Faculty of Science and Technology

Department of Mathematics and Statistics

Modelling and analysis of health care services using regression and Markov models

—

Lars Bakke Hindenes

STA-3900 Master's Thesis in Statistics May 2017

”Remember to breathe and then take one step at a time.”

— Unknown

”The greatest moments are those when you see the result pop up in a graph or in your statistics analysis - that moment you realise you know something no one else does and you get the pleasure of thinking about how to tell them.”

— Emily Oster

Abstract

Using data from electronic health records this thesis aims to model and analyse health care services provided to adult patients with chronic conditions. Two aspects of health care services, with unique aims, have been examined.

The first aspect is related to the aim of investigating factors affecting the patients' self experienced quality of the health care encounters with regards to satisfaction, personalized help and general information received. Significant factors were determined by odds ratios resulting from either logistic or multinomial regression, combined with generalized boosted regression.
The main findings included: Better self perceived health, increased age, the absence of long lasting illness and not having experienced debasing, accounted positively for the odds of being satisfied. Factors implying a sicker patient increased the odds of receiving help and information; though higher age reduced the odds. Specifically regarding receiving personal help, higher level of education showed an increase in the odds. There were also indications that satisfaction could be negatively correlated with the amount of help and information received.

The second task has been to construct discrete-time patient trajectories, consisting of unique states or events that describe health service usage. Using such patient trajectories this aspect's aim is to model and describe changes and stability in health service usage, and predict future health care events using discrete-time Markov chain and hidden Markov models. Estimation was performed by maximum likelihood and trained by the Baum-Welch algorithm. Both Markov models were justified to describe certain perspectives of health care utilization. Prediction of future health events was only theoretically adequate using hidden Markov models, but its accuracy was unsatisfactory. Also the hidden states of the hidden Markov model, with unknown physical interpretation in a patient trajectory setting, can be induced to represent complex health levels or indices for patients.

Acknowledgements

I would like to express my deepest thanks and gratitude to my supervisor Sigrunn Holbek Sørbye for providing invaluable advice, guidance and support, virtually whenever I required it. You have certainly gone the extra mile to help realize this thesis, while also being a great source of inspiration. Without everything that you have provided from day one, I'm not sure if this thesis in statistics would ever have existed.

A huge thanks to my co-supervisors Gro Berntsen and Aslak Steinsbekk for lending us the project data, reviewing the text and spending their spare moments with hours of discussion while sharing insight. The medical knowledge you learned me, combined with the shower of enthusiasm towards my findings have most definitely been fuelling my motivation.

To all my friends, both new and old: The times with you all have been the best and hopefully they will continue to persist. Especially the times that took place during rainy days.

Finally, to my closest family: Words cannot describe how much I appreciate you for always being there with advice, encouragement and unconditional support.

Contents

Chapter 1

Introduction

1.1 Background

Ensuring that as many patients as possible receive the best health care service and assistance they require is important. One possible way to improve the health care provided, is by using health or medical data from registers. It is therefore unfortunate that large amounts of data about patients from electronic patient journals, or electronic health records (Tang and McDonald, 2006), are not used efficiently enough with all their potential to improve the health services provided (Jensen et al., 2012). Electronic patient journals, if allowed to be collected across different services, can provide medical information about every patient that has been treated.
Most electronic health record systems (Tang and McDonald, 2006), that visualize information in electronic patient journals, do not offer advanced yet easily interpretable information that could support clinical tasks (Rind et al., 2013). There exists earlier attempts using different techniques trying to improve health services through the use of electronic health record data. One example is using natural language processing (Chowdhury, 2003) to classify clinical text notes (Perlis et al., 2012), and another is identifying types of clinical note sections in written notes by using hidden Markov models (Li et al., 2010). Jensen et al. (2012) describe more general examples in addition to providing visualizations and mentioning data management considerations, with regards to the underused source of medical information.

1.2 Overview of PAsTAs' data

The analyses in this thesis will be using and analyzing data that describes how adult patients with chronic diseases interact with the health care system. The data is part of and have been borrowed from a project named PAsTAs (Patient Trajectories). See 'https://www.ntnu.no/wiki/display/pastas/HOME' for the project's home page with more details. The data includes visits to somatic health care services among all inhabitants in a geographical area from 2012 to 2013, and is collected from electronic patient journals or patient administrative systems.

PAsTAs' data include three fully detailed datasets with information across different services, and one aggregated dataset created from the detailed datasets. We do also have a set of data from a questionnaire (questionnaire included in section 10.2.2), which a subset of patients were selected to answer. The questionnaire data combined with the aggregated data provide qualitative and quantitative frequency information about health care utilisation.

The three detailed sets of data contain information from three separate sources that complement each other in terms of explaining the health service usage. One of the sets contains records of somatic patients that received care at or were admitted to the St. Olavs hospital during 2012 and 2013. The dataset is named after the hospital. Another set of data, named Kuhr dataset, contains the information about patients' use of general practitioners and other health care specialists outside of hospital care. The last of the three sets, named PLO (i.e. "pleie og omsorg") dataset, have information about patients that received services from the municipalities, for example receiving extra help at home or being admitted to a nursing home. The three sets of detailed patient data will be referred to as: detailed dataset, fully detailed data or equivalently. The datasets contain more than nine million entries in total, and the only common factor is the unique anonymized identification number constructed for each patient. This identification number is important as it is used to select and extract the chronic patients' events to be used from these sets of data.

In addition to the detailed datasets, PAsTAs include other smaller sets of data, but these have been used to a lesser degree and only to manage and restructure the larger datasets.

1.3 Aims and motivations of the thesis

Based on the large amount of data across different health services from PAsTAs, our main objective is to analyze and model health service data as a case study, illustrating and using statistical techniques that are assisted by machine learning approaches. The main objective includes different tasks:

- A central task is to evaluate significant factors or covariates associated with three different measures of experienced quality from the health care services the patients have received. The qualities measured are with respect to the experienced satisfaction and the experienced degree of personalized and general information help received. Personalized and general information help can together be thought of as a collective measure of guidance received, but we will often treat them separately.

- Another task is to model, describe and present what we will refer to as patient trajectories, which illustrate patients' health care utilization. Based on the trajectory models we also want to figure out if these can be used to predict future health care events or states at least one step ahead.

Knowing which significant factors that are affecting experienced quality, health care personnel could then with the help from administration be able to improve interpersonal contact and the treatment of patients. Similarly, if realistic trajectory models can be created, then it would be possible to integrate trajectory models into a clinical decision support or risk identification systems. Such a decision support or risk identification system should then again provide better support to clinical and medical administrative tasks. It is reasonable to expect an improvement since the use of health information technology has shown to improve certain aspects of medical care (Himmelstein et al., 2010), though a successful implementation may heavily rely on human factors or elements (Buntin et al., 2011).

1.4 Patient trajectories and an illustration

From a conceptual and theoretical point of view, any individual will have different trajectories that describe certain aspects of their life per time. What data trajectories from a population can explain, is only bounded to what is measurable from the individuals. In other words, the term trajectories could be interpreted differently depending on previous experiences. Let's introduce

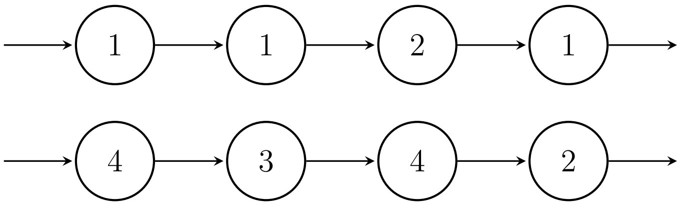

Figure 1.1: Example: One way to visualize two individuals' discrete-time patient trajectory or sequence.

how the trajectories in this thesis are structured and how these can be understood.
We will specifically look at patient trajectories in the health care services. Our patient trajectories will be created with information from electronic patient journals, since they contain health care events registered to a patient. These trajectories will then be explaining and visualizing what kinds of health services that a patient has received at different times. We will limit these times to be discrete-time events for certain modelling purposes. In other words we will not use continuous time and the trajectories can be thought of as discrete sequences or vectors of events per patient. The length and precision of those sequences are only limited to the amount of patient data available to us. As we have access to two years worth of data, we can for instance create trajectories that are 24 months long. Larger sets of data will allow extra flexible trajectories or sequences to be created for even more optimal and finely tuned models. Every detail regarding the states or events and how the patient trajectories are to be constructed will be provided later in chapter 5.
Now consider an example of discrete-time patient trajectories. Suppose that we have four unique disjoint events of health care that a patient can receive, numbered *1*, *2*, *3* and *4*. Assume also for simplicity that we have observations corresponding to three time units from the past. Then the patient trajectories, of discrete time, can be visualized as beads of events (Figure 1.1) on a line representing the transition to the right per time unit. A matrix or a two dimensional line plot can also be used to visualize the trajectories for one or many trajectories.

1.5 Outline of the thesis

This thesis has many aims, and several and different statistical methods were required to perform the analyses. Chapter 2 to 4 will present all the necessary baseline theory used later, chapter 5 contains an intermediate preprocessing chapter presenting preliminary considerations and constructs before the analyses. Including both regressions and Markov models the analyses with results, conclusions and interpretations are presented in chapter 6 to 8. Discussion of the results and interpretations are given in chapter 9 and the appendix is in chapter 10. The specific outline is as:

Chapter 2 describes the logistic regression model to be used within generalized linear models and within the generalized boosting regression model framework. Inferential measures, for example Akaike information criterion and odds ratios, that are to be used to perform inference about the regression are also defined here.

Chapter 3 describes the multinomial regression model. The multinomial regression will be presented within two different frameworks, namely integrated nested Laplace approximations and neural networks. Additional inference measures used with Bayesian statistics will also be included.

Chapter 4 is the last chapter with theory and describes two different Markov models, namely discrete-time Markov chain models and hidden Markov models. Theory about how to estimate or train models, and how to handle and predict states from these models will be presented such that the analyses in chapter 8 can be executed.

Chapter 5 focuses on the data itself, its properties and how it has been modified or transformed to fit into the models presented in earlier chapters. This chapter can be thought of as a preliminary analysis before chapter 6, 7 and 8. There will also be presented an overview of all the core elements, i.e. predictors and states, that are used in the analyses later on. New variables constructed based on the data will also be explained in detail this chapter.

Chapter 6 applies the methods from chapter 2 in an analysis of the first quality measure. In this analysis we will try to find significant effects relatively to whether the patients were above averagely satisfied or not. We will mainly use generalized linear models and then try to validate or debunk these results by using boosting as an extra verification measure.

Chapter 7 puts the methods from chapter 3 into practice in two analyses of the second and third quality measure. Here we look at how much help each patient received from the health care and then try to find significant effects in the multinomial regression setting. Instead of using boosting to validate or debunk the results, we will in this chapter use boosting to find the most significant predictors to the multinomial regression models.

Chapter 8 bases its analyses on the theory from chapter 4 to provide detailed descriptions or models of the patient trajectories. Both Markov models will here be estimated or trained and then interpreted. The capabilities of the models will also be assessed. If a model is deemed theoretically appropriate, the model will be used to try and predict the future states one step ahead in time.

Chapter 9 discusses the most important indications, conclusions and effects from the earlier analyses in chapter 6, 7 and 8. After the discussion of the key points at the end, a section about possible future work based on the results and discussion in this thesis is also included.

Part I

Theory and methods

Chapter 2

Logistic regression using GLM and boosting

The methods presented in this chapter are used to perform logistic regression and provide inference for the resulting logistic regression models. The two different main frameworks used to perform regression is generalized linear models from a classical statistical setting, and boosting or generalized boosted regression models that is a machine learning technique (Murphy, 2012). Two frameworks are presented, because we want to use them both to find an optimal model.

2.1 Generalized linear models (GLM)

Generalized linear models (Nelder and Wedderburn, 1972) is a framework used to perform linear regression over a broader range of distributions. The main idea is to link observations to a linear predictor using a transformation. Mathematically the generalized linear model can be expressed as:

$$E(Y_i) = \mu_i = g^{-1}\left(\sum_{j=1}^{m} x_{ij}\beta_j\right), \; i = 1, \ldots, n \tag{2.1}$$

where $g^{-1}(\cdot)$ is the transformation that is used on the linear predictor. μ_i represents the value an individual or observation, i, is expected to have given measured covariates, x_{ij}, and estimated model parameters, β_j. m is defined to be the number of unique covariates included. n is here the number of observations. Y_i is the random response that has yet to realise a value y_i.

Generalized linear models can be thought of as being a regression method that is more general than ordinary least squares linear regression. To apply these models there are three main criteria that have to be fulfilled by the problem or data one is working with. They have to be satisfied in order to utilize the generalized linear models properly from a theoretical point of view. The three criteria are as follows:

1. The first criterion is that the distribution of the response, Y, can be written in a general fashion, namely as a member of the exponential family. This implies that the density of the response and parameters (for instance the mean and standard deviation in the Gaussian distribution) is defined as follows:

   $$f(y;\theta) = c(y)d(\theta)e^{a(y)b(\theta)}, \tag{2.2}$$

   where and θ is one parameter. Also the θ terms can be of higher dimension depending on the actual distribution. The $c(y)$, $d(\theta)$, $a(y)$ and $b(\theta)$ are functions of either the response or the model parameters, not both. Examples of common distributions that belong to the exponential family are the Gaussian, binomial, Poisson and gamma distributions.

2. The second criterion is related to the linear predictor. The linear predictor is typically on the form:

   $$\eta_i = \sum_{j=1}^{m} x_{ij}\beta_j, \tag{2.3}$$

   where x_{ij} is the j'th measured predictor variable for the i'th individual and the $\beta_j, j = 1, \ldots, m$ are the unknown model parameters to be estimated through for example weighted least squares.

3. The third criterion requires the presence of a link function. The link function is a transformation that relates the linear predictor to the mean of the response Y, $E(Y) = \mu$. The link function is often denoted as $g(\cdot)$, defined by,

   $$g(\mu_i) = \eta_i. \tag{2.4}$$

   Alternatively, the expectation is expressed by,

   $$\mu_i = g^{-1}(\eta_i). \tag{2.5}$$

   There are a variety of different link functions to be used for various situations and distributions. For example the identity link has the property that the generalized linear model becomes the specialized case of ordinary least squares regression. For each distribution in the generalized linear model framework there exists a link function that has been named a canonical link. It is named canonical if the linear predictor, $\eta_i = \theta_i$ in (2.2). For example we have that the logit link function is the canonical link function for a binomial distribution and the identity link is canonical for the Gaussian distribution.

2.1.1 Model selection: An automated generalized linear model procedure

In the regression analysis we wish to select covariates to be included in the generalized linear model, and since we are in the case of having a lot of different covariates we need automated procedures. There exists automated procedures that attempts to fit an optimal generalized linear model with regards to a criterion; for instance the Akaike information criterion (Venables and Ripley, 2002).

The Akaike information criterion (AIC) is a measurement of the loss of information using a fitted versus the true model and is used to compare different regression models. Intuitively one looks for the models with relatively small or the smallest AIC value, to minimize the loss of information. The AIC is defined as

$$\mathrm{AIC} = -2\log(L) + k \cdot m_p, \tag{2.6}$$

and can be thought of as a measurement of the trade-off between goodness-of-fit and model complexity. The likelihood is represented as L, while the number of parameters in the model is m_p. The positive constant k in (2.6) is usually set equal to 2, and it could be considered a penalizing constant. In R, the generalized linear model regression fit utilizes $k = 2$.

Automated model selection procedures are handy when there are a lot of different combinations of predictors or covariates that could be included in the model. An automated procedure will thus help save a lot of time fitting different models, instead of doing it manually. Specifically, in R there is a function named *step* (R Core Team, 2016), that given an outset or base model formula, a (saturated) scope model formula and a direction, will attempt to find the best model by minimizing the Akaike information criterion. At each step it evaluates how the addition or removal of a predictor will affect the criterion in question. Thus when the AIC is sufficiently minimized the procedure will stop and return the supposedly optimal model.

2.1.2 The logistic regression case

Logistic regression is a special case of the generalized linear model in which the response variable only takes two different values, often described as success versus failure.
Suppose that Y_i is the response of one Bernoulli trial in a binomial distribution. We can then define the probability of a success and a failure as:

$$p = P(Y_i = 1), \text{ and } 1 - p = P(Y_i = 0), \tag{2.7}$$

where p is defined as the probability of a success happening. Based on the Bernoulli trials then $\sum_{i=1}^{n} Y_i$ is binomially distributed when the response variables $Y_1, \ldots, Y_n$ are independent. The binomial distribution belongs to the exponential family and the second criterion is satisfied since we can use the logit link function, among other link functions,

$$g(p) = \log\left(\frac{p}{1-p}\right). \tag{2.8}$$

The linear predictor can then be constructed and we have a special case within the generalized linear model framework. Setting equation (2.8) equal to η (or replace $g(p)$ with η) and taking the inverse of it we end up with the logistic function used to calculate the probability of success, given the linear predictor,

$$p = \frac{1}{1+e^{-\eta}}. \tag{2.9}$$

The logistic function will ensure that the estimated probability for success, p, is no greater than one or less than zero, making the estimate viable.

2.1.3 Odds ratio

A common way to interpret results from logistic regression is in terms of odds ratios. The odds of an event is defined as:

$$\text{odds} = \frac{P(Y_i = 1)}{P(Y_i = 0)} = \frac{p}{1-p}. \tag{2.10}$$

Having defined the odds, we can define odds ratio (OR) as:

$$\mathrm{OR} = \frac{\mathrm{odds}_1}{\mathrm{odds}_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}. \tag{2.11}$$

The subscript notation with odds_1 and odds_2 is used to illustrate the ratio between two distinct odds of events occurring. The logit link relationship,

$$\eta = \log\left(\frac{p}{1-p}\right), \tag{2.12}$$

makes convenient use of the odds ratio by exponentiating it,

$$e^{\eta} = \frac{p}{1-p}. \tag{2.13}$$

Then finally,

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)} = e^{\eta_1}/e^{\eta_2} = e^{\beta_j \cdot (x_{1j} - x_{2j})}. \tag{2.14}$$

The difference $x_{1j} - x_{2j}$ comes from the difference between the two linear predictors where we have made a practical assumption. Let us say for example that we have two linear predictors that are defined as

$$\eta_q = \sum_{j=1}^{m} \beta_j \cdot x_{qj}, \text{ with } q = 1, 2, \tag{2.15}$$

and we assume that $x_{1j} \neq x_{2j}$ for the j'th covariate we want to examine. The expression for the odds ratio in (2.14) thus makes it simple to calculate the effect of each unique covariate in the model relative to some change in units or category $\Delta x_j = x_{1j} - x_{2j}$. When comparing two unique categories the expression becomes even more simplified, as the norm is to use a reference category as $x_{2j} = 0$ while the other category is $x_{1j} = 1$, implying:

$$e^{\beta_j \cdot (x_{1j} - x_{2j})} = e^{\beta_j \cdot \Delta x_j} = e^{\beta_j}. \tag{2.16}$$

Specifically, this gives the following three interpretations of the odds ratio:

- $\mathrm{OR} = 1 \Rightarrow$ No effect on the outcome occuring.
- $\mathrm{OR} > 1 \Rightarrow$ Higher odds of the outcome occuring.
- $\mathrm{OR} < 1 \Rightarrow$ Lower odds of the outcome occuring.

