Helmholtz Zentrum München

Bachelor's Thesis in Bioinformatics

Identification of protein biomarkers for early detection of neonatal chronic lung disease in preterm infants

Simon Weck

Helmholtz Zentrum München

Bachelor's Thesis in Bioinformatics

Identification of protein biomarkers for early detection of neonatal chronic lung disease in preterm infants

Identifikation von Protein-Biomarkern für die Früherkennung von neonataler chronischer Lungenkrankheit in Frühgeborenen

Supervisor:	Prof. Dr. Dr. Fabian Theis
Advisors:	Dr. Stefen Sass
Submission date:	15.09.2015

I confirm that this bachelor's thesis is
my own work and I have documented
all sources and material used.

15.9.2015 ——————————

Simon Weck

Zusammenfassung

Chronische Lungenkrankheit, auch bekannt als Bronchopulmonale Dysplasie (BPD), ist eine Lungenkrankheit die meist Frühgeborene betrifft. Die Lungenkrankheit hat einen signifikanten Einfluss auf die Krankhaftigkeit und Sterblichkeit von betroffenen Patienten. Bronchopulmonale Dysplasie wird in Woche 36 nach der Geburt diagnostiziert. Eine frühere Diagnose könnte die Prävention und Behandlung von BPD verbessern. Wie versuchen eine frühere Diagnose von BPD zu finden, indem wir mit Hilfe einer neuen groß angelegten Studie des Proteoms von Frühgeborenen nach frühen Protein-Biomarker suchen. In Zusammenarbeit mit dem Klinikum der Ludwig-Maximilians Universität (LMU) München führen wir eine Studie über Proteomik in Blut, Urin und Trachealsekret durch.

Wir erstellen mit den Proteom Daten und den klinischen Daten statistische Modelle um die Beziehung von jedem Protein mit dem Grad der BPD zu analysieren. Als Resultat finden wir eine Menge von möglichen Protein-Biomarker. Wir versuchen außerdem eine neue verbesserte Diagnose von BPD zu finden. Dabei konstruieren wir statistische Modelle bestehend aus Daten von Magnetresonanztomographie, der Lungenfunktion und der Diagnose nach 36 Wochen. Mit der neuen Diagnose bestehend aus MRT Lungendaten suchen wir wiederum nach Biomarkern. Schlussendlich sind die Proteine, die in mehreren Resultaten vorkommen, die Protein-Biomarker Kandidaten. Die Literatur der Kandidaten OSM und CFH zeigt, dass sinnvolle Biomarker in den Resultaten vorhanden sind.

Frühe Protein-Biomarker sind wichtig für die Frühdiagnose von BPD und für eine Verbesserung der Behandlung der Patienten. Die Suche nach Protein-Kandidaten ist ein essenzieller Schritt für die Identifizierung von Protein-Biomarkern.

Abstract

Chronic lung disease (CLD), also known as Bronchopulmonary Dysplasia, is a lung disease affecting predominately preterm infants and has a significant contribution to morbidity and mortality of the affected infants. Bronchopulmonary Dysplasia is first diagnosed in week 36 after birth. An earlier diagnosis would be a way to improve the prevention and treatment of BPD. We try to find an early diagnosis of BPD by searching for protein biomarkers with a new comprehensive study of the proteome of preterm infants. In cooperation with the hospital of the Ludwig-Maximilians-University (LMU) Munich we performed a study of large scale proteomics in blood, tracheal and urine secretion.

With the available proteome data and clinical data we set up statistical models to analyse the association of protein expression with the grade of BPD. We were able to identify a set of candidate early protein biomarkers. Furthermore we attempted to create an improved novel diagnosis of BPD by building statistical models with the Magnetic resonance imaging, lung function data and the diagnosis at week 36 after birth. With this diagnosis consisting of MRI lung data we again search for biomarkers. Finally we denoted proteins present in the results of multiple models as candidate protein biomarkers. The literature of the candidates OSM and CFH show that reasonable biomarkers are existing in the results.

Early protein biomarkers are important for the early diagnosis of BPD and for the improvement of the treatment of the patients. Searching for candidates is a essential step for the identification of protein-biomarkers.

Contents

1 Introduction

Bronchopulmonary Dysplasia (BPD) (also known as Chronic lung disease (CLD)) is a lung disease affecting mostly preterm infants and has a significant contribution to morbidity and mortality of the affected infants [12]. Commonly BPD occurs in infants treated with mechanical ventilation and oxygen therapy to counteract their respiratory distress syndrome or an other severe lung diseases [9]. BPD is mostly defined as the disruption of the growth of the lung [9]. Preterm infants weighting less than 1000 g at birth have a 75% risk to get BPD and the risk increases with decreasing weight [12].

The lung disease has prenatal and postnatal factors and long term effects that can persist into adolescence and early adulthood [12].

Clinical practices like prenatal steroid use, improved ventilation strategies and improved nutrition have resulted in considerable improvements of the clinical outcomes of BPD. But until today there are no really safe and effective therapies to prevent or to reverse BPD. So the overall incidence of the disease has not improved in the last 10 years [9, 12]. In 2005 there have been more than 16000 cases of BDP which corresponds a rate of 3.9 per 1000 infants in the USA [2]. Overall "BPD remains a heavy burden on health care resources" [9].

Bronchopulmonary Dysplasia is first diagnosed in week 36 after birth with a set of criteria including days of oxygen supply and days of mechanical ventilation. Since early treatment of the preterm infants can counteract the development of BPD, it would be desirable to find an alternative early diagnosis of BPD. Such a diagnosis can be created with early protein biomarkers, but one of the unsolved problems is that until today there is no comprehensive large scale study to identify these early protein biomarkers for BPD. So the question is to find candidate early protein biomarkers of BPD in preterm infants with protein expression data.

A possibility to find protein biomarkers are large scale measurement methods. These are still a relative new way to study traits and diseases but they become more and more common with the improvement of the different measurement technologies. For the study of a proteome there are different methods such as mass spectrometry or microarray-based custom regents binding to specific proteins like in the new method SOMAscan™.

In order to address the problem of identifying early protein biomarkers we worked together with the hospital of the Ludwig-Maximilians-University (LMU) Munich on a study of large scale proteomics in blood, tracheal and urine secretion with these two methods of a cohort of 40 preterm patients at different time points after birth. With this new protein data our goal in the thesis is to identify early protein biomarkers to

diagnose the disease earlier and more accurate. However the big dataset consisting of large-scale proteome data, clinical data, MRI and lung function measurements is complex and therefore demands so for statistical analysis techniques to determine the possible biomarkers.

We set up statistical models to analyse the relationships of each protein with BPD so that we found a set of candidate early protein biomarkers. We also attempt to create an improved diagnosis of BPD by incorporating Magnetic resonance imaging (MRI) data of the preterm infants in the cohort (compare Figure 1). With new models including our new diagnosis composed of T1 and T2 lung data we find a set of additional protein biomarkers. The newly found possible biomarkers can now be further analysed.

In chapter 2 we initially describe regression analyses and biological background informations, which are important for the understanding of the methods introduced later. Subsequently in chapter 3 we give an overview about detailed information about the diagnosis of BPD. In the next chapter we characterize the different protein and MRI datasets. We analyse the structure of the datasets and particularly the large proportion of missing values and null values in the protein datasets measured with mass spectrometry and in the MRI data. In chapter 5 we describe the methods used throughout this work. We here first address the problem of missing data by imputation with the value distributions in each dataset. With the complete data we further describe the structure and computation of the different models to get the protein biomarkers. In the first approach we compute models with the protein expression and the diagnosis of Bronchopulmonary Dysplasia with linear regression analysis. In the second approach we compute the models of the first method with a new diagnosis determined by elastic net regression and lasso of a model containing the common diagnosis and the MRI dataset. Finally in Section 6 we analyse and interpret the resulting potential protein biomarkers. We determine a low percentage of overlapping proteins between the different models, which we can explain with the different assessments of the diagnoses to the disease. In the outlook we explain improvements and possible future directions of the study.

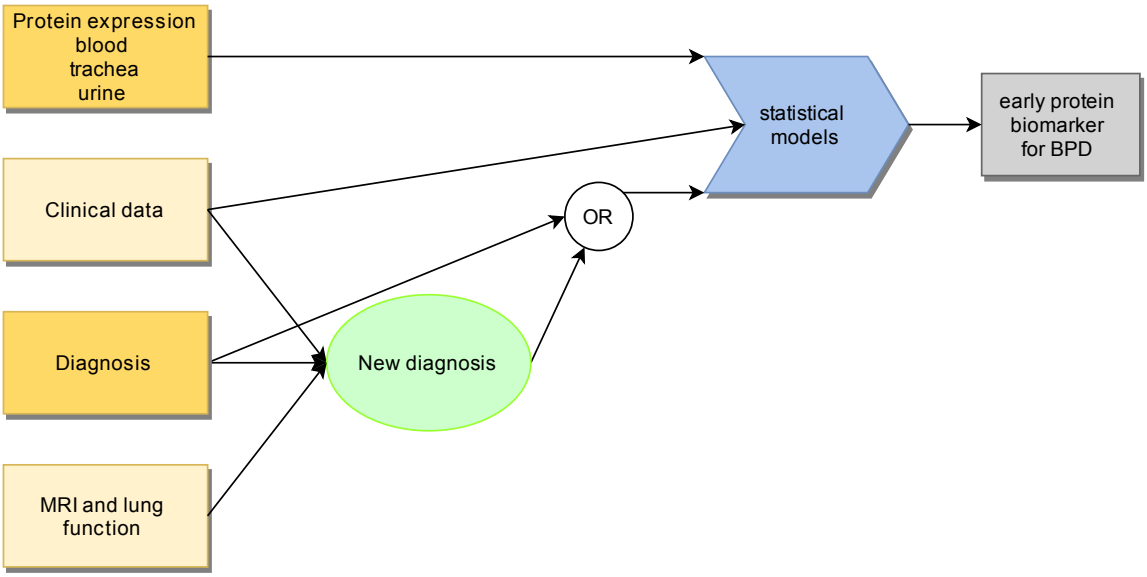

Figure 1: Protein expression data, clinical data and a diagnosis are used for the statistical models to attempt to find early protein biomarkers. We search for a new improved diagnosis by creating a model with the clinical data, the diagnosis, magnetic resonance imaging (MRI) and lung function data.

2 Background

Here we introduce statistical methods like linear regression and the shrinkage methods lasso and elastic net. After the statistical part we will give an overview about the biological background of proteomics and magnetic resonance imaging.

2.1 Regression analysis

2.1.1 Linear regression

In statistics *regression* is the modeling of an output $\mathbf{y} = y_1 + ... + y_d$ with the quantitative inputs variables $\mathbf{X} = \mathbf{x}_1 + ... + \mathbf{x}_n$. The model describes how the output is affected by the input. The *linear regression model* has the assumption that $\mathbf{X}$ and $\mathbf{y}$ are approximately linear related to each other. If n is greater than 1, and so there are multiple inputs, the process is called *multiple linear regression*. The formula of the linear regression model is

$$\mathbf{y} = \beta_0 + \sum_{j=1}^{n} \mathbf{x}_j \beta_j + \epsilon \tag{1}$$

where the Gaussian random variable ϵ is the error $\epsilon \sim \mathcal{N}(0, \sigma^2)$, β_0 the intercept and $\beta_{1,...,n}$ are unknown parameters for each input variable. The *least squares method* is a common method to estimate the coefficients $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_n)^T$ to predict an outcome $\hat{y}$. The method estimates these coefficients by minimizing the residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{d} (y_i - \beta_0 - \sum_{j=1}^{n} \mathbf{x}_{ij} \beta_j)^2 \tag{2}$$

with the data $\mathbf{X}_{1...n} = (\mathbf{x}_{11}, ..., \mathbf{x}_{nd})^T \mid d = |\mathbf{y}|$ and $\mathbf{y}$. We can now calculate β as a closed-form solution [11].

