, Netherlands: Tilburg University. 168 pp.

[143] Bergsma, W. P. and Rudas, T. (2002). Marginal models for categorical data. *The Annals of Statistics* 30 (1), pp. 140–159.

[144] Liang, K.-Y., Zeger, S. L., and Qaqish, B. F. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B. (Methodological)* 54 (1), pp. 3–40.

[145] Glonek, G. (1996). A class of regression models for multivariate categorical responses. *Biometrika* 83 (1), pp. 15–28. DOI: 10.1093/biomet/83.1.15.

[146] Bonney, G. E. (1987). Logistic regression for dependent binary observations. *Biometrics* 43 (4), pp. 951–973.

[147] Joe, H. and Liu, Y. (1996). A model for a multivariate binary response with covariates based on compatible conditionally specified logistic regressions. *Statistics & Probability Letters* 31 (2), pp. 113–120. DOI: 10.1016/S0167-7152(96)00021-1.

[148] Nikoloulopoulos, A. K. and Karlis, D. (2008). Multivariate logit copula model with an application to dental data. *Statistics in Medicine* 27 (30), pp. 6393–6406. DOI: 10.1002/sim.3449.

[149] Rao Chaganty, N. and Joe, H. (2004). Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66 (4), pp. 851–860. DOI: 10.1111/j.1467-9868.2004.05741.x.

[150] Agresti, A. (2014). Two Bayesian/frequentist challenges for categorical data analyses. *METRON* 72 (2), pp. 125–132. DOI: 10.1007/s40300-014-0036-1.

[151] Bonat, W. H. and Jørgensen, B. (2016). Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 65 (5), pp. 649–675. DOI: 10.1111/rssc.12145.

[152] Bonat, W. H. (2018). Multiple response variables regression models in R: The mcglm package. *Journal of Statistical Software* 84 (4). DOI: 10.18637/jss.v084.i04.

[153] Lesaffre, E. and Molenberghs, G. (1991). Multivariate probit analysis: A neglected procedure in medical statistics. *Statistics in Medicine* 10 (9), pp. 1391–1403. DOI: 10.1002/sim.4780100907.

[154] Ashford, J. R. and Sowden, R. R. (1970). Multi-variate probit analysis. *Biometrics* 26 (3), p. 535. DOI: 10.2307/2529107.

[155] Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* 85 (2), pp. 347–361. DOI: 10.1093/biomet/85.2.347.

[156] Chen, M.-H. and Dey, D. K. (1998). Bayesian modeling of correlated binary responses via scale mixture of multivariate normal link functions. *Sankhyā: The Indian Journal of Statistics, Series A* 60 (3), pp. 322–343. URL: https://www.jstor.org/stable/25051211.

[157] Balakrishnan, N. (1992). Introduction and historical remarks. In: *Handbook of the logistic distribution*. Ed. by N. Balakrishnan. Statistics 123. New York: Dekker, pp. 1–16. ISBN: 0824785878.

[158] Lesaffre, E. and Kaufmann, H. (1992). Existence and uniqueness of the maximum likelihood estimator for a multivariate probit model. *Journal of the American Statistical Association* 87 (419), p. 805. DOI: 10.2307/2290218.

[159] Todem, D. and Kim, K. (2007). Existence and uniqueness conditions for the maximum likelihood solution in regression models for correlated Bernoulli random variables. *Journal of Mathematics and Statistics* 3 (3), pp. 134–141. DOI: 10.3844/jmssp.2007.134.141.

[160] Pearson, K. (1900). Mathematical contributions to the theory of evolution: VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 195, pp. 1–47+405.

[161] Ekström, J. (2011a). A generalized definition of the polychoric correlation coefficient. Department of Statistics, UCLA. https://escholarship.org/uc/item/583610fv.

[162] Ekström, J. (2011b). The Phi-coefficient, the tetrachoric correlation coefficient, and the Pearson-Yule debate. Department of Statistics, UCLA. https://escholarship.org/uc/item/7qp4604r.

[163] Arnold, B. C. (1992). Multivariate logistic distributions. In: *Handbook of the logistic distribution*. Ed. by N. Balakrishnan. Statistics 123. New York: Dekker, pp. 237–262. ISBN: 0824785878.

[164] Kotz, S., Balakrishnan, N., and Johnson, N. L. (2005). *Continuous multivariate distributions: Models and applications*. 2nd ed. Vol. 1. New York, NY: Wiley. ISBN: 0471183873.

[165] Gumbel, E. J. (1961). Bivariate logistic distributions. *Journal of the American Statistical Association* 56 (294), pp. 335–349. DOI: 10.2307/2282259.

[166] Li, J. and Wong, W. K. (2011). Two-dimensional toxic dose and multivariate logistic regression, with application to decompression sickness. *Biostatistics* 12 (1), pp. 143–155. DOI: 10.1093/biostatistics/kxq044.

[167] Nikoloulopoulos, A. K. (2012). Letter to the editor. *Biostatistics* 13 (1), pp. 1–3. DOI: 10.1093/biostatistics/kxr014.

[168] Molenberghs, G. and Lesaffre, E. (1994). Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association* 89 (426), pp. 633–644. DOI: 10.2307/2290866.

[169] Forcina, A. and Dardanoni, V. (2008). Regression models for multivariate ordered responses via the Plackett distribution. *Journal of Multivariate Analysis* 99 (10), pp. 2472–2478. DOI: 10.1016/j.jmva.2008.02.037.

[170] Swihart, B. J., Caffo, B. S., and Crainiceanu, C. M. (2010). A unified approach to modeling multivariate binary data using copulas over partitions. Johns Hopkins University. https://biostats.bepress.com/jhubiostat/paper213.

[171] Panagiotelis, A., Czado, C., and Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association* 107 (499), pp. 1063–1072. DOI: 10.1080/01621459.2012.682850.

[172] Stöber, J., Hong, H. G., Czado, C., and Ghosh, P. (2015). Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses. *Computational Statistics & Data Analysis* 88, pp. 28–39. DOI: 10.1016/j.csda.2015.02.001.

[173] Nooraee, N., Abegaz, F., Ormel, J., Wit, E., and van den Heuvel, E. R. (2016). An approximate marginal logistic distribution for the analysis of longitudinal ordinal data. *Biometrics* 72 (1), pp. 253–261. DOI: 10.1111/biom.12414.

[174] Davis, P. J. (1959). Leonhard Euler's integral: A historical profile of the Gamma function: In memoriam: Milton Abramowitz. *The American Mathematical Monthly* 66 (10), pp. 849–869. DOI: 10.2307/2309786.

[175] Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88 (422), pp. 669–679. DOI: 10.2307/2290350.

[176] Hirk, R., Hornik, K., and Vana, L. (2018). Multivariate ordinal regression models: An analysis of corporate credit ratings. *Statistical Methods & Applications* 88 (3), pp. 507–539. DOI: 10.1007/s10260-018-00437-7.

[177] Hirk, R., Hornik, K., and Vana, L. (2019). mvord: An R package for fitting multivariate ordinal regression models. R package version 0.3.5. https://CRAN.R-project.org/package=mvord.

[178] Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis* 92 (1), pp. 1–28. DOI: 10.1007/s10182-008-0060-7.

[179] Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* 21 (1), pp. 5–42. URL: www3.stat.sinica.edu.tw/statistica/oldpdf/A21n11.pdf.

[180] Gao, X. and Song, P. X.-K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association* 105 (492), pp. 1531–1540. DOI: 10.1198/jasa.2010.tm09414.

[181] Caubet Fernandez, M., Drouin, S., Samoilenko, M., et al. (2019). A Bayesian multivariate latent t-regression model for assessing the association between corticosteroid and cranial radiation exposures and cardiometabolic complications in survivors of childhood acute lymphoblastic leukemia: A PETALE study. *BMC Medical Research Methodology* 19 (1). DOI: 10.1186/s12874-019-0725-9.

[182] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* 90 (430), pp. 773–795. DOI: 10.1080/01621459.1995.10476572.

[183] Jeffreys, H. (1998). *Theory of probability*. 3rd ed. Oxford: Clarendon Press. 459 pp. ISBN: 0198503687.

[184] Wald, A. and Mann, H. B. (1943). On stochastic limit and order relationships. *The Annals of Mathematical Statistics* 14 (3), pp. 217–226.

[185] Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45 (4), pp. 427–437. DOI: 10.1016/j.ipm.2009.03.002.

[186] Altman, D. G. and Bland, J. M. (1994a). Diagnostic tests 1: Sensitivity and specificity. *BMJ* 308 (6943), p. 1552. DOI: 10.1136/bmj.308.6943.1552.

[187] Altman, D. G. and Bland, J. M. (1994b). Diagnostic tests 2: Predictive values. *BMJ* 309 (6947), p. 102. DOI: 10.1136/bmj.309.6947.102.

[188] Japkowicz, N. (2006). Why question machine learning evaluation methods? An illustrative review of the shortcomings of current methods. In: *Evaluation methods for machine learning: Papers from the AAAI workshop* (July 17, 2006). Ed. by C. Drummond. 6. Menlo Park, CA: AAAI Press. ISBN: 978-1-57735-288-4.

[189] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* 27 (8), pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

[190] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval* 1 (1/2), pp. 69–90. DOI: 10.1023/A:1009982220290.

[191] Kazawa, H., Izumitani, T., Taira, H., and Maeda, E. (2004). Maximal margin labeling for multi-topic text categorization. In: *Advances in Neural Information Processing Systems 17*. Ed. by L. K. Saul, Y. Weiss, and L. Bottou. Neural Information Processing Systems Foundation, Inc. Vancouver, BC: MIT Press.

[192] Tsoumakas, G., Katakis, I., and Vlahavas, I. (2010). Mining multi-label data. In: *Data mining and knowledge discovery handbook*. Ed. by O. Z. Maimon and L. Rokach. 2nd ed. New York: Springer. DOI: 10.1007/978-0-387-09823-4_34.

[193] Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology* 25, pp. 111–163. DOI: 10.2307/271063.

[194] Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* 90 (431), pp. 928–934.