2.1.4 Wald test

An important step in fitting a logistic regression model to a dataset is to identify significant covariates. One approach to evaluate the statistical significance of a model parameter is by performing a Wald test. The null hypothesis is defined to be:

$$H_0 : \beta_j = 0, \tag{2.17}$$

while the alternative hypothesis is defined as.

$$H_1 : \beta_j \neq 0. \tag{2.18}$$

Using the obtained estimates from an iterative weighted least squares procedure (Charnes et al., 1976) for the coefficients, $\hat{\beta}_j$, and the corresponding standard deviations, $\hat{s}_{\beta_j}$, it is then possible to calculate the Gaussian Wald Z-statistics,

$$Z_j = \frac{\hat{\beta}_j}{\hat{s}_{\beta_j}}, \tag{2.19}$$

or the equivalent two-sided confidence intervals,

$$[\hat{\beta}_j - z_{\alpha/2}\hat{s}_{\beta_j}, \hat{\beta}_j + z_{\alpha/2}\hat{s}_{\beta_j}]. \tag{2.20}$$

Here $z_{\alpha/2}$ is defined as the Gaussian quantile at a significance level determined by α. Regardless of which approach is used, the conclusion about the significance of each coefficient is the same.

The H_0 can be discarded if the calculated p-values are lower than the significance level α; selected by the researcher or typically set to $\alpha = 0.05$. Lower p-values are associated with smaller chances of discarding a correct null hypothesis. A p-value is calculated based on the absolute value of Z_j compared to the Gaussian quantiles. If we use the $1 - \alpha$ confidence interval instead, we can discard H_0 if zero isn't included in the interval.

2.1.5 Multicollinearity

An important aspect related to model selection is to check for multicollinearity. It affects a test's ability to accurately determine the p-value corresponding to a covariate. Multicollinearity can be explained as the dependence between covariates. It can occur either between individual covariates or from more complicated effects such as large groups of covariates in a model. Further, the degree of multicollinearity can be subtle as well as easily noticeable. Presence of multicollinearity can, for example among other methods, be detected by fitting two models that differ only by a single covariate and then check if any p-values have been sufficiently changed by the absence or presence of a covariate. If we want to be cautious of the possible multicollinearity in our models, we have to be especially careful to mindlessly accept the models provided by automatic procedures. Especially, the automatic generalized linear model procedure does not take this multicollinearity into account.

2.2 Boosting

We will now consider a different approach to perform or assist logistic regression, referred to as boosting. Boosting or more specifically for this thesis' use, generalized boosted regression models (GBM) is a machine-learning technique for instance used to perform logistic regression on binomial data.

Our motivation to use boosting in a regression setting arose from the uncertainty or question which goes as follows: How can we be more certain that the researcher or the automated model selection algorithm do find the optimal or best predictors and the number of covariates in a given model from the data? Sure, there exist established factors and guidelines which are followed within different fields of science, but they leave little room for innovation if followed to a fault. This makes it more difficult to discover, select and use other perhaps experimental predictors. In other words it is of great interest to explore and utilize GBM as a tool to perform and help with and hopefully improve the procedure of an analysis. Within medicine, we have examples of established lifestyle risk factors being smoking, inactivity and alcohol consumption (Schuit et al., 2002). The first way GBM could help improve an analysis is to check if the analysis done and the final conclusions reached, are reasonable. The analysis in question could for example be with regards to regression performed with generalized linear models, as will be done later. The second way is by using GBM itself to select predictors. Put shortly, GBM could be used to datamine predictors for other regression frameworks; like the generalized linear model framework; since boosting tends to have more credible inferences about models than strictly linear approaches (Schonlau et al., 2005). This should reduce the amount of model combinations required to fit with respect to multicollinearity and relevant covariates, and speed up the analysis as a consequence. How GBM can help improve regression analyses is presented more precisely in section 2.2.1 and 2.2.3.

The main idea or underlying foundation to boosting is the use of learners, weak or strong. A weak learner can be any type of simple model or function that says something about the problem at hand which we want to know more about. To be more specific a weak learner could be, as is relevant to the logistic regression, a regression fit related to a distribution. When adding together many weak learners the result should be a strong learner. The strong learner is then assumed to be better than each of the weak learners by themselves. The idea behind boosting may be comparable to the concept of indirect democracy where many different people with limited or less than optimal knowledge come together to agree upon a single more precise decision as a group. The representatives that have better arguments will naturally have more influence, as is with the weak learners within boosting. Thus in our case the representatives will be logistic model fits.

There are many different methods or algorithms to perform boosting and one of them is GBM. Variations of boosting algorithms share the similar concept by adding together many weak learners, even though certain parts between the algorithms do differ. This similarity should be apparent when looking at the general mathematical formulations. Section 2.2.1 together with section 2.2.2 provides two algorithms which are not exactly similar, that will shed some light on possible differences.

Assume now that we have a regression setting with observed covariates $\boldsymbol{x}_i = \{x_{i1}, \ldots, x_{im}\}$, $i = 1, \ldots, n$ and a response y_i for each individual. Ideally we want to find the relationship which maps the covariates to the response with the least amount of error. This relationship can be represented as $\boldsymbol{y} = F(\boldsymbol{x})$, but the problem is that we do not know $F(\cdot)$. Generally speaking, boosting will help us to find an approximation, $F^*(\cdot)$, through some iterative and additive scheme of weighted weak learners. Specifically, according to Friedman (2002), gradient boosting (complete algorithm in section 2.2.2) is trying to minimize the expectation of a loss function $\Psi(y, F(\boldsymbol{x}))$ with respect to $F(\boldsymbol{x})$ over the joint distribution of both y and $\boldsymbol{x}$. This can mathematically be expressed as:

$$F^*(\boldsymbol{x}) = \arg\ \min_{F(\boldsymbol{x})} E_{y,\boldsymbol{x}}(\Psi(y, F(\boldsymbol{x}))). \tag{2.21}$$

The loss function $\Psi(y, F(\boldsymbol{x}))$ could for example be the sum of squares:

$$\Psi(y, F(\boldsymbol{x})) = \sum_i (y_i - F(\boldsymbol{x}_i))^2. \tag{2.22}$$

In order to better understand how the weak learners, defined as $h(\boldsymbol{x}; \boldsymbol{a}_m)$, affect the minimization routine in (2.21), it is necessary to look at how they relate to the strong learner, $F(\boldsymbol{x})$. The effect the weak learners have on $F(\boldsymbol{x})$ is:

$$F(\boldsymbol{x}) = \sum_{m=0}^{M} \beta_m h(\boldsymbol{x}; \boldsymbol{a}_m). \tag{2.23}$$

According to Friedman (2002) the β_m are called expansion coefficients and the $\boldsymbol{a}_m$ are parameters. The $\boldsymbol{a}_m$ could, if thinking in the context of generalized linear models, be the coefficients of a linear predictor. Equation (2.23) can then be interpreted as the main idea, i.e. that a strong learner, $F(\boldsymbol{x})$, is a sum of many weak learners. We could also drop the summation notation in (2.23) and end up with an equivalent recursive expression,

$$F_m(\boldsymbol{x}) = F_{m-1} + \beta_m h(\boldsymbol{x}; \boldsymbol{a}_m), \tag{2.24}$$

which is to be used to update the approximation in an iterative algorithm.

In order to update the strong learner, values of $\boldsymbol{a}_m$ and β_m need to be determined. They can in general be determined by another minimization routine:

$$(\beta_m, \boldsymbol{a}_m) = \text{arg min}_{\beta, \boldsymbol{a}} \sum_{i-1}^{N} \Psi(y_i, F_{m-1}(\boldsymbol{x}_i) + \beta h(\boldsymbol{x}_i, \boldsymbol{a})). \tag{2.25}$$

To initialize the boosting algorithm, the equation (2.25) require an initial guess or value of $F(\cdot)$, $F_0(\cdot)$. This initial guess has to be either constructed or specified.

2.2.1 Generalized boosted regression models (GBM)

Similarly to generalized linear models, if one is boosting the exponential family regression models it relies on and requires almost all the same underlying assumptions (Ridgeway, 1999). For instance, the underlying distribution must belong to the exponential family and utilizing a link function to accommodate different distributions. Note that boosting does not necessarily contain a linear predictor η_i as in the generalized linear model case, but there is a function that is rather similar. Instead boosting have the summation of every weak learner that eventually becomes a strong learner, $F(\boldsymbol{x}_i)$. Thus rather than using (2.4) the following relationship is required instead:

$$g(\mu_i) = F(\boldsymbol{x}_i), \tag{2.26}$$

which is a relaxation of the linear assumption (Ridgeway, 1999). The relaxation implies that GBMs may be considered to be more general than generalized linear models and even generalized additive models, depending on the construction of the strong learner. Though the complexity of a generalized additive model will exceed our use and what is illustrated here. The following algorithm from Ridgeway (1999), slightly rewritten to match the notation of section 2.2, shows a boosting algorithm for the exponential family regression models by using Fisher scoring:

Algorithm: Boosting by fisher scoring

```
Initialize $F_0(\boldsymbol{x}) = g(\bar{y}) \;\; \forall \boldsymbol{x}$
For $m$ in $1, \dots, M$ do
    Compute the current "$m$'th" working response $z_i$,
    $z_i = (y_i - \mu_i)g'(\mu_i)$,
    where $\mu_i = g^{-1}(F_{m-1}(\boldsymbol{x}_i))$ using (2.26) since $F_{m-1}(\boldsymbol{x}_i)$ is known.
    Fit a regression model, $h(\boldsymbol{x})$, predicting $z_i$ using $\boldsymbol{x}_i$ with weights
    $w_i = \frac{1}{g'(\mu_i)^2 V(\mu_i)}$, where $V(\mu_i)$ is the variance function.
    Update the boosted regressor as
    $F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + h(\boldsymbol{x})$
End loop
```

In this specific algorithm, the parameters $\boldsymbol{a}_m$ is absorbed into the $h(\boldsymbol{x})$ expression, while the expansion coefficients β_m are set to equal one. GBMs have as generalized linear models, many distributions (for instance Gaussian, binomial, Laplace and Poisson) which opens up a wide span of different boosting regressions to be performed. Since there is a plethora of different options within GBM, it is referred to the manual regarding the exact specifics behind the GBM function in the R package `gbm` (Ridgeway et al., 2015). In the code's documentation it is stated that the implementation of GBM "(...) closely follows Friedman's Gradient Boosting Machine (Friedman, 2001)." Therefore it is absolutely necessary that we take a closer look at *Friedman's Gradient Boost algorithm* that GBM is built upon.

2.2.2 Gradient boosting algorithm

This section will present an example of an algorithm that can be used to perform gradient boosting, and there will also be introduced two key parameters or inputs that are present in GBM. These two parameters are rather important and they can affect the result heavily depending on how they are chosen. Knowing how these two parameters should be chosen is a necessity to use GBM properly and to have some control over the boosting. The gradient boosting algorithm chosen to presented here is *Friedman's Gradient Boost algorithm* (Ridgeway, 1999), and the algorithm is similar to "Algorithm 1" in Friedman (2001). This algorithm will be rewritten slightly such that the notation match with the previous notation used with boosting.

Algorithm: Friedman's Gradient Boost algorithm

```
1 Initialize $F_0(\boldsymbol{x}) = \arg\min_\gamma \sum_{i=1}^N \Psi(y_i, \gamma)$.
2 For $m$ in $1, \dots, M$ do
3     Compute the negative gradient as the working response
      $z_i = -\frac{\partial}{\partial F(\boldsymbol{x}_i)} \Psi(y_i, F(\boldsymbol{x}_i)) \mid_{F(\boldsymbol{x}_i)=F_{m-1}(\boldsymbol{x}_i)}$
4     Fit a regression model, $h(\boldsymbol{x})$, predicting $z_i$ from the covariates $\boldsymbol{x}_i$.
5     Choose a gradient descent step size as
      $\beta_m = \arg\min_\beta \sum_{i=1}^N \Psi(y_i, F_{m-1}(\boldsymbol{x}_i) + \beta \cdot h(\boldsymbol{x}_i))$
6     Update the estimate of $F(\boldsymbol{x})$ as
      $F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \beta_m \cdot h(\boldsymbol{x})$
7 End loop
```

As is apparent, this algorithm is more general and more complicated than the Fisher scoring algorithm. A picture of how the algorithm in `gbm` functions will now be completed by discussing two key parameters.

The first parameter to look at is often called the learning rate or "shrinkage", with range $0 < \nu \leq 1$. The learning rate comes into effect within boosting algorithms when the procedure is to update the strong learner with another weak learner. In the *Friedman's Gradient Boost algorithm* this will result in some changes to step 6 in the above algorithm, which is replaced by:

$$F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \nu \cdot \beta_m \cdot h(\boldsymbol{x}). \tag{2.27}$$

This implies that ν controls the rate of how much a new weak learner should affect the strong learner $F_m(\boldsymbol{x})$ when updated. How should the value of ν be chosen then? Choosing ν too small results in a really slow learning rate, while too large values would mean giving every weak learner a lot of importance or influence. From a strictly empirical standpoint it has been found (Friedman, 2002) that quite small values are the ones which yield the best accuracies. It is suggested that $\nu \leq 0.1$ to minimize error. By decreasing ν, one should then achieve greater accuracy, but reducing the learning rate too much will come at the cost of more time spent finalizing the boosting.

The second parameter to look at is called bag fraction and may be defined as

$$\text{Bag Fraction} = \frac{\tilde{N}}{N}, \tag{2.28}$$

where $\tilde{N} \leq N$. N is the total number of observations, while $\tilde{N}$ is the size of a random subsample of the total sample, N. Bag fraction can introduce randomness into the boosting if one choose to bag (or use) less than a 100% of the total observations. A bag fraction set equal to one will thus remove the randomness introduced by the bag fraction. In the *Friedman's Gradient Boost algorithm*, the randomness introduced will take place between step 2 and 3, creating a new step in the algorithm. What the new step does is executing a resampling without replacement from indices and then using these indices to select the corresponding parts of the covariates $\boldsymbol{x}$ and the responses $\boldsymbol{y}$.

A natural question to ask now is why one would even consider to use less than the total sample size without having to do so. A very direct and perhaps obvious consequence of using less observations is that computation will be performed faster. On the other hand and more interestingly, there are empiric evidence (Friedman, 2002) suggesting that using a bag fraction less than one may actually reduce error to a relative best measurement and increase accuracy. This reduction in error depends on the unique problem one is dealing with. Thus it is not guaranteed to be the same reduction in error per bag fraction value. Without knowing how this randomness can affect the problem at hand it might be reasonable to keep the bag fraction relatively large since there is also evidence of too small bag fraction values generating worse precision than the case containing no randomness at all.

2.2.3 Model selection: Automated GBM procedure

The GBM, like the generalized linear model, does also have an automated procedure, but it does not function exactly like the automated model selection in generalized linear model framework. The function that performs this procedure in R is named "gbm.step" and is from the `dismo` package (Hijmans et al., 2016). The `dismo` package builds further upon the `gbm` package which performs the actual GBM fitting.

In general the automated procedure tries to find the optimal number of boosting trees using k-fold cross validation (section 10.1.1) to minimize the loss of information; see Hijmans et al. (2016) for details. When the optimal number of boosting trees have been found it then fits the resulting GBM with its model information.

A particularly interesting part of information that is gained from this fit is the measurement of relative influence (Friedman, 2001) each predictor has in the final model calculated by the automated GBM procedure. Thus, even though you initially specify that you want to fit the whole model with all possible predictors you would end up with a measurement of importance of each predictor in that setting. Predictors that have little influence or importance won't affect the final model so it would be as if one would fit only the important parts. This information about influence is what makes us able to interpret which predictors that could have a great impact in the model as a whole. The measurements of relative influence for each predictor is what will be used to check the appropriateness of the final conclusion or to mine predictors.

Chapter 3

Multinomial regression using INLA and neural network

In questionnaires there are often questions that have more than two categories as answers. Using a logistic regression model would then result in a loss of perhaps vital information. To prevent this loss of information the logistic model can be extended, such that it becomes a multinomial regression model instead. This chapter contains theory about two frameworks that can be used to perform multinomial regression and inference about the multinomial models fitted. The main framework to be applied is integrated nested Laplace approximations (Rue et al., 2009, 2017). The second method used is neural networks (Murphy, 2012). The idea is to use the neural networks method to serve as a double check of the other experimental method that is implemented by myself from looking at a specific example.

3.1 Multinomial regression

Multinomial regression can be used to model any response that consists of three or more nominal factors, categories, types or species. The conceptual relationship between the multinomial regression and logistic regression is simple. The logistic regression is just a special case of the multinomial regression, where the response is binary. Alternatively, the multinomial regression can be said to be a general case of the logistic regression.

Let us say that the number of different factors or categories in a multinomial response-variable is defined as K. Then pick one of the categories to be a reference category to all the remaining $K-1$ categories, for example category 0. Then for each category $k = 0, 1, \dots, K-1$, including the reference category, we can define a linear predictor,

$$h_k(\boldsymbol{x}_i) = \alpha_k + \sum_{j=1}^{m} \beta_{kj} x_{ij} \tag{3.1}$$

relative to the reference category. Unique for the reference category, using $k = 0$, is that the linear predictor is $h_0(\boldsymbol{x}_i) = 0$. Here m is defined as the number of unique covariates. α_k and β_{kj} are the model parameters that are calculated relative to the pivot around the reference category. A change of the reference category will therefore most likely result in entirely new parameter estimates.

According to Ripley (1996) the Softmax function, defined as

$$P(Y_i = k \mid \boldsymbol{x}_i) = \frac{\exp(f_k(\boldsymbol{x}_i))}{\sum_{j=1}^{K} \exp(f_j(\boldsymbol{x}_i))}, \tag{3.2}$$

is appropriate if it is desirable to group each observation into one of the K different, already specified, groups. This is because the Softmax function (3.2) is a more general logistic function that makes sure that the probability calculated is restricted between zero and one. The components, $f_j(\boldsymbol{x}_i)$, can be non-linear or they can be linear. We can therefore set the components, $f_j(\boldsymbol{x}_i) = h_j(\boldsymbol{x}_i)$. Replacing the components will then yield the following function that can be used to calculate the probability of a category k, given the covariates:

$$P(Y_i = k \mid \boldsymbol{x}_i) = \frac{\exp(h_k(\boldsymbol{x}_i))}{1 + \sum_{j=1}^{K-1} \exp(h_j(\boldsymbol{x}_i))}. \tag{3.3}$$

3.1.1 Odds and odds ratio in multinomial regression

Since logistic regression is a special case of the multinomial regression, a measure closely related to odds and odds ratios can be used to infer about the models. The following expression is a general formulation of the odds in the multinomial setting:

$$\frac{P(Y_i = k \mid \boldsymbol{x}_i)}{P(Y_i = 0 \mid \boldsymbol{x}_i)} = \exp(h_k(\boldsymbol{x}_i)). \tag{3.4}$$

Similarly as before we can use the odds measure to create an odds ratio in the multinomial setting:

$$\mathrm{OR} = \frac{\exp(h_k(\boldsymbol{x}_i))}{\exp(h_k(\boldsymbol{x}_j))},\ \boldsymbol{x}_i \neq \boldsymbol{x}_j. \tag{3.5}$$

The measure may be referred to as an odds ratio, but we will use the name relative odds ratio instead to differentiate it from the odds ratio measure in the logistic setting. These K different responses will in turn yield a minimum of $K-1$ relative odds ratio estimates, calulated relative to the reference category. Remembering this part regarding the reference category is important, since it constitutes the interpretational difference from simple logistic regression. If it is desirable to see every relative odds ratio relative to each category, then it is simply required to change the reference category and fit more models. Though checking many models with different reference category will result in many relative odds ratios, which may not be that useful and messy if strict notation isn't applied.

The reference category can be chosen based on convenience, practicality or theoretical consideration. Careful selection of the reference category may yield easier interpretations of the relative odds ratios. Finally, each relative odds ratio estimate in a multinomial regression can be interpreted similarly to the other odds ratios, as in section 2.1.3. The similarity regarding interpretation is because we can change the values of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in (3.5) to investigate the effects of each estimated coefficient by themselves.