2.1.2 Shrinkage methods

The least squares estimates of linear regressions can have a high prediction error [11]. Furthermore the interpretation of the results of a linear regression gets difficult if the model contains a huge number of predictors. This is especially true for determining the subset of predictors with the strongest effects [11] . One possibility to improve these problems is *subset selection* of a model. It selects a subset of the predictors $\mathbf{x}_1, ..., \mathbf{x}_a$ and discards the other to simplify the model and to reduce possibly prediction error. The problem with subset selection is that it is a discrete method and

therefore has often high variance. Shrinkage methods like *lasso* and *elastic net* are more continuous and have so less problems with high variance [11]. Shrinkage methods shrink the coefficients by for example setting a penalty on the size of them like the regression method *ridge regression*. Another possibility of shrinking coefficients is to set the fraction of the coefficients exerting no influence to the prediction of the output variable to zero like the method Least Absolute Shrinkage and Selection Operator (lasso) [11].

Lasso and elastic net shrink the set of input variables and coefficient to only the variables which exert influence to the prediction of the output. Elastic net solves the problem

$$\min_{\beta} \left[\frac{1}{2d} \sum_{i=1}^{d} (\mathbf{y}_i - \beta_0 - \sum_{j=1}^{n} \mathbf{x}_{ij}\beta_j)^2 + \lambda P_\alpha(\beta) \right] \tag{3}$$

$$P_\alpha(\beta) = (1-\alpha)\frac{1}{2}||\beta||^2_{L2} + \alpha|\beta|_{L1} \tag{4}$$

$$= \sum_{j=1}^{n} \left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right] . \tag{5}$$

Here $P_\alpha(\beta)$ is called the elastic-net penalty consisting of the $L1$ lasso penalty and the $L2$ ridge penalty [8, 17]. The special case of the method $\alpha = 0$ is equal to ridge regression, which shrinks the size of the coefficients of correlated input variables towards each other so that these variables appear in the result. Coefficients get never zero. The method lasso corresponds to elastic net with $\alpha = 1$. In contrast to ridge regression lasso tends to take only one of the correlated input variables, which we can consider as a loss of information on highly correlated variables. Between $\alpha = 0$ and $\alpha = 1$ elastic net is a compromise between ridge regression and lasso. With an increasing $\alpha = [0 : 1]$ the method increases the number of coefficients $\beta_j = 0$ monotonically from 0 to the lasso solution [8]. The parameter λ is a factor how much the coefficients are shrunken. λ is varied to compute a path of solutions [11].

The solution of elastic net cannot be solved in closed-form. The R package *glmnet* [8] tackles this problem with cyclical coordinate descent algorithms to find optional solutions. The glmnet algorithm calculate the coefficients $\hat{\beta}$ with a coordinate decent approach for a set of λ values. With cross validation of this set of λ values we can find the optimal solution [17].

2.2 Proteomics

The term *proteome* describes all proteins in a cell or a organism at specific point of time. Omics-techniques in molecular biology comprise genomics, transcriptomics, proteomics and metabolomics and are defind as the characterisation and quantification of the different "omes". Proteomics is a large-scale study of a *proteome* in a specific organism. The objective of proteomics is to characterise the protein expression quantitatively as well as the changes in protein expression caused by perturbations like diseases and drugs [1]. There are different methods to measure the expression of proteins. In the next two sections we introduce the measurement methods Mass spectrometry with MaxQuant and SOMAscan™[14]. These methods all are for the measurement of proteins.

2.2.1 Mass spectrometry

We can use *mass spectrometry* for the identification and measurement of proteins. This method uses *untargeted proteomics* which does not preselect proteins. Untargeted proteomics attempts to measure and identify all proteins in a sample commonly with shotgun liquid chromatography coupled with tandem mass spectrometry [10]. The first step in the workflow is to separate the different proteins for example with liquid chromatography, where in contrast to gas chromatography the mobile phase of the separation is a liquid. Mass spectrometers break the peptides of a pure protein sample into ions to measure the mass-charge ratio of them. With the spectrum of this ions the mass of the whole peptide is calculated (compare Figure 2). The protein is now identified by the sequence of the measured ions. The extent of the expression of a protein is measured by a computational comparison of their peaks [18]. There can be problems with the identification and quantification of proteins. Protein peaks can overlap so that they can not be identified and quantified. Here in our thesis the identification of the proteins and the calculations of the expression values were computed by MaxQuant [7].

2.2.2 SOMAscan

SOMAscan™ is a quantitative tool to measure proteins with *targeted proteomics*. In contrast to untarged proteomics targeted proteomics selectively isolates, identifies and quantifies proteins [13]. The targeted approach can overcome fundamental limitations of the untargeted protein measurement like the limited sensitivity [13].

The targeted method SOMAscan™ can measure up to 1129 proteins by using a reagent called *SOMAmer* for each protein. The SOMAmers consist of modified

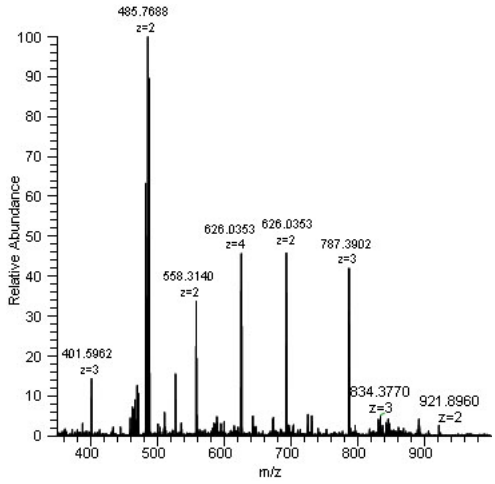

Figure 2: Example of resulting spectrum with the relative abundance on y-axis and the mass-charge ratio m/z on x-axis. Figure from [15].

nucleic acids which bind to specific protein tertiary structures. The proteins P_i bind to the specific reagents S located on beads. The residual secretion and SOMAmers are removed by washing. The process performs a biotinylation with the protein and the reagent to bind them to each other. Photocleavable linker break the binding of the beads to the regents. After the release, the nucleic acid reagents are separated from the protein by capturing first the complex with a SA bead and than detach the SOMAmer. The resulting nucleic acid sequences are quantitatively measured using DNA microarrays (see Figure 3) [14].

Problems can result from the bias that only proteins which were chosen beforehand are measured. Hence for example biomarkers can be missed. A sensible selection is very important to get reasonable results.

2.3 Magnetic resonance imaging

Proton magnetic resonance imaging (MRI) is a imaging technique to probe the anatomy of the body. The method uses radio waves and magnetic fields to get high quality 2D or 3D images. With the MRI it is possible to study dynamic processes like for example the respiratory motion of the lung. We can use it to get combined morphological and functional information. The adjustment of the MRI is constantly changed during a MRI scan to get different views and also to compensate failed images. For example the terms T1 and T2 are time constants for signal decay. Each of them can be used to analyse the body differently. Such as a high signal

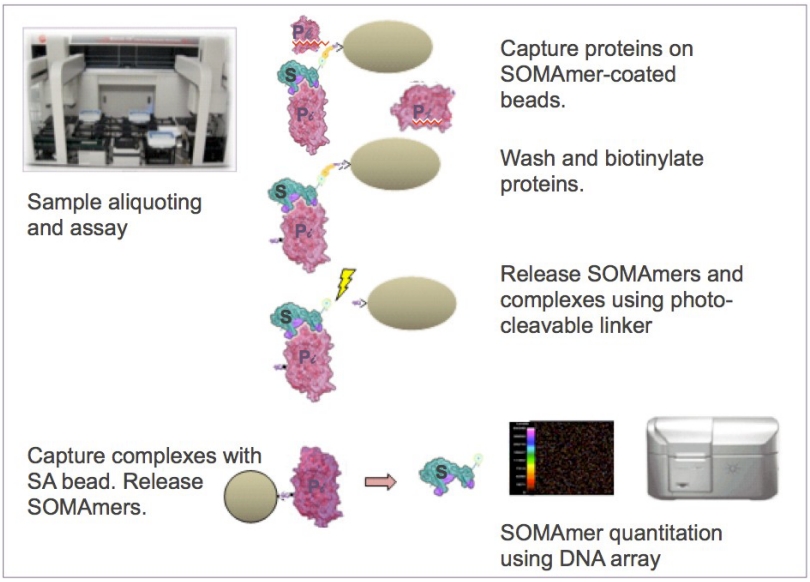

Figure 3: An overview of the process of the SOMAscan method. SOMAmers consist of modified nucleic acids and can bind to specific protein tertiary structures. In the workflow the proteins are captured on SOMAmer-coated beads. Then the method washes the residues away and biotinylates the proteins and the SOMAmers. Photocleavable linker separate regents from the beads. SA beads capture protein complexes and the SOMAmers are split from the protein. The quantification of the SOMAmers is performed by using DNA microarrays. (from [14])

with T1 we can interpret as a inflammatory activity. Because there is no radiation exposure of the patients in MRI it is specifically attractive for the usage in infants [4, 5].

3 Diagnosis of neonatal chronic lung disease

The diagnosis of bronchopulmonary dysplasia is split into 3 grades, from grade 1 mild to 3 severe BPD. In the following we use the grade 0 as no disease. Table 1 shows the NHI diagnostic criteria for BPD.

	Gestational age	
	<32 weeks	>32 weeks
Timepoint of assessment	36 weeks post-menstrual age or discharge *	>28 days but <56 days postnatal age or discharge *
Treatment with oxygen	>21% for at least 28 days	>21% for at least 28 days
Bronchopulmonary dysplasia		
Mild	Breathing room air at 36 weeks post-menstrual age, or discharge*	Breathing room air at 36 weeks post-menstrual age, or discharge*
Moderate	Need for <30% O2 at 36 weeks post-menstrual age, or discharge*	Need for <30% O2 to 56 days postnatal age, or discharge*
Severe	Need for >30% O2, with or without positive pressure ventilation or continuous positive pressure at 36 weeks post-menstrual age, or discharge*	Need for >30% O2, with or without positive pressure ventilation or continuous positive pressure at 56 days postnatal age, or discharge*

* Whichever comes first.

Table 1: NHI diagnosic critieria for bronchopulmonary dysplasia. Table adapted from [12].

A multitude of factors own a influence to the disease, but they are not fully studied and understood. The following paragraphs about factors and consequences of BPD are based on [12] and [9]. Premature infants have weaknesses in the antioxidant enzyme systems and they also own insufficient numbers of antioxidants like

the vitamins C and E. With an oxidation therapy the lung has a excessive exposure to oxygen and such hyperoxia can occur, leading to increased production of cytotoxic oxygen free radicals. The radicals can overcome the weak defend system of the preterm infants and can so induce lung injury.

The lung damage with mechanical ventilation by Pulmonary volutrauma, which is an overdistention of the lung, is also possibly a cause of the development of the lung disease in infants.

An other factor is the vascular endothelial growth factor (VEGF) signalling contributing to vascular disease through hyperoxia. Endothelial-epithelial interactions particularly with VEGF signalling, has a critical role for a healthy lung growth. Reduced or disrupted VEGF signalling leads to hindered vascular growth and alveolarization.

Other risk factors are prenatal and postnatal inflammations of the lung in the infants. Proinflammatory citokines like interleukins Interleukin-1beta (IL-1β) and Interleukin-6 (IL-6) have a raised expression from the birth to 6 months after birth. For example IL-1β causes release of inflammatory mediators and the activation of inflammatory cells [12].

Furthermore genetic factors are contributing to Bronchopulmonary Dysplasia. So for example polymorphisms in the genes tumor necrosis factor alpha (TNF), Toll like receptor 10 and VEGF possibly have a important role in the development of the disease. In summary inflammations, hyperoxia, a week defence system, mechanical ventilation and genetic predispositionshave all a contribution to the injury and the prenatal and postnatal growth of the lung.

Implications of Bronchopulmonary Dysplasia are increased respiratory rates (tachypnoea) with shallow breathing and retractions and also on ausculation wheezes can be heard. These breathing characteristic with the increased rates increase dead space ventilation. The dynamic lung compliance is reduced because of small airway narrowing, fibrosis, oedema resulting of large collateral vessels shunting blood flows to the lung, and atelectasis [12].