[195] Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* 83 (2), pp. 251–266.

[196] Cavanaugh, J. and Neath, A. (1999). Generalizing the derivation of the Schwarz information criterion. *Communications in Statistics - Theory and Methods* 28 (1), pp. 49–66. DOI: 10.1080/03610929908832282.

[197] Kass, R. E., Tierney, L., and Kadane, J. B. (1990). The validity of posterior expansions based on Laplace's method. In: *Essays in Honor of George Bernard*. Ed. by S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner. Amsterdam: North-Holland, pp. 473–488.

[198] Pauler, D. K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika* 85 (1), pp. 13–27.

[199] Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B. (Methodological)* 49 (1), pp. 1–39.

[200] Huzurbazar, V. S. and Jeffreys, H. (1950). Probability distributions and orthogonal parameters. *Mathematical Proceedings of the Cambridge Philosophical Society* 46 (02), pp. 281–284. DOI: 10.1017/S0305004100025743.

[201] Zehna, P. W. (1966). Invariance of maximum likelihood estimators. *The Annals of Mathematical Statistics* 37 (3), p. 744. DOI: 10.1214/aoms/1177699475.

[202] Königsberger, K. (2002). *Analysis 2*. 4th ed. Berlin and Heidelberg: Springer. 461 pp. DOI: 10.1007/978-3-662-05699-8.

[203] Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. 2nd ed. New York, NY: Cambridge University Press. 643 pp. ISBN: 9781139020411.

[204] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6 (2), pp. 461–464.

[205] Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* 91 (435), pp. 1343–1370.

[206] Raftery, A. E. (1999). Bayes factors and BIC. *Sociological Methods & Research* 27 (3), pp. 411–427. DOI: 10.1177/0049124199027003005.

[207] Neath, A. A. and Cavanaugh, J. E. (2012). The Bayesian information criterion: Background, derivation, and applications. *WIREs Computational Statistics* 4 (2), pp. 199–203. DOI: 10.1002/wics.199.

[208] Pauler, D. K., Wakefield, J. C., and Kass, R. E. (1999). Bayes factors and approximations for variance component models. *Journal of the American Statistical Association* 94 (448), pp. 1242–1253. DOI: 10.1080/01621459.1999.10473877.

[209] Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics* 56 (1), pp. 256–262. DOI: 10.1111/j.0006-341X.2000.00256.x.

[210] Debray, T. P. A., Koffijberg, H., Nieboer, D., et al. (2014). Meta-analysis and aggregation of multiple published prediction models. *Statistics in Medicine* 33 (14), pp. 2341–2362. DOI: 10.1002/sim.6080.

[211] Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science* 14 (4), pp. 382–417. DOI: 10.1214/ss/1009212519.

[212] Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science* 19 (1), pp. 81–94. DOI: 10.1214/088342304000000035.

[213] Raftery, A. E. and Zheng, Y. (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association* 98 (464), pp. 931–938. DOI: 10.1198/016214503000000891.

[214] Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* 95 (3), pp. 759–771. DOI: 10.1093/biomet/asn034.

[215] Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica* 22 (2). DOI: 10.5705/ss.2010.216.

[216] Żak-Szatkowska, M. and Bogdan, M. (2011). Modified versions of the Bayesian information criterion for sparse generalized linear models. *Computational Statistics & Data Analysis* 55 (11), pp. 2908–2924. DOI: 10.1016/j.csda.2011.04.016.

[217] Drton, M. and Plummer, M. (2017). A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79 (2), pp. 323–380. DOI: 10.1111/rssb.12187.

[218] Watanabe, S. (2013). A widely applicable Bayesian information criterion. *The Journal of Machine Learning Research* 14 (1), pp. 867–897. URL: http://www.jmlr.org/papers/volume14/watanabe13a/watanabe13a.pdf.

[219] Barber, R. F. and Drton, M. (2015). High-dimensional Ising model selection with Bayesian information criteria. *Electronic Journal of Statistics* 9 (1), pp. 567–607. DOI: 10.1214/15-EJS1012.

[220] Foygel, R. and Drton, M. (2010). Extended bayesian information criteria for Gaussian graphical models. In: *Advances in Neural Information Processing Systems 23*. Ed. by J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta. Neural Information Processing Systems Foundation, Inc. Curran Associates, Inc., pp. 604–612.

[221] Konishi, S., Ando, T., and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika* 91 (1), pp. 27–43. DOI: 10.1093/biomet/91.1.27.

[222] Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine* 30 (25), pp. 3050–3056. DOI: 10.1002/sim.4323.

[223] Mehrjou, A., Hosseini, R., and Nadjar Araabi, B. (2016). Improved Bayesian information criterion for mixture model selection. *Pattern Recognition Letters* 69, pp. 22–27. DOI: 10.1016/j.patrec.2015.10.004.

[224] Lee, E. R., Noh, H., and Park, B. U. (2014). Model selection via Bayesian information criterion for quantile regression models. *Journal of the American Statistical Association* 109 (505), pp. 216–229. DOI: 10.1080/01621459.2013.836975.

[225] Kawano, S. (2014). Selection of tuning parameters in bridge regression models via Bayesian information criterion. *Statistical Papers* 55 (4), pp. 1207–1223. DOI: 10.1007/s00362-013-0561-7.

[226] Luo, S., Xu, J., and Chen, Z. (2015). Extended Bayesian information criterion in the Cox model with a high-dimensional feature space. *Annals of the Institute of Statistical Mathematics* 67 (2), pp. 287–311. DOI: 10.1007/s10463-014-0448-y.

[227] Bayarri, M. J., Berger, J. O., Jang, W., et al. (2019). Prior-based Bayesian information criterion. *Statistical Theory and Related Fields* 3 (1), pp. 2–13. DOI: 10.1080/24754269.2019.1582126.

[228] Hannart, A. and Naveau, P. (2012). An improved Bayesian information criterion for multiple change-point models. *Technometrics* 54 (3), pp. 256–268. DOI: 10.1080/00401706.2012.694780.

[229] Zhang, N. R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* 63 (1), pp. 22–32. DOI: 10.1111/j.1541-0420.2006.00662.x.

[230] Bogdan, M., Ghosh, J. K., and Doerge, R. W. (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* 167 (2), pp. 989–999. DOI: 10.1534/genetics.103.021683.

[231] Bogdan, M., Frommlet, F., Biecek, P., et al. (2008). Extending the modified Bayesian information criterion (mBIC) to dense markers and multiple interval mapping. *Biometrics* 64 (4), pp. 1162–1169. DOI: 10.1111/j.1541-0420.2008.00989.x.

[232] Frommlet, F., Ruhaltinger, F., Twaróg, P., and Bogdan, M. (2012). Modified versions of Bayesian information criterion for genome-wide association studies. *Computational Statistics & Data Analysis* 56 (5), pp. 1038–1051. DOI: 10.1016/j.csda.2011.05.005.

[233] Renard, B. Y., Xu, B., Kirchner, M., et al. (2012). Overcoming species boundaries in peptide identification with Bayesian information criterion-driven error-tolerant peptide search (BICEPS). *Molecular & Cellular Proteomics* 11 (7). DOI: 10.1074/mcp.M111.014167.

[234] Lu, Z. and Zoubir, A. M. (2013). Generalized Bayesian information criterion for source enumeration in array processing. *IEEE Transactions on Signal Processing* 61 (6), pp. 1470–1480. DOI: 10.1109/TSP.2012.2232661.

[235] Huang, L., Xiao, Y., Liu, K., So, H. C., and Zhang, J.-K. (2016). Bayesian information criterion for source enumeration in large-Scale adaptive antenna array. *IEEE Transactions on Vehicular Technology* 65 (5), pp. 3018–3032. DOI: 10.1109/TVT.2015.2436060.

[236] Chen, S. S. and Gopalakrishnan, P. S. (1998). Clustering via the Bayesian information criterion with applications in speech recognition. In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing* (May 12–15, 1998). ICASSP (Conference) and IEEE Signal Processing Society. Piscataway, NJ: IEEE Service Center, pp. 645–648. DOI: 10.1109/ICASSP.1998.675347.

[237] Tian, Y., Wu, J., Wang, Z., and Lu, D. (2003). Fuzzy clustering and Bayesian information criterion based threshold estimation for robust voice activity detection. In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Apr. 6–10, 2003). ICASSP (Conference) and IEEE Signal Processing Society. Piscataway, NJ: IEEE, pp. 444–447. DOI: 10.1109/ICASSP.2003.1198813.

[238] Zhao, Q., Xu, M., and Fränti, P. (2008). Knee point detection on Bayesian information criterion. In: *2008 20th IEEE International Conference on Tools with Artificial Intelligence* (Nov. 3–5, 2008). IEEE, pp. 431–438. DOI: 10.1109/ICTAI.2008.154.

[239] Chou, W. and Reichl, W. (1999). Decision tree state tying based on penalized Bayesian information criterion. In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Mar. 15–19, 1999). ICASSP (Conference) and IEEE Signal Processing Society. Piscataway, NJ: IEEE, 345–348 vol.1. DOI: 10.1109/ICASSP.1999.758133.

[240] Stafylakis, T., Katsouros, V., and Carayannis, G. (2009). Redefining the Bayesian information criterion for speaker diarisation. In: *Proceedings of Interspeech 2009* (Sept. 6–10, 2009). Vol. 2009. International Speech Communication Association, pp. 1051–1054.

[241] Stafylakis, T., Katsouros, V., and Carayannis, G. (2010). The segmental Bayesian information criterion and its applications to speaker diarization. *IEEE Journal of Selected Topics in Signal Processing* 4 (5), pp. 857–866. DOI: 10.1109/JSTSP.2010.2048656.

[242] National Center for Biotechnology Information and U.S. National Library of Medicine (2019). *PubMed*. URL: https://www.ncbi.nlm.nih.gov/pubmed/ (visited on 09/13/2019).

[243] Venables, W. N. and Ripley, B. D. (2010). *Modern applied statistics with S*. 4th ed. New York, NY: Springer. 495 pp. ISBN: 978-0-387-21706-2.