3.2 Integrated nested Laplace approximation (INLA)

INLA is a computational method used to perform Bayesian inference for a large class of regression models, referred to as latent Gaussian models. To gain a general understanding regarding what happens when using INLA, we will cover latent Gaussian models (section 3.2.1) and Gaussian Markov random fields (section 3.2.2). The main motivation behind using INLA is, besides being able to perform multinomial regression, the possibility to extend analyses beyond the regression models with only fixed effects and add random effects. Another reason to use a Bayesian framework is that the model fits come with the marginal distributions to each coefficient. This makes the inference about the model coefficients intuitive and straightforward by using credible intervals (section 3.2.5).

3.2.1 Latent Gaussian models (LGM)

The class of latent Gaussian models includes a vast variety of models, which suggest a great amount of flexibility in fitting models to a given dataset. For example, LGM includes: Generalized additive and mixed models, time series and spatial models. A generalized linear model (GLM) is actually a simple special case of the latent Gaussian models. The predictor $g(\mu_i) = \eta_i$ of the LGMs is defined as:

$$g(\mu_i) = \eta_i = \alpha + \sum_{j=1}^{m_\beta} \beta_j z_{ji} + \sum_{l=1}^{m_f} f_l(c_{li}) + \epsilon_i,\ i = 1, \ldots, n \tag{3.6}$$

and it is called a structured additive predictor (Rue et al., 2009). This predictor takes into account linear (β_j) and non-linear ($f_k(\cdot)$) effects of the covariates z_{ji} and c_{ki} while ϵ_i is the error representing an iid effect. Within the R-INLA program (Rue et al., 2009) ('http://www.r-inla.org'), the linear effects are referred to as fixed effects while the non-linear effects are referred to as random effects. The LGMs utilizes this structured additive predictor to create a latent field $\boldsymbol{x} = \{\alpha, \boldsymbol{\beta}, \{f_k(\cdot)\}, \boldsymbol{\eta}\}$, which is done by collecting all the terms in (3.6).

The latent field is part of a three-stage hierarchical model formulation that is used as a computational framework to analyse LGMs in a unified way. This hierarchical formulation bases itself first on the observations $\boldsymbol{y}$, second the latent field $\boldsymbol{x}$ and lastly the hyperparameters $\boldsymbol{\theta}$. The hyperparameters control the latent field and the likelihood for the data. They are also used as precision parameters for the Gaussian priors that will be assigned to the latent field.

There are three assumptions (Rue et al., 2017) related to this formulation which are highly advantageous computationally when satisfied:

1. The first one is that the observations, $\boldsymbol{y}$, are mutually conditionally independent given a set of latent field and the hyperparameters, $\boldsymbol{\theta_1}$,

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta_1} \sim \prod_{i=1}^{n} \pi(y_i|x_i, \boldsymbol{\theta_1}). \tag{3.7}$$

2. The second assumption is that the distribution of the latent field, $\boldsymbol{x}|\boldsymbol{\theta}$, is assumed to be Gaussian and that the field needs to be a Gaussian Markov Random Field (GMRF). When this is the case, then the dimension of the field can be large, for example 10^4 to 10^5.

3. The third assumption is that the dimension of the hyperparameters is small, preferably less than a two digit number, typically two to five.

3.2.2 Gaussian Markov random fields (GMRF)

What really distinguishes a Gaussian vector from a GMRF is a simple addition of conditional independence properties. The latent field $\boldsymbol{x}$ needs to be structured such that some x_i and x_j, for $i \neq j$, are conditionally independent, given every other $\boldsymbol{x}_{-ij}$.

The benefits of having such a field is that we gain a lot of computational time in factorization due to sparsity. For instance the precision matrix, $\boldsymbol{Q}$ (inverse of covariance matrix), for the latent field $\boldsymbol{x}$ would be simplified and only changes in the hyperparameters would require us to recalculate it. The simplification manifests itself in the precision matrix such that for each x_i and x_j satisfying the conditional independence property, the precision matrix then takes the value $Q_{ij} = 0$, where the Q_{ij} corresponds to an element the precision matrix, $\boldsymbol{Q}$.

3.2.3 Multinomial to Poisson transformation

Performing multinomial regression using INLA is not completely straightforward. The multinomial regression is in fact not supported directly by INLA. The workaround to make multinomial regression possible requires a transformation of the data matrix and a very specific formula. The transform is named *multinomial-Poisson*, and transforms the data, both predictors and response, which is on a multinomial form to data which is applicable with the Poisson regression. Mathematically the transformation can be represented as (Baker, 1994), from the multinomial likelihood

$$L_M(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{k=0}^{K-1} \left(\frac{\exp(h_k(\boldsymbol{x}_i, \boldsymbol{\beta}_k))}{\sum_{j=0}^{K-1} \exp(h_j(\boldsymbol{x}_i, \boldsymbol{\beta}_j))} \right)^{y_{ik}}, \tag{3.8}$$

to the Poisson likelihood

$$L_P(\boldsymbol{\phi}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{k=0}^{K-1} \left(\frac{(\exp(h_k(\boldsymbol{x}_i, \boldsymbol{\beta}_k)) \cdot \exp(\phi_i))^{y_{ik}}}{\exp(\exp(h_k(\boldsymbol{x}_i, \boldsymbol{\beta}_k)) \cdot \exp(\phi_i))} \right). \tag{3.9}$$

The notation is slightly changed to take into account each individual, i, and such that an emphasis is put on the parameters to be optimized, here $\boldsymbol{\beta} = \{\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_{K-1}\}$ and $\boldsymbol{\beta}_k = \{\alpha_k, \beta_{k1}, \ldots, \beta_{km}\}$. Here y_{ik} is either valued one or zero depending on which category that was observed for the corresponding individual, so $\sum_k y_{ik} = 1$. $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_n\}$ is introduced during the transformation, see Baker (1994) for further details. Thus it is possible to perform multinomial regression through the use of Poisson regression. The

ID	Response:	Value of predictor:
1	A	15
2	B	20
3	C	78

Table 3.1: Simple multinomial example. $K = 3$

ID	Response:	Value:	ResponseShift:	Observed:	ValueB:	ValueC:
1	A	15	A	1	0	0
2	B	20	B	1	20	0
3	C	78	C	1	0	78
1	A	15	B	0	15	0
2	B	20	C	0	0	20
3	C	78	A	0	0	0
1	A	15	C	0	0	15
2	B	20	A	0	0	0
3	C	78	B	0	78	0

Table 3.2: Multinomial to poisson transformed example data.

transform is intuitive when considering the two distributions: Multinomial and Poisson.

To illustrate this, consider a simple example (Table 3.1) with $K = 3$ categories for the response, with only one predictor of some kind. In order to perform multinomial regression in a Poisson distributed setting the data is extended or transformed, as shown in Table 3.2. Here, the number of rows were extended or duplicated $K - 1$ times. Then a new variable was created based on the multinomial response variable, to create a shift in categories. In the example we have shifted $A \rightarrow B$, $B \rightarrow C$ and $C \rightarrow A$, $K - 1$ times.

The next step is to create the variable that keep track of which observations that were actually observed in the original data. In the example the variable is named "Observed" and it is equal to one if the data is observed, while it is equal to zero otherwise. This is the new variable that is to be used as the response in the Poisson regression. It makes sense, because the Poisson distributions can only deal with count observations as a response.

The next thing to do is to create $K - 1$ dummy variables which should be equal to the true value only if the shifted response contains the same category as the dummy, otherwise it is set to be zero or as having no contributions. There are only $K - 1$ dummy variables since one of them is to be used as a reference (here we used the first category A) in order to be able to derive meaningful interpretations of the final model. These kinds of dummy variables would have to be created for each new predictor that one would want to use in the multinomial regression, so the greater the K and the more predictors to be used the larger the data-matrix will become post transform. For the multinomial regression to be performed in R-INLA there are a few more specifications that need to be done, but this sums up the multinomial to Poisson transformation in itself.

3.2.4 Deviance information criterion (DIC)

The deviance information criterion (Spiegelhalter et al., 2002) is the Bayesian equivalent to AIC. It is used to compare different models and a smaller value in a relative setting would indicate a better model, similar to AIC. The core differences are how DIC is computed and that DIC is used in a Bayesian setting. DIC is defined as

$$\text{DIC} = D(\bar{\theta}) + 2p_D \tag{3.10}$$

or equivalently as

$$\text{DIC} = \bar{D} + p_D. \tag{3.11}$$

Here $\bar{D}$ is defined as the posterior expectation of the deviance. The parameter p_D is defined as the effective number of parameters. $D(\bar{\theta})$ is defined as the deviance evaluated at the posterior mean, that is desired to be as small as possible. Finally the relationship between the two equivalent definitions, in equation (3.10) and (3.11), is $p_D = \bar{D} - D(\bar{\theta})$. The definition in equation (3.10) is rather similar to the definition of the AIC, and it is not surprising why DIC is the Bayesian equivalent to AIC.

3.2.5 Credible intervals

A credible interval is the Bayesian equivalent to a confidence interval in frequentist statistics. The credible intervals are calculated from the posterior distribution. Creating the credible interval can thus be done by taking the quantiles of the posterior distribution corresponding to a coefficient. There are different ways to create credible intervals, including highest posterior density (HPD) interval and equal-tailed intervals. When using credible intervals with R-INLA it is possible to request and use the marginal distributions of each coefficient. Then by using these marginals it is possible to create a function which will measure the significance at any level or the same levels GLM would provide; i.e. $\alpha = (5\%, 1\%, 0.1\%)$. The criterion for significance can be implemented as: If the product of the upper and lower $\alpha/2$ quantile is positive, then the coefficient is significant at a α level. If the marginal distribution at hand is non-skewed and unimodal, then this will basically be a HPD interval.

3.3 Neural network

Neural network is a method that originally was heavily inspired by biology (McCulloch and Pitts, 1943). The method can be seen as an attempt to mimic how a brain would process information, for instance how signals travel between neurons in the brain to process meaningful information. Today most neural networks are artificial and have no direct correspondence with biology. As with boosting there is a vast amount of different ways to use neural networks to solve problems. Neural networks can for instance be used to perform pattern recognition, regression and classification. It is the regression aspect that is of interest to me in this context.

3.3.1 Basic neural network structure elements

The most generic and simple neural networks are those that only feed information forward, or is feedforward. Intuitively, this would mean that the input information is only pushed towards the direction of the output. Figure 3.1 illustrates the flow in a linear feed forward neural network. Mathematically this is described by a linear combination,

$$x_j^{(l)} = \sum_i w_{ij} \cdot x_i^{(l-1)}, \tag{3.12}$$

where $x_j^{(l)}$ is a neuron or node in the network and the superscript (l) is to be interpreted as such that the node $x_j^{(l)}$ is in the l'th layer. The subscript j is referring to the j'th node in the layer while w_{ij} is the weight creating a connection between the two nodes. A non-linear and more general version of (3.12) can be specified as

$$x_j^{(l)} = f_j \left(\sum_i w_{ij} \cdot x_i^{(l-1)} \right), \tag{3.13}$$

where $f_j(\cdot)$ is an activation or transfer function (Murphy, 2012). Setting the transfer function equal to the identity function would yield the special case in (3.12). Equation (3.13) can be interpreted such that the node in the next

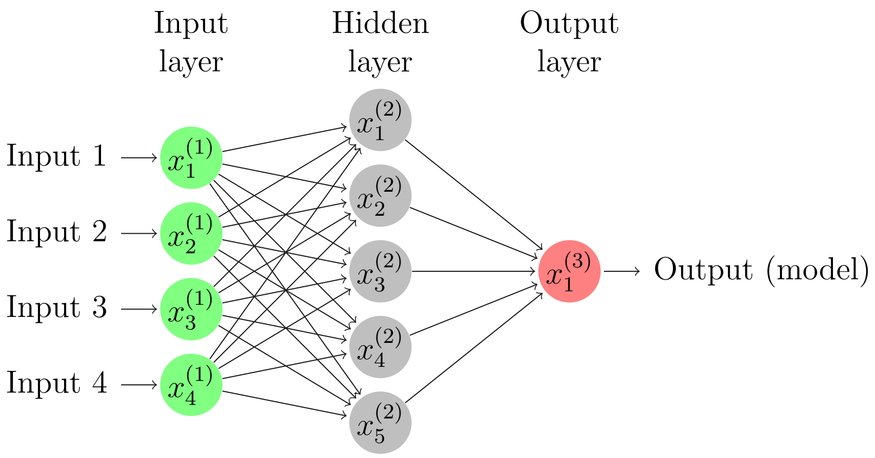

Figure 3.1: Illustration of a simple feedforward neural network, with a single hidden layer and no skip-layer connections.

layer is a function of a linear combination based on the nodes from the previous layer or previous layers if skip-layer connections are present.

Hidden layers are sets of intermediate nodes between the two layers containing the input and the output nodes. An increase in the number of hidden layers will increase the network's ability to perform more complex computations. Though having more than necessary layers could possibly slow down the computation, compared to if less layers could be used. Adding many hidden layers will result in the neural network to transition over to become a deep neural network.

The number of nodes at each layer do seem to be problem specific. By layer it is referred to the input layer, output layer and hidden layers. For example having only one node at each of the layers except the input layer and assuming one hidden layer, the neural network could easily solve a problem of the form $y = g(f(\boldsymbol{x}))$. Where $f(\cdot)$ is the function from layer one to two and $g(\cdot)$ is the function from layer two to the final output layer. This example illustrates just how trivial and simple the neural networks can be understood and thought of, even though the example by itself probably could do without a neural network.

Skip-layer connections are as the name itself implies connections that skip certain layers and go directly to another node further ahead. They may be used to preserve some information, for example measures of linearity, through a neural network were the standard feed forward path would lose that information through the transformations between the layers.

3.3.2 Multinomial regression with neural network

Performing multinomial regression in R can also be done through the use of neural networks and this can be achieved by using the package `nnet` (Venables and Ripley, 2002). In this package there is a function named "multinom", which is a lot easier to use compared to in R-INLA with the data transformation that had to be done. The level of complexity to use this package for multinomial regression is comparable to the level of difficulty that comes with performing logistic regression within GLM in R. To name a few examples as to why it is easy to use: The model is specified through a standard simple formula and the resulting model returns the AIC and the deviance, which ease the task of performing inference. Thus, this an ideal package to help double check the estimates received by using R-INLA. In the manual to the package there are descriptions suggesting that this implementation utilizes feedforward neural networks, a single hidden layer and skip-layer connections.

The exact amount of nodes at each layer and the number of skip-layers is not directly or readily available in the documentation. Though it could be possible that if we use the Softmax function (3.2) in the non-linear flow function (3.13) with calculated linear predictors (3.1) as inputs, neural networks should without a doubt be able to perform multinomial regression.

Chapter 4

Two discrete-time Markov models

Using different Markov models we will try to model, describe, present and predict patient trajectories. In general, Markov models appear to be theoretically suitable to complete these objectives. The two well known Markov models to be described in this chapter have been thoroughly described in the literature. A large part of the theory in this chapter will therefore be based on Ross (2010) and Murphy (2012). This chapter will start out by introducing discrete-time Markov chains in section 4.1, while section 4.2 will present the closely related, but more advanced hidden Markov model.

4.1 Discrete-time Markov chains

Let a finite sequence of random variable be defined as $\{X_n\}_{n=1}^N$. We assume that $\{X_n\}_{n=1}^N$ is a stochastic process defined on a finite, discrete sample space. Specifically, $X_n = j$ can be interpreted as that the process or patient is in state j at time n.

The Markov property is a core part of what defines a discrete-time Markov chain. A mathematical formulation of the Markov property is

$$P\{X_{n+1} = k \mid X_n = j, \ldots, X_1 = j_1\} = P\{X_{n+1} = k \mid X_n = j\} \tag{4.1}$$

and its interpretation is straightforward. The Markov property tells us that the probability to enter the next state in a process only depends on the current state the process is inhibiting.

If the Markov property (4.1) is not fulfilled by the situation at hand and there is a n-step dependence (where $n > 1$), it is possible to work around the lack of one-step dependence. This can be done by defining a single state as a sequence of the original single states. The approach will of course increase the number of states since any permutation and ordering of the individual states will be included in the new transition matrix which will comply better with the Markov property. Consider for example a three by three transition matrix. If we assume that there is a two-step dependence, then instead of the three states $\{1, 2, 3\}$, there will be a state space as $\{11, 12, 13, 21, 22, 23, 31, 32, 33\}$, with nine states.

As we are interested in modelling possible transitions within and between all the states in a process at any time, we need to define a probability of transitioning, denoted as A_{jk}. The transition probability,

$$A_{jk} = P\{X_{n+1} = k \mid X_n = j\}, \tag{4.2}$$

can be read as the probability of moving from state j to state k. A figure illustrating the transition probabilities between three states can be found in Figure 4.1. There are two conditions that are required to make the transition probabilities valid. First we have that $A_{jk} \geq 0 \ \forall j, k$. Also, $\sum_{k=1}^{Q} A_{jk} = 1$, where $Q =$ Number of unique states.

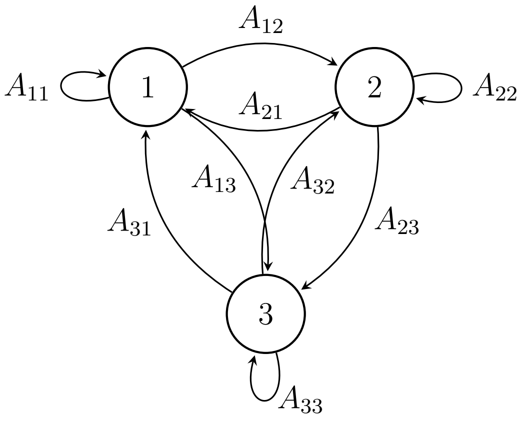

Figure 4.1: Illustration of a discrete-time Markov process with the transition probabilities between all three states.

4.1.1 Transition matrix and stationarity

Let $\boldsymbol{A}_{Q \times Q}$ define a transition matrix. Q is still referring to the number of unique states. To follow Figure 4.1 and to illustrate a transition matrix take a look at the three by three transition matrix, which has three unique states with corresponding transition probabilities:

$$\boldsymbol{A}_{3\times 3} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}. \tag{4.3}$$

The transition matrix (4.3) is to be interpreted such that, at each row j ($j = 1, 2$ or 3) there will on the column k ($k = 1, 2$ or 3) be the probability of going from state j to state k. Thus for this transition matrix to be valid, $A_{j1} + A_{j2} + A_{j3} = 1$ has to be the case. This follows directly from the defined transition probabilities in (4.2).

A useful property with Markov chain processes is that they can be stationary. For a Markov chain to be stationary and have existing limiting probabilities, the process is required to be irreducible and ergodic (Ross, 2010). If the process is stationary a stationary convergence will happen fast or slow depending on the transition matrix. A general three by three transition matrix that converges, can be calculated and formulated as:

$$\boldsymbol{A}_{lim} = \lim_{n\to\infty} \boldsymbol{A}^n_{3\times 3} = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 \\ \pi_1 & \pi_2 & \pi_3 \\ \pi_1 & \pi_2 & \pi_3 \end{pmatrix}. \tag{4.4}$$

Here π_j is defined as a limiting probability of going to state j, taking no consideration about the current state. The π_j can explain the proportion of time that will be spent in state j in the long run, depending on how fast it converges. Similarly as with the transition probabilities, the limiting probabilities have to follow this relationship:

$$\sum_{j=1}^{Q=3} \pi_j = \pi_1 + \pi_2 + \pi_3 = 1.$$

4.1.2 Estimation for a Markov chain

In real data examples, the transition matrix usually needs to be estimated. First of all it is desirable to transform the data to portray the transitions per discrete time unit. When the data is on the desired form it is then reasonable to use the maximum likelihood (ML) estimator

$$\hat{A}_{jk} = \frac{N_{jk}}{\sum_k N_{jk}} \tag{4.5}$$

for each possible transition, $j \to k$, to estimate the transition matrix (Murphy, 2012). N_{jk} is defined as the total number of observed transitions from state j to state k.