In the first 2 years the preterm infants often need to go once again to the hospital especially because of respiratory syncytial virus infections [12] . Long and comprehensive studies of patients into adulthood are not present and so long term outcomes of patience with BPD are not well known [12] .

4 Materials

4.1 Protein data

For the identification of protein biomarkers of BPD we require protein expression data and clinical data for a defined set of patients. We use the clinical data to eliminate environmental effects from the computations. As clinical data we include the weight, the gender, if the infants had an early onset infection, the degree of Respiratory Distress Syndrome (RDS), the usage of steroids and the gestational age (Table 4.1). We have 3 different protein expression datasets available. These are measured in blood plasma, tracheal segregation and urine by the group of Anne Hilgendorff at the hospital of the Ludwig-Maximilians-University (LMU) Munich. They measured the blood protein expression data with the SOMAscan™method (compare Section 2.2). Here 18 patients have each mostly 3 blood protein measurements at different time points between day 1 and 224 after birth.

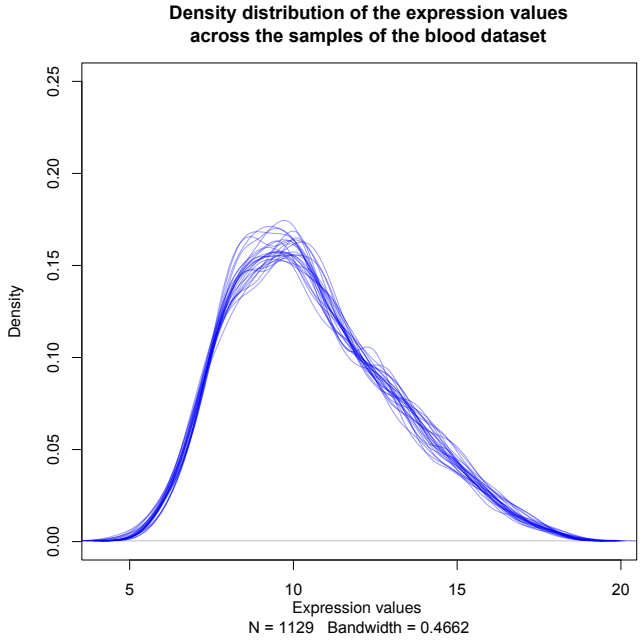

Figure 4: Density distribution of the logarithmic expression values of the first 7 days across the samples/patients of the blood dataset. Here we observe that the $\log_2$ protein expressions are approximately normally distribution. The distributions have a long tail at the higher expression values.

Patient	Gestational.age	Weight	Gender	Early.onset.infection	BPD	RDS	mechVent_days	Steroide	0xygen_days
1	24	580	f	no	1	2	55	yes	53
2	27	885	m	no	0	1	33	no	0
11	30	1630	f	no	0	1	21	no	7
12	30	1125	f	no	0	2	19	yes	1
14	30	1630	m	no	0	0	0	no	0
15	30	930	m	no	0	0	12	no	1
19	30	1770	m	no	0	0	3	no	0
20	26	780	m	yes	1	2	56	no	63
21	26	575	m	yes	3	4	70	no	176
22	25	750	f	yes	0	2	53	yes	1
23	28	1040	f	no	2	3	40	no	20
24	27	760	f	no	3	3	78	yes	186
25	28	700	m	no	0	2	33	no	0
28	30	1370	m	no	0	2	12	yes	8
34	27	930	f	no	1	4	44	yes	32
35	27	760	f	no	2	4	56	yes	67
37	29	1510	m	no	1	1	17	no	35
38	26	820	m	no	1	2	49	yes	44
39	27	815	f	yes	3	3	53	yes	72
40	25	720	m	no	1	2	56	yes	59
41	30	995	f	no	0	1	9	no	0
42	30	1440	f	no	0	1	0	no	0
44	29	1090	m	yes	1	3	34	no	40
46	31	1100	f	no	0	1	6	yes	0
47	28	950	f	yes	1	2	31	no	31
49	25	700	m	yes	1	3	60	yes	36
56	30	1300	f	no	0	2	5	no	3
57	30	1440	m	no	0	1	4	no	4
58	27	655	f	no	1	1	45	no	48
59	25	730	f	no	1	2	56	yes	66
60	27	915	m	no	0	2	35	no	9
61	24	530	f	no	2	3	73	yes	105
62	27	1200	m	yes	0	0	15	no	0
63	30	1415	f	no	0	2	2	no	1
64	24	415	f	no	2	3	73	yes	87
65	28	930	f	no	0	0	38	yes	6
66	28	730	f	no	0	2	43	yes	9
72	24	850	f	no	3	2	68	yes	91
73	28	1500	m	no	0	2	28	no	6
74	26	920	m	yes	1	4	57	no	49

Table 3: Clinical data of the 40 patients in the cohort. Columns in the table: mechVent_days = days of mechanic ventilation; 0xygen_days = days of oxygen supply; Gestational age in weeks; weight in grams; Respiratory Distress Syndrome (RDS) grades

We have also protein expression data measured in urine and tracheal secretion with mass spectrometry and analyzed with MaxQuant. For normalization we take the logarithm $\log_{10}$ of protein expression values. A large fraction of these two datasets consist of missing values (see Figure 9 and 10). The expression values are measured at multiple time points after the birth of the preterm infant.

4.2 MRI data

The diagnosis in week 36 is divided into 4 disease grades ranging from 0 to 3. This categorization in groups is not very precise and also the nuances between the patients in one group are here not considered. So since the diagnoses with BPD grades do not describe the disease well, Magnetic resonance imaging (MRI) data of the patients is included in one approach.

The MRI data consist of lung, heart function, the size of the lung and of T1 and T2 values. Some values are missing due to some difficulties by the measurement of the different variables (Figure 5). The lungs of the infants are very small and for that reason it is difficult to determine the different measurement values. An other problem is that the infants can move in the MRI scanner and so the quality of the picture can be bad (Section 2.3).

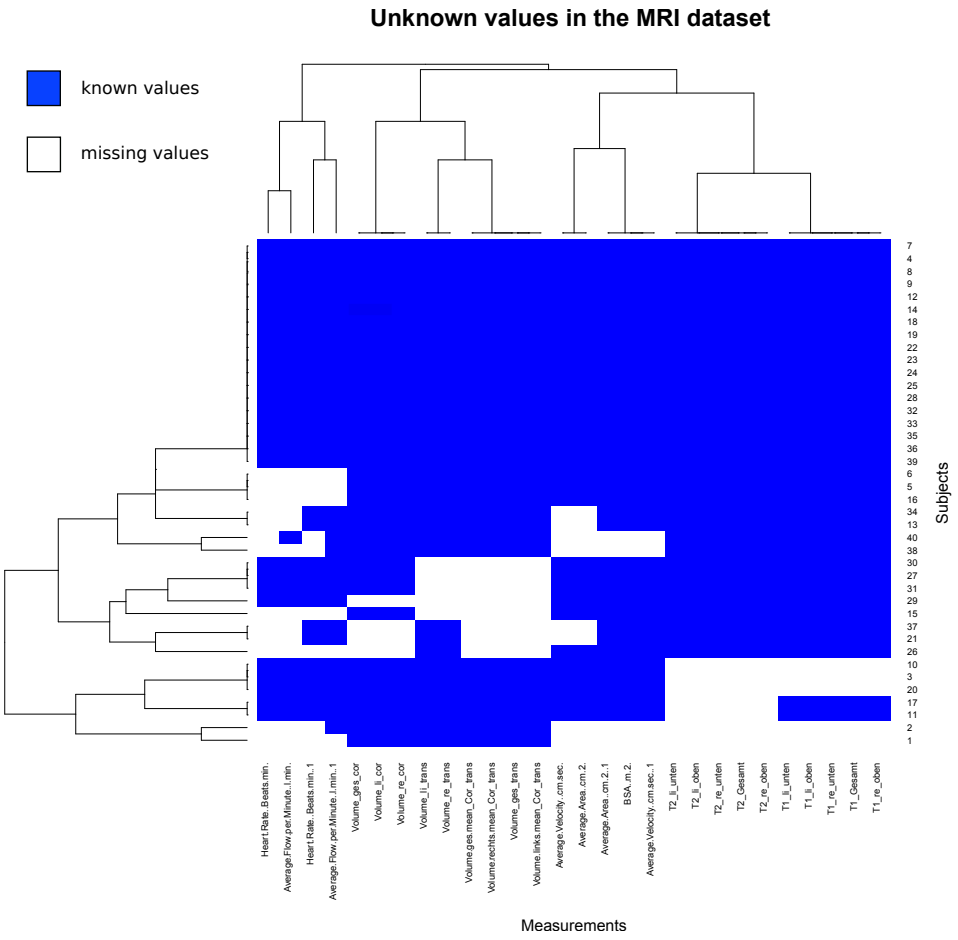

Figure 5: Missing values in the MRI data. White are missing values and blue are known values.

5 Methods

We performed multiple differing approaches to identify early protein biomarkers whereby we find different sets of biomarkers. We only use protein expression measurements of the first 7 days to account for our problem to find early protein biomarkers (Section 4.1). Generally we perform the computation of lasso and elastic net with leave-one-out cross-validation implemented in the R package *glmnet* [8]. In the following we will first describe the imputation of missing values by sampling out of a normal distribution which is parametrized according to the data. After we present two different workflows to find early protein biomarkers. The approach A works with lasso and linear regression applied to the diagnosis of BPD and the large scale protein data. In contrast the workflow B creates a new prediction of BPD with the help of the MRI data and than uses this prediction to perform the method introduced in the first approach.

5.1 Imputation of unknown values

If we want to apply a method to a dataset with missing values, there is no ideal way to handle them. We can either try to remove these values or we can try to impute the missing data. Here we choose simple but unbiased methods to compute these values.

5.1.1 Protein dataset

As mentioned in Section 4 in the urine and tracheal protein dataset are lots of missing values (Figure 9 and 10). We do not know if these are not measured due to technical difficulties of mass spectroscopy or are not present in the body fluid. We decided to impute these values by assuming that the protein expression is normally distributed. First we remove the proteins with more than 30% missing measurements. With the removing of proteins we can possibly loose proteins with significant effects. But in the following imputations reasonable protein expression means are important and these are only possible with a large fraction of non null values. The 0 values NA_x of a protein p is calculated with the formula

$$NA_x = \mathcal{N}((0.3 \; quantil \; of \; \mathcal{N}(mean_p, 1.5)), 1) \quad \cdot \tag{6}$$

The 30% quantile of the normal distribution with the mean of the respective protein expression values is here applied to account the explanation of 0 means no expression.

Prediction	Meaning
BPD	multinomial BPD with 0,1,2,3
BPD0	binary with BPD 0 and combined BPD 1,2,3
BPD1	binary with combined BPD 0,1 and combined BPD 2, 3
Days of mechanical ventilation	length of the mechanical ventilation in days
Days of oxygen supply	length of the supply with oxygen in days

Table 4: Overview of the different diagnoses used in Section 5.

5.1.2 MRI dataset

We have also a problem with unknown values in the MRI dataset (Figure 5. Here the measurements are missing and not measured and so we have to impute these values. We use the R package *MICE* [6] to impute the missing data.

5.2 Identification of disease-associated proteins

The general idea of our first approach is to explain the expression of the protein with a prediction variable by excluding the available cofactors consisting of clinical data (Table 4.1) and the general influence of the patients. Hence the protein biomarkers are the proteins which are significantly influenced by the prediction. We model this concept by using linear regression analysis with the clinical data as predictor variables and the expression of one protein as the output. But because we restrict the protein measurements to the first 7 days and we have several input variables we have a big model with few information. So we first shrink the model with lasso to reduce the number of variables and than we calculate the linear regression of the smaller model to obtain p-values for the prediction and and account for multiple testing.