[244] R Core Team (2019). R: A language and environment for statistical computing. Version 3.6.1. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

[245] van Belle, G. and Fisher, L. (2004). *Biostatistics: A methodology for the health sciences*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc. 871 pp. DOI: 10.1002/0471602396.

[246] Grogan, T. R. and Elashoff, D. A. (2017). A simulation based method for assessing the statistical significance of logistic regression models after common variable selection procedures. *Communications in Statistics - Simulation and Computation* 46 (9), pp. 7180–7193. DOI: 10.1080/03610918.2016.1230216.

[247] Akaike, H. (1982). Prediction and entropy. University of Wisconsin-Madison.

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

[249] Hastie, T. J., Wainwright, M., and Tibshirani, R. J. (2015). *Statistical learning with Sparsity: The lasso and generalizations*. Vol. 143. Monographs on statistics and applied probability. Boca Raton: CRC Press LLC. 362 pp. ISBN: 978-1498712163.

[250] DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 44 (3), p. 837. DOI: 10.2307/2531595.

[251] Wood, A. M., White, I. R., and Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine* 27 (17), pp. 3227–3246. DOI: 10.1002/sim.3177.

[252] Meng, X.-L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* 79 (1), p. 103. DOI: 10.2307/2337151.

[253] Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley series in probability and mathematical statistics Applied probability and statistics. New York, NY: Wiley. ISBN: 047108705X.

[254] Schafer, J. L. (1999). *Analysis of incomplete multivariate data*. 1st ed. Vol. 72. Boca Raton, Fl: Chapman and Hall/CRC. 430 pp. ISBN: 0412040611.

[255] Robin, X., Turck, N., Hainard, A., et al. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12, p. 77. DOI: 10.1186/1471-2105-12-77.

[256] Lotan, T. L. and Epstein, J. I. (2010). Clinical implications of changing definitions within the Gleason grading system. *Nature reviews. Urology* 7 (3), pp. 136–142. DOI: 10.1038/nrurol.2010.9.

[257] Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating*. 2nd ed. 2019. Statistics for biology and health. Cham, Switzerland: Springer International Publishing. 558 pp. DOI: 10.1007/978-3-030-16399-0.

[258] Carpenter, B., Gelman, A., Hoffman, M. D., et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software* 76 (1). DOI: 10.18637/jss.v076.i01.

[259] Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B* 195 (2), pp. 216–222. DOI: 10.1016/0370-2693(87)91197-X.

[260] Neal, R. M. (1994). An improved acceptance procedure for the hybrid Monte Carlo algorithm. *Journal of Computational Physics* 111 (1), pp. 194–203. DOI: 10.1006/jcph.1994.1054.

[261] Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* 15, pp. 1593–1923. URL: http://jmlr.org/papers/v15/hoffman14a.html.

[262] Gelman, A., Carlin, J. B., Stern, H. S., et al. (2014). *Bayesian data analysis*. 3rd ed. Chapman and Hall / CRC. 667 pp. ISBN: 9781439840955.

[263] Gronau, Q. F., Sarafoglou, A., Matzke, D., et al. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology* 81, pp. 80–97. DOI: 10.1016/j.jmp.2017.09.005.

[264] Gronau, Q. F. and Singmann, H. (2019). bridgesampling: Bridge sampling for marginal likelihoods and Bayes factors. R package version 0.7-2. https://CRAN.R-project.org/package=bridgesampling.

[265] Stan Development Team (2019). RStan: the R interface to Stan. R package version 2.19.2. http://mc-stan.org/.

[266] Sanderson, C. and Curtin, R. (2016). Armadillo: A template-based C++ library for linear algebra. *The Journal of Open Source Software* 1 (2), p. 26. DOI: 10.21105/joss.00026.

[267] Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. Vol. 64. Use R! New York, NY: Springer. 236 pp. DOI: 10.1007/978-1-4614-6868-4.

[268] Eddelbuettel, D. and Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data Analysis* 71, pp. 1054–1063. DOI: 10.1016/j.csda.2013.02.005.

[269] Li, Y. and Ghosh, S. K. (2015). Efficient sampling methods for truncated multivariate normal and Student-t distributions subject to linear inequality constraints. *Journal of Statistical Theory and Practice* 9 (4), pp. 712–732. DOI: 10.1080/15598608.2014.996690.

[270] Genz, A., Bretz, F., Miwa, T., et al. (2019). mvtnorm: Multivariate normal and t. R package version 1.0-11. http://CRAN.R-project.org/package=mvtnorm.

[271] Wilhelm, S. and Manjunath, B. G. (2015). tmvtnorm: Truncated multivariate normal and Student-t distribution. R package version 1.4-10. http://CRAN.R-project.org/package=tmvtnorm.

[272] Geweke, J. (1991). Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints and the evaluation of constraint probabilities.

In: *Computing science and statistics: Proceedings of the 23rd symposium on the interface.* (Apr. 21–24, 1991). Ed. by E. Keramidas. Fairfax Station, VA: Interface Foundation of North America.

[273] Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing* 5 (2), pp. 121–125. DOI: 10.1007/BF00143942.

[274] Ma, T. F., Ghosh, S. K., and Li, Y. (2018). tmvmixnorm: Sampling from truncated multivariate normal and t. R package version 1.0.2. https://CRAN.R-project.org/package=tmvmixnorm.

[275] Mehrotra, S., Li, Y., and Ghosh, S. K. (2018). tmvn: Generate random samples from truncated univariate and multivariate normal. https://github.com/suchitm/tmvn.

[276] Botev, Z. I. (2017). The normal law under linear restrictions: Simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79 (1), pp. 125–148. DOI: 10.1111/rssb.12162.

[277] Hasselman, B. (2018). nleqslv: Solve systems of nonlinear equations. R package version 3.3.2. https://CRAN.R-project.org/package=nleqslv.

[278] Botev, Z. and Belzile, L. (2019). TruncatedNormal: Truncated multivariate normal and Student distributions. R package version 2.1. https://CRAN.R-project.org/package=TruncatedNormal.

[279] Koch, H. and Bopp, G. P. (2019). Fast and exact simulation of multivariate normal and Wishart random variables with box constraints. https://arxiv.org/abs/1907.00057 (visited on 11/24/2019).

[280] Koch, H. (2019). cdist. R package version 1.0. https://github.com/hillarykoch/cdists (visited on 11/24/2019).

[281] Hardt, N. and Ahmadou, D. (2013). Faster multivariate normal densities with RcppArmadillo and OpenMP. https://gallery.rcpp.org/articles/dmvnorm_arma/.

[282] Selig, K. (2019). multlogreg: Multivariate Logistic Regression Models. R Version 1.0. https://github.com/kathsel/multlogreg.

[283] Wang, L. E., Shaw, P. A., Mathelier, H. M., Kimmel, S. E., and French, B. (2016). Evaluating risk-prediction models using data from electronic health records. *The Annals of Applied Statistics* 10 (1), pp. 286–304. DOI: 10.1214/15-AOAS891.

Appendix

A Data description and exploratory data analysis

A.1 List of variables

Pathological outcome	Variable name
ECE	ece
SVI	svi
LVI	lvi
LNI	lni
PGG	pgg

Table A.1 Pathological outcomes and corresponding variable name in data.

Covariates	Variable name
Age at prostatectomy	age_at_surgery
BMI at last biopsy	bx_last_bmi
Number of biopsies	bx_count
Last biopsy result	bx_last_result
Percentage positive cores at last biopsy	bx_last_cores_pos_perc
Number of positive cores at last biopsy	bx_last_num_cores_pos
Number of cores at last biopsy	bx_last_num_cores_total_with_mm
Primary Gleason grade at last biopsy	bx_last_gleason_max_prim
Secondary Gleason grade at last biopsy	bx_last_gleason_max_sec
Prostate volume at last biopsy	bx_last_volumen_prostate

Table A.2 Patient characteristics and biopsy related covariates with the corresponding variable name in data.

Covariates	Variable name
Egawa (PSADT)	dt_slope_egawa
MSKCC (PSADT)	dt_slope_mskcc
Sengupta (PSADT)	dt_slope_sengupta
Stephenson (PSADT)	dt_slope_stephenson
D'Amico (PSAV)	v_damico
MSKCC (PSAV)	v_mskcc
Sengupta (PSAV)	v_sengupta
Thompson (PSAV)	v_thompson
D'Amico (PSAV cut off)	damico_cut

Table A.3 PSA dynamics covariates with the corresponding variable name in data.

Covariates	Variable name
Number of PSA values	psa_count_$*$
last $\log(\text{PSA})$	psa_ln_pre_op
mean $\log(\text{PSA})$	psa_ln_mean_$*$
standard deviation $\log(\text{PSA})$	psa_ln_sd_$*$
minimum $\log(\text{PSA})$	psa_ln_min_$*$
maximum $\log(\text{PSA})$	psa_ln_max_$*$
last PSA	psa_pre_op
mean PSA	psa_mean_$*$
standard deviation PSA	psa_sd_$*$
minimum PSA	psa_min_$*$
maximum PSA	psa_max_$*$

Table A.4 Simple summary statistics applied to PSA values available before prostatectomy and the corresponding variable name in the data. $*$ can be replaced by all, 1y, 2y, 3y indicating all, values within 1 year, 2 years or 3 years before prostatectomy are used for calculation.