The initial distribution can be estimated by the maximum likelihood estimator

$$\hat{\pi}_j^1 = \frac{N_j^1}{\sum_j N_j^1} \tag{4.6}$$

where N_j^1 represents the number of chains in state j at the first possible time in the data (Murphy, 2012).

If the stationary distribution exists, it can be estimated from the transition matrix as described in section 4.1.1. Also, if the Markov chain is rapidly converging, compared to the length of the time dimension in the data, it can also be approximated from another maximum likelihood estimator similar to the initial distribution estimator (4.6). The estimator

$$\hat{\pi}_j = \frac{N_j}{\sum_j N_j} \tag{4.7}$$

provides the second way of calculating the stationary distribution. N_j is defined as the number of times state j is entered from any of the other states.

4.2 Hidden Markov Model

A hidden Markov model can be thought of as an extension of the discrete time Markov chain in section 4.1. Suppose we still have the sequence of random variables, $\{X_n\}_{n=1}^N$, exactly as defined in section 4.1, then the new addition is that at each time, n, each state will have a probability to emit a signal denoted as S_n. The different unique signals or observations, S_n, the states can emit have to be from a finite set of signals. Therefore a hidden Markov model basically consist of an underlying Markov chain that emits observable signals. Figure 4.2 illustrates how such a process could be visualized for a single individual or patient. Implicitly, the Markov property as described and the transition matrix (section 4.1.1) with its properties are present as part of the underlying Markov chain or hidden process.

As the states can emit signals, we need a formulation that dictates and describes which signal that is to be emitted by a state. Such a formulation is

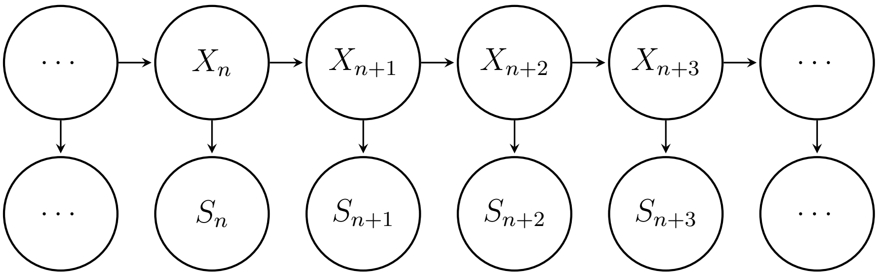

Figure 4.2: Illustration of a hidden Markov model for one individual.

$$B_{sj} = P(S_n = s \mid X_n = j) \tag{4.8}$$

where B_{sj} is the emission probability. It should be interpreted as the probability of a state j, to emit a signal s, at any time n. In addition the emitted signal at time n, only depends on the current hidden state, X_n, and no other previous values from either the observed signals or the hidden Markov chain states. The property can be formulated, similarly to the Markov property, as:

$$P(S_n = s \mid X_n = j, X_{n-1}, S_{n-1}, \dots, S_1) = P(S_n = s \mid X_n = j). \tag{4.9}$$

These emission probabilities can be represented as a matrix. For the sake of an example, consider three unique hidden states and two possible emissions per state, in which the emission probability matrix is,

$$\boldsymbol{B}_{2\times 3} = \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \end{pmatrix}. \tag{4.10}$$

For the emission probability matrix to be valid it is required to fulfil these two criteria. First that $B_{sj} \geq 0$. Also that $\sum_{s=1}^{Q_s} B_{sj} = 1$, where $Q_s =$ Number of signals or symbols.

4.2.1 Estimating or training the model

Depending on the relevant application, it may be necessary to estimate or train the model. We distinguish between two different types of scenarios, where the main difference is whether the underlying Markov chain is known or hidden. To be specific, the underlying process would be known if we have data that could be used to estimate the transition matrix. One would think that the name of the model implied that the Markov chain had to be hidden, but that is not always the case. In both of these cases the signals or patient trajectories are directly observable to the researcher. We will therefore distinguish between a full-information scenario and a partial-information scenario.

The full-information scenario uses maximum likelihood to estimate the model completely. The complete model is represented by the initial distribution, the estimated transition matrix and the emission matrix. The estimator used to estimate the transition matrix should already be known from equation (4.5), which was also used with the discrete-time Markov chain model. The second maximum likelihood estimator required for the hidden Markov models is the one that estimates the emission probability matrix. It is defined by Murphy (2012) as:

$$\hat{B}_{sj} = \frac{N_{sj}}{N_j}. \tag{4.11}$$

Intuitively, the estimator counts how many times a state, j, and a symbol, s, has been observed at the same time, N_{sj}, and then divides by the total number of times the process is in a state, N_j. N_{sj} and N_j are not to be confused with the definition of the estimator (4.5). These estimators assume that the underlying distribution is multinomial. There are also other distributions that could be used instead, for instance the Gaussian distribution.

The partial-information scenario is slightly more complicated. The scenario becomes more complex since it does not use estimators directly and it is required to train a model based on observations. One known and established training algorithm is the Baum-Welch algorithm (Baum et al., 1970), which in essence is a specialization of the Expectation Maximization algorithm (Dempster et al., 1977). Here is a pseudo-algorithm describing the Baum-Welch algorithm that train hidden Markov models based on available observed sequences :

Pseudo-algorithm: Baum-Welch algorithm

1 Provide initial estimates to $\boldsymbol{\pi}^1$, $\boldsymbol{A}_{Q\times Q}$ and $\boldsymbol{B}_{Q_s\times Q}$.

2 Until convergence or termination by iteration or convergence, do and repeat 3 then 4:

3 Use the most recent estimates of $\boldsymbol{\pi}^1$, $\boldsymbol{A}_{Q\times Q}$ and $\boldsymbol{B}_{Q_s\times Q}$ with observed sequences to calculate the expected number of times each relevant combination of events are happening.

4 Re-estimate $\boldsymbol{\pi}^1$, $\boldsymbol{A}_{Q\times Q}$ and $\boldsymbol{B}_{Q_s\times Q}$ using the calculated expectation values and set these as the most recent log-likelihood maximizing estimates.

The Expectation Maximization algorithm within the Baum-Welch algorithm can be used to optimize the model since the hidden Markov model contains latent, hidden or missing data. Consequently, problems that are related to optimization will follow. A central issue is finding the global maxima due to some initialization. If we have information or an idea about how the best model could look like, then we can initialize with that information. Though, that information about where to start may not be available to us. Then the best option would be to utilize the random start technique enough times and choose the desired model based on a criterion. Due to the two matrices that need to be initialized it makes the process to find a good initialization difficult. As if this level of difficulty wasn't enough we also have to specify the number of hidden Markov chain states within the model. Here, again, information about the current problem at hand is a must and an expert opinion is to be desired when choosing the number of states. Especially since an increase in the number of hidden states affect the initialization as well as the computational cost required to train the model. If an expert opinion is not available, then there exist workarounds to choose the number of hidden states (Murphy, 2012).

4.2.2 Assessing the probability of hidden states

After a model has been determined it can be necessary to find a probable sequence of hidden states. This is especially important in the partial-information scenario, since we have no exact data that can specify the hidden sequences. A posterior distribution defined as

$$\text{Posterior probability} = P(X_n = j \mid \boldsymbol{S}^N), \tag{4.12}$$

will help us determine the probability of a hidden state, j, given the whole sequence of observed signals. $\boldsymbol{S}^N = (S_1, \ldots, S_N)$ is defined as the random sequence of the observed signals of length N. To obtain the probability for all discrete times, go through all the time values, $n = 1, \ldots, N$. Likewise, if we want the posterior distribution for each unique hidden state we just go through all the values corresponding to a state, $j = 1, \ldots, Q$. This posterior distribution do offer itself as a tool to create an estimate similar to the Viterbi algorithm (see section 10.1.5), i.e. an estimate of the overall most probable sequence of hidden states given the observations. To receive this similar estimate, we can just take the maximum of each posterior probability at a state, j, at each time, n. How to find this most probable hidden state sequence estimate is formulated as:

$$\text{For each "}n\text{" do: arg max}_j P(X_n = j \mid \boldsymbol{S}^N). \tag{4.13}$$

Although do note that the core difference between equation (4.13) and the Viterbi algorithm is that the estimate based on the posterior only uses one state, while the Viterbi algorithm uses the whole sequence of hidden states.

The forward and backward algorithms are two probability measures that can simplify calculations within the hidden Markov model setting. By themselves they have meaningful interpretations, but they can also help calculate the posterior probability (4.12). If we define the forward algorithm as

$$F(j, n) = P(X_n = j, \boldsymbol{S}^n), \tag{4.14}$$

and the backward algorithms as

$$B(j, n) = P(\boldsymbol{S}^{n+1:N} \mid X_n = j), \;\; \boldsymbol{S}^{n+1:N} = (S_{n+1}, \ldots, S_N), \tag{4.15}$$

then we can calculate the posterior probability by:

$$P(X_n = j \mid \boldsymbol{S}^N) = \frac{F(j, n)B(j, n)}{\sum_k F(k, n)B(k, n)}. \tag{4.16}$$

4.2.3 Prediction of the next signal

The hidden Markov models can also be used to predict future signals. At first glance it may be reasonable to suggest that we could use the property in equation (4.9), but that would only give the correct answer if we know for sure what value the next hidden state in the process will take. In the partial-information scenario it is unlikely to know the last hidden state with a 100% certainty. This section will show how to perform one-step prediction using the hidden Markov model and its properties in the partial-information scenario.

Assume that we have realizations of the observed sequence, $\boldsymbol{S}^N$, where N is the last event. We want to the calculate the probability of
$P(S_{N+1} = s \mid \boldsymbol{S}^N)$ and predict S_{N+1}. By using the definition of the conditional probability the hidden Markov model properties in equation (4.1) and (4.9) imply that

$$P(X_{N+1} = i \mid X_N = j) = \frac{P(X_{N+1} = i, X_N = j)}{P(X_N = j)}, \tag{4.17}$$

and

$$P(S_N = s \mid X_N = j) = \frac{P(S_N = s, X_N = j)}{P(X_N = j)}. \tag{4.18}$$

The left-hand side in equation (4.17) and (4.18) is known to us from matrices in the trained or estimated model. It is therefore necessary to rewrite both (4.17) and (4.18) by multiplying with $P(X_N = j)$, and then condition on the complete observed sequence, $\boldsymbol{S}^N$, which is known. Do note that the hidden Markov model properties still hold and will simplify the conditioning. Equation (4.17) will then become:

$$\begin{gathered} P(X_{N+1} = i, X_N = j \mid \boldsymbol{S}^N) = \\ P(X_N = j \mid \boldsymbol{S}^N) \cdot P(X_{N+1} = i \mid X_N = j). \end{gathered} \tag{4.19}$$

Equation (4.18) will have a similar reconstruction as (4.17), but we also have to change or increment the time notation such that it can find the probability of S_{N+1}. The rewritten equation (4.18) will thus become:

$$\begin{gathered} P(S_{N+1} = s, X_{N+1} = i \mid \boldsymbol{S}^N) = \\ P(S_{N+1} = s \mid X_{N+1} = i) \cdot P(X_{N+1} = i \mid \boldsymbol{S}^N). \end{gathered} \tag{4.20}$$

The right-hand side of equation (4.19) now contains only known measurements, namely the posterior probability, $P(X_N = j \mid \boldsymbol{S}^N)$, and the transition probability, $P(X_{N+1} = i \mid X_N = j)$. In equation (4.20) we only know the emission probability, $P(S_{N+1} \mid X_{N+1})$, given a state.The unknown part in equation (4.20) is simply a marginalization of equation (4.19) over X_N. In fact both (4.19) and (4.20) need only to be marginalized to finish the proof to find the probability of the next signal given the observed sequence. Marginalizing equation (4.19) over X_N will then result in

$$P(X_{N+1} = i \mid \boldsymbol{S}^N) = \sum_k P(X_{N+1} = i, X_N = k \mid \boldsymbol{S}^N), \tag{4.21}$$

which can be put directly into equation (4.20). After this we marginalize (4.20) over X_{N+1} to get

$$P(S_{N+1} = s \mid \boldsymbol{S}^N) = \sum_k P(S_{N+1} = s, X_{N+1} = k \mid \boldsymbol{S}^N), \tag{4.22}$$

which gives us the probability of the next signal being s, given all the known signals. To finally predict a single signal a selection criterion is necessary, and it should be constructed or chosen depending on the context to get the most optimal results. A simple example of such a selection criterion could be the maximum of the calculated probabilities. Though before the selection criterion can be applied, every possible s for $P(S_{N+1} = s \mid \boldsymbol{S}^N)$ has to be calculated.

Part II

Analyses, results, discussion and future work

Chapter 5

Preprocessing the datasets to be analysed

This chapter describes the sets of data with its variables and how they have been transformed or modified to be used in the analyses. The transformation and modification can be thought of as preliminary changes or analyses that will serve as a common ground for the analyses in chapter 6, 7 and 8. Thus, the purpose of this chapter is to provide means to replicate the analyses and deliver enough information to interpret the results. Since we have two distinct types of data, aggregated and questionnaire versus fully detailed, this chapter will be sectioned into two parts.

5.1 Preprocessing the aggregated and questionnaire data used in regression model analyses

Before the aggregated data could be used in any type of analytical setting, several preprocessing steps were required. For instance, missing values in the form of *NA* (not available) had to be taken care of. Missing values of variables with a logical zero was set to zero, while the other missing values would have the respective patient to be removed from the data to be analysed. The selection of participants to answer the questionnaire is our source that the data should be representative; it will still be assumed even though patients or data entries are removed due to missing values.

Removing the whole row (or patient) in the dataset given one missing value in any of the categorical variables is the same kind of procedure the GLM-function does in R. It should be sufficiently valid to remove them, especially considering that we have no means of constructing with 100% accuracy the missing values available; though intelligent guesses are a possibility if we wanted to settle for it. The action of removing missing values will also make sure that the AIC or DIC estimates will be correct in a relative setting, with a constant number of observations. In order to prevent removal of too much data, as it is desirable to have data close to the original data and the sample size as large as possible, relevance of variables was checked with regards to their properties (section 5.1.3). Properties here refer to their correlation between covariates, skewness or possible outliers. The variables, or possible predictors checked were chosen based on whether they were assumed or guessed to have influence. If there is any doubt about a covariate, it is to be included if it can potentially explain anything new or uncorrelated; either from a theoretical or empirical perspective.

5.1.1 Properties of the aggregated data

As time is involved in the data collected, one should always consider the fact that the data gathered directly from the hospital has taken place before the data from the questionnaire was gathered. The point is that there is a causality which needs to be taken into consideration when using any of the variables from the questionnaire to predict a response contained in the hospital data. Using data from the questionnaire as a response should not breach this causality.

In general the data collected range over a span of two years, but there is no actual time axis in this aggregated data which can be used to determine further relationships in time; as it was lost in the aggregation process of the detailed sets of data. Therefore the only time dependency in this aggregated data is the distinction between the aggregated hospital data and the questionnaire data. Consequently, a potential weakness of the aggregated data is that it is not possible to look at a specific time a patient received a diagnosis or when an event occurred. As such some relationships between the covariates may be hidden and overlooked. In order to distinguish the two types of covariates we will index (later in Table 5.1, 5.3 and 5.4) the health care predictor-covariates with an H and the questionnaire predictor-covariates with a Q. Covariates or variables that do not fit into either of these two categories will be indexed G; for general.

5.1.2 Constructing the two main response-variables

In order to perform both logistic and multinomial regression we require categorical response-variables with at least two unique categories, $K \geq 2$. One unique response for each of the analyses is also necessary. In order to not breach the causality we have chosen two questions or covariates from the questionnaire as a basis for our response-variables.

The first response-variable to be analysed is a variable containing eight subquestions that measures the quality ratings from a patient regarding the satisfaction of self-perceived healthcare services received: The variable is named *S7* and its original formulation can be found under question 7 in the questionnaire, section 10.2.2. The variable represents 8 questions, each having integer-values within the range from one to five. We can define the simplified *S7* as:

$$\boldsymbol{y}_i = (y_{i1}, \dots y_{i8}),\ y_{ij} \in \{1, 2, 3, 4, 5\},\ j = 1, \dots, 8,\ i = 1, \dots, N.$$

A higher value of y_{ij} means being more satisfied with the j'th health service aspect. Before the simplification removing the irrelevant answer and performing correction of negative or contradictive phrasing this was not necessarily the case. For it to be used in the logistic analysis, a mean is first applied

$$\overline{y}_i = \frac{1}{8} \sum_{j=1}^{8} y_{ij},$$

resulting in the response-variable named *MeanSpm7*. Then the resulting response-variable, *MeanSpm7*, is dichotomized around its empirical mean, $\overline{y} = \frac{1}{N} \sum_{i=1}^{N} \overline{y}_i$:

$$y_i^{(d)} = \mathbb{1}_{\{\overline{y}_i \geq \overline{y}\}} \tag{5.1}$$

Creating a dichotomized or binary response-variable, $y_i^{(d)}$, that represents those who are above averagely satisfied, with $y_i^{(d)} = 1$, versus those who are less than averagely satisfied with the care they have received, with $y_i^{(d)} = 0$.

The second response-variable to be analysed is a variable with ten subquestions that has to be transformed somehow in order for it to be used properly with multinomial regression. The response-variable in question is named *S10*; and its original phrasings can be found under question ten in the questionnaire, section 10.2.2. The variable represents 10 questions, each having integer values ranging from zero to one. It can be illustrated as:

$$\boldsymbol{y}_i = (y_{i1}, \dots, y_{i10}),\ y_{ij} \in \{0, 1\},\ j = 1, \dots, 10,\ i = 1, \dots, N.$$

Each subquestion, represented by y_{ij}, describes whether or not a patient have experienced that the health service helped them with an unique challenge. If help is received $y_{ij} = 1$ and if it is not received then $y_{ij} = 0$. The choice of transformation was decided to be a sum score of the subquestions. Thus depending on different groupings (which is specified later in chapter 7), it will be a response-variable containing more than two categories up to a maximum of eleven categories, when including zero. For instance if three of the subquestions are used ($K = 4$) to create a sum-score response-variable, $y_i^{(s)}$, we will have:

$$y_i^{(s)} = \sum_{k=1}^{K-1=3} y_{ik},\, y_i^{(s)} \in \{0, 1, 2, 3\}\ \forall i.$$

5.1.3 Processing and preliminary selection of predictor variables

The aggregated dataset with the questionnaire covariates, contains a total of 101 variables which are possible to include as predictors in a regression analysis. Naturally, all of these predictors are not relevant and a variable-selection procedure was needed. The predictors in Table 5.1 were chosen as an outset for our preliminary analysis here and for the regression analyses (in chapter 6 and 7) from the whole aggregated data. When further selecting the predictor-variables, measures and tests play a central part.

First the predictor-variables were checked with a histogram to see if any variables were too skewed. In other words we want to check if covariates have too few observations for certain values, making the corresponding covariate less viable; since the effects of certain values will be poorly represented compared to the other values. Using a histogram, the predictor-variables *Occupation* and *HasSupport* were omitted from further examination in the analyses. Since they were both bi-valued and one of their categories had too few observations, resulting in almost homogeneous covariates.