Figure 6 shows an overview of the following workflow. We first match the samples of the imputed protein expression data and the clinical data including the diagnosis. Then we perform lasso in step 3 for all proteins $a \in \{1 : N\}, N = |proteins|$ with

$$\mathbf{y}_a = \beta_0 + \beta_{diagnosis} * \mathbf{x}_{diagnosis} + \boldsymbol{\beta}_{clinical} * \mathbf{X}_{clinical} + \sum_{j=1}^{s} \beta_j * \mathbf{X}_j + \varepsilon \qquad (7)$$

$\mathbf{y}_a$ is defined as the protein expressions of protein a. In the formula $\mathbf{x}_{diagnosis}$ is either BPD0, BPD1, days of mechanical ventilation or days of oxygen supply (defined in Table 4) and $\mathbf{X}_{clinical}$ are the clinical variables of Table 4.1. s is the number of subjects, which have more than one measurement and we define $\mathbf{X}_j$ as a Boolean vector marking the subject j. These variables indicate the influence of each subject.

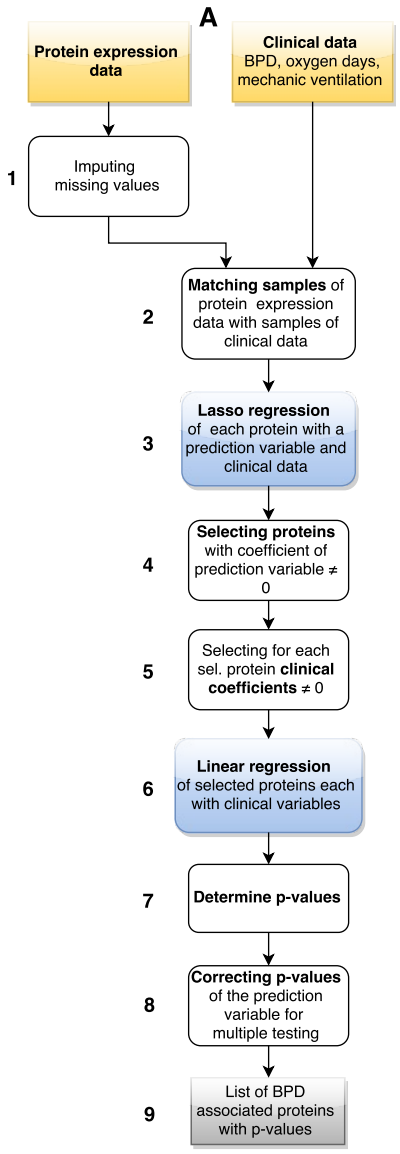

Figure 6: Overview of the workflow A described in Section 5.2. Workflow steps for the identification of proteins associated with BPD.

With this computation we obtain for each protein a shrunken model. We pick in step 4 in Figure 6 only the models which select a non-zero coefficient for the diagnosis. This model selection is the starting basis of the multiple linear regressions which have the form

$$\mathbf{y}_b = \beta_0 + \beta_{diagnosis} * \mathbf{x}_{diagnosis} + \boldsymbol{\beta}_{selected} * \mathbf{X}_{selected} + \varepsilon \tag{8}$$

Here protein $b \in \{1 : m\}$ with $m = |selected\ proteins|$. $\mathbf{X}_{selected}$ are the selected input variables (compare step 5) in the shrunken model without the prediction and $\beta_{diagnosis}$ are new coefficients associated to each of this variables. Next in Figure 6 step 7 we determine the associated p-value of the prediction for all proteins of the shrinkage models and correct these with the Benjamini and Hochberg procedure to correct the multiple testing error. The results are presented in Section 6.

5.3 Identification of proteins associated with MRI patterns

In the following approach we want to improve the prediction of BPD and identify proteins associated with the new prediction. Figure 7 B shows an overview of the workflow. The problems of the predictions in Table 4 are that the gradation of the strength of Bronchopulmonary Dysplasia are determined very late at week 36 and the classification in 4 grades is vague. Also days of mechanical ventilation and days of oxygen supply are only unsatisfying characteristics of BPD (Section 3). We want to accomplish this improvement by including the MRI dataset. In the new workflow we first search for a new diagnosis in the MRI data with Lasso and elastic net and apply it to a new model similar to Section 5.2.

5.3.1 Identification of disease-associated MRI patterns

The Correlation plot Figure 8 shows the strong correlation of 4 sets of measurement variables to each other. Especially T1, T2 and lung volume variables have each a strong positive correlation in their group. Lasso takes only one variable of these highly correlated sets into account, so that information is possibly lost (compare Section 2.1.2). Elastic net includes these strong correlations [17] and for this reason we also perform the following method with elastic net. Figure 7 describes the following workflow. Similar to A we match the samples of the imputed MRI data and the clinical data. In step 3 we calculated elastic net with an $\alpha \in]0 - 1[$ and lasso $\alpha = 1$ with the equation

$$\mathbf{y}_{diagnosis} = \beta_0 + \boldsymbol{\beta}_{MRI} * \mathbf{X}_{MRI} + \boldsymbol{\beta}_{clinical} * \mathbf{X}_{clinical} + \varepsilon \quad \cdot \tag{9}$$

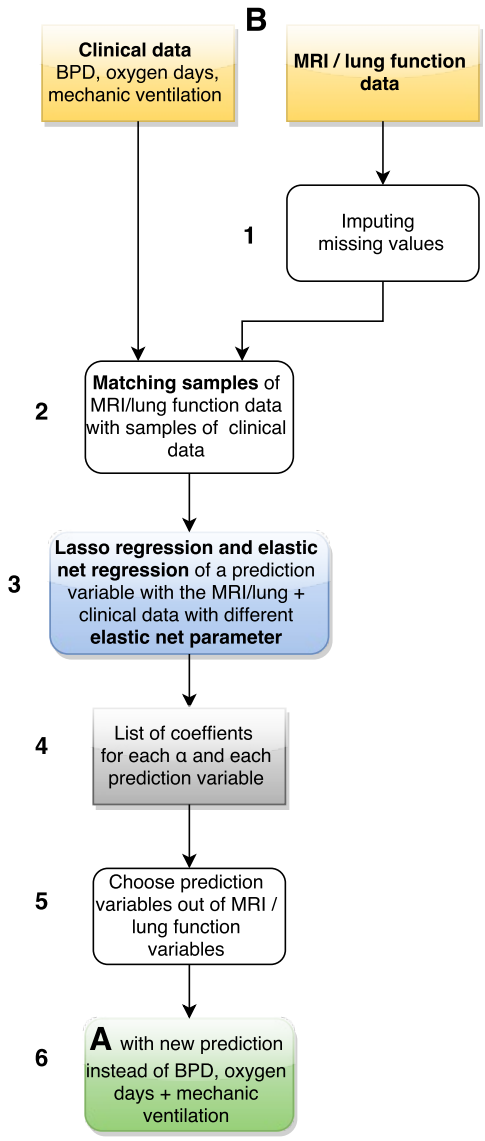

Figure 7: Overview of the workflow B described in Section 5.3. Searching for a new improved prediction of BPD in the MRI and lung function data. With the new prediction we perform workflow A (Section 5.2).

The *diagnosis* above is a prediction of Table 4 and $\mathbf{X}_{MRI}$ are the variables of the MRI dataset. The results are coefficients for each MRI parameter (see step 4). With this reduced sets of parameters we can define new biological meaningful diagnoses by choosing prediction variables of the MRI data.

5.3.2 Identification of MRI-associated proteins

We calculate the same steps like in 5.2 (Figure 6) but with the improved prediction $X_{prediction}$ from approach B step 5 in the formula 10 and 11. In the models we exclude missing data in the new prediction and hence we exclude the associated measurements in step 2. Like in step 3 in A we calculate a lasso regression with the model

$$\mathbf{y}_a = \beta_0 + \beta_{prediction} * \mathbf{x}_{prediction} + \boldsymbol{\beta}_{clinical} * \mathbf{X}_{clinical} + \sum_{j=1}^{s} \beta_j * \mathbf{x}_j + \varepsilon \qquad (10)$$

and the linear regression in step 6 with

$$\mathbf{y}_b = \beta_0 + \beta_{prediction} * \mathbf{x}_{prediction} + \boldsymbol{\beta}_{selected} * \mathbf{X}_{selected} + \varepsilon \cdot \qquad (11)$$

The other steps are equal to the steps in workflow A. Finally we get again a set of proteins, which we can now use for further analyses.

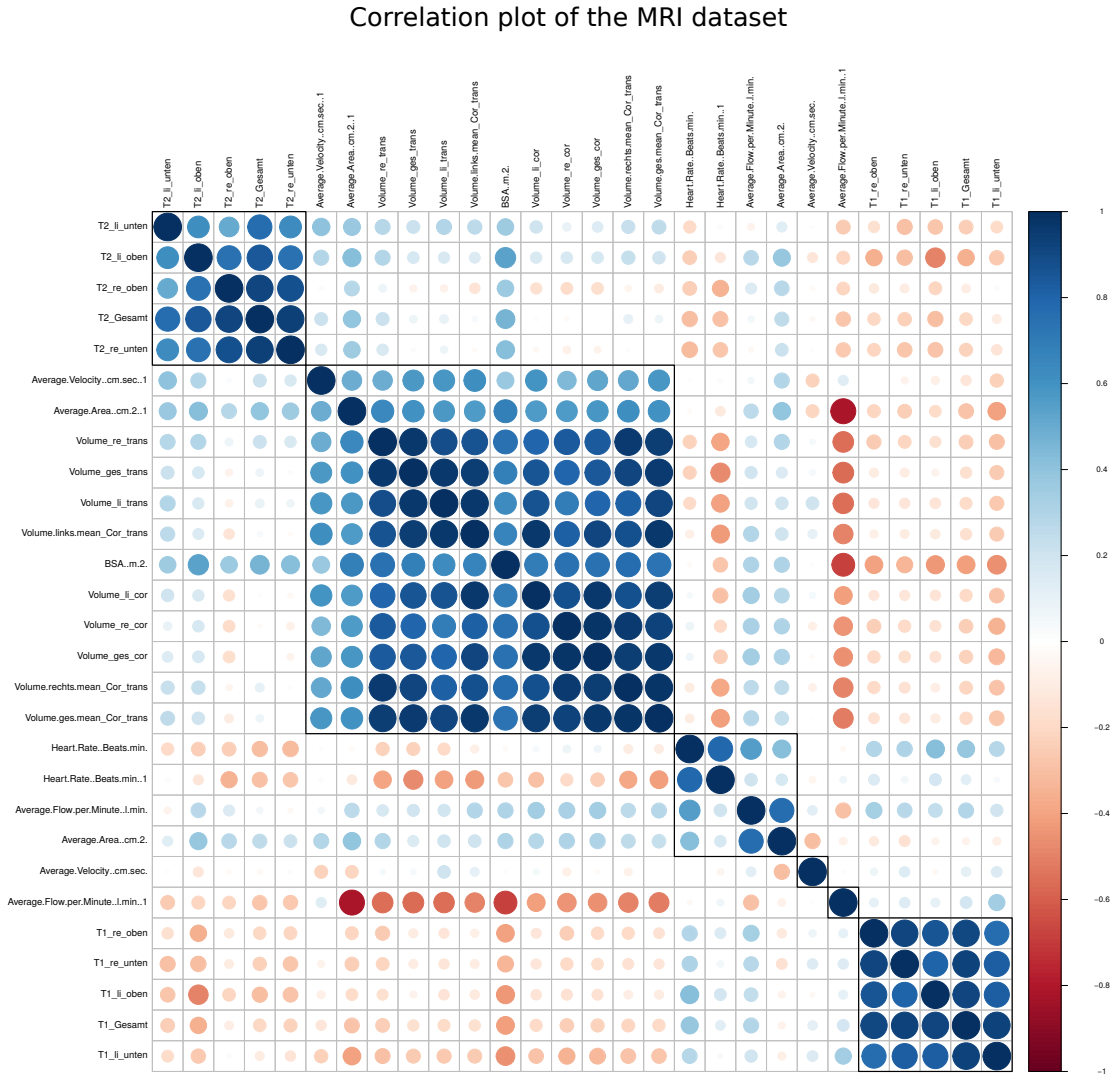

Figure 8: Correlation plot of the MRI dataset. Blue means positive and red negative correlation. The boxes are sets of closely correlated measurement pairs. The size of the circles show the strength of the correlation. The correlations are computed without imputed values. The missing values for the computation of the correlations are omitted.