Covariates	Variable name
$\log$(PSA) linear regression, 1st coefficient	psa_lm_ln_coeff1_$*$
$\log$(PSA) linear regression, 2nd coefficient	psa_lm_ln_coeff2_$*$
$\log$(PSA) polynomial regression degree 2, 1st coefficient	psa_lm_ln_time2_coeff1_$*$
$\log$(PSA) polynomial regression degree 2, 2nd coefficient	psa_lm_ln_time2_coeff2_$*$
$\log$(PSA) polynomial regression degree 2, 3rd coefficient	psa_lm_ln_time2_coeff3_$*$
$\log$(PSA) polynomial regression degree 3, 1st coefficient	psa_lm_ln_time3_coeff1_$*$
$\log$(PSA) polynomial regression degree 3, 2nd coefficient	psa_lm_ln_time3_coeff2_$*$
$\log$(PSA) polynomial regression degree 3, 3rd coefficient	psa_lm_ln_time3_coeff3_$*$
$\log$(PSA) polynomial regression degree 3, 4th coefficient	psa_lm_ln_time3_coeff4_$*$
PSA linear regression, 1st coefficient	psa_lm_coeff1_$*$
PSA linear regression, 2nd coefficient	psa_lm_coeff2_$*$
PSA polynomial regression degree 2, 1st coefficient	psa_lm_time2_coeff1_$*$
PSA polynomial regression degree 2, 2nd coefficient	psa_lm_time2_coeff2_$*$
PSA polynomial regression degree 2, 3rd coefficient	psa_lm_time2_coeff3_$*$
PSA polynomial regression degree 3, 1st coefficient	psa_lm_time3_coeff1_$*$
PSA polynomial regression degree 3, 2nd coefficient	psa_lm_time3_coeff2_$*$
PSA polynomial regression degree 3, 3rd coefficient	psa_lm_time3_coeff3_$*$
PSA polynomial regression degree 3, 4th coefficient	psa_lm_time3_coeff4_$*$

Table A.5 Covariates extracted from regression models applied to PSA values available before prostatectomy. $*$ can be replaced by all, 1y, 2y, 3y indicating all, values within 1 year, 2 years or 3 years before prostatectomy are used for calculation.

A.2 Additional exploratory data analysis

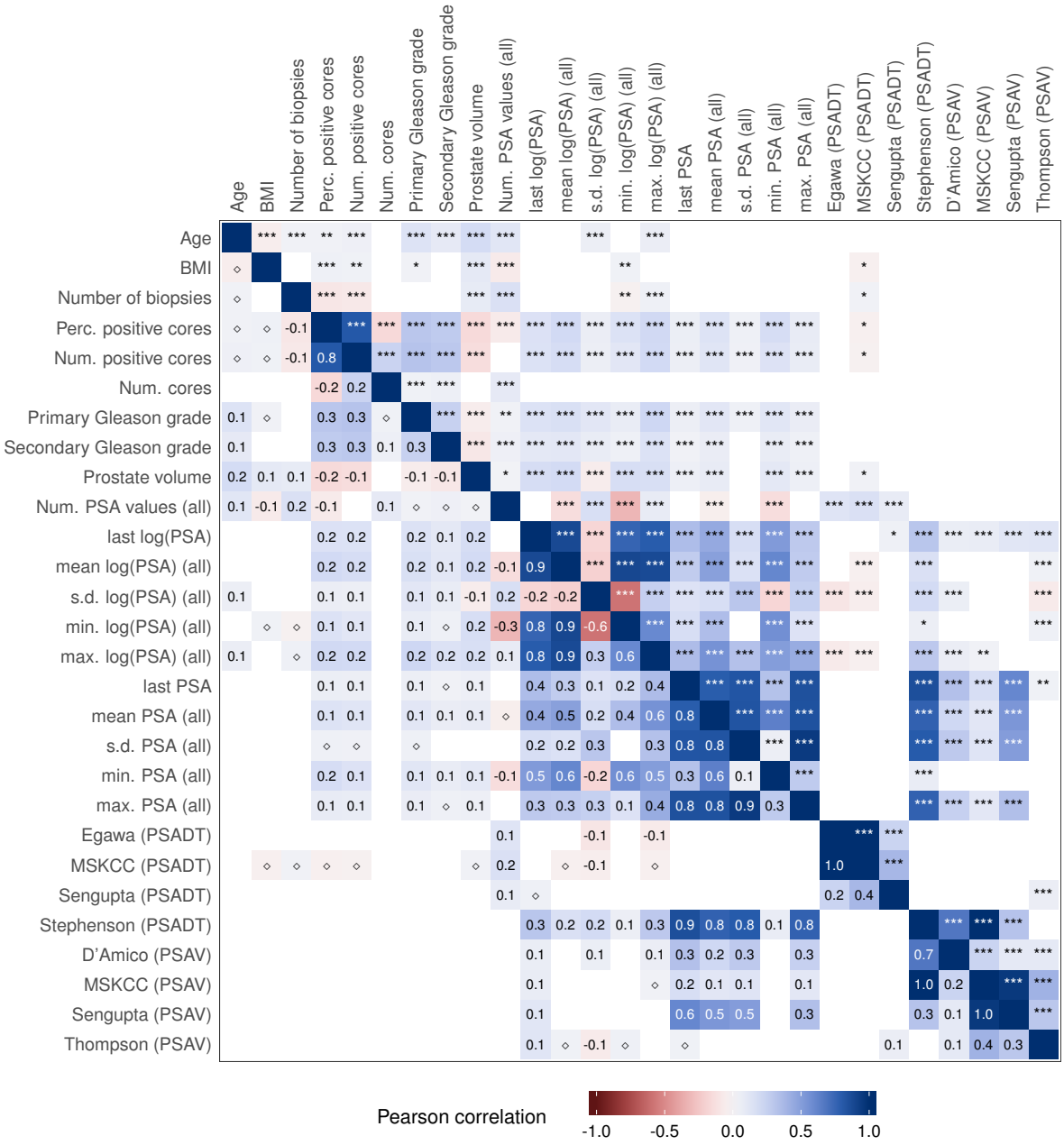

Figure A.1 Pearson correlation coefficients for continuous covariates. (all) indicates all PSA values before prostatectomy were used, $***$ indicates a p-value < 0.001, $**$ < 0.01, and $*$ < 0.05 after adjusting for 8,385 multiple comparisons with the Bonferroni method. White color indicates no significant correlation, $\diamond$ correlation coefficients with an absolute value smaller than 0.05.

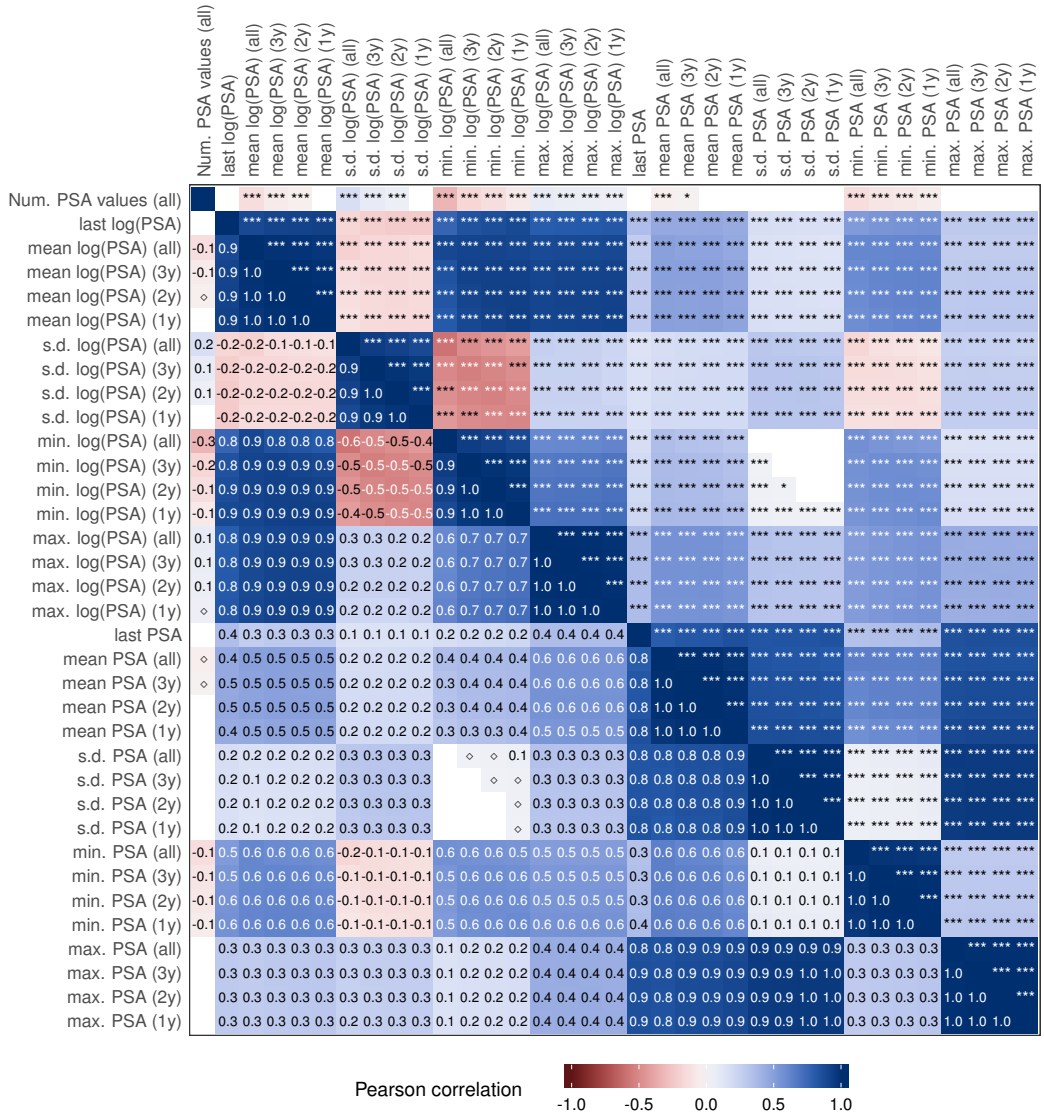

Figure A.2 Pearson correlation coefficient of PSA related covariates for different time periods. (all), (1y), (2y), or (3y) indicate all PSA values or those within one, two or three years of prostatectomy are used. $***$ indicates a p-value < 0.001, $** < 0.01$, and $* < 0.05$ after adjusting for 8,385 multiple comparisons with the Bonferroni method. White color indicates no significant correlation, $\diamond$ correlation coefficients with an absolute value smaller than 0.05.