Secondly, the range of the different variables were checked, i.e. the minimum and maximum values. This was an important part in figuring out whether to prefer to use *DiagnosesICD* or *DiagnosesICPC* since they essentially describe the same property with a patient, namely the number of diagnoses that patient has. *DiagnosesICPC* has a slightly wider range, which makes it more preferred in a possible final model. The same logic can be applied to the number of diagnose chapter and diagnose category variables relating to the diagnoses, thus making these less preferable. The number of

Variable name:	Explanation:	Type:
Age	Integer age	G
Gender	Gender of a patient	G
Services	Number of contacts with hospital services	H
TimeInHospital	Length of hospital stays	H
TypeServices	Number of different types of hospital services	H
Readmissions	Number of readmissions to hospital	H
Departments	Number of departments visited	H
Wards	Number of wards visited	H
Procedures	Number of procedures undergone	H
DiagnosesICD	Number of ICD diagnoses(1)	H
CategoriesICD	Number of ICD categories(1)	H
ChaptersICD	Number of ICD chapters(1)	H
DiagnosesICPC	Number of ICPC diagnoses(1)	H
ChaptersICPC	Number of ICPC chapters(1)	H
SelfRateHealth	Self perceived or rated health	Q
NoLongTermSick	Absence of long term or lasting illness	Q
Education	Level of education	Q
Occupation	Occupation - Employed, student, etc	Q
LightExercise	Degree of light exercise	Q
ToughExercise	Degree of tough exercise	Q
Smoking	Degree of smoking	Q
Alcohol	Degree of alcohol consumption	Q
NoDebased	Never experienced debasing	Q
HasSupport	Relationships with others that can assist	Q
Income	Household income	Q

(1) ICD and ICPC (or ICD-10 and ICPC-2) are different systems that provide diagnose codes to conditions on different levels. There is no strict one-to-one correspondence between the two systems.

Table 5.1: Overview of the predictors initially chosen from the data to be examined further.

wards is then also to prefer above the number of departments a patient has been through, as these two are also dependent on each other.

Cramer's V (see section 10.1.4) was used to find a measure of the correlation between the categorical predictor-variables in the questionnaire. No notably correlation was found, but naturally this does not imply that there is no correlation or dependencies at all between the variables.

In order to find more information regarding the adequacy of hospital or health care covariates, Welch two sample t-tests (see section 10.1.2) were utilized. This could be done since the logistic response ($y_i^{(d)} \in \{0, 1\}$) makes for a reasonable grouping of the values registered per patient into two samples. This test assumes that the data or covariate values are approximately normally distributed. Since the sample size is rather large (over 1000 observations in each) the assumption should be ok, but is not infallible. The other assumption demands that the variance within the two groups are different, which they were confirmed to be empirically. Theoretically one can also argue that the assumption about different variances holds since we do not know the proportion of patients having different degree of severity with regards to sickness that are in the two groups. Table 5.2 shows the p-values for the grouped health care covariates.

Variables "$0 - 1$" grouped:	p-value:
Services	0.19
TimeInHospital	0.051
TypeServices	0.019
Readmissions	0.036
Departments	0.0013
Wards	0.000412
Procedures	0.032
DiagnosesICD	0.0008984
CategoriesICD	0.0005366
ChaptersICD	0.0001015
DiagnosesICPC	$2 \cdot 10^{-15}$
ChaptersICPC	$6 \cdot 10^{-12}$

Table 5.2: P-values from the Welch two sample t-tests.

A p-value less than 0.05 should indicate that there is a significant difference between the two groups within the respective variables. The significance level here may indicate that the variables will be more significant in a logistic regression setting; which also happened to be the case in chapter 6. With this in mind it could be even more reasonable to believe (based on Table 5.2) that *DiagnosesICPC* should have some higher degree of significance than the rest. Based on these results combined with the measures of range, the diagnose chapter and categories are omitted from further analysis.

5.1.4 Simplification of interesting predictor-variables

Since the models to be used in the analyses are only logistic or multinomial, it is desirable to transform some of the variables in order to simplify the interpretation of the models fit. That is because the interpretation of these models relies on the measures of odds ratio. Having poly-categorical predictors present is not impossible, but could yield a large amount of unnecessary coefficients. Also, similar multi-categorical answers would make the interpretation of the different odds ratios more challenging if they aren't sufficiently distinct from a theoretical perspective. The simplification, or transformation used was dichotomization and it was applied on every multi-categorical variable that was found to be of any interest. Another reason as to why dichotomization was applied is that subgroups of categorical variables became too small and skewed. Thus splitting the poly-categorical groupings into two groups would absorb the smaller groups, such that new variables could explain somewhat more distinctively with fewer outliers. The categorical variables were already grouped such that not a lot of information would be lost by doing the dichotomization. Thus these dichotomized predictors will be used instead of their poly-categorical counterparts. The overview of the new variables resulting from the transformation are in Table 5.3, while the details are as follows:

- The variable concerning the level of education, *Education*, was split into two parts and the new dichotomized variable is named *HigherEducation*. *HigherEducation* is constructed such that if a patient had any higher education from university it would equal to one. On the other side, a zero would indicate that a patient had no education from a university or of equivalent degree.

- *LightExercise*, describing the amount of light exercise a patient performs was also dichotomized. The new variable is named *ActiveLight*. If its value is a one it is defined to represent above three hours of activity per week, while if it is zero less than three hours per week is measured. The natural interpretation is that less than three hours could be interpreted as inactive, while the other as active.

- *ToughExercise*, describing tough exercise has been dichotomized in the same manner as *LightExercise*, since they were similarly structured. The key difference lies in the interpretation which should be slightly different as *ToughExercise* explains more intense workout or exercise than *LightExercise*. The new variable is named *ActiveTough*.
- *Smoking* is also dichotomized and the new variable *SmokedOnce* is created. *SmokedOnce* could be interpreted such that if the patient has ever smoked, this will be represented by a one. While a zero will represent that the patient never has smoked.
- *Alcohol* was dichotomized into the covariate *NoAlcoholBinge*. The dichotomization was done as such that a one would represent no excessive intake of alcohol, while a zero would represent the opposite.
- *Income*, describing the household income, was split such that a one would represent above the mean income (around 500.000 NOK according to Statistisk Sentralbyrå (2017), and assuming that the patient lives alone) while a zero would represent an income that is below the average. The new variable is named *HighIncome*.

Variable name:	Explanation, if equal to one:	Type:
HigherEducation	Patient has higher education	Q
ActiveLight	Patient is physically active (light)	Q
ActiveTough	Patient is physically active (tough)	Q
SmokedOnce	Patient has smoked at least once	Q
NoAlcoholBinge	Patient has no excessive alcohol usage	Q
HighIncome	Patient has income above mean	Q

Table 5.3: Overview of the additional and simplified dichotomized predictor-variables created from the interesting variables.

A final overview of possible predictors of interest in the regression analyses are presented in Table 5.4. Their simplifications are in Table 5.3. The exact values and interpretation of all variables are found in section 10.2.

Variable name:	Explanation:	Type:
Age	Integer age	G
Gender	Gender of patient	G
Services	Number of contacts with hospital service	H
TimeInHospital	Length of hospital stays	H
TypeServices	Number of different types of hospital services	H
Readmissions	Number of readmissions to hospital	H
Wards	Number of wards visited	H
Procedures	Number of procedures undergone	H
DiagnosesICD	Number of ICD diagnoses	H
DiagnosesICPC	Number of ICPC diagnoses	H
SelfRateHealth	Self perceived health	Q
NoLongTermSick	Absence of long term or lasting illness	Q
Education	Level of education	Q
LightExercise	Degree of light exercise	Q
ToughExercise	Degree of tough exercise	Q
Smoking	Degree of smoking	Q
Alcohol	Degree of alcohol consumption	Q
NoDebased	Never experienced debasing	Q
Income	Income of household	Q

Table 5.4: Overview of the predictors determined to be fit to use in the regression analyses.

5.2 Preprocessing the fully detailed dataset used for trajectory analysis

As the fully detailed sets of data were not inherently intended or fit to use directly with any kinds of Markov models, preprocessing was necessary in order to utilise the data with a Markov models. These fully detailed datasets contain timestamps for each event which makes it possible to use different types of Markov models, compared to the aggregated data which did not have a time axis. The time span is the same as with the aggregated data and it will therefore be possible to utilise events from the beginning of 2012 to the end of 2013. Patient events starting or ending outside of this time interval were included, but were censored or cut off at both sides to conform with the remaining observations.

As already established we are interested in modelling and analysing patient trajectories with discrete-time models, therefore a time resolution has to be specified. The time resolution was set to be at a monthly level to better satisfy the assumptions of the models to be used and to reduce a possible issue with increased discrete dimensionality per amount of data and increased computational cost. A monthly time resolution will result in a maximum length of 24 months per patient trajectory.

5.2.1 The case of simultaneous events per month

One problem with a monthly time resolution is to code events happening within the same month. This problem has to be solved. As the discrete-time Markov models and hidden Markov models require one unique state at each time, it was thought to be sufficient to use the urgency or severity degree ordering to decide which events were allowed to overwrite other events in the trajectory vector. The more expensive or urgent types of care are also of greater importance to bring forth in a model. See an illustration of the severity, cost or urgency ordering principle in equation (5.2) and Table 5.5.

$$\mathrm{NO} < \mathrm{GP} < \mathrm{OP} < \mathrm{IP} \tag{5.2}$$

Thus we have that at each time unit a *NO* event is not allowed to write over any state at all. A *GP* event is only allowed to overwrite the *NO* event. The *OP* event can overwrite both the *NO* and *GP* event, while the *IP* event can overwrite all the other events. With this kind of ordering the patient trajectories created can be considered and interpreted to be a minimalistic measure of the most severe or costly state per month.

Symbol:	Description:	Urgency/cost priority:
IP	Inpatient	First
OP	Outpatient	Second
GP	General practitioner	Third
NO	None of the others	Fourth

Table 5.5: Overview of the four constructed states with description and urgency.

5.2.2 Training and test set

The patients that were used to create the patient trajectories, as described in section 5.2.3 for those with special interest, are a large group of adult patients with chronic conditions. A group of 12714 patients that had been invited to partake in the questionnaire was used as a training set, while another group of 67210 patients was used as a testing set. Using these patients opens up the possibility to compare any result with the questionnaire since about 3000 of the about 12714 patients in the training set that were invited had answered the questionnaire.

5.2.3 Details in constructing patient trajectories

The four unique events *NO*, *GP*, *OP* and *IP* are not directly present in the data as they are simplifications of other more detailed events that the patients have undergone. In other words, they are defined by values from other variables in the detailed sets of data. Specifically, these four events can be described as follows.

1. The *IP* or *inpatient* event is based on a variable named *StdCareCat* and a state is *IP* if this variable is equal to one for a patient at a certain time. The *StdCareCat* variable is found in the St. Olav dataset. This event or state is equivalent to receiving inpatient care at a hospital.

2. The *OP* or *outpatient* event is also based on the *StdCareCat* variable, but in this case it has to equal two instead of one. Also, the *OP* event is based on values from a variable named *PRAKSIS_REFUSJONSGRUNNLAG*, found in the Kuhr dataset. See Table 5.7 for an overview of the exact values or events included. This state is equivalent to receiving some outpatient care or treatment at a hospital.

3. The *GP* or *general practitioner* event is only based on the values from the *PRAKSIS_REFUSJONSGRUNNLAG* variable from Kuhr data, and the details regarding the values can be found in Table 5.6. *GP* thus represents care at a general practitioner level or equivalently.

4. The *NO* event is special since it is set to be the default value when creating the trajectories. This means that if none of the other events happen, the patients' state is *NO* for this particular month. In other words, this event is the complementary empty set of every other health care events.

A design flaw with these patient trajectories is that if the health care services aren't sufficiently mapped to one of the four events, then the *NO* event will contain these services by default in the patient trajectories. For example services related to receiving help at home or being admitted to a nursing home will fall into the *NO* event, as they haven't been taken into consideration. This will be a source of error and creating inaccuracies.

Value:
"Fastlege"
"Fastlønnet"
"Turnuslege fastlønnet"
"Legevakt"
"Legevakt kommunal"

Table 5.6: Overview of values in the variable *PRAKSIS_REFUSJONSGRUNNLAG* from Kuhr data that has been used as a basis for *GP*.

Value:
”Spesialist”
”Spesialist anestesiologi
”Spesialist barnesykdommer”
”Spesialist fysikalsk medisin og rehabilitering”
”Spesialist gynekologi”
”Spesialist hudlege”
”Spesialist indremedisin”
”Spesialist kirurgi”
”Spesialist nevrologi”
”Spesialist revmatologi”
”Spesialist øre-nese-hals”
”Spesialist øyelege”

Table 5.7: Overview of values in the variable *PRAKSIS_REFUSJONSGRUNNLAG* from Kuhr data that has been used as a basis for *OP*.

Chapter 6

Logistic analysis: Factors affecting satisfaction with health care

In this chapter the analysis, results, interpretation and conclusion from the logistic regression analysis will be presented. The main aim of this analysis is to figure out what affects how satisfied the patients are with the health-care they have received. This is to be achieved by logistic regression. It is thus desirable to find predictors that have significant influence, which could explain which elements in the hospital or personal sphere that have the greatest impact on the satisfaction levels. The second aim of this analysis is to try to find empiric evidence or indications of whether boosting can provide assistance in deciding the relevant predictors or not.

A description of how the results were found will be provided in section 6.1. Section 6.2 present the results gained from the procedures described, and finally interpretations and comments based on the results.

6.1 Procedure of the logistic analysis

During this chapter the response, in any model mentioned, is the dichotomized response $y_i^{(d)}$ in (5.1) which describes whether a patient was above averagely satisfied with the health care received. Based on the basic investigation performed in section 5.1.3 on the properties of the covariates, many different small models with one to four or five predictors from Table 5.4, were fit. This was done in order to figure out if there were any obvious confounding effects or multicollinearity in the logistic models. A large amount of such logistic models were fit in GLM. Therefore, the sheer amount of models fit could bring about random errors and difficulty figuring out which predictors are consistently significant. To correct for such possible errors it was decided to use an automated predictor selection procedure (see section 6.1.1). From the heap of models fit it was found that some predictors were more interesting and significant than others. For instance many of the hospital covariates were insignificant at a 5% level, except for *DiagnosesICPC*. Many of the questionnaire covariates were significant at a 5% level or lower in different models, though not in all of the models. One exception was *Alcohol*, which was the least significant of the covariates over all the models fit. This manual check of about 150 models thus only resulted in reasons to omit *Alcohol* and *NoAlcoholBinge* in further analysis.

6.1.1 Using the automated GLM procedure

Since the manual check only provided enough information to exclude one predictor, it is therefore reasonable to use an automated procedure in trying to find the predictors that are the most consistently significant. The automated procedure mentioned in section 2.1.1 was then used, where direction input was set to be *both*. This option was seen to yield the most accurate and finely tuned results, as it allows the algorithm to both remove and add predictor-covariates to the model.

Table 6.1 shows an overview of the two different scopes constructed. A scope contains the covariates to be tried included by the automatic model selection procedure. *SelfRateHealth* was included in the second scope in order to see if it changes the automatic procedure significantly. This was done since *SelfRateHealth* was found to be consistently significant, at a 0.1% level, when checking the manually fit models. *SelfRateHealth* is assumed to possibly have a great deal of influence in the logistic model.

In order to receive some reliable findings whether any of the predictors were consistently significant, the data was split into different groups. Eight different data groupings were constructed, to be used to fit nine different models using the automated GLM procedure based on either one or both of the scopes in Table 6.1. The data groups and the corresponding models became as follows:

- Two models utilized the whole sample, except there were no entries with missing data. Then only changing between the two different scopes, this resulted in two different models fit.

Scope without *SelfRateHealth*:	Scope with *SelfRateHealth*:
Gender	Gender
Age	Age
Services	Services
TimeInHospital	TimeInHospital
TypeServices	TypeServices
Readmissions	Readmissions
Wards	Wards
Procedures	Procedures
DiagnosesICPC	DiagnosesICPC
DiagnosesICD	DiagnosesICD
NoLongTermSick	NoLongTermSick
HigherEducation	HigherEducation
ActiveLight	ActiveLight
ActiveTough	ActiveTough
SmokedOnce	SmokedOnce
NoDebased	NoDebased
HighIncome	HighIncome
	SelfRateHealth

Table 6.1: Overview of the two scopes constructed to use with the automatic GLM selection procedure.

- The last seven groups were based strictly upon how the covariate *SelfRateHealth* could be split. Thus only using the scope without *SelfRateHealth* in the GLM procedure, since *SelfRateHealth* was implicitly used when partitioning the data.
  - Five of the seven data-groups were created based on the respective five subcategories of *SelfRateHealth*. Thus resulting in five new models fit.
  - The two remaining models fit were based upon the data where the empirical mean of *SelfRateHealth* was used to split the whole sample into two samples.

Using the last seven groups and models, the idea was that if a unique covariate ended up being consistently significant within many of these automated model fits then they could and should be regarded as actually significant for our final logistic model in question. The previously stated can be thought of as a comparison routine. The two first models within the first groups were only created to determine once and for all whether *SelfRateHealth* truly is consistently significant.

It could be argued that the subgroups should have been randomly split or divided. Though since *SelfRateHealth* showed undeniably consistent significance through the models fit before the automated procedures, it was thought that the groupings of values in this covariate could hide information. This is the main reason behind the segmentations of the data done.

6.1.2 Using the automated GBM procedure

After a handful of consistently significant predictors have been chosen by this comparison routine (described in section 6.1.1), GBM is used to validate or debunk whether the covariates are as significant as they have been found to be. We can do this since the logistic regression is built into its automated selection procedure (described in section 2.2.3). If the measures of relative influence yielded from GBM are similar to what was gained through the means in section 6.1.1, then it could be argued that the predictors are indeed significant.

6.2 Results and models

First of all the two different scopes yielded different logistic models when run on the undivided sample. The main difference is that *SelfRateHealth* was included by the automated procedure when it had the opportunity to do so. Thus *SelfRateHealth* can be said to have influence and notable significance.

6.2.1 Results of comparison routine

In the remaining seven models to be used in the comparison routine (described in section 6.1.1), many predictors were found to be necessary in at least one of the seven, but only a few ended up being significant in four or more models selected by the automatic GLM procedure.

The predictors that were significant in more than half of the models are listed in Table 6.2. The predictors in Table 6.2 are the ones which could be said to be consistently significant. There are some evidence of two other predictors which may be regarded as consistently significant. These two predictors are *Gender* and *DiagnosesICPC*. Potentially final models that have significant predictors, relatively low AIC and as few predictors as possible

Name of predictor x_j:	Proportion of models significant in:
NoLongTermSick	4/7
NoDebased	5/7
Age	4/7

Table 6.2: Summary of the comparison routine.

may then be the logistic models presented in Table 6.3, 6.4 and 6.5.

Name of predictor x_j:	Estimate of coefficients β_j:	P-value:
Intercept	-0.458	0.039
Age	0.015	$3.75 \cdot 10^{-10}$
SelfRateHealth.2	-0.296	0.107
SelfRateHealth.3	-0.824	$6.05 \cdot 10^{-6}$
SelfRateHealth.4	-1.194	$4.09 \cdot 10^{-9}$
SelfRateHealth.5	-1.856	$3.75 \cdot 10^{-13}$
NoLongTermSick	0.489	$9.12 \cdot 10^{-7}$
NoDebased	0.479	$3.26 \cdot 10^{-7}$

Table 6.3: Logistic model #1 with its coefficient - Only the most significant predictors included.

Name of predictor x_j:	Estimate of coefficients β_j:	P-value:
Intercept	-0.313	0.177
Age	0.015	$2.34 \cdot 10^{-10}$
SelfRateHealth.2	-0.276	0.133
SelfRateHealth.3	-0.785	$1.82 \cdot 10^{-5}$
SelfRateHealth.4	-1.133	$3.24 \cdot 10^{-8}$
SelfRateHealth.5	-1.783	$4.57 \cdot 10^{-12}$
NoLongTermSick	0.471	$2.38 \cdot 10^{-6}$
NoDebased	0.456	$1.39 \cdot 10^{-6}$
DiagnosesICPC	-0.025	0.022

Table 6.4: Logistic model #2 with its coefficients, including *DiagnosesICPC* as predictor.