6 Results and Discussion

6.1 Imputation

In the protein expression datasets measured in urine and tracheal secretion and in the MRI data there are large percentages of unknown values. With the methods in Section 5.1 we imputed these values so that they are reasonably in the comparison to the known data. The resulting density plots of the protein expression data are showed in Figure 12 and 11. Here the imputed values of the tracheal and the urine protein expression data do not contain zeros anymore now. Also these resulting imputed protein expressions approach normal distributions. We also imputed the values of the MRI dataset for the method in Section 5.1.2. Although there is a relative high variance with this imputation between different imputations it is sufficient enough to draw the conclusions in Section 6.3.

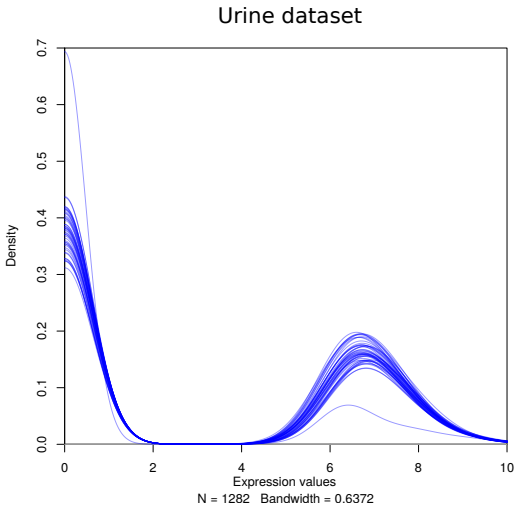

Figure 9: Density distribution of the logarithmic expression values ($\log_{10}$) of the first 7 days across the samples of the **urine** dataset.

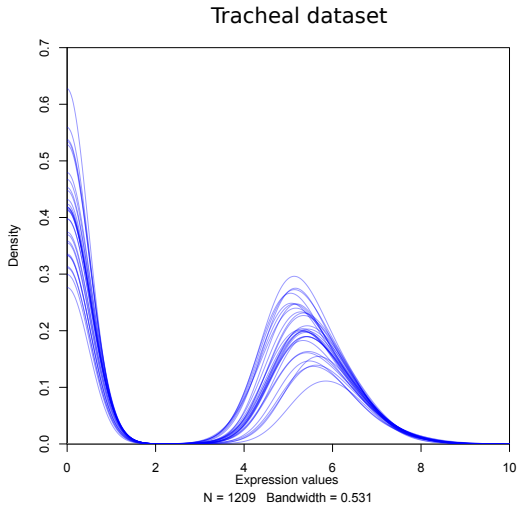

Figure 10: Density distribution of the logarithmic expression values ($\log_{10}$) of the first 7 days across the samples of the **tracheal** dataset.

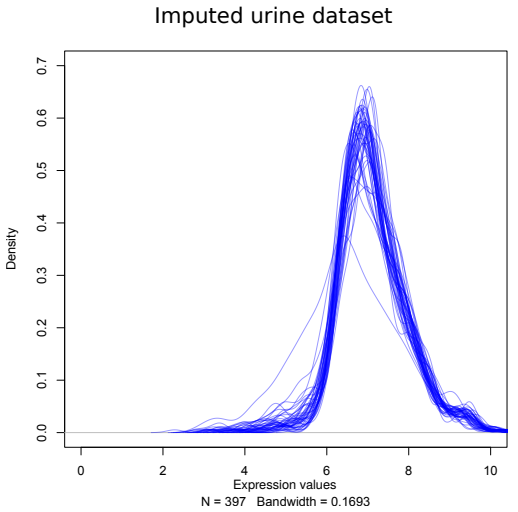

Figure 11: Density distribution of the logarithmic expression values ($\log_{10}$) of the first 7 days across the samples of the **urine** dataset with the imputation of the 0 values with the method in Section 6.1.

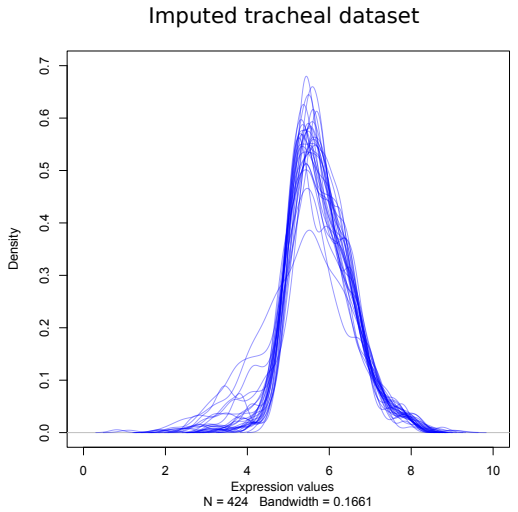

Figure 12: Density distribution of the logarithmic expression values ($\log_{10}$) of the first 7 days across the samples of the **tracheal** dataset with the imputation of the 0 values with the method in Section 6.1.

6.2 Results of the identification of disease-associated proteins

Using the method described in Section 5.2 we calculated the coefficients with 4 different types of diagnosis BPD0, BPD1, days of mechanical ventilation and days of oxygen supply (see Table 4) and with the blood and the imputed urine and tracheal protein expression data.

In addition we computed the method by including days of oxygen supply and days of mechanical ventilation at the same time. Instead of using only one variable as diagnosis we here used two. If lasso selected minimum one of the two variables , we performed linear regression with the non-zero variables. After these regressions we used again the proteins where one or both of the coefficients of days of oxygen supply and mechanical ventilation are non-zero. We created two lists, one where the coefficients of days of oxygen supply are non-zero and the same with mechanical ventilation. So we got two list of proteins each with 2 columns containing the p-values of days of oxygen supply and mechanical ventilation. For each list one of the columns is complete and we can corrected the p-values in this column for multiple testing.

As the result of this method we obtained tables of proteins for the prediction variables for each secretion type (see appendix A.1). Furthermore all proteins own a p-value corrected for multiple testing. To further interpret the results we apply a p-value threshold. Here we set the significance level of the p-values to $p < 0.05$. The resulting lists of significant proteins are candidates to be protein biomarkers of BPD.

In Table 5, showing the number of significant proteins for each computation, we observe that blood and tracheal secretion have the highest number of significantly associated proteins. In contrast to these in the urine sets there are only significant p-values with the prediction BPD0. The reason can be the imputation of the missing values. Therefore we discarded a large part of the proteins and we set the constants of the imputation fixed. With the discarding we can loose significant proteins and the same can happen with a suboptimal imputation. The fundamental issues are the missing values due to the problems of mass spectrometry.

We computed days of mechanical ventilation and days of oxygen supply together, because we wanted to find out if they complement each other. We observed rarely cases where both variables have coefficients not equal zero. The results are so very similar to the list of proteins with only one prediction variable. Some differences can occur if lasso selected slightly different variables.

To further study the results, we especially analysed the proteins which are significant in multiple tables. These proteins have probably a close relationship to BPD. So our general idea to interpret these tables is that we analysed the intersections of

	Number of proteins in the result sets of		
	Blood	**Tracheal**	**Urine**
BPD0	21	28	2
BPD1	45	4	0
Mechanical ventilation	4	10	0
Mechanical ventilation + oxygen supply	2	0	0
Oxygen supply	13	16	0
Oxygen supply + Mechanical ventilation	13	18	0
T1	24	-	-
T2	0	-	-

Table 5: Number of significant proteins of the different secretion types and the different prediction variables. Mechanical ventilation = days of mechanical ventilation; Oxygen supply = days of oxygen supply. Days of mechanical ventilation + days of oxygen supply means that we both include in the model and select and correct days of mechanical ventilation. Days of oxygen supply + days of mechanical ventilation works the other way around.

the protein lists to find possible new protein biomarkers.

The Venn diagram in Figures 13 and 19 show the intersection of the blood protein sets. We can see that mostly only a small amount of the proteins are contained in 2 or more lists. The tracheal result in Figure 14 has also this feature.

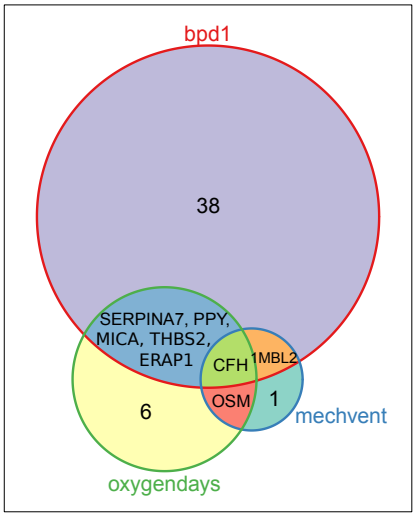

Figure 13: Venn diagram of the significant proteins of the blood dataset with the predictors BPD1, days of oxygen supply (oxygen days) and mechanical ventilation. The gene complement factor H (CFH) is the intersection between the 3 sets of proteins.

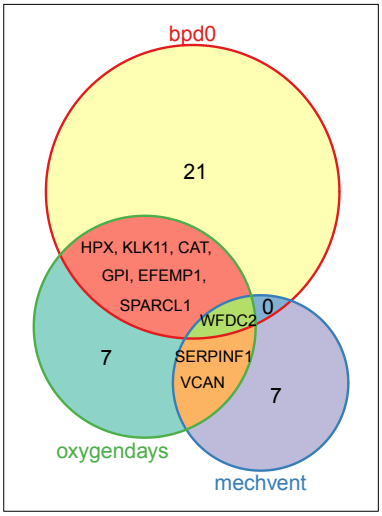

Figure 14: Venn diagram of the significant proteins of the tracheal dataset with the predictors BPD0, days of oxygen supply (oxygen days) and mechanical ventilation. The intersection of the 3 sets is the gene WAP four-disulfide core domain 2 (WFDC2).

The gene complement factor H (CFH) is present in BPD0, BPD1, days of oxygen supply and days of mechanical ventilation of the significant blood secretion results. The p-value of the coefficient BPD1 is $3.277 * 10^{-3}$. CHF is significantly associated with the disease when no and mild BPD is compared to medium and severe BPD. In Figure 15 we can see that after the elimination of the side effects of the clinical variables there is a distinct increase of the expression values by infants having a severe disease with a BPD degree of 2 and 3 to infants with a BPD degree of 0 and 1. The gene CFH is overexpressed in preterm infants with BPD. The encoded protein of the gene CFH has "an essential role in the regulation of complement activation, restricting this innate defense mechanism to microbial infections" [16]. Preterm infants have often infections so CFH as a biomarker can be reasonable.

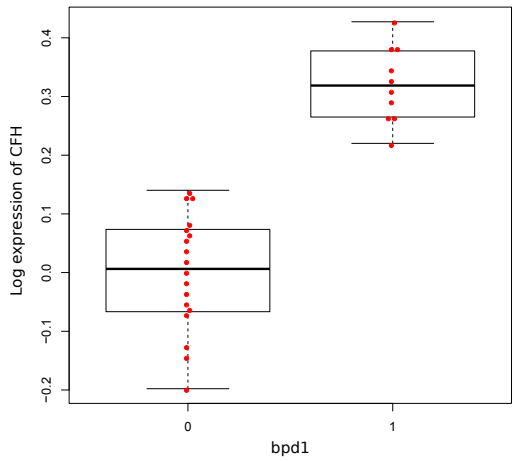

Figure 15: Boxplot of the logarithmic gene expression of complement factor H (CFH) and BPD1. In the plot we eliminate the effects of the clinical data by subtracting $\boldsymbol{\beta}_{clinical} * \mathbf{X}_{clinical}$ and the intercept of the linear regression model from the logarithmic gene expression. So the plot only shows the relationship of BPD with the protein expression. The corrected p-value of BPD1 in the linear regression is 0.003277.