B Approximating the Bayes Factor for univariate logistic regression models

B.1 Simulation study results for the second scenario

In this section, we discuss the results of the second scenario for the simulation studies performed in Section 4.2. In Scenario 2 we use one covariate that is sampled from the $\log(\text{PSA})$-values of the EMR data.

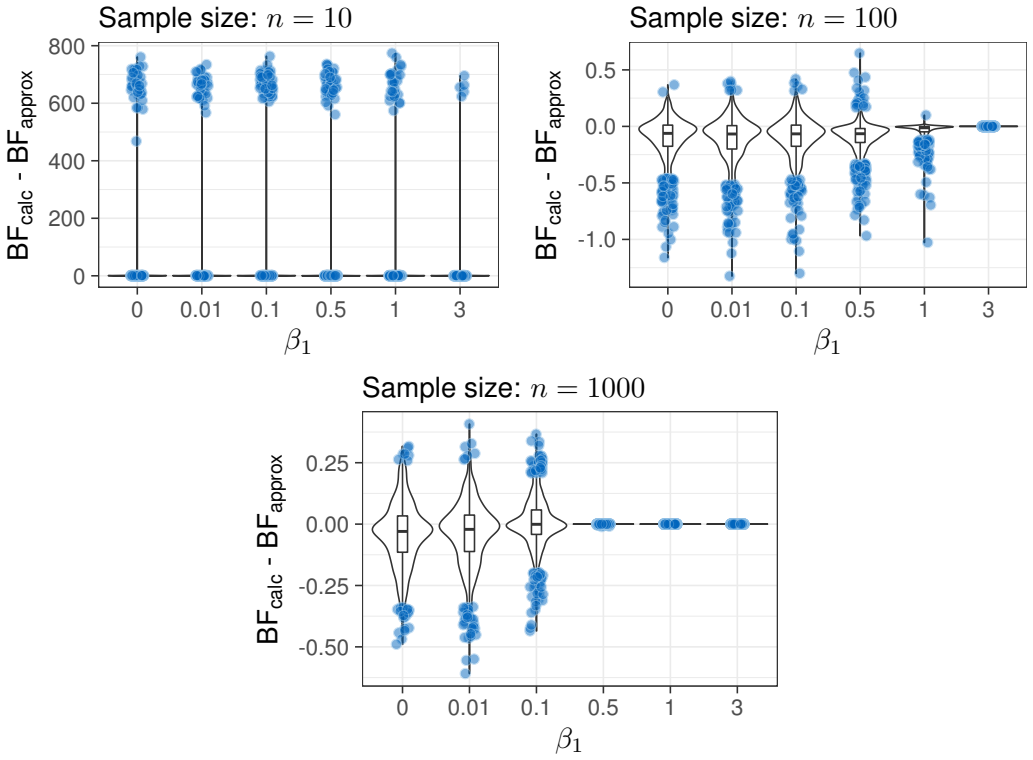

Figure B.1 Differences between BF_{calc} and $\text{BF}_{\text{approx}}$ for the second scenario of the simulation study for logistic regression models with sample sizes $n = 10$, $n = 100$, and $n = 1000$ based on 6,000 simulations each.

Similar to Scenario 1 we see in Figure B.1 that for an increasing sample size the difference between the $\text{BF}_{\text{approx}}$ and BF_{calc} decreases. Overall it is larger than for the discrete case (Fig. 4.3). For a small sample size of $n = 10$, we detect 202 outlying values (3.4%) with an absolute difference between $\text{BF}_{\text{approx}}$ and BF_{calc} greater than 2. These outliers occur when $Y_i = 1$ for all $i \in \{1, \ldots, 10\}$ and we exclude them in Figure B.2 and the following analysis.

Figure B.2 shows that in general the absolute difference between BF_{approx} and BF_{calc} is smaller than 0.5, but there are several outlying values up to 1 for all cases of β_1. The distribution of the difference between BF_{approx} and BF_{calc} is symmetric, whereas in Scenario 1 it is left skewed

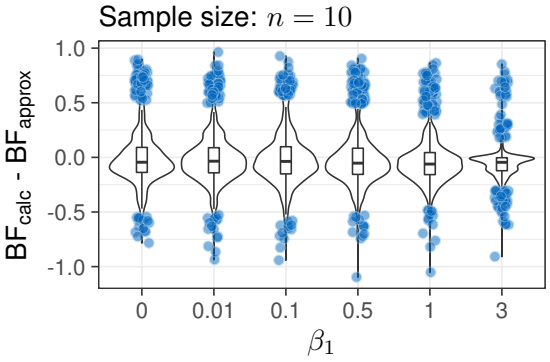

Figure B.2 Difference between calculated BF_{approx} and BF_{calc} for the second scenario of the simulation study for logistic regression models with a sample size of $n = 10$ excluding 202 outlying cases.

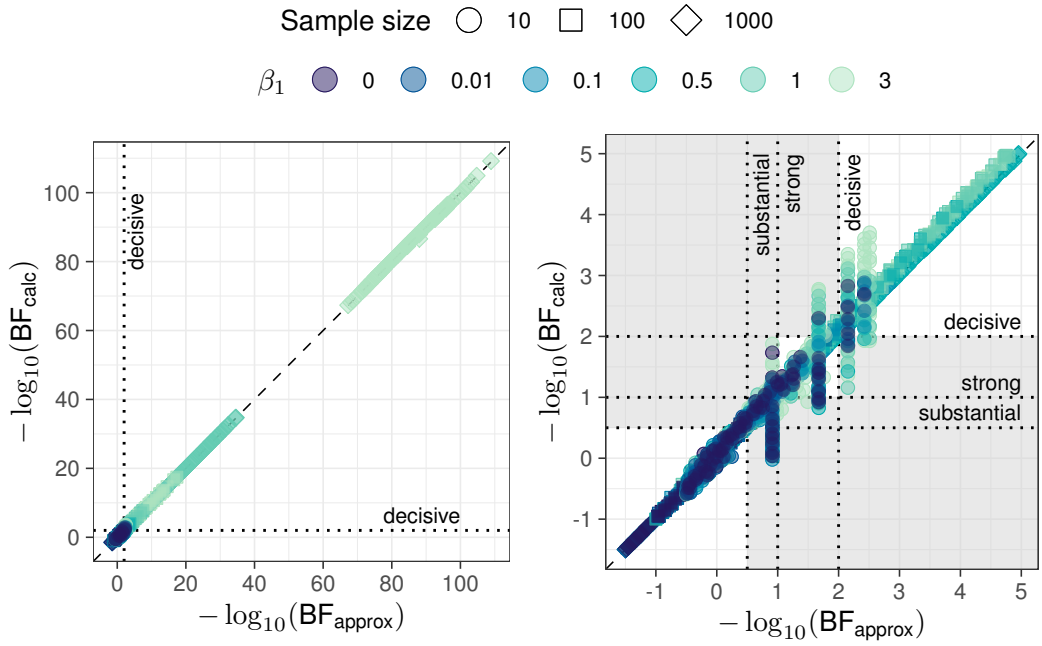

Figure B.3 Comparison of BF_{approx} to BF_{calc} on a negative $\log_{10}$ - scale using Jeffreys' rule for the second scenario. On the left side all simulation runs are shown, whereas on the right side the focus is on the range between -1.5 and 5 on a negative $\log_{10}$-scale.

We plot BF_{approx} against BF_{calc}, on a negative $\log_{10}$-scale and classify the evidence against the null hypothesis according to Jeffreys' rule into substantial, strong, and decisive (Fig. B.3). Overall $-\log_{10}(BF_{approx})$ and $-\log_{10}(BF_{calc})$ agree well, but the classification differs in 633 cases located in one of the gray areas. Stratified by sample size the misclassification rates are 8.3% (484/5,798) for $n = 10$, 2.4% (146/6,000) for $n = 100$, and 0.1% (3/6,000) for $n = 1000$.

When investigating the 484 misclassified cases for $n = 10$ in more detail, we find that in 200 cases the minimum $\log(\text{PSA})$-value of one group is larger than the maximum $\log(\text{PSA})$-value of the other group and the data are completely separated as defined in Section 3.1. For 192 misclassified cases for $n = 10$,we observe an quasicomplete separation in the data in the sense that the two groups barely overlap and only one observation of one group has a $\log(\text{PSA})$-value within the range of the other group. Figure B.4 displays this quasicomplete separation exemplified for five simulations. The assumptions for Laplace regularity are not fulfilled and the approximation fails in these cases. With an increasing sample size, no such outliers occur in the simulations. Similar to Scenario 1, we exclude these 392 cases, as in medical applications, statistical analysis would not proceed in the presence of small samples with unstable effects.

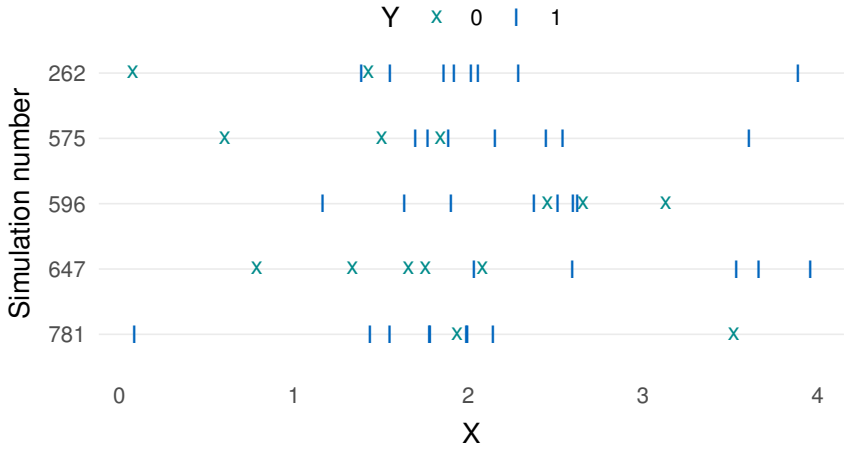

Figure B.4 Quasicomplete separation of the continuous data with a binary outcome, where only one observation of on group falls within the range of the observations of the second group. Five simulations of $n = 10$ are displayed.