Name of predictor x_j:	Estimate of coefficients β_j:	P-value:
Intercept	-0.493	0.027
Age	0.014	$3.81 \cdot 10^{-9}$
SelfRateHealth.2	-0.286	0.119
SelfRateHealth.3	-0.823	$6.24 \cdot 10^{-6}$
SelfRateHealth.4	-1.188	$5.04 \cdot 10^{-9}$
SelfRateHealth.5	-1.843	$5.70 \cdot 10^{-13}$
NoLongTermSick	0.475	$1.94 \cdot 10^{-6}$
NoDebased	0.460	$1.04 \cdot 10^{-6}$
Gender	0.222	0.010

Table 6.5: Logistic model #3 with its coefficients, including *Gender* as predictor.

Odds ratio may be the most intuitive and reasonable method to interpret logistic models. Due to low confounding when adding *Gender* and the number of diagnoses (*DiagnosesICPC*) into the model, one single table containing the odds ratios (Table 6.6) from the different predictors should be sufficient to explain how the covariates affect the response in these three models; i.e. affect the odds of being satisfied or not. See section 6.2.3 for an interpretation.

Name of predictor x_j:	Corresponding OR:
Age	1.015
SelfRateHealth.2	0.743
SelfRateHealth.3	0.438
SelfRateHealth.4	0.303
SelfRateHealth.5	0.156
NoLongTermSick	1.631
NoDebased	1.615
DiagnosesICPC	0.974
Gender	1.248

Table 6.6: Odds ratio per unit/category increase, i.e. from x_j to $x_j + 1$.

6.2.2 Results of GBM procedure

Now that possible final core models (Table 6.3, 6.4 and 6.5) are decided upon it is desirable to double check with results from generalized boosted regression models whether we have actually found the most significant predictors. The information, in Table 6.7, about variables and relative influence was gained from running a summary of the object returned from the automated GBM procedure (section 2.2.3). When running the procedure the learning rate was set to equal 10^{-4} and the bag fraction was set equal to one in order to minimize the random elements. A sufficient number of maximum trees was also set such that the procedure did not reach this maximum.

Predictor x_j:	Relative influence:
NoLongTermSick	41.107
SelfRateHealth	36.652
Age	15.374
NoDebased	6.810
Gender	0.053
Services	0.000
TypeServices	0.000
Readmissions	0.000
TimeInHospital	0.000
Wards	0.000
DiagnosesICD	0.000
Procedures	0.000
DiagnosesICPC	0.000

Table 6.7: Summary from the automatic boosting procedure of the logistic regression.

6.2.3 Interpretation of the odds ratios

Based upon the results and the calculated odds ratios (in Table 6.6) it would now be reasonable to infer the following about the predictors:

- The higher the *Age* a patient has will increase the odds of being satisfied with the health care. The increase in odds is not by much per unit increase in age, but when a larger age gap is present the odds will be increased even further.

- Regarding self perceived health, *SelfRateHealth*, the notation of its odds ratio is different since it is grouped into five different categories and all the ratios presented in the table are relative to if the patient perceives its own health as excellent. Thus all corresponding ratios from *SelfRateHealth.2* to *SelfRateHealth.5* will indicate a decrease in self-perceived health or feeling worse than excellent. Keeping this ordering in mind, we can now clearly state that the worse a patient perceives its own health the lesser the odds of being satisfied with the health care received.

- Now *NoLongTermSick*, here the odds of being satisfied with the health care will increase if the patient haven't experienced long lasting illness.

- The same interpretation or conclusion made about *NoLongTermSick* can be applied to *NoDebased* as well since they have such a similar odds ratio value and are both only two categories. Thus not having experienced debasing will increase the odds of being satisfied with the health care received.

- When looking at the odds ratio for *DiagnosesICPC* it is clear that a unit increase in the number of diagnoses will result in a decrease in the odds of being satisfied. As with age, a greater increase in number of diagnoses per patient will cause a more noticeable effect on the odds of being satisfied.

- Based on how *Gender* was defined then being a man will increase the odds of being satisfied with the health care.

6.2.4 Comparing the GLM results with the GBM results

As one can see from the measurements of relative influence (Table 6.7) only *NoLongTermSick*, *SelfRateHealth*, *Age* and *NoDebased* do clearly have some influence on a logistic model with the dichotimized *MeanSpm7* ($y_i^{(d)}$) as response. *Gender* has some slight influence, but not to the same extent as the other influential predictors. This predictor should therefore only be considered under doubt. The results from the automatic boosting procedure did end up with about the same conclusion as from when running multiple automatic procedures within the generalized linear models framework and using the comparison routine. This suggests that generalized boosted regression models could be used as an assisting tool to select or mine predictors.

With boosting backing up the decision and logic from the GLM analysis it would be rather reasonable to say that *NoLongTermSick*, *SelfRateHealth*, *Age* and *NoDebased* are consistently significant. These predictors, with their corresponding interpretation, will be the elements which explains any change in satisfaction levels.

Chapter 7

Multinomial analyses: Factors affecting help received from health care

In this chapter the response-variables, analyses, results, interpretation and conclusion from the multinomial regression analyses on the sum-score responses will be presented. The main aim of this chapter is similar to the main aim in chapter 6. In this chapter though, the goal is specifically to determine which predictors or effects that affect the odds of receiving medical related help from the healthcare services. It is also of interest to note throughout this analysis if the two types of responses based on question 7 and 10 in the questionnaire, introduced in section 5.1.2, could explain the same elements or not. Lastly, again, it is of interest to find evidence or indications that boosting could assist in the selection of predictors.

The response-variables to be used, is completely specified in section 7.1. The procedure of the analyses and results with interpretations are presented in section 7.2 and 7.3, respectively. Finally, section 7.4 contains comments regarding the results.

7.1 Specifying the response-variables in detail

The multinomial response $y_i^{(s)}$, that is to be used throughout this chapter, was defined in section 5.1.2 to be a sum score to summarize up to 10 sub-questions. In each sub-question the patients could answer with either yes or no whether they had received help in an aspect from the health care.

Due to these sub-questions the sum-score response can range from zero to a maximum of ten. The maximum depends on what kind of groupings of the sub-questions we decide to look at. There are two groupings of *S10*'s sub-questions that are of interest to use with multinomial regression. The two unique groupings are presented in Table 7.1. These groups end up leaving out sub-question seven, nine and ten from both responses. The groups, personalized help (TPH) and general information help (GIH), are constructed purely on how the questions about help in *S10* were categorised (Table 7.2) and interpreted by myself. The responses as sum-scores of these groups would thus essentially represent how many different challenges a patient did receive help with. Mathematical formulations of the sum-score responses based on the groupings are:

$$\text{TPH response: } y_i^{(s)} = \sum_{k=1}^{K-1=3} y_{ik},\ y_i^{(s)} \in \{0,1,2,3\}\ \forall i,$$

and

$$\text{GIH response: } y_i^{(s)} = \sum_{k=1}^{K-1=4} y_{ik},\ y_i^{(s)} \in \{0,1,2,3,4\}\ \forall i,$$

while still $y_{ik} \in \{0,1\}$ for every k.

Since the response variables are constructed as such they can neither be classified as strictly nominal or ordinal. It would be more correct to classify them as something between nominal and ordinal. As such a multinomial regression model should be sufficiently appropriate, though an ordinal regression model could be equally adequate.

Personalized help (TPH)	Information help (GIH)
S10_2	S10_1
S10_5	S10_3
S10_8	S10_4
	S10_6

Table 7.1: Overview of the multinomial sum score responses.

7.2 Procedure of the multinomial analyses

In order to avoid checking many different models as done in the logistic analysis, we will attempt to utilize the generalized boosting regression models to

Name:	Question: Did you receive help with the following challenges (. . .)
S10_2	listen to you what matters the most regarding your health problems.
S10_5	your priorities are taken into account when creating plans.
S10_8	develope a diet and/or exercise plan.
S10_1	help you understand your health condition.
S10_3	be able to explain complications from your medication usage.
S10_4	inform you of your future healthcare plans.
S10_6	advice on how you can pursue and persevere a healthy life.

Table 7.2: Description of the subquestions used in Table 7.1. Translated to english by the author of this thesis.

pick predictors for the analysis. This should be possible by using the information about relative influence to make a decision about which predictors or covariates that definitely need to be examined further. The results and conclusion from the logistic regression analysis (section 6.2.4) do also suggest that this may be an alternative way to find significant predictors. Metaphorically speaking it is our intention to use GBM as a guidebook to ensure that the most interesting attractions are visited. There is one problem regarding this approach though, and that is that the automatic GBM procedure does not support the multinomial distribution. To solve this problem, we now propose a way to avoid the issue while also using the GBM framework as it is.

7.2.1 Solving a limitation with GBM using several logistic submodels

As mentioned in chapter 3, the multinomial regression is a generalization of the logistic regression using $K > 2$ categories. This opens up a possibility, since the multinomial regression has only one reference category in regards to measuring the coefficients the same way the logistic regression does. If considering these elements and the fact that GBM does support the logistic regression then it is reasonable enough to simply choose a global reference category, for example category zero, and then divide the multinomial problem into $K - 1$ logistic subproblems that can be run in the GBM framework.

The logistic subproblems will then be the cases when one is always comparing and treating category zero as a reference category against the other remaining $K - 1$ values that will be treated as a one. The logistic subproblems are then supposed to act as an approximation to the main multinomial problem.

In the GBM algorithm I will again set bag fraction equal to one in order to minimize the random effects and ensuring near replicability without a seed. The learning rate will be set to 10^{-4}, to make sure that accuracy won't be a problem and it will not take too long. The maximum number of trees are set to twenty thousand and if any sub problem exceed this limit it could be necessary to increase it, unless it is deemed sufficient for our purpose. The responses used are the sum scores of the groupings in Table 7.1, while the initial predictors the boosting can choose from are the ones in Table 5.4.

7.2.2 Procedure using GBM to find predictors

Based on the $K-1$ logistic subproblems to be run in GBM it is reasonable to use the same logic as in the comparison routine (section 6.1.1). In this case it would translate to looking at how many times the GBM finds any predictor to be of relatively high influence across the different $K-1$ subproblems. If it is desirable to be cautious of being too strict in the selection, one could look at how many times the predictors are included and that the relative influence differ from zero. The idea and method should transfer well to this case since we are dealing with subgroups of a group as before.

After gaining knowledge of how influential each predictor is relative to each response, different multinomial models are to be fitted and examined with diagnostic tools. The diagnostics will include the use of credible intervals to check significance level, since we have access to them from the Bayesian framework, and DIC to measure the loss of information. To make sure my approach to the multinomial distribution through INLA using a Bayesian framework is correct, we will control the results with another method of choice that can fit multinomial regression to validate the model. That other method utilizes neural networks (section 3.3) to perform multinomial regression. Thus any noticeable differences will be remarked.

7.3 Results, models and interpretations

As we have two unique sum-score responses to perform multinomial regression with, this section is structured accordingly. In section 7.3.1 we use the personalized help grouped response (TPH) as a response-variable in the models, while section 7.3.3 provides models using the general information help grouped response (GIH) as a response-variable.

Regarding the notation in most of the tables in this section, the suffix at the end of each variable name (*variablename.k*) is designed to index which category that is referenced against the reference category, zero. Again, since we are working with sum-scores of the sub-questions, category k of the j'th variable will relate to having received k of $K - 1$ possible types of help. In other words, the greater the k the more help has the patient received of that type.

Do note that, as we are interested in the predictors' effects, the $K - 1$ intercepts have been omitted from the models presented in the tables. This is done since the intercepts will cancel themselves when we are checking odds ratios.

7.3.1 Personalized help group response

The GBM procedures yielded that the predictors in Table 7.3, had some relative influence within the three logistic subproblems. The number of diagnoses (*DiagnosesICPC*) and *Age* clearly have great relative influence. *NoLongTermSick* is probably the strongest contender of the ones that only appeared once, this is because this predictor had the greatest relative influence of the ones appearing once. Predictors that have influence in two of the models have about the same relative influence when comparing the information between the subproblems.

Name of predictor x_j:	Proportion of sub-models significant:
Age	3/3
DiagnosesICPC	3/3
HigherEducation	2/3
Procedures	2/3
TimeInHospital	1/3
NoLongTermSick	1/3
SelfRateHealth	1/3

Table 7.3: Summary of the three (TPH) logistic GBM subproblems.

Regarding the models fit in a multinomial setting, most of the predictor combinations from the Table 7.3 was checked. As a security measure in order to be certain that the boosting actually gave the desired result, the remaining predictors with zero relative influence was fitted to check if they had some interest at all. Most predictors with zero influence had little to no interest at all, and if any it was only because they correlated or introduced multicollinearity with some of the predictors with high relative influence. Out of all the predictors in the Table 7.3 *TimeInHospital* and *SelfRateHealth* could have been included by some randomness or approximation fault as they were not even close to being satisfyingly significant. In the case when approximating the multinomial model using logistic models, the relative influence that is measured to be below a certain threshold may not yield consistently significant predictors.

By satisfyingly significant, we define that three out of three and at the very least two out of three coefficients related to one predictor is significant at a 5% level. The remaining predictors in Table 7.3 are those that would be able to build the optimal model in this setting. Based on the models checked there are some models that are suitable to be considered as final models which only contains predictors that are satisfyingly significant. Table 7.4, 7.5 and 7.6 contain these final models.

Name of predictor x_j:	Estimate of β_{kj}:	CI$_L$:	CI$_U$:
Age.3	-0.016^{***}	-0.026	-0.006
Age.2	-0.022^{***}	-0.031	-0.013
Age.1	-0.013^{*}	-0.023	-0.002
DiagnosesICPC.3	0.136^{***}	0.086	0.187
DiagnosesICPC.2	0.103^{***}	0.055	0.151
DiagnosesICPC.1	0.073^{**}	0.019	0.127
HigherEducation.3	0.583^{***}	0.243	0.926
HigherEducation.2	0.626^{***}	0.311	0.942
HigherEducation.1	0.510^{**}	0.150	0.871
NoLongTermSick.3	-0.959^{***}	-1.318	-0.607
NoLongTermSick.2	-0.333^{*}	-0.671	-0.002
NoLongTermSick.1	-0.534^{**}	-0.912	-0.159

The credible intervals (CI$_L$ & CI$_U$) cover 95% of the distribution.
* Significant at a 0.05 level.
** Significant at a 0.01 level.
*** Significant at a 0.001 level.

Table 7.4: Model 1 (TPH) - With all predictors significant - DIC = 12803

Name of predictor x_j:	Estimate of β_{kj}:	CI$_L$:	CI$_U$:
Age.3	-0.018^{***}	-0.027	-0.007
Age.2	-0.022^{***}	-0.032	-0.014
Age.1	-0.013^{**}	-0.023	-0.003
DiagnosesICPC.3	0.119^{***}	0.068	0.171
DiagnosesICPC.2	0.088^{***}	0.040	0.138
DiagnosesICPC.1	0.067^{*}	0.012	0.122
HigherEducation.3	0.575^{***}	0.233	0.918
HigherEducation.2	0.620^{***}	0.305	0.937
HigherEducation.1	0.509^{**}	0.149	0.870
NoLongTermSick.3	-0.912^{***}	-1.272	-0.557
NoLongTermSick.2	-0.294	-0.634	0.038
NoLongTermSick.1	-0.518^{**}	-0.899	-0.143
Procedures.3	0.075^{**}	0.025	0.128
Procedures.2	0.064^{**}	0.015	0.115
Procedures.1	0.031	-0.023	0.087

The credible intervals (CI$_L$ & CI$_U$) cover 95% of the distribution.

* Significant at a 0.05 level.

** Significant at a 0.01 level.

*** Significant at a 0.001 level.

Table 7.5: Model 2 (TPH) - With the lowest DIC value - DIC = 12792

Name of predictor x_j:	Estimate of β_{kj}:	CI$_L$:	CI$_U$:
Age.3	-0.016^{***}	-0.026	-0.007
Age.2	-0.022^{***}	-0.031	-0.0132
Age.1	-0.013^{*}	-0.0234	-0.003
DiagnosesICPC.3	0.158^{***}	0.109	0.209
DiagnosesICPC.2	0.110^{***}	0.063	0.158
DiagnosesICPC.1	0.085^{**}	0.032	0.139
HigherEducation.3	0.457^{**}	0.123	0.794
HigherEducation.2	0.585^{***}	0.274	0.898
HigherEducation.1	0.442^{*}	0.086	0.798

The credible intervals (CI$_L$ & CI$_U$) cover 95% of the distribution.

* Significant at a 0.05 level.

** Significant at a 0.01 level.

*** Significant at a 0.001 level.

Table 7.6: Model 3 (TPH) - With all coefficients significant and little confounding - DIC = 12842

Relative odds ratios based on the model in Table 7.5 can be found in Table 7.7. The ratios were based only on values from this model since the coefficient estimates did not differ much between the three models.

Name of predictor x_j:	Relative odds ratios:
Age.3	0.982
Age.2	0.977
Age.1	0.987
DiagnosesICPC.3	1.126
DiagnosesICPC.2	1.091
DiagnosesICPC.1	1.069
HigherEducation.3	1.777
HigherEducation.2	1.858
HigherEducation.1	1.663
NoLongTermSick.3	0.401
NoLongTermSick.2	0.745
NoLongTermSick.1	0.595
Procedures.3	1.077
Procedures.2	1.066
Procedures.1	1.031

Table 7.7: (TPH) Relative odds ratios (1 unit increase) calculated from Table 7.5.

7.3.2 Interpretation of relative odds ratios (TPH)

Based upon the relative odds ratios in Table 7.7, one could draw the following conclusions about how the parameters affect the response:

- First the age of a patient (*Age*) has a significant effect. The older a patient is relatively to another patient causes a greater decrease in odds of receiving personalized help. The magnitude change in odds is almost equal over all the categories.

- Second, we have that the more diagnoses a patient has (*DiagnosesICPC*) the higher the odds of that patient receiving personalized help. The increase in odds is steady when comparing the three different subgroups. Thus a patient will be slightly more likely to have a 3 in help score rather than 0, compared to a score of 1 than 0; when having many diagnoses.

- Third, the level of education (*HigherEducation*) has an effect. The three different categories do not to agree on strictly increasing odds, but there is still a tendency present. The tendency is that having a higher degree of education will over all increase the odds of receiving personalized help.

- Fourth, the absence of long lasting illness (*NoLongTermSick*) affect the odds negatively with regards to receiving personalized help. The absence of chronic illness is especially affecting the odds of receiving "maximal" personalized help.

- Last we have indications that, as similarly with the number of diagnoses, that the number of procedures (*Procedures*) performed on the patient will affect the odds of having received personalized help positively. A patient with relatively many more procedures will thus have higher odds of receiving help.

Personally some of these sounds like reasonable results. For instance that patients that have more diagnoses, that is sick over a longer period of time and that have undergone more procedures are more likely to have received more personalized help. The education level could also be reasonable since doctors may have an easier time communicating with patients at the same mental level as them; or that patients have a higher ability to access, understand and apply information regarding their condition in general. The age factor could be too complex, when considering predictors, to decide whether it is reasonable or not; without relying on other sources.

7.3.3 General information help group response

As with the other multinomial response, this response yielded the summary in Table 7.8 from the four GBM procedures. There are two candidate predictors indicated to have much relative influence. They are the *Age* of a patient and the number of procedures (*Procedures*) performed on a patient.

Again the boosting has found the most significant and valuable predictors. Though, there were some predictors that did not have any relative influence that was significant in some of the models fit, but they had low explanation value in terms of reducing the DIC. Correlation did also hide potentially different predictors, but from the models fit they were less important or had less impact.