An other candidate biomarker is the gene oncostatin M (OSM). Figure 13 shows that the gene exists in the lists of days of oxygen supply and days of mechanical ventialtion. In the plot in Figure 16 we can see a reasonable line of the data points. With increasing days of oxygen supply of the infants the expression of OSM increases. The p-value of days of oxygen supply is $9.035 * 10^{-3}$ in the linear regression with OSM. The encoded protein of OSM has a regulation function of the cytokine production,

like interleukin 6 (IL-6), colony stimulating factor 3 (CSF3) and colony stimulating factor 2 (CSF2) from the endothelial cells [3]. The pro-inammatory cytokine IL-6 induces broblast and collagen production in BPD (compare Section 3) [12]. This suggests that CFH can have a influence on BPD and can work as a protein biomarker.

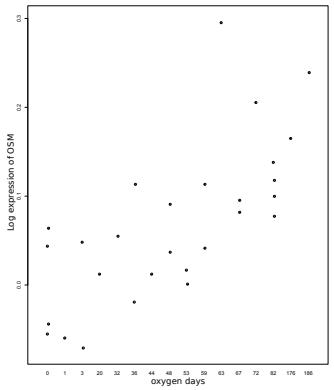

Figure 16: Plot of the logarithmic gene expression of oncostatin M (OSM) on the y-axis and days of oxygen supply (oxygen days) on the x-axis. In the plot we eliminate the effects of the clinical data by subtracting $\boldsymbol{\beta}_{clinical} * \mathbf{X}_{clinical}$ and the intercept of the linear regression model from the logarithmic gene expression. So the plot only shows the relationship of OSM with the protein expression. The corrected p-value of days of oxygen supply in the linear regression is 0.009035.

6.3 Results of the identification of proteins associated with MRI patterns

To possibly improve the previous results we perform the extended approach of Section 5.3. There we first search a new prediction out of the MRI data and than perform lasso and a linear regression with this new variable and the protein expression. We computed the first step with elastic net $\alpha = (0.1, 0.2, 0.5, 0.8)$ and with lasso $\alpha = 1$. We use the outputs days of mechanical ventilation, days of oxygen supply, BPD0, BPD1 and the grade of BPD. The results of this method are for each α a set of lists of coefficients for all variables in the MRI dataset.

In the heatmap in Figure 18 we observe that with lasso mostly clinical variables like gestational age, Respiratory Disease Syndrome (RDS) and gender are present. Variables like heart beat rate, which are universal and present in lots of diseases, do not really produce a new prediction. The coefficients in Figure 17 and 18 of the lung volume data are zero expect for the volume of the right lung.

The T1 and T2 variables stand out in the result. In the lasso result in Figure 18 there are positive T2 coefficients like T2_left_down and T2_right_top and a T1 variable T1_right_down. Here the terms left/right, down/top determine the position of the measurement in the lung. With lower α values the number of non-zero coefficients of T1 and T2 variables clearly increases in contrast to for example the lung volume variables (compare Figure 17 and appendix A.2). Because the T1 and T2 variables are closely correlated (Figure 8) this indicates that T1 and T2 have an strong relationship with BPD.

The T1 coefficients are generally negative and the T2 coefficients are positive. The coefficients of "BPD0" are an exception, because the variable represent the infants with no BPD. So these coefficients are refers to the coefficients of the other prediction variables. The negative coefficients of T1 and the positive coefficients of T2 denotes that with increasing severity of BPD the values of the T1 variables decrease and the T2 variables increase.

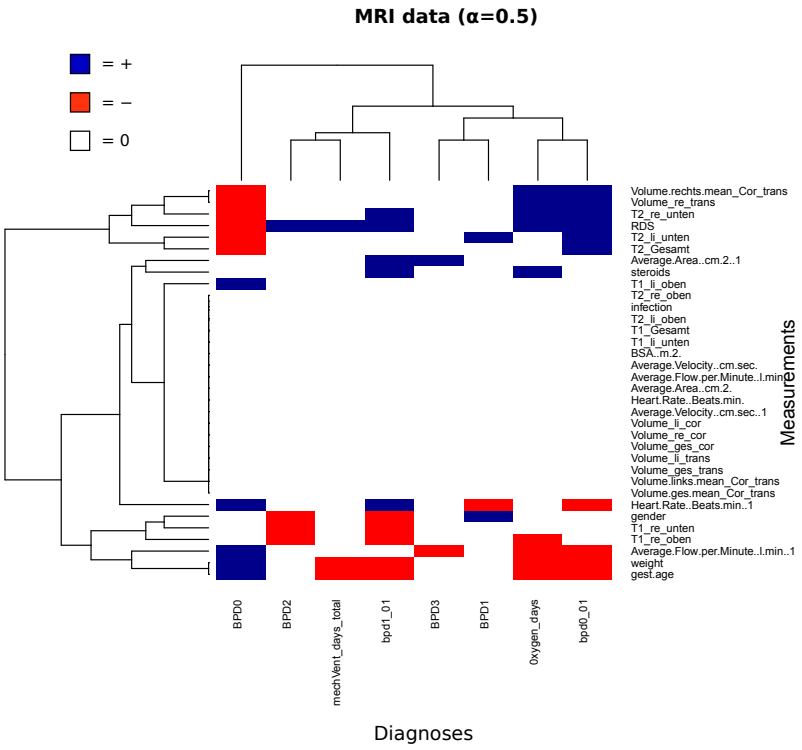

Figure 17: Heatmap of the elastic net results of the method in Section 5.3 with $\alpha = 0.5$. The heatmap shows the coefficients for each output variable. BPD0, BPD1, BPD2, BPD3 are the results of the output with the grades of BPD as multinomial variable. bpd0_01 corresponds to BPD0 in Table 4 and bpd1_01 corresponds to BPD1 in Table 4. mechVent_days_total = days of mechanical ventilation, Oxygen_days = days of oxygen supply. All negative coefficients are red and all positive coefficients are blue to increase the visibility of small coefficients. Coefficients with the value zero are white in the plot.

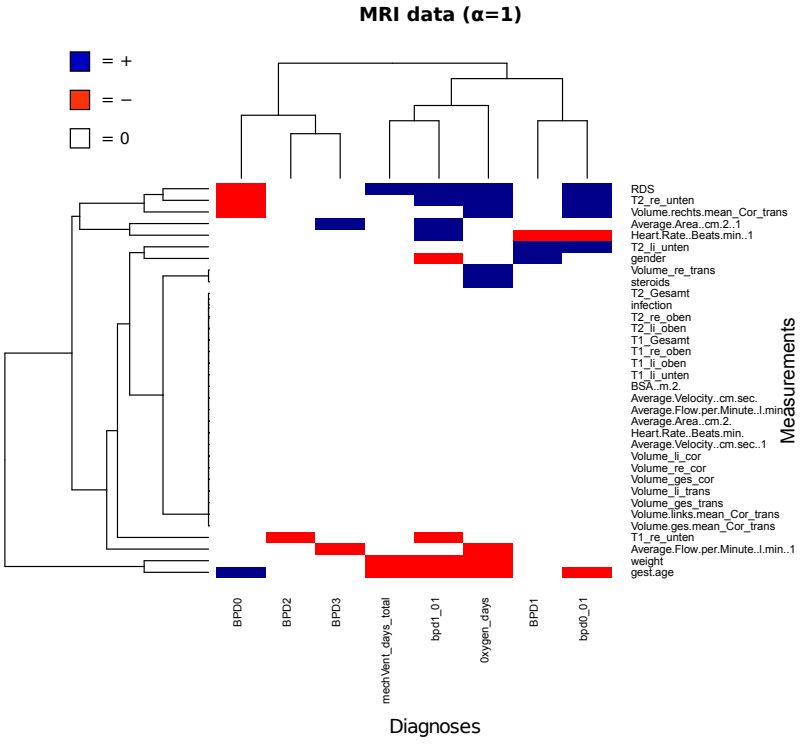

Figure 18: Heatmap of the lasso results ($\alpha = 1$) of the method in Section 5.3. The heatmap shows the coefficients for each output variable. BPD0, BPD1, BPD2, BPD3 are the results of the output with the grades of BPD as multinomial variable. bpd0_01 corresponds to BPD0 in Table 4 and bpd1_01 corresponds to BPD1 in Table 4. mechVent_days_total = days of mechanical ventilation, Oxygen_days = days of oxygen supply. All negative coefficients are red and all positive coefficients are blue to increase the visibility of small coefficients. Coefficients with the value zero are white in the plot.

With these observations and the clinical knowledge of our cooperation parters of the hospital of the LMU we decide to take T1 and T2 as our new predictions. We choose the T1 total (T1 Gesamt) and T2 total (T2 Gesamt). T1 and T2 total represent the sum of all T1 or T2 variables and are so a reasonable combination of these variables.

We calculate the second step of the method with T1 and independently with T2. The results are once again protein tables. By computing the method with the blood dataset we only get significant results with T1 (Table 5). The T1 set of proteins has no intersections with the other blood protein lists. This suggests that T1 of the MRI data is a distinctly other classifier than the other predictors. However we can consider these proteins as possible protein biomarkers of BPD. These proteins possibly represent an other feature of Bronchopulmonary displasia. But further statistical and biological investigation have to be done to get certainty.

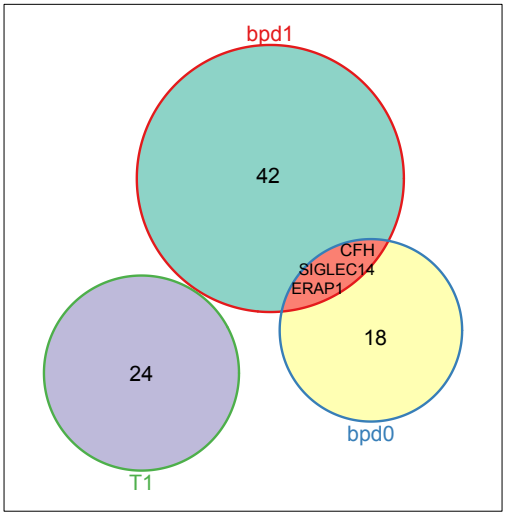

Figure 19: Venn diagram of the significant proteins of the blood dataset with the predictors BPD1, BPD0 and T1.

7 Summary and outlook

We created two different approaches to find protein biomarkers in Bronchopulmonary Displasia with statistical methods. The general idea of the approaches is to perform linear regression with the protein expression data, the clinical data and a diagnosis of BPD. In the first approach we use the conventional diagnosis, the days of oxygen supply and days of mechanical ventilation as the prediction. The second approach searches for a new diagnosis to use in the linear regression with a statistical model using elastic net and lasso with MRI data and the predictions of the first method.

We performed the first approach with blood, tracheal and urine protein expression data and obtained as results lists of proteins of each segregation type and of each diagnosis. To further analyse these results we examined intersections of the lists. As examples CFH and OSM seem to be reasonable candidate protein biomarkers of BPD.

By the extended approach with the MRI data we decided based on the results of elastic net to use T1 and T2 as the new prediction of BPD. In the resulting lists of the computation of the linear regression with T1 and T2 and the blood protein data we obtain only significant results with the T1 variable.

There are a good amount of blood and tracheal protein biomarkers but we have hardly no biomarkers of the urine dataset. The absence of significant proteins in the urine dataset and the prediction with T2 suggest that improvements of the methods are required or that there are no protein biomakers.

We can for example perform modifications on the input of the methods. Furthermore we can increase the amount of protein data by increasing the cutoff days for the measurements of the proteins. Generally the imputation of the missing values is a big issue. We can collect more data to address this problem. Also an improvement of the imputation of the missing protein expression values can be by changing the constants of the imputation method. It is possible that this can improve the resulting protein lists. To validate the protein biomarkers a validation on a control cohort is necessary.

With methods described here and their results we can create a set of reasonable early protein biomarkers. With the set we can possibly build a new model that performs as a classifier to predict BPD early after birth. So the goal is that the clinicians only have to measure protein expression values of a distinct set of proteins in a preterm infant and than can predict if this infant is prone to develop a BPD. This information can help to adjust the treatment and to take preventive steps against BPD. Also the information that some proteins have a influence on the disease can help to study the factors of the disease.