For the remaining 92 misclassified cases for $n = 10$ as well as the 146 and 3 misclassified cases for $n = 100$ and $n = 1000$, misclassification occurs on the boundaries of Jeffreys' rule. The median, minimum, and maximum difference $-\log_{10}(\text{BF}_{\text{approx}}) - (-\log_{10}(\text{BF}_{\text{calc}}))$ for these cases are -0.12, -0.29, and -0.01 respectively. Thus, $\text{BF}_{\text{approx}}$ underestimates BF_{calc} in all misclassified cases for Scenario 2. We do not exclude these 241 misclassified cases.

We compare $\text{BF}_{\text{approx}}$ to the p-value obtained from the Wald test for the coefficient of the covariate on a negative $\log_{10}$-scale. Again, we use a negative $\log_{10}$-scale and classify $-\log_{10}(\text{BF}_{\text{approx}})$ according to Jeffreys' rule. We consider 3 different significance levels for the p-value, 0.05, 0.01, and 0.001, which correspond to $\approx$ 1.3, 2, and 3 on the negative $\log_{10}$-scale.

On the left side in Figure B.5, we see for a large sample size and large β_1-value both the p-value and $\text{BF}_{\text{approx}}$ are highly significant ($p < 0.001$) and decisive, respectively. In

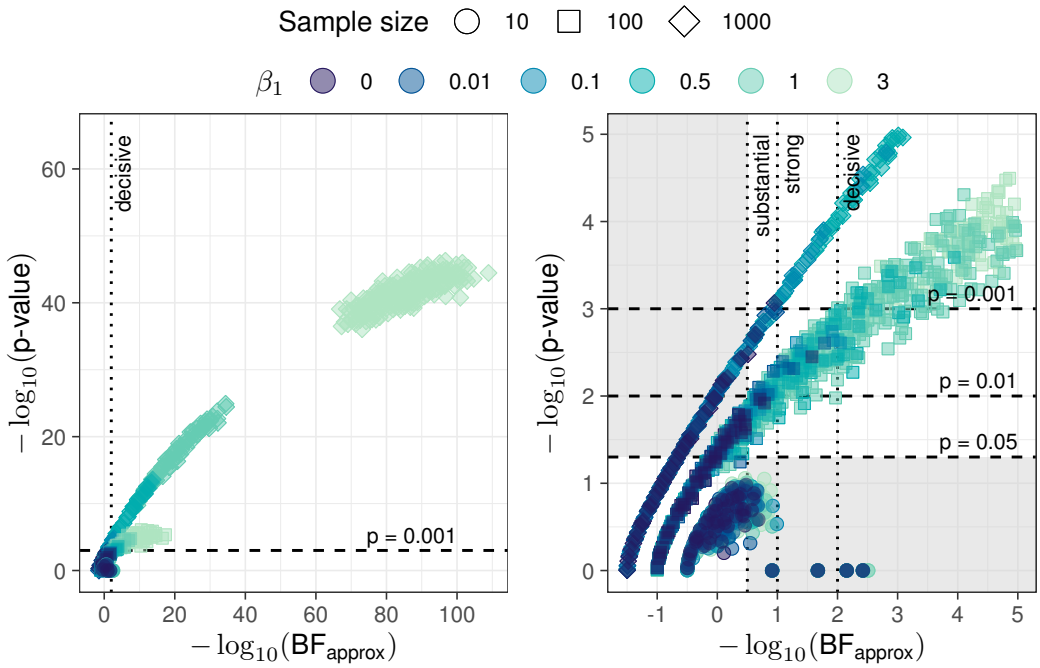

Figure B.5 Comparison of $\text{BF}_{\text{approx}}$ and the p-value on a negative $\log_{10}$ - scale for the second scenario. On the left side all simulation runs are shown, whereas on the right side the focus is on the range between -1.5 and 5 on a negative $\log_{10}$-scale.

contrast to Scenario 1 we do not detect outlying cases for $n = 100$, probably due to the continuous covariate. On the right side in Figure B.5 we observe cases for $n = 10$, where the p-value is large, but the $\text{BF}_{\text{approx}}$ provides substantial, strong or decisive evidence against the null-hypothesis. Those 334 cases out of 4,525 remaining simulations for $n = 10$ (7.4 %) occur when complete separation of the data is possible as described above.

We detect that for larger sample sizes of $n = 100$ or $n = 1000$ the BF is more conservative than the p-value, whereas for a sample size of $n = 10$ in some cases the BF provides substantial or strong evidence against the null hypothesis, but the p-value is not significant (lower right gray area).

For $n = 100$ and $n = 1000$ we split the right side in Figure B.5 and display the results separated by samples size and β_1-value in Figure B.6 for better visibility.

For $n = 1000$ and $\beta_1 = 1, 3$ no values fall within the displayed range and for $n = 100$ and $\beta_1 = 3$ all p-values are less than 0.01 and $\text{BF}_{\text{approx}}$-values indicate at least strong evidence against the null hypothesis (Fig.B.6, bottom row). We expect these results as for these parameter setting evidence against the null hypothesis should be provided.

For $\beta_1 \leq 0.5$ we observe a similar shift in p-values as described for Scenario 1 in Section 4.2. For a sample size of $n = 1000$ compared to $n = 100$ for similar $-\log_{10}(\text{BF}_{\text{approx}})$-values the p-value tends to be smaller, and thus larger on a negative $\log_{10}$-scale for the larger

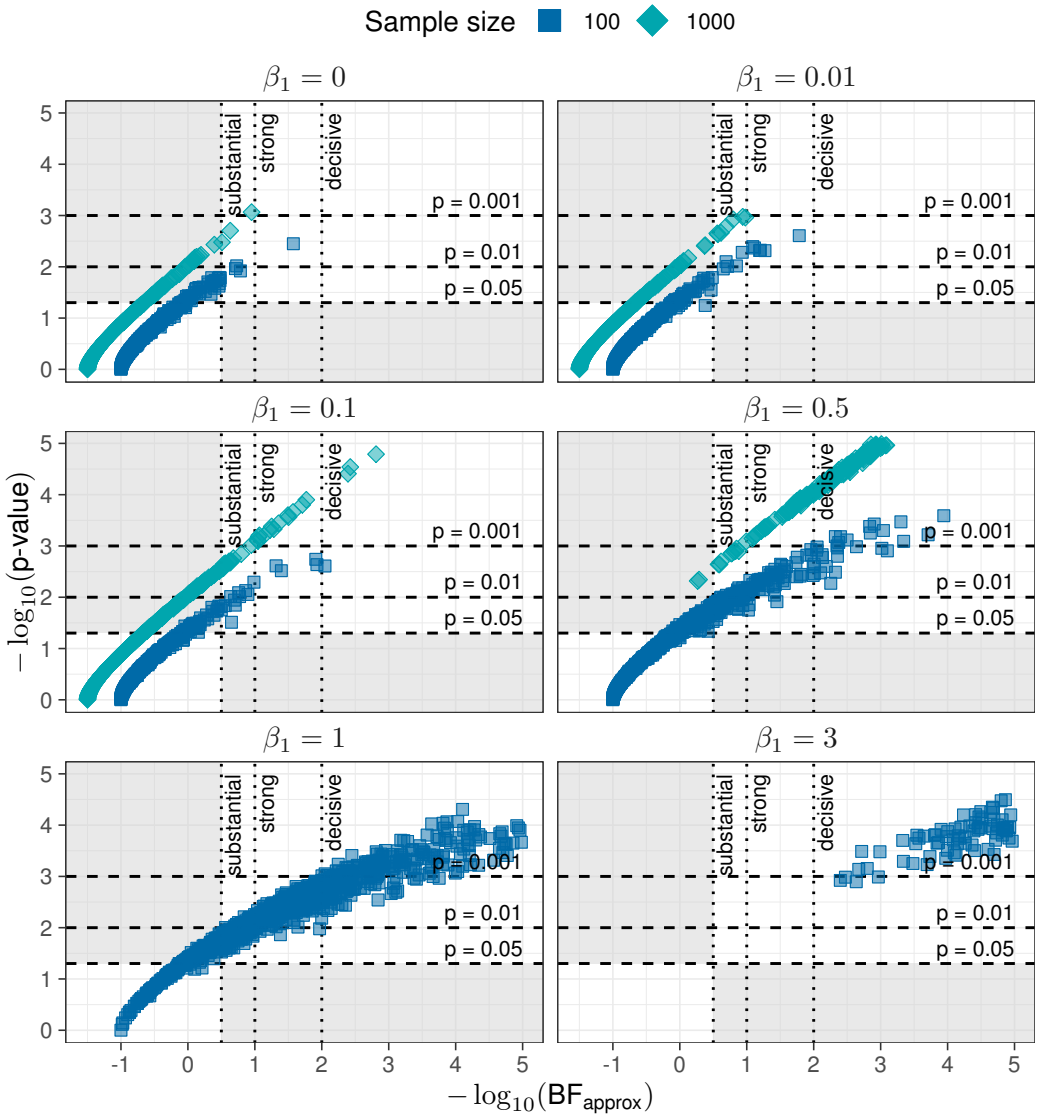

Figure B.6 Comparison of $\text{BF}_{\text{approx}}$ and the p-value on a negative $\log_{10}$ - scale for the second scenario for $n = 100$ and $n = 1000$ with focus on the range between -1.5 and 5 on a negative $\log_{10}$-scale.

sample size. Therefore, for comparable evidence against the null hypothesis based on $\text{BF}_{\text{approx}}$, the p-value yields more significant results for a larger sample size. Moreover, we detect several cases, where $\text{BF}_{\text{approx}}$ does not provide evidence against the null hypothesis with $-\log_{10}(\text{BF}_{\text{approx}}) < 0.5$, but the Wald test yields p-values at least smaller than 0.05. These disagreeing cases are located in the upper gray areas. Stratified by sample size the disagreement proportions are 7.3% (441/6,000) for $n = 100$ and 4.7% (281/6,000) for $n = 1000$. For $n = 10$ in 81 cases out of 4,191 (1.9%) $\text{BF}_{\text{approx}}$ indicates evidence against the null hypothesis, but the p-value is larger than 0.05.