When deciding on the possible good and final models to predict the response the idea about predictors needing to be satisfyingly significant (from

Name of predictor x_j:	Proportion of sub-models significant:
Age	4/4
Procedures	4/4
DiagnosesICD	2/4
DiagnosesICPC	1/4
Services	1/4
TimeInHospital	1/4

Table 7.8: Summary of the four (GIH) logistic GBM subproblems.

section 7.3.1) had to be relaxed from 2/3 to 2/4. This relaxation was seemingly necessary since the subcoefficients to each predictor were less significant. Another interesting element that made itself truly apparent when fitting the different models, was that the number of procedures were highly correlated or confounding with the other predictors found to have relative influence. Out of every predictor that was not deemed to have relative influence by the GBMs, only *ActiveTough* has a significant effect. Though a possible replacement of *DiagnosesICPC* with *ActiveTough* in the third model (Table 7.11) did not affect the DIC enough, which only makes it worth to mention for future references. I thus propose the final multinomial models, see Table 7.9, 7.10 and 7.11.

In model one (Table 7.9) and three (Table 7.11) the subcoefficients of the number of procedures and number of diagnoses (*DiagnosesICPC*) that reference category one and two were not significant. Therefore model one (Table 7.9) has two coefficients that is not significant while model three (Table 7.11) has four coefficients with the same lack of significance.

Name of predictor x_j:	Estimate of β_{kj}:	CI$_L$:	CI$_U$:
Age.4	-0.019^{***}	-0.030	-0.009
Age.3	-0.025^{***}	-0.035	-0.014
Age.2	-0.019^{***}	-0.030	-0.008
Age.1	-0.013^{*}	-0.026	$-4 \cdot 10^{-4}$
Procedures.4	0.119^{***}	0.064	0.179
Procedures.3	0.095^{**}	0.036	0.156
Procedures.2	0.054	-0.008	0.120
Procedures.1	0.057	-0.012	0.128

The credible intervals (CI$_L$ & CI$_U$) cover 95% of the distribution.
* Significant at a 0.05 level.
** Significant at a 0.01 level.
*** Significant at a 0.001 level.

Table 7.9: Model 1 (GIH) - containing the two predictor with the most relative influence - DIC = 13478

Name of predictor x_j:	Estimate of β_{kj}:	CI$_L$:	CI$_U$:
Age.4	-0.022^{***}	-0.033	-0.012
Age.3	-0.026^{***}	-0.037	-0.015
Age.2	-0.021^{***}	-0.032	-0.009
Age.1	-0.015^{*}	-0.027	-0.001
DiagnosesICD.4	0.129^{***}	0.073	0.190
DiagnosesICD.3	0.092^{**}	0.033	0.155
DiagnosesICD.2	0.066^{*}	0.003	0.132
DiagnosesICD.1	0.072^{*}	0.003	0.143

The credible intervals (CI$_L$ & CI$_U$) cover 95% of the distribution.
* Significant at a 0.05 level.
** Significant at a 0.01 level.
*** Significant at a 0.001 level.

Table 7.10: Model 2 (GIH) - with only significant predictors - DIC = 13475

Name of predictor x_j:	Estimate of β_{kj}:	CI$_L$:	CI$_U$:
Age.4	-0.021^{***}	-0.031	-0.010
Age.3	-0.026^{***}	-0.036	-0.014
Age.2	-0.020^{***}	-0.031	-0.008
Age.1	-0.014^{*}	-0.026	$-7 \cdot 10^{-4}$
Procedures.4	0.096^{***}	0.040	0.156
Procedures.3	0.078^{**}	0.019	0.141
Procedures.2	0.045	-0.018	0.111
Procedures.1	0.052	-0.018	0.124
DiagnosesICPC.4	0.094^{***}	0.041	0.149
DiagnosesICPC.3	0.066^{*}	0.010	0.123
DiagnosesICPC.2	0.041	-0.019	0.101
DiagnosesICPC.1	0.019	-0.048	0.088

The credible intervals (CI$_L$ & CI$_U$) cover 95% of the distribution.

* Significant at a 0.05 level.

** Significant at a 0.01 level.

*** Significant at a 0.001 level.

Table 7.11: Model 3 (GIH) - with the least relatively DIC value - DIC = 13457

In order to illustrate possible interpretations, we will take the unique values from the predictors first appearance in the models from one to three and calculate relative odds ratios, Table 7.12. This can be done without much loss since the different coefficient estimates corresponding to the same predictor are not that different between the models.

Name of predictor x_j	Relative odds ratios:
Age.4	0.981
Age.3	0.975
Age.2	0.981
Age.1	0.987
Procedures.4	1.126
Procedures.3	1.099
Procedures.2	1.055
Procedures.1	1.058
DiagnosesICD.4	1.137
DiagnosesICD.3	1.096
DiagnosesICD.2	1.068
DiagnosesICD.1	1.074
DiagnosesICPC.4	1.098
DiagnosesICPC.3	1.068
DiagnosesICPC.2	1.041
DiagnosesICPC.1	1.019

Table 7.12: (GIH) Relative odds ratios (1 unit increase), based on the top-down coefficient estimates from model one to three.

7.3.4 Interpretation of relative odds ratios (GIH)

Again, basing my reasoning on the relative odds ratios calculated in Table 7.12, one could draw the following conclusions about the parameters:

- The age of the patients (*Age*) reduce the odds of receiving general information help. This applies to all the subgroups.
- The remaining predictors in this case, *DiagnosesICD*, *DiagnosesICPC* and *Procedures*, all affect the odds of receiving help positively and also in an increasing fashion.

Most of these results do sound reasonable, as they are similar to the measures regarding the personalized help group response. Though, due to a lot of correlation it is possible to find many different models which basically could predict almost the same values. For instance since the number of procedures appeared to correlate strongly with the number of ICD diagnoses the first two models may in practice be extremely similar. It is interesting that age is significant with this response as well, which might be an indicator that the complexity introduced by age may be common.

7.4 Comments and concluding remarks

Comparing the results from the responses including the results from the logistic analysis in section 6.2, we can try to say something about the similarity between the two questions the three responses were constructed from. Based on the results it should then be safer to think that the multinomial responses and logistic response do explain something rather different; in addition to that they were from two uniquely constructed set of questions in the questionnaire. It becomes apparent by looking at the difference in predictors significant and the sign at the corresponding coefficients between the two analyses. The fact that there have been used different distributions may speak against such a conclusion, but as already established the multinomial and logistic distribution are not that different; when considering the differences between the most common distributions.

From the results in the two multinomial cases (section 7.3.1 and 7.3.3) boosting did without a doubt find the most valuable or significant predictors to use in that exact setting. The simplification by using $K-1$ logistic distributed boosting instead of a single multinomial distributed boosting did for the most part yield the desired result, but may have yielded a few variables with relative influence that only would have been relevant within the actual subproblem. It was noted that the boosting could have missed certain predictors that could be of interest, but again the explanation (DIC) value of those were weaker than those chosen by the boosting. Those predictors not noticed by the boosting procedure did tend to be correlated with other predictors that had more relative influence. It could or should be possible to find these optional predictors by either using theoretical or empirical measures of correlation.

Thus based on the empiric evidence from the previous cases analyzed we could assume or conclude that boosting could in fact have potential to be used to mine significant predictors with relative high certainty. As expected the certainty became better the lower the learning rate and the accuracy of the distribution, for instance when using a logistic distribution when there is a true logistic problem present and not using the logistic distribution when the true distribution is multinomial. Of course many more regression cases should be explored with boosting to better and further determine its accuracy when used to mine predictors.

Chapter 8

Patient trajectory analysis of health care utilization

In this chapter the main goal is to apply the Markov model framework, and use models to investigate patient trajectories. The two types of models which we will be using are the discrete-time Markov chain models and the hidden Markov models introduced in chapter 4. Discrete-time Markov chain models are used to describe any system that can be contained within the model's restrictions, with regards to the Markov property and having a finite number of states. Hidden Markov models are mostly used to model systems where we have underlying information, missing or not, that is assumed to directly affect an observed sequence of events or values. It is for example used within cryptanalysis (Karlof and Wagner, 2003), speech recognition (Rabiner, 1989), finance (Hassan and Nath, 2005), medicine (Ohlsson et al., 2001) and many other areas. As the model has inherited restrictions from the discrete-time Markov model, its applications do sometimes require generalizations of the model. Ideally one or both of these models should be able to explain, without further generalizations, a trajectory for a patient or trajectories for a population of patients and be able to predict upcoming events in the trajectories.

8.1 Deciding on the finite and observable states

In the analyses to be done, both the discrete-time Markov chain models and the hidden Markov models have a finite state space. The states considered here correspond to four groups of health care services, which can be seen to reflect illness severity, urgency or cost. These four states are, again, summarized in Table 8.1

Symbol:	Description:	Urgency/cost priority:
IP	Inpatient	First
OP	Outpatient	Second
GP	General practitioner	Third
NO	None of the others	Fourth

Table 8.1: Overview of the four constructed states with description and urgency.

The most urgent event group is *IP* or the *inpatients* that represent all the health care events that require a patient to be admitted to a hospital. Second we have the event *OP* or *outpatients* that represent all those that receive outpatient care, for instance at polyclinics. Third is the *GP* event and it represents all those who receive consultations at a general practitioner or medical assistance of equivalent level. Last is the *NO* or *None of the others* event, which is the event of not being in any of the other three events.

The detailed information about the preprocessing of the data, creation of the patient trajectories and events in this analysis have been covered in section 5.2.

8.2 Trajectory analysis using discrete-time Markov chain

In this section we fit the discrete-time Markov chain model with states *NO, GP, OP* and *IP* to the data. The initial distribution and the transition probabilities are estimated by using the estimators in (4.4) – (4.7).

8.2.1 Estimated Markov chain model

Using the training set of chronic patient trajectories (section 5.2.2) in the estimators, a transition probability matrix,

$$\hat{\boldsymbol{A}}_{4\times 4} = \begin{array}{c c} & \begin{array}{cccc} NO & GP & OP & IP \end{array} \\ \begin{array}{c} NO \\ GP \\ OP \\ IP \end{array} & \begin{pmatrix} 0.662 & 0.243 & 0.084 & 0.011 \\ 0.361 & 0.476 & 0.140 & 0.023 \\ 0.314 & 0.300 & 0.344 & 0.042 \\ 0.150 & 0.310 & 0.268 & 0.272 \end{pmatrix} \end{array} \tag{8.1}$$

a transition count matrix,

$$\hat{\boldsymbol{A}}_{C,4\times 4} = \begin{array}{c} \\ NO \\ GP \\ OP \\ IP \end{array} \begin{array}{c} \begin{array}{cccc} NO & GP & OP & IP \end{array} \\ \begin{pmatrix} 96521 & 35459 & 12223 & 1592 \\ 34745 & 45903 & 13475 & 2234 \\ 13355 & 12767 & 14639 & 1780 \\ 1156 & 2399 & 2070 & 2104 \end{pmatrix} \end{array} \tag{8.2}$$

an initial probability distribution vector and a stationary distribution vector was calculated. See Table 8.2 for the initial- and stationary distribution estimates.

Distribution type/state:	NO	GP	OP	IP
Initial - $\hat{\boldsymbol{\pi}}^1$	0.506	0.319	0.148	0.026
Stationary (ML) - $\hat{\boldsymbol{\pi}}$	0.499	0.330	0.145	0.026
Stationary (Limit) - $\hat{\boldsymbol{\pi}}$	0.499	0.330	0.145	0.026

Table 8.2: Overview of the initial distribution and the stationary distribution measures, rounded up to three digits.

Regarding the robustness of the model and estimates, there is only a slight difference between the maximum likelihood estimate and the limit estimate, suggesting that they are both reliable in our case. Also, looking at the counts of the different transitions (8.2) we can see that there should be enough transitions between each state to provide a reliable estimate of the transition probability matrix. As a simplification we have assumed that the Markov property is valid, when it probably is not completely valid. We assume this because the discrete-time Markov chain model is supposed to serve as a simple outset and pave the way for the hidden Markov model.

8.2.2 Interpretation of the Markov chain model and conclusions

The first thing to note about this model is that the initial distribution is quite similar to the stationary distribution. This may imply that the Markov model is initially stationary, but on the other hand it took at least six months for the Markov process to converge; based on estimator (4.4). The patients definitely had some events before the time the data was gathered, thus it may not be that far fetched to claim that the Markov model is stationary from the first observed state.

By looking at the transition probability matrix, $\hat{\boldsymbol{A}}_{4\times 4}$, and assuming that the Markov property holds for our process, we can deduce how the current state is going to affect the probability of entering another state. A rather general remark that is possible to draw from this is how the severity or financial cost of the current state affect the probability of entering a more or less severe state. For instance the more severe a current state is, then the more probable is it that the next state is also rather severe. On the other side the less severe the current state is, the more probable is it that the next state is less severe. Another observation is that the diagonal, $\hat{A}_{jj}$ for $j = 1, 2, 3, 4$, carry the maximum probability value, three out of four times, which means that a patient in a given state most likely remains in the same state in the next month. The one exception is if the current state is *IP*, in which the probability of not being admitted to a hospital again the next month is 72.8%.

Using the transition matrix, we can easily calculate the expected time in months, spent in a state given that a patient already is in that state, $E(\text{Time in state } j \mid X_{\text{current}} = j)$. Using the geometric distribution and its expectation together with the diagonal values of the transition matrix,

$$\hat{E}(\text{Time in state } j \mid X_{\text{current}} = j) = \frac{1}{1 - \hat{A}_{jj}}$$

we then get the following estimated expectations

$$\hat{E}(\text{Time in state } j \mid X_{\text{current}} = j) = (2.959, 1.910, 1.525, 1.374).$$

If we round to the closest monthly integer we can thus say that a patient is not expected to be at the hospital the next month, if already in state *IP* that month. While a patient that has no need of health care at the current month will be expected to not need it for the next two months. Lastly those who receive help from a general practitioner or outpatient treatment will be expected to receive the same type of help next month.

In general the transition matrix can be used to answer many questions regarding the trajectories, especially it can be useful to calculate the most probable sequence of states or patient trajectories. The sequences can be of any length, but since this is a Markov chain model it will only be interesting to check a short amount of the N most probable states after a unique initial state. It is desired to keep the sequences short, or not long, since the Markov property will always play a key role when calculating the transitions. For example if we start in state $j = 1$, then when the process enters another state $j \neq 1$ we can instead look at the most probable states when beginning in that other state. Table 8.3 provides a short excerpt of the most probable sequences of states of length four, i.e. a quarter of a year into the future.

Starting state:	Probability:	Sequence of states:
$NO = 1$	0.145	$1 \to 1 \to 1 \to 1$
$NO = 1$	0.053	$1 \to 1 \to 1 \to 2$
$NO = 1$	0.038	$1 \to 1 \to 2 \to 2$
$NO = 1$	0.029	$1 \to 1 \to 2 \to 1$
$NO = 1$	0.029	$1 \to 2 \to 1 \to 1$
$GP = 2$	0.052	$2 \to 1 \to 1 \to 1$
$GP = 2$	0.038	$2 \to 2 \to 1 \to 1$
$GP = 2$	0.036	$2 \to 2 \to 2 \to 2$
$GP = 2$	0.027	$2 \to 2 \to 2 \to 1$
$GP = 2$	0.019	$2 \to 1 \to 1 \to 2$
$OP = 3$	0.020	$3 \to 1 \to 1 \to 1$
$OP = 3$	0.010	$3 \to 2 \to 1 \to 1$
$OP = 3$	0.010	$3 \to 3 \to 1 \to 1$
$OP = 3$	0.010	$3 \to 2 \to 2 \to 2$
$OP = 3$	0.007	$3 \to 2 \to 2 \to 1$
$IP = 4$	0.002	$4 \to 2 \to 1 \to 1$
$IP = 4$	0.002	$4 \to 2 \to 2 \to 2$
$IP = 4$	0.002	$4 \to 1 \to 1 \to 1$
$IP = 4$	0.001	$4 \to 3 \to 1 \to 1$
$IP = 4$	0.001	$4 \to 2 \to 2 \to 1$

Table 8.3: Overview of the five most probable sequences of length four, when starting in the different states. $NO = 1$, $GP = 2$, $OP = 3$, $IP = 4$.

The stationary distribution also offers other interpretations than the transition matrix to further enrich the explanation of the patient trajectories. Especially, the stationary distribution can be interpreted as the proportion of time spent in the states or the probability of being in the states in the long run. Table 8.2 illustrates that a patient is expected to spend about 50% of the time in state *NO*, 33% in state *GP* and slightly less than 15% in state *OP*. Finally, only two to three percent of the time is spent in state *IP*.

Since the stationary distribution provides limiting estimates of the probability to enter a state regardless of the current state, its probabilities can be used with the binomial and multinomial distribution. By using, for instance, the binomial distribution we can calculate the probability of never entering one of the states during a period of time. Table 8.4 provide the probabilities of never entering a certain state for twelve months.

State:	NO	GP	OP	IP
Probability:	0.00025	0.008	0.153	0.729

Table 8.4: Overview of the probability to never enter a certain state for 12 months, calculated using the binomial distribution and stationary distribution with its probabilities.

In conclusion, the discrete-time Markov chain model can be used to model patient trajectories. Its strengths is that it can be estimated from 100% known and accurate data. However, we have assumed that the Markov property is valid, an assumption which is most likely not valid. It is easy to think of a situation that breaches this property. An example could be that we have a reference to a specialist from a doctor where it may take more than two months to receive the specialist health service. Due to the requirement of the Markov property, the discrete-time Markov chain model is not fit to be used to predict future states. Even though prediction isn't this model's strongest ability, it still has a way of providing useful information through patient trajectory descriptions. In this section there have been a few examples of ways the model can describe the trajectories for both patients and whole populations, but there are still more possible measurements left to be calculated; depending on what is required to know. As such, at a lack of better options the discrete-time Markov chain models can definitely be used to illustrate tendencies of health service usage. Furthermore if we approximately know the strain and stress the health services are put through, we can assume that it will provide incentives to delegate more funds to those services.

8.3 Trajectory analysis using hidden Markov model

This section presents another model from the Markov model framework, since the discrete-time Markov chain model showed some promise with its descriptions. Here, we will be using a hidden Markov model instead, whose model assumptions are more valid compared to the previous analysis. Our patient trajectory analysis will be in the partial-information scenario, and will use the four same states (*NO, GP, OP* and *IP*) as emission or observable states. The patient trajectories extracted from the data, will thus in this case represent observable sequences instead of Markov chain sequences. Working in the partial-information scenario implies that we will have a hidden state space of some order. The order of hidden states will be set to five, and it is assumed by a professional that this number of unique hidden states can provide sufficient complexity to explain, describe and predict the observed states properly.