Furthermore we can also study the course of the disease on the protein level. It is possible to perform our methods with protein measurements in different periods of time. So we get sets of proteins which are characteristic for each period. A comparison between these sets can now help to describe the development of the disease on the protein level over time.

A Appendix

A.1 Resulting protein lists of the linear regressions with the protein expression and the predictions

A.1.1 Blood proteins

gene name	bpd0
BCAM	0.003553
FGF19	0.003941
FGFR1	0.006498
ICOS	0.006498
CNTN5	0.006498
CFH	0.006498
FCER2	0.007862
TNFRSF9	0.007908
EPHA5	0.008463
SIGLEC14	0.013000
OSM	0.016622
TOP1	0.018339
PRSS2	0.018360
IL13RA1	0.023378
LEPR	0.024194
NBL1	0.033236
ERAP1	0.033867
CFC1	0.035949
SERPINA1	0.035949
CST2	0.035949
ICOSLG	0.039646
CCL18	0.050286
CKM	0.069105
CNTN4	0.070016
SSRP1	0.070035
MICA	0.096067
GRN	0.116817
BPI	0.129134
ULBP1	0.133655
PIM1	0.133655

ASAH2	0.134459
THBS2	0.159510
PRSS1	0.183995
CDON	0.183995
SERPINA4	0.235944
IGFBP4	0.236809
CST1	0.238864
AGR2	0.246445
MPO	0.275685
SERPINA7	0.289186
DKK1	0.327124
PRTN3	0.335321
HAT1	0.341383
SERPIND1	0.380885
SERPINF2	0.411445
HIST1H1C	0.456441
RGMB	0.556689
CASP2	0.557140
EIF5	0.619423
PLA2G1B	0.681525
CKB	0.741141
RPS7	0.842886
TNFRSF13B	0.875571
HGFAC	0.875571
COLEC12	0.973419

Table 6: In the blood dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here with bpd0.

gene name	bpd1
CASP2	0.001637
COL18A1	0.003020
CHEK1	0.003020
MRC2	0.003020
FUT5	0.003020
ERAP1	0.003020
GPNMB	0.003020

TGFB2	0.003277
CFH	0.003277
JAG1	0.004544
ADAMTS13	0.007066
GREM1	0.008578
PROS1	0.009044
CLEC7A	0.009044
TNFRSF19	0.009044
CA1	0.009469
SERPINA7	0.009689
SPINT2	0.010268
AGT	0.010887
F7	0.011614
TNFRSF17	0.019574
PDGFRB	0.023035
CAST	0.025203
IL18BP	0.025203
MATN2	0.025203
SIGLEC14	0.026108
CSNK2A2	0.031703
IL11RA	0.033340
F11	0.035998
MB	0.035998
THBS2	0.035998
LTA	0.035998
CRP	0.035998
SELL	0.035998
CDNF	0.035998
PLA2G7	0.035998
EPHB2	0.035998
MMP8	0.036391
ROR1	0.036941
CDH15	0.038348
MICA	0.038936
LILRB1	0.038936
MBL2	0.040379
PPY	0.045014
SIGLEC9	0.049469

TNC	0.053354
PRSS22	0.062417
CA6	0.064378
CHIT1	0.074787
ULBP1	0.075218
CHST15	0.080015
IGHE	0.102354
IL13RA1	0.104567
CD209	0.104567
CD33	0.104567
FOLH1	0.104567
GPC5	0.104567
PTPN2	0.117012
HGFAC	0.140069
FGR	0.140069
FCN3	0.140069
FGFR1	0.162800
FLT4	0.182944
TNFRSF25	0.182944
SERPING1	0.184359
LEPR	0.184359
SERPINF2	0.195371
CCL27	0.197973
CST5	0.197973
CD163	0.197973
RTN4R	0.217730
IL16	0.224683
KLRK1	0.229903
PRSS2	0.230026
LSAMP	0.260471
BMX	0.277976
TPO	0.279114
CCL18	0.301885
ESAM	0.333536
CST2	0.342058
ICAM5	0.352583
ICAM1	0.362472
HMGCR	0.453365

MICB	0.508347
CXCL8	0.899009

Table 7: In the blood dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here with bpd1.

gene name	mechanic ventilation
MBL2	0.009102
CNTN5	0.009102
CFH	0.017959
OSM	0.025105
RTN4	0.135552
SIGLEC14	0.137908
MICA	0.217823
KLK12	0.287561
LEPR	0.287561
ULBP1	0.296855
ASAH2	0.296855
SIGLEC9	0.308802
CKM	0.490616
TOP1	0.490616
EPHA5	0.490616
CASP2	0.490616
CCL18	0.521258
SERPINA7	0.572529
CNTN4	0.572529
HGFAC	0.572529
FGFR1	0.574097
CKB	0.659614
CST2	0.717920
CST1	0.741420
ICAM1	0.902748

Table 8: In the blood dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here with mechanic ventilation oxygen days.

gene name	oxygen days
SERPINA7	0.002990
C3	0.003264
LEPR	0.003264
HGFAC	0.006540
PPY	0.006540
OSM	0.009035
MICA	0.027693
THBS2	0.041904
ICAM5	0.045104
EPHA5	0.045486
CFH	0.045486
ERAP1	0.045486
ICOSLG	0.047422
CCL18	0.055841
MST1	0.064956
CA6	0.075285
BPI	0.086402
IL13RA1	0.089845
SIGLEC9	0.089845
TPO	0.089845
SIGLEC14	0.089845
CGA	0.091659
BCAM	0.094721
FGFR1	0.096614
GPNMB	0.096614
LRIG3	0.104605
TNFRSF25	0.104605
PLAUR	0.109087
RETN	0.118543
CNTN4	0.126206
ACP1	0.144381
PRSS2	0.191960
MBL2	0.208418
CDON	0.208418
PRSS1	0.240323
PDGFRB	0.262640

FGFR2	0.267745
APOE	0.267781
CASP2	0.286299
CKM	0.300352
CST2	0.348311
RPS7	0.458322
TOP1	0.603322
FGR	0.603322
PIM1	0.603322
CXCL8	0.865736

Table 9: In the blood dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here with oxygen days.

row	mechanic ventilation	oxygen days
MBL2	0.004733	
CNTN5	0.004733	
ULBP1	0.339602	
KLK12	0.339602	
ASAH2	0.339602	
CKB	0.417287	0.363065
TOP1	0.583133	
ADIPOQ	0.817044	0.718324
SIGLEC14	0.817044	0.080561
CST2	0.841175	
CST1	0.841175	
FGFR2	0.902748	0.718324
ICAM1	0.902748	

Table 10: In the blood dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here we compute it with mechanic ventilation and oxygen days and mechanic ventilation is selected. Only the p-values of mechanic ventilation are corrected.

gene name	mechanic ventilation	oxygen days
SERPINA7		0.002795
C3		0.003051
LEPR		0.003051
HGFAC		0.006113
PPY		0.006113
OSM		0.008445
MICA		0.025887
THBS2		0.039171
ICAM5		0.042162
EPHA5		0.044329
CFH		0.044329
ERAP1		0.044329
ICOSLG		0.044329
CCL18		0.052199
MST1		0.060719
CA6		0.070375
BPI		0.080767
SIGLEC9		0.087098
IL13RA1		0.088185
TPO		0.088185
CGA		0.089762
FGFR1		0.098167
GPNMB		0.098167
LRIG3		0.102488
TNFRSF25		0.102488
SIGLEC14	0.573575	0.102488
PLAUR		0.105749
RETN		0.114770
CNTN4		0.124832
PRSS2		0.182612
CDON		0.208581
PRSS1		0.245711
PDGFRB		0.267830
APOE		0.279766
CASP2		0.298213
CKM		0.311960

CKB	0.207039	0.357026
RPS7		0.473530
FGR		0.634472
PIM1		0.634472
ADIPOQ	0.579114	0.719867
FGFR2	0.849510	0.735427
CXCL8		0.865736

Table 11: In the blood dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here we compute it with mechanic ventilation and oxygen days and oxygen days is selected. Only the p-values of oxygen days are corrected.

gene name	T1_Gesamt
MPO	0.00347385597573767
STAB2	0.00652314972580249
C3	0.00652314972580249
LIFR	0.00652314972580249
CKM	0.00880294869810828
TDGF1	0.00891632948850196
HAT1	0.00891632948850196
PRTN3	0.0100580060292269
TGFB1	0.0115938201212218
HMGB1	0.0115938201212218
NAAA	0.0118641409989654
SERPINA4	0.0118641409989654
BSG	0.0118641409989654
MAP3K7	0.0127150371009887
TYK2	0.0127150371009887
EPHA1	0.0173038413430662
SLC25A18	0.0173038413430662
F9	0.0173038413430662
ASAH2	0.0178694391090733
FCGR2A	0.0178694391090733
PTGS2	0.0178694391090733
CTSA	0.0271180801313128
AGR2	0.0356361623048741

SERPIND1	0.0362192195351234
AGER	0.0591499538238937
ITGA1	0.0832256745056885
HIST1H1C	0.0879480945028115
TYMS	0.0937223068609075
F9	0.112501609107036
EPHB4	0.169657671345404
CST1	0.169657671345404
ADAM12	0.170547456016056
AHSG	0.194219873232884
PCNA	0.214282703491488
FCN3	0.222258238526902
ERAP1	0.33951421841372
MST1	0.361342793570273
LTA	0.451330662580831
PDGFRB	0.480624942387435
IGHG1	0.547891671495786
TNFRSF12A	0.617013159106812
TPO	0.621112349129242
GSTA3	0.762319192491334
OLR1	0.81087395700346
PSME3	0.829115201034299
CCL18	0.841931000912589
KIR3DL2	0.84553983917542
KDR	0.869749751077632
CFH	0.884660814429189

Table 12: In the blood dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here with T1 Gesamt.

	T2_Gesamt
ULBP2	0.145685100583298
NOTCH3	0.145685100583298
SIRT2	0.206197825506219
IL27	0.218884839617593
MMP8	0.218884839617593
CDH2	0.218884839617593

ESD	0.218884839617593
HMGB1	0.250411831357224
MMEL1	0.305911651800059
OLR1	0.305911651800059
C5	0.307696464591949
LRRTM3	0.311873246931524
PTGS2	0.392220104819684
ICAM1	0.394445888394011
CST1	0.394445888394011
SHBG	0.606907585824095
BIRC3	0.606907585824095
REN	0.80426650531997
CASP2	0.80426650531997
RETN	0.863112010567837
CFH	0.863112010567837

Table 13: In the blood dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here with T2 Gesamt.

A.1.2 Urine proteins

gene name	bpd0
CDH11	0.039741
S100A9	0.039741
PGLYRP2	0.710985

Table 14: In the urin dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here with bpd0.

gene name	oxygen days
HEXB	0.548172
S100A9	0.548172

Table 15: In the urin dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here with oxygen days.

gene name	mechanic ventilation	oxygen days
HEXB		0.548172
S100A9		0.548172

Table 16: In the urin dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here we compute it with mechanic ventilation and oxygen days and oxygen days is selected. Only the p-values of oxygen days are corrected.