B.2 Results for AIC and LASSO$_{max}$

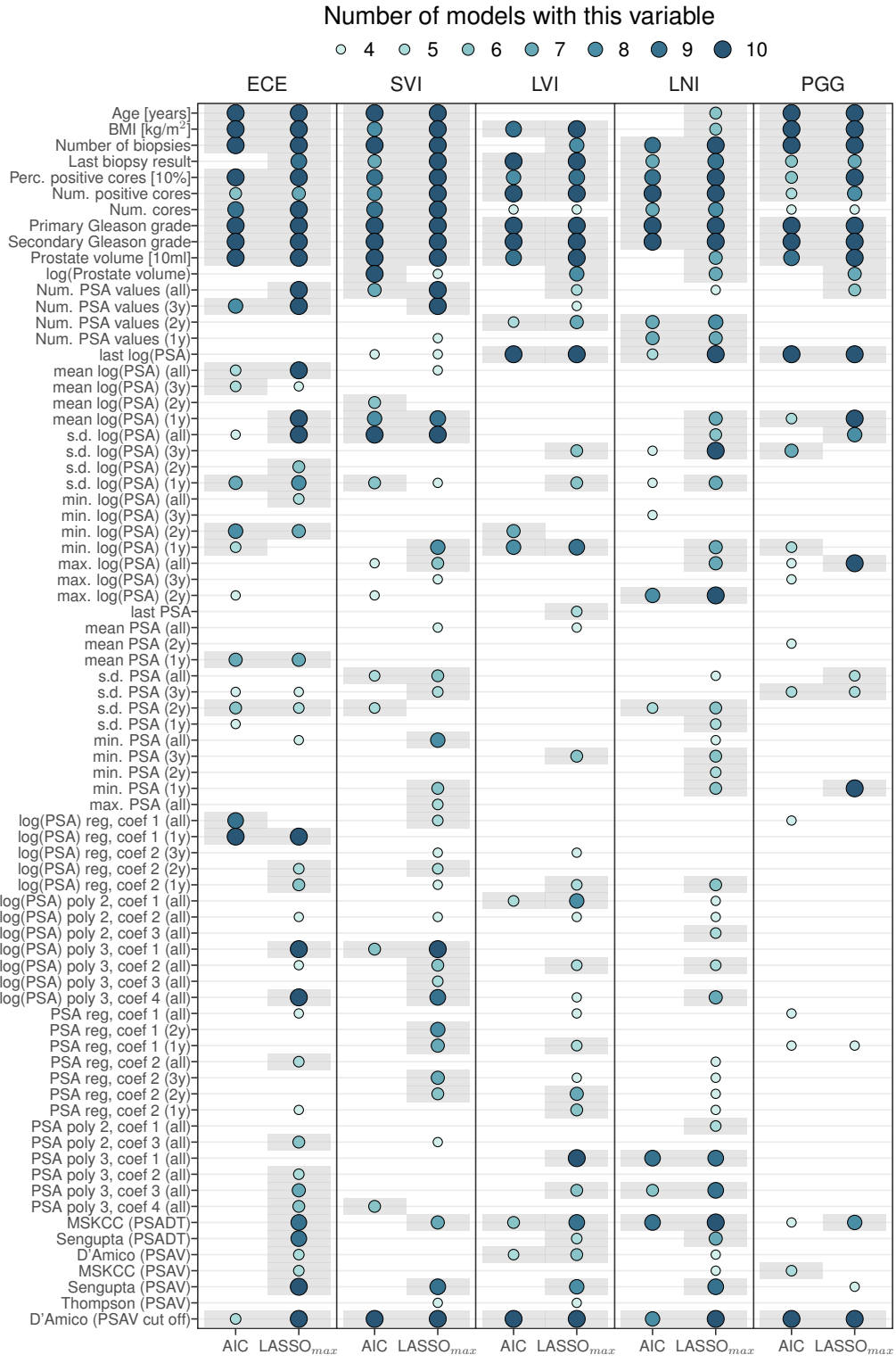

Figure B.7 Variables selected in at least 4 of the 10 imputation sets ($n = 16,427$) using AIC or the LASSO$_{max}$ method. Size and color indicate the number of imputation sets on which the specific variable was selected. Grey backgrounds indicate variables selected in at least 50% of the imputed sets.

Figure B.7 displays the covariates selected by AIC and LASSO$_{max}$. Due to the large number of variables selected, we only show those that were included in a model at least 4 times. Overall the number of included covariates is much larger compared to the BIC and LASSO$_{1se}$.

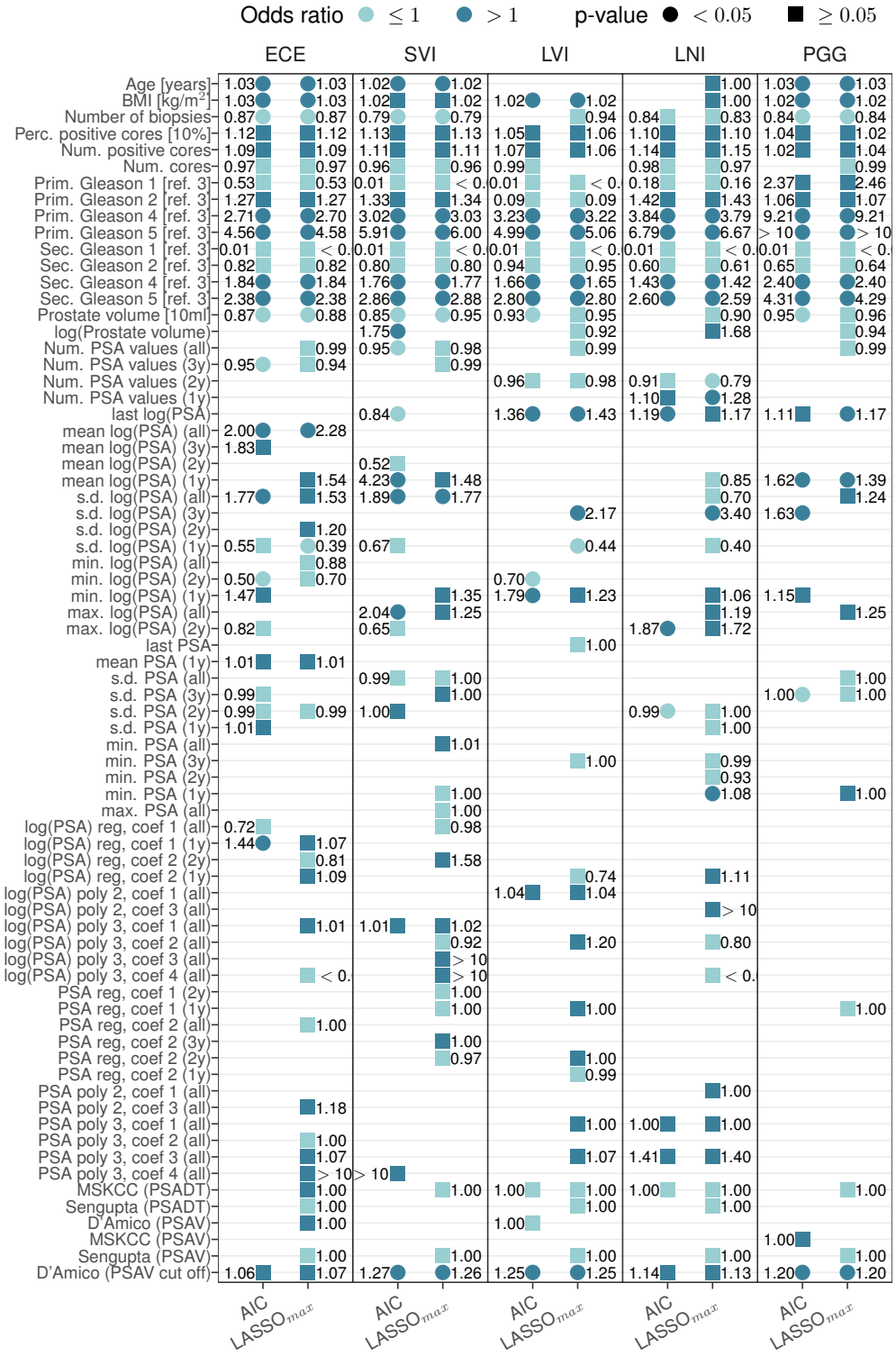

Figure B.8 Odds ratios for models selected by the AIC and LASSO$_{max}$, with magnitude greater than 1 indicated by dark blue and significance at a 0.05 level, by circles.

B.3 Stan model specification

The exact specification in Stan for models M_0 and M_1 for LNI introduced Section 4.3 are given below.

model0.stan

```
data {
  // Number of observations
  int<lower=0> N;
  // Number of parameters
  int<lower=0> K;
  // Responses
  int<lower=0,upper=1> lni[N];
  // Design matrix
  matrix [N, K] X;
  // Priors on regression coefficients
  vector[K] scale_beta;
}

parameters {
  // Coefficients
  vector[K] beta;
}

model {
  // priors
  beta ~ normal(0., scale_beta);
  // likelihood
  lni ~ bernoulli_logit(X * beta);
}
```

model1.stan

```
functions {
  // Calculate the inverse Hessian for
  // the unit information prior
  real inv_hes(int N, vector W, vector dat_vec) {
   real hessian_inverse;
   hessian_inverse = 0;
   for (n in 1:N) {
    hessian_inverse += - dat_vec[n] * W[n] * dat_vec[n];
   }
   return sqrt(- N * inv(hessian_inverse));
  }
}

data {
  // Number of observations
  int<lower=0> N;
  // Number of parameters
  int<lower=0> K;
  // Responses
  int<lower=0,upper=1> Ini[N];
  // Design matrix
  matrix [N, K] X;
  // Priors on regression coefficients
  vector[K] scale_beta;
  // Vector for number of positive cores
  vector[N] num_pos_cores;
}

parameters {
  // Coefficients
  vector[K] beta;
  real b_num_pos_cores;
}
```

```
model {
  vector[N] xb;
  vector[N] W;
  // priors
  beta ~ normal(0., scale_beta);
  xb = X * beta;
  for (n in 1:N) {
   W[n] = inv_logit(xb[n]) * (1 - inv_logit(xb[n]));
  }
  b_num_pos_cores ~ normal(0, inv_hes(N, W, num_pos_cores));
  // likelihood
  lni ~ bernoulli_logit(X * beta +
                        b_num_pos_cores * num_pos_cores);
}
```

C Bayesian multivariate logistic regression model

C.1 Sampling values for the coefficients related to LVI, LNI, and PGG

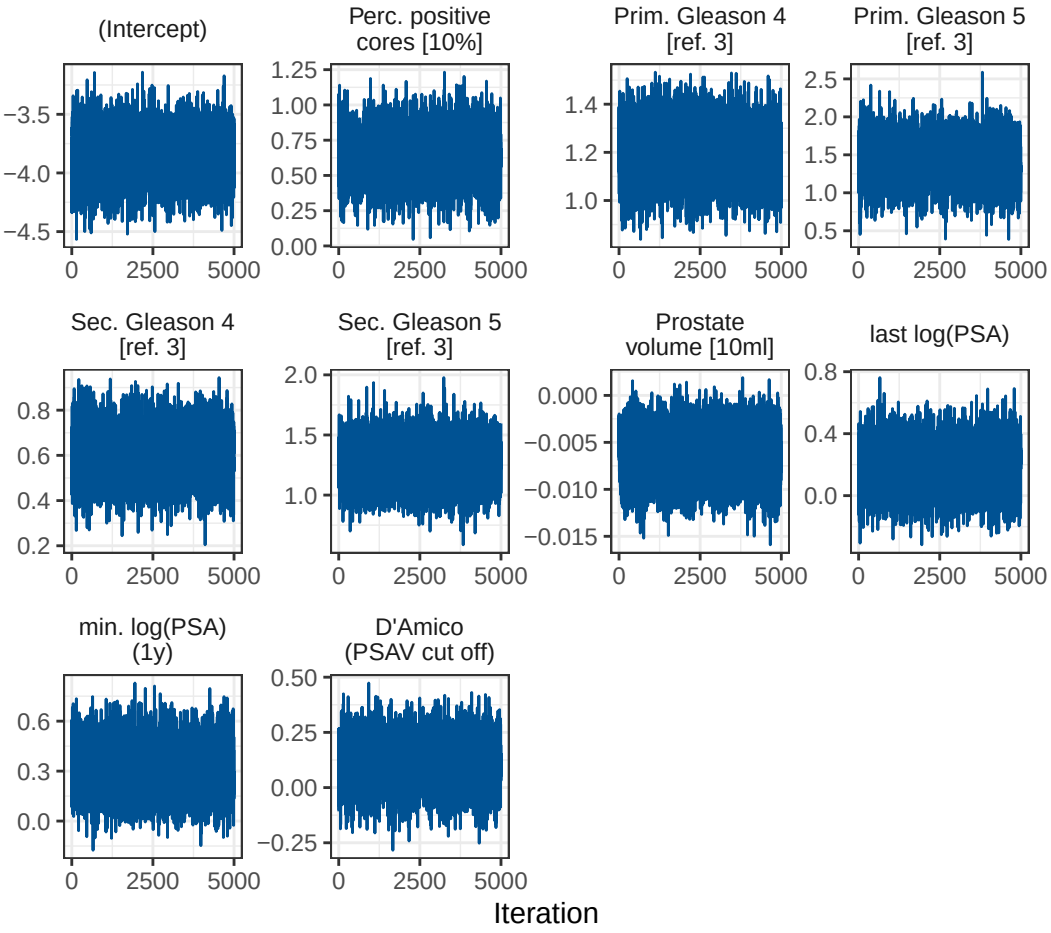

Figure C.1 MCMC samples for the log OR for LVI based on 5,000 samples after thinning of 20 and 1,000 bun-in samples.

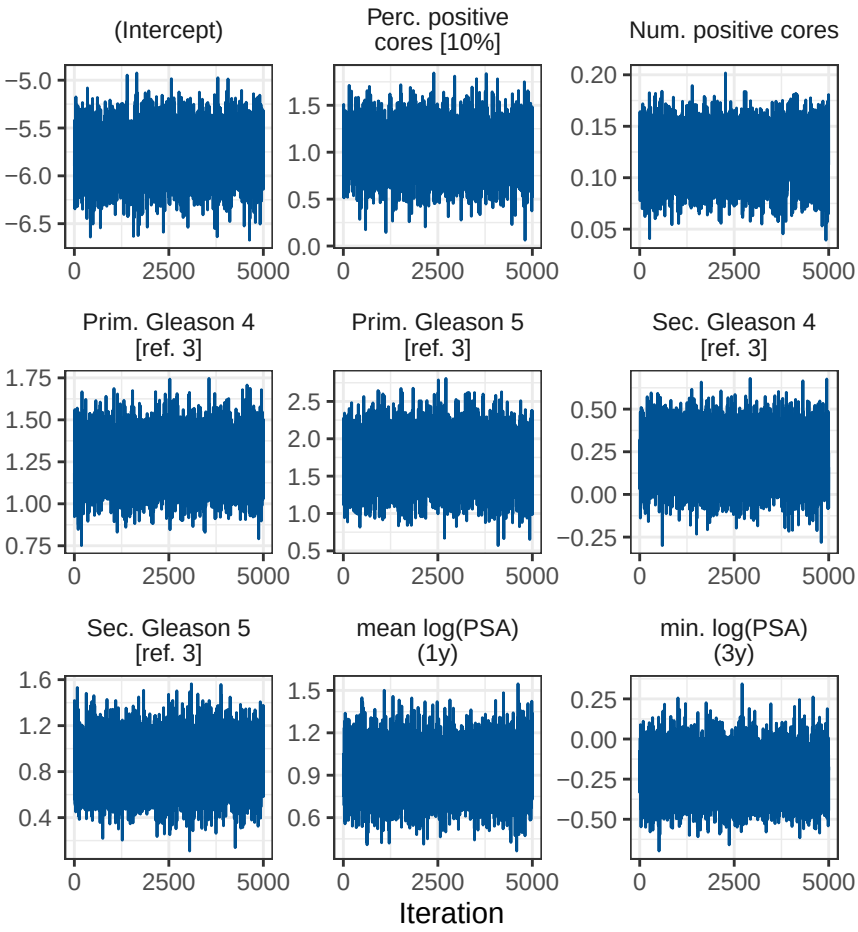

Figure C.2 MCMC samples for the log OR for LNI based on 5,000 samples after thinning of 20 and 1,000 bun-in samples.

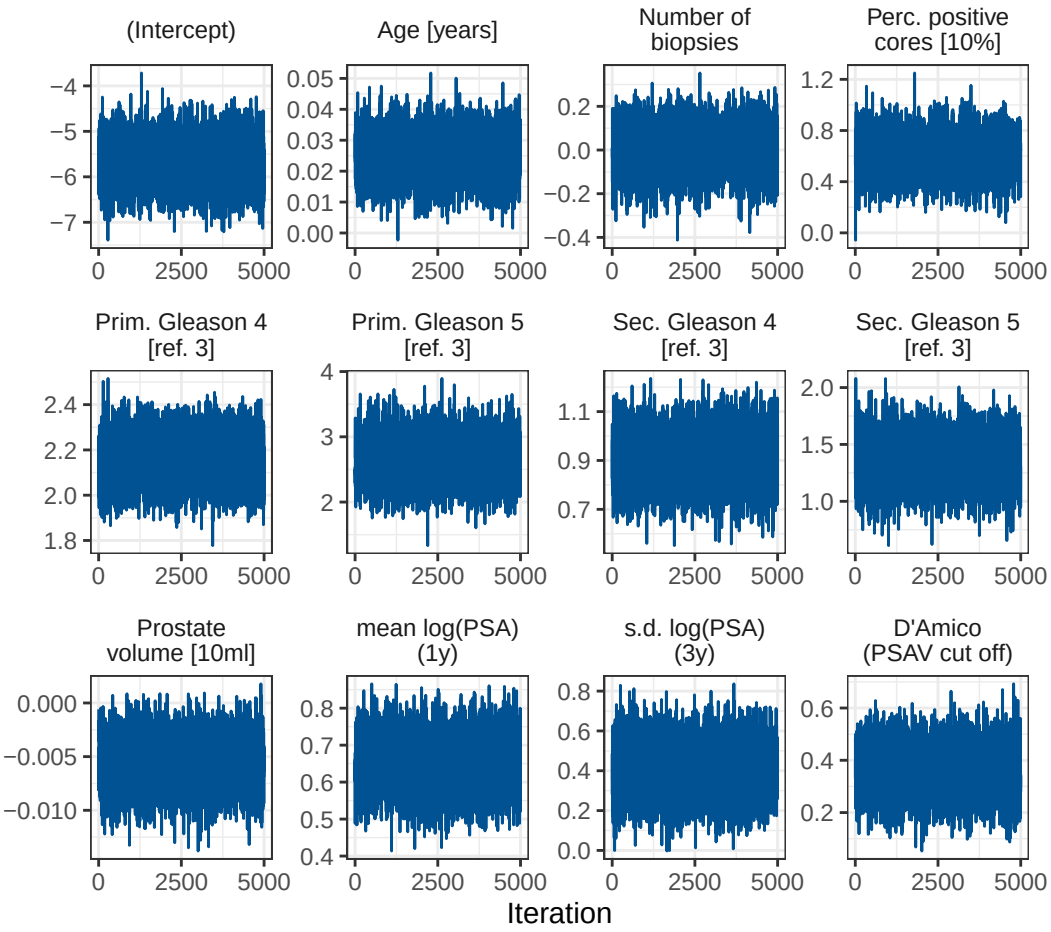

Figure C.3 MCMC samples for the log OR for PGG based on 5,000 samples after thinning of 20 and 1,000 bun-in samples.