A.1.3 Tracheal proteins

gene name	bpd0
HPX	0.000107
ANXA5	0.000107
DAG1	0.000163
MYH9	0.000271
ANXA2	0.000502
BASP1	0.000958
ANXA4	0.000968
ATP1A1	0.002654
COL6A3	0.004358
AZGP1	0.004812
KLK11	0.006116
CLIC1	0.006308
LCP1	0.007785
CAT	0.007785
ANXA1	0.010546
SPTAN1	0.010978
GPI	0.010978
ANXA3	0.010978
TUBB	0.010978
EFEMP1	0.010978
CD36	0.011121
WFDC2	0.014275
WDR1	0.017994
PSAP	0.021352

SPARCL1	0.025690
COL6A1	0.025690
PRDX2	0.026645
FABP5	0.028271
ITIH4	0.052163
SERPINF1	0.052163
LDHB	0.060999
CST3	0.075599
MSLN	0.076077
CTSB	0.100665
PGD	0.166084
FOLR1	0.166084

Table 17: In the trachial dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here with bpd0.

	bpd1
TPPP3	0.026880
FLNA	0.026880
ANXA3	0.026880
ARHGDIA	0.026880
PGD	0.175664
MSLN	0.257452
ANXA4	0.257452
HNRNPK	0.301009
SERPINF1	0.523306

Table 18: In the trachial dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here with bpd1.

gene name	mechanic ventilation
SCGB3A1	0.008852
CD59	0.008852
ANXA6	0.008852
C6	0.008852

HNRNPK	0.008852
VCAN	0.008852
SERPINF1	0.008852
MUC16	0.014263
WFDC2	0.020104
C2	0.024424
CIB1	0.096242
CLIC6	0.096242
EFEMP1	0.096242
TPPP3	0.100086
KLK11	0.195193
ALDH3B1	0.340423
SCGB1A1	0.477492
PGD	0.683010
PROM1	0.979342

Table 19: In the trachial dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here with mechanic ventilation.

gene name	oxygen days
B3GNT1	0.000704
CAT	0.001380
HPX	0.005248
FLNA	0.005248
VCAN	0.005248
EFEMP1	0.005248
SERPINF1	0.005248
KLK11	0.006370
GPI	0.006370
WFDC2	0.006370
S100A4	0.006370
FAM3C	0.006370
CTSB	0.006370
SPARCL1	0.010099
PGD	0.042071
LDHA	0.043611
HNRNPK	0.055025

MUC16	0.127404
CD59	0.128460
WDR1	0.137277
TMC5	0.243735
ALDH3B1	0.303693
MSLN	0.349257
CST3	0.394476
C2	0.437655
TPPP3	0.458581

Table 20: In the trachial dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here with oxygen days.

gene name	mechanic ventilation	oxygen days
SCGB1A1	0.104310	
CD59	0.108527	0.991835
PIP	0.144988	
CD14	0.726054	0.099512
PROM1	0.979342	
KLK11	0.979342	0.212106
SDF4	0.979342	
VCAN	0.979342	0.012117

Table 21: In the trachial dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here we compute it with mechanic ventilation and oxygen days and mechanic ventilation is selected. Only the p-values of mechanic ventilation are corrected.

gene name	mechanic ventilation	oxygen days
C6		0.000264
FAM3C		0.002465
RNASE4		0.002941
DAG1		0.003331
EFEMP1		0.004367

SERPINF1		0.004367
CAT		0.004587
HPX		0.004710
FLNA		0.004710
PLTP		0.005088
WFDC2		0.006743
S100A4		0.010386
SPARCL1		0.012030
CD36		0.012030
MUC16		0.012880
VCAN	0.657078	0.014200
ATRN		0.017876
PGD		0.040453
GPI		0.052753
LMNB1		0.076350
ITGB2		0.087378
SERPINA1		0.096407
CD14	0.388957	0.097349
EEF1A1		0.100336
CIB1		0.154495
KLK11	0.906278	0.214145
MSLN		0.343286
CST3		0.390141
AHNAK		0.533835
CD59	0.030147	0.991835

Table 22: In the trachial dataset the measurement results in the first week were applied to Lasso and after to a normal linear regression. Here we compute it with mechanic ventilation and oxygen days and oxygen days is selected. Only the p-values of oxygen days are corrected.

A.2 Results of the regressions of the MRI data and the predictors

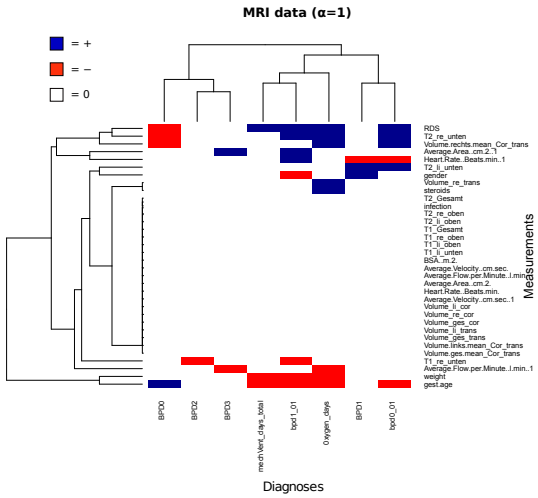

Figure 20: Heatmap of the lasso results $\alpha = 1$ of the method in Section 5.3. The heatmap shows the coefficients for each output variable. BPD0, BPD1, BPD2, BPD3 are the results of the output with the grades of BPD as multinomial variable. We set all negative coefficients to -1 and all positive coefficients to +1 to increase the visibility of small coefficients. Coefficients with the value zero stay zero in the plot.

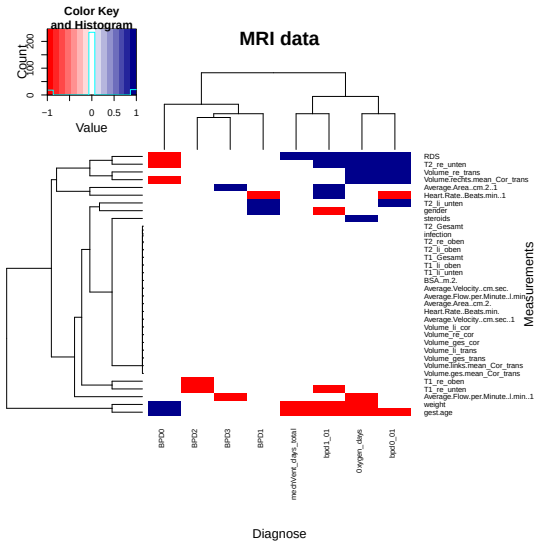

Figure 21: Heatmap of the elastic net results $\alpha = 0.8$ of the method in Section 5.3.

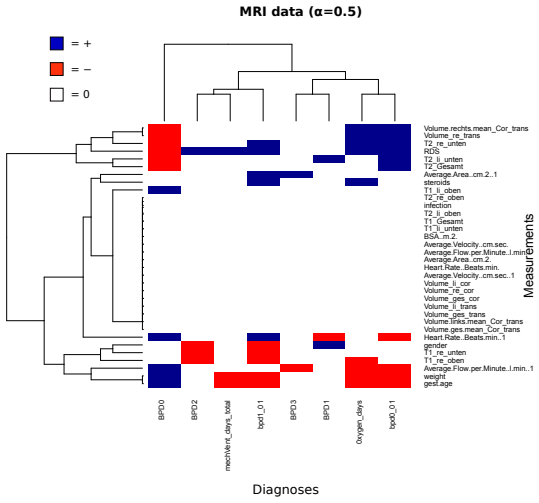

Figure 22: Heatmap of the elastic net results $\alpha = 0.5$ of the method in Section 5.3.

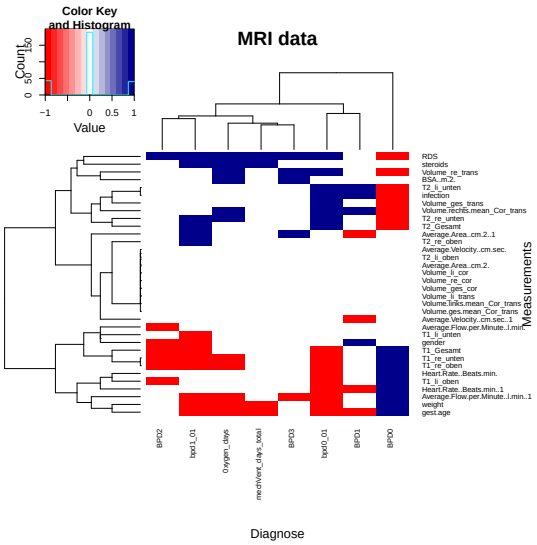

Figure 23: Heatmap of the elastic net results $\alpha = 0.2$ of the method in Section 5.3.

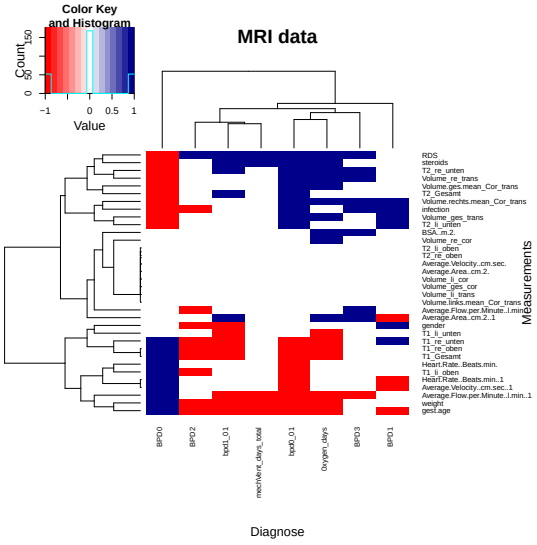

Figure 24: Heatmap of the elastic net results $\alpha = 0.1$ of the method in Section 5.3.

References

[1] N. Leigh Anderson and Norman G. Anderson. "Proteome and proteomics: New technologies, new concepts, and new words". In: *ELECTROPHORESIS* 19.11 (1998), pp. 1853–1861. ISSN: 1522-2683. DOI: 10.1002/elps.1150191103. URL: http://dx.doi.org/10.1002/elps.1150191103.

[2] American Lung Association et al. "American Lung Association lung disease data: 2008". In: *Retrieved April* 21 (2008), p. 2010.

[3] K. Berlin et al. *Methods and nucleic acids for the analysis of gene expression associated with tissue classification*. US Patent App. 12/088,384. 2009. URL: https://www.google.com/patents/US20090191548.

[4] Jürgen Biederer et al. "MRI of the lung (2/3). Why when how?" In: *Insights into imaging* 3.4 (2012), pp. 355–371.

[5] Jürgen Biederer et al. "MRI of the lung (3/3)current applications and future perspectives". In: *Insights into imaging* 3.4 (2012), pp. 373–386.

[6] Stef Buuren and Karin Groothuis-Oudshoorn. "mice: Multivariate imputation by chained equations in R". In: *Journal of statistical software* 45.3 (2011).

[7] Jürgen Cox and Matthias Mann. "MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification". In: *Nature biotechnology* 26.12 (2008), pp. 1367–1372.

[8] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "glmnet: Lasso and elastic-net regularized generalized linear models". In: *R package version* 1 (2009).

[9] Jason Gien and John P Kinsella. "Pathogenesis and treatment of bronchopulmonary dysplasia". In: *Current opinion in pediatrics* 23.3 (2011), p. 305.

[10] Wilson Wen Bin Goh and Limsoon Wong. "Computational proteomics: Designing a comprehensive analytical strategy". In: *Drug discovery today* 19.3 (2014), pp. 266–274.

[11] Trevor Hastie, Rob Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Springer, 2011.

[12] John P Kinsella, Anne Greenough, and Steven H Abman. "Bronchopulmonary dysplasia". In: *The Lancet* 367.9520 (2006), pp. 1421–1431.

[13] Johan Malmström, Hookeun Lee, and Ruedi Aebersold. "Advances in proteomic workflows for systems biology". In: *Current opinion in biotechnology* 18.4 (2007), pp. 378–384.

[14] *SOMAscan Proteomic Assay, Technical White Paper*. Accessed: 2015-09-01. 2014. URL: http://www.somalogic.com/somalogic/media/Assets/PDFs/SSM-002-Rev-2-SOMAscan-Technical-White-Paper-3-7-15.pdf.

[15] *What is Mass Spectrometry?* Accessed: 2015-09-01. URL: https://www.broadinstitute.org/scientific-community/science/platforms/proteomics/what-mass-spectrometry.

[16] Robert YL Zee, Kirsti A Diehl, and Paul M Ridker. "Complement factor H Y402H gene polymorphism, C-reactive protein, and risk of incident myocardial infarction, ischaemic stroke, and venous thromboembolism: a nested case–control study". In: *Atherosclerosis* 187.2 (2006), pp. 332–335.

[17] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.

[18] Marketa Zvelebil and Jeremy Baum. *Understanding bioinformatics*. Garland Science, 2008.