8.3.1 Trained hidden Markov model

In this analysis a hidden Markov model will be trained by using the Baum-Welch algorithm (section 4.2.1). The hidden Markov model will be trained by using a training set (section 5.2.2), containing 12714 patients with all 24 months of data. It is assumed that these patients in the training data are representative for the others in this group of chronic patients. They should be representative in such a fashion that the distribution of chronicity are supposed to be the same within the training and test set. The optimal hidden Markov model trained on the data, selected based on the best log-likelihood value, has a transition probability matrix:

$$\hat{\boldsymbol{A}}_{5\times 5} = \begin{array}{c|ccccc} & HS1 & HS2 & HS3 & HS4 & HS5 \\ HS1 & 0.9504 & 0.0340 & 0.0102 & 0.0001 & 0.0053 \\ HS2 & 0.0266 & 0.9378 & 0.0171 & 0.0115 & 0.0070 \\ HS3 & 0.0337 & 0.0297 & 0.9058 & 0.0110 & 0.0198 \\ HS4 & 0.0001 & 0.0290 & 0.0069 & 0.9418 & 0.0222 \\ HS5 & 0.0269 & 0.0415 & 0.1610 & 0.1049 & 0.6657 \end{array}. \tag{8.3}$$

An emission probability matrix:

$$\hat{\boldsymbol{B}}_{4\times 5} = \begin{array}{c|ccccc} & HS1 & HS2 & HS3 & HS4 & HS5 \\ NO & 0.849 & 0.499 & 0.231 & 0.093 & 0.019 \\ GP & 0.105 & 0.433 & 0.233 & 0.745 & 0.149 \\ OP & 0.043 & 0.063 & 0.515 & 0.141 & 0.285 \\ IP & 0.003 & 0.006 & 0.021 & 0.021 & 0.547 \end{array}. \tag{8.4}$$

The "*HSX*" label, in the matrices (8.3) and (8.4), represent the hidden state number "X", and is provided to ease interpretation of the matrices. Lastly an estimate of the initial/stationary distribution of the hidden states is:

$$\hat{\boldsymbol{\pi}} = (0.310, 0.341, 0.162, 0.156, 0.032). \tag{8.5}$$

This model is trained under the assumption that the sequences constructed are from a process that is stationary at the first point in time the data covers, i.e. January 2012. This can be argued for as all patients have existed and lived before the collection of the data, since they are all adults. Estimates from the discrete-time Markov chain model analysis (Table 8.2) also indicate that this can be a rather feasible assumption. The definition of chronic illnesses backs up this assumptions as well, suggesting that a chronic condition should be relatively unchanged or stable over a larger period of time. Though different degrees of chronic illnesses may be plausible.

8.3.2 Interpretation of the hidden Markov model

Using the transition probability matrix (8.3) in combination with the emission probability matrix (8.4) we have the opportunity to calculate similar estimates as with the discrete-time Markov chain model. We focus now on other aspects which we did not achieve using only the Markov model. Before anything else it is necessary to check what this hidden Markov model actually does tell us, or else we cannot really interpret or draw conclusions from the model at all. We do know from section 5.2.1 the severity and how to interpret the observable signal states, but how are we going to give meaning to the hidden states since we are in a partial-information scenario?

One possibility is to interpret the hidden states as levels of chronic complexity patients inhibit. This implies that the hidden states are representing a complexity measure that have an effect on the health services a patient is receiving or is given at the time.

Our inductive reasoning behind this interpretation is as follows. Knowing that the observable states have a degree of severity, we can then reason backwards using only the emission probability matrix (8.4) to reach the inductive conclusion. From element $\hat{B}_{11}$ in column one in the emission matrix, patients in the hidden state one ($HS1$) will have a 85% chance of not receiving health care. If we look at the NO row we can see that it has the highest probability in column one. This suggests that the hidden state one ($HS1$) has the least severe level of chronicity. Patients in the second hidden state ($HS2$), column two, has a greater chance of receiving health care, but overall less chance than hidden state three, four and five. Following this logic it is then reasonable to believe the hidden state three and four ($HS3$ and $HS4$) describe about the same level of severity, though they describe different types of chronic situations. Patients inhibiting hidden state five ($HS5$) at a time will have more than a 50% chance of requiring hospital admission, making it the most severe hidden state. As 80% of hospital admissions are due to emergency, the hidden state five may represent chronic patients not being able to cope with recurring and worsening symptoms. As such, an apparently valid interpretation is that we have five different groups or levels of chronic severity within the hidden states that are supposed to explain health service usage.

A property of the trained hidden Markov model is that the hidden states can be interpreted to be relatively permanent. This can be seen from the diagonal in the transition matrix (8.3), which is without doubt diagonally dominant. The only case which can be called an exception is the most critical hidden state five ($HS5$). This hidden state has about a 33% chance to move away from this state in the next month. It is still expected that a patient changes the hidden state he or she inhibits at least once every twentieth month. This property may also result in the model being slow to register actual changes in the hidden state sequence, and subsequently in the observed state sequences, when it is correct to notice a change. This will most definitely yield inaccuracies for patients that in practice have rapidly changing hidden states.

8.3.3 Prediction accuracy results

Applying the given model, a first aim is to predict future states of a patient trajectory. Prediction was then performed according to section 4.2.3 while using the hidden Markov model in (8.3) – (8.5). A simple max criterion was used to select the most probable state from the resulting vector of probabilities. To verify whether the model is representable and generalizable, the model was tested on 67210 patients not included in training the model. As we have access to a rather large amount of prior trajectory information per patient it was of interest to use different lengths of prior data. The three prediction cases primarily checked had prior information of length two, six and twelve and a summary of the success chance can be seen in Table 8.5. It is worthwhile to note that the number of prediction cases used to calculate the mean success depend on the amount of prior information used, since we only have 24 months of data.

Prior information length:	Mean success:
2	0.594
6	0.604
12	0.605

Table 8.5: Summary of the prediction accuracy.

On a side note it should be mentioned that a prediction test was also performed on the patients that were used to train the model in section 8.3.1, but then another similar model using less training data was utilised instead. This prediction managed at worst 56% and at best 60% success, which is interesting compared to the results in Table 8.5 as it suggests that the training set is representative for the testing set or vice versa.

8.3.4 Conclusions about of the results

The model trained did get on average about 60% chance of success when trying it out on the testing set. Compared to completely random selection of four alternatives it is a rather significant improvement from 25% to 60%. It is not entirely unexpected that the amount of prior information had little effect considering that the Markov property is part of the model, but on the other hand we do have the posterior probability which is conditioned on the whole observed sequence from a patient. Thus, it is possible that the trained model itself do play a part and that this non-necessity of long sequences of prior information may be a case specific property unique to this model.

The accuracy test did not measure which unique patients that was predicted successfully with different prior information. This could imply that a shorter amount of prior information predicted a certain group of patients correctly, while a larger amount of prior information predicted the same group incorrectly, but another group correctly. All in all this is perhaps not that important, as we simply need to have a model that can predict correct with the least amount of failures. Though it could be important to be aware of this possibility in the future, and in retrospect it is something that should have been taken into account.

Another element to the prediction that is crucial to discuss is how the hidden Markov model tends to underestimate or predict a lot of false negatives, see Tables 8.6 – 8.9 for the chosen example to illustrate the unfortunate, but recurring trend when predicting. Of course the model does not only give us false negatives, false positives are also present, but the amount of false negatives clearly exceeds the number of false positives, see Table 8.7. The large quantity of these failed predictions do provide some notion of how problematic this issue is, but if we also look at the matrix of conditional probabilities in Table 8.9 we can see how bad it is. For instance if we look at the second, third and fourth columns, then we have respectively 57.9%, 76% and 90.1% chance of producing false negatives. To compare with the chance of overestimating in the first, second and third columns we have the respective probabilities to be 14.8%, 4.7% and 1.3%. Thus it is reasonable to say that this model is heavily biased toward the less severe observable states, as the example is representative for the other prediction cases.

Number of successes:	Success percentage:
39486	0.587

Table 8.6: Prediction excerpt example part 1/4 - Summary of successes. Also illustrated by summing the diagonal in Table 8.8

Number of overestimates:	Number of underestimates:
6175	21549

Table 8.7: Prediction excerpt example part 2/4 - Summary of the number of over- and underestimates. These are also referred to as false positives and false negatives. It is also illustrated by the off-diagonal in Table 8.8

	A:NO	A:GP	A:OP	A:IP
P:NO	28699	13284	3778	555
P:GP	4223	8596	3031	605
P:OP	711	977	2031	296
P:IP	55	89	120	160

Table 8.8: Prediction excerpt example part 3/4 - A count matrix describing the number of times a state is predicted (rows) versus the number of times a state is actually occurring (columns)

There are a few probable reasons as to why the model used in predicting is this biased. One of them could be that there are too many chronic patients that are too healthy compared to a less healthy minority. Another reason could be that the model has too few hidden states. This second reason is based on the fact that the transition probability matrix is diagonally dominant. Since a patient has a rather high probability of entering the same state, this could imply that each of the hidden states may include more than one degree of complexity. A degree of complexity within the hidden states that this hidden Markov model possibly hasn't been able to take into account with the number of unique hidden states being set equal to five. The fault can also lie within the selection criterion post prediction, which was set to be the maximum criterion. The selection criterion in the prediction can be

	A:NO	A:GP	A:OP	A:IP
P:NO	0.852	0.579	0.422	0.343
P:GP	0.125	0.375	0.338	0.374
P:OP	0.021	0.043	0.227	0.183
P:IP	0.002	0.004	0.013	0.099

Table 8.9: Prediction excerpt example part 4/4 - A probability matrix describing the probability of predicting a state given the true state, i.e. the columns sum to one

imprecise, since the maximum criterion does not take into consideration any similar probability values or other probability values that are above a certain threshold. Other smarter criteria than the maximum could then hypothetically be very desirable to use instead, though there is no apparently better alternative than the maximum criterion at this point.

In summary, we have now found a hidden Markov model whose base assumptions are not violated by the data at hand, which was the case with the discrete-time Markov chain model. At the same time we have performed prediction that had about 60% accuracy on average, an increase from the theoretical random selection with four alternatives. The presented hidden Markov model is not perfect, but it is better than the discrete-time Markov chain model since we have the hidden states as a structural backbone to the observations. The imperfections related to this model as of now are definitely not impossible to handle, but they would require time and computational resources to solve. Compared to the discrete-time Markov chain model, the hidden Markov model has achieved an additional objective other than being able to describe the patient trajectories. That additional goal is that it can perform predictions while the model's assumptions are sufficiently satisfied. Though some tuning is still required and absolutely necessary before it is put into practical situations, since it does have an unsatisfactory amount of failed predictions. An interesting aspect resulting from this hidden Markov model analysis is the interpretation and information gained from the hidden states. A suggested interpretation with an inductive proof were provided, and it is reasonable to believe as of now that those hidden states could be able to serve a larger purpose than only being the underlying structure in the model. Even though our analyses have shown that hidden Markov models are better fit at modelling patient trajectories than discrete-time Markov chain models, it would probably be wise to check more advanced models that relaxe the one-step dependence. Though, at a monthly level the hidden Markov model is sufficient to analyse patient trajectories from a theoretical point of view.

Chapter 9

Discussions and future work

9.1 Discussion of key points in the regression analyses

Two different regression models, supplemented by boosted regression, were applied on the aggregated data to find significant predictors. Some predictors found to be significant were recurring between the models, but would tend to have different interpretations. This difference in interpretation suggests that the questions used as response-variables (*S7* and *S10*) in the logistic and multinomial regression do represent and capture different aspects of the health service that was provided to the patients.

9.1.1 Significant logistic model predictors

The predictors that were deemed to have an increasing or decreasing effect on the odds of being satisfied with the health care received were varied, but often intuitive. Two predictors were unclear regarding why they should be significant, and they were representing age (*Age*) and gender (*Gender*). Being a man increases the odds of being satisfied, likewise higher age of a patient would increase the odds as well. Apparently age is a common and well known factor for explaining the level of satisfaction, and satisfaction typically increases with higher age (Rahmqvist, 2001; Thi et al., 2002; Cohen, 1996). A slightly unusual choice to our age variable is that it is continuous rather than grouped into groups of ages, but even such a choice yielded known results. This could imply that if enough data is present, continuous age can be preferable to age intervals in groups; in terms of degree of freedom. Rahmqvist

(2001) found that gender did not correlate with satisfaction, but males were somewhat more satisfied. While another study found men to have higher odds of being more satisfied with health care than women, though not consistently (Thi et al., 2002). These findings are in correspondence with ours, since gender was not the most consistently significant predictor compared to the others. Any differences between gender could for instance be attributed to other underlying factors, for example differing expectations within groups (Hsieh and Kagle, 1991) or group specific treatment, and it might have been coincidental that we found it to be significant. Put shortly, whether gender truly is significant is undecided.

The remaining significant predictors to this response have effects that are more intuitive. For instance we found that not having experienced long lasting illness (*NoLongTermSick*) will increase the odds of being satisfied and the same applies for never having experienced being debased (*NoDebased*). It is reasonable that not having a long lasting illness affects this way, because it implies that the health care services have succeeded at their job. On the other hand having experienced being debased, bullied or harassed over a period of time can affect a patient's mental health for the worse (Hinduja and Patchin, 2010). Mental health is previously found to correlate with satisfaction, especially within psychiatric health services (Rahmqvist, 2001). As our data only contain somatic care data and not psychiatric care data, mental health factors can then be said to affect satisfaction levels even outside of psychiatric health care.

Somewhat reasonably, and similarly to the effect of long lasting illness, the number of unique diagnoses (*DiagnosesICPC*) on a patient did account negatively on the odds of being satisfied. It could therefore be inferred that frequent contact with the health service, perhaps due to unsuccessful treatments or being attached to many new diagnoses, can reduce the odds of being satisfied.

Lastly, similarly to the number of diagnoses, the worse a patient felt regarding his or her own self perceived health (*SelfRateHealth*) the lesser the odds of being satisfied. It is not unheard of that self perceived health affects satisfaction with the health care (Hall et al., 1993; Thi et al., 2002). Different practices by measuring self perceived health do on the other hand create slightly differing ways of interpreting how self perceived health is affecting satisfaction levels. With respect to how this covariate has been measured in our data, such measured odds regarding self perceived health is not unreasonable from a logical perspective. As there would definitely be fewer reasons for a patient in general to be dissatisfied with the care received if the patient is feeling well after a health care encounter.

9.1.2 Significant multinomial model predictors

When looking for significant predictors with respect to the amount of personalized help received (TPH), five unique predictors were found. The odds to receive more personalized help increased if a patient had higher education (*HigherEducation*). This effect could be explained by that the ones with higher education are more literate and thus are able to act with greater empowerment when being in contact with the health care services. Literacy is related to the term health literacy, which is used to explain how able a patient is in finding and utilising health care information in a beneficial way (Nutbeam, 2000). However, being able to empower oneself is no guarantee that the individual will and using odds to describe potential action of empowerment can be seen as appropriate.
The number of diagnoses (*DiagnosesICPC*) and number of procedures (*Procedures*) performed on a patient had about the same positive effect on the odds of receiving more personalized help. Increased odds of help per diagnose or procedure do make sense since the patient in question is more likely to have more contact with health personnel. If we also assume that patients contact health services when they require the help.
Compared to when age had a positive effect on the odds of being satisfied, here higher age does represent a decrease in odds of receiving more help. With respect to age and the number of diagnoses, it is surprising that patients that are more likely to have received more personalized help is less likely to be satisfied with the health care.
At last, we have that the absence of long lasting illness (*NoLongTermSick*) does imply reduced odds of receiving more help. Logically this predictor could in some cases mean that the patients have less procedures or diagnoses, since they are less ill, and this type of odds could make sense in that case.

In the last case we were looking for significant predictors that explain the amount of general information help (GIH) received. Higher age did again imply a decrease in the odds of receiving help. The remaining predictors that described the number of procedures (*Procedures*) and diagnoeses (*DiagnosesICPC* and *DiagnosesICD*) all implied an increased odds of receiving general information help. The odds would increase more the higher a value of these covariate. These predictors can be argued to make sense as done in the personalized help (TPH) model, since both models' response stems from *S10*. By using this logic though, it is worth noting that the education level (*HigherEducation*) was not significant for the general information help response variable. This can suggest that the two multinomial regression anal-

yses did model sufficiently distinct responses to yield additional information. We could for instance interpret this such that an individual is more likely to receive personalized help than other types of help, if he or she has higher education.
Similarly to receiving more personalized help, patients that are more likely to receive general information help is less likely to be satisfied with the health care. We've already established that the two response-variables (*S7* and *S10*) are describing two different aspects of health care, but from the analyses performed there are not enough evidence to conclude more about any correlation or relationship safely. Concluding further now would be to resort to speculation. As such, further research on the connection between satisfaction and help is encouraged.

9.1.3 Other remarks regarding the regression analyses

Overall some of the significant predictors or effects could be found in other studies where they have been found to influence satisfaction levels to a lesser or greater amount. A few of them were also significant in explaining degree of help received, but had often contradicting interpretations. On the other hand there are studies reporting factors not found significant in our study, or not even measured in our data at all.

Rosenheck et al. (1997) conducted a study in an inpatient environment for mental health care and found that longer length of stay was linked with greater levels of satisfaction. This length of stay measure should be comparable to our *TimeInHospital* covariate, which we did not find to be significant at all.

Another study found some support that if prior expectation of the primary care they were to receive were met, it would affect the level of satisfaction (Linder-Pelz, 1982). Different groups of people may even have differing expectations toward health services (Hsieh and Kagle, 1991). This suggests, together with the effect of never having experienced being debased (*NoDebased*), that some psychological factors can play a role. Due to when the questionnaire was sent out, it was not possible for this study to have a reliable prior expectation measure.

Even though education level wasn't significantly affecting the satisfaction levels, only degree of help, a meta-analysis by Hall and Dornan (1990) found that less education gave greater satisfaction. This does not directly contradict our findings, but rather it builds up under an idea about the relationship between satisfaction and the help received. The relationship that greater satisfaction could be somewhat negatively correlated by how much information or help a patient receives. Since we already have age and number of diagnoses providing contradictory effects to support such a relationship. *NoLongTermSick* is also common between the analyses and in each separate analysis its effect makes sense, but between the analyses the covariate does not contradict. It does not contradict as it both makes sense to need less information and personalized help and also be satisfied with the care received, if a patient per definition haven't had a long lasting illness.

To sum it up, these relationships are complicated, but we can argue that there are indications of a possible negatively correlated connection between the two quality measures. Further research is definitely required to determine the link between satisfaction of health care and degree of help received. We have also found both more known and unknown significant predictors. Here the more known predictors are most definitely age and self perceived health, while more unknown or unexpected predictors were for example *NoDebased*, explaining mental health since we aren't strictly in a psychiatric setting.

As a final note regarding the regression analyses, we should emphasis the following: As most of the many studies are heterogeneous with regards to location, size, method, sample, subgroups and predictors, it makes direct comparison between their studies and our, difficult. Therefore our focus have been on similar or dissimilar effects as it is manageable. Taking this into consideration, it is remarkable that some predictors are recurrently significant and with similar interpretation.

9.1.4 Comments about boosting regression predictors

From the few cases of boosted regression examined, there have been positive results. The results do provide indications that this type of boosted regression definitely has the potential to serve as an automatic predictor mining procedure for other more known standard regression models, for instance within the generalized linear model framework. In one of the cases, boosted regression was used to validate logistic regression, with success. While in the remaining cases the boosted regression was used to select the predictors for the two multinomial regression model frameworks, also with relative success.

9.2 Discussion of key points in the trajectory analyses

Initially, it was unclear what types of statistical models that could efficiently explain patient trajectories, but here we have found indications that models in the Markov model framework does appear to be justified. In the patient trajectory analyses we used two different models within the Markov model framework trying to model, describe and predict these patient trajectories.

9.2.1 Remarks about the discrete-time Markov chain model

The presented discrete-time Markov chain model is a rather naive attempt at modelling and describing the patient trajectories, due to the Markov property's one-step dependence assumption. Our estimated discrete-time Markov chain model did provide many ways of describing the trajectories, more ways than what is reasonable to include in this thesis. As such the model can be used to illustrate tendencies in the health service usage by patients, and serve as a source of information to aid administrative decision making.

9.2.2 Remarks about the hidden Markov model

Since the first model was able to describe the patient trajectories well, we decided to also use the hidden Markov model on the same data. Due to similarities between the two models, the hidden Markov model should inherit the discrete-time Markov chain model's ability to describe patient trajectories. From a theoretical and logical perspective this model can be assumed to solve the problem of the one-step dependency. That is because the hidden states possibly can model the dependency and other unknown necessary complexities present in the data. The addition of the hidden states implies that this hidden Markov model is more flexible than the discrete-time Markov model, and it should as a consequence therefore fit better to our data. How to interpret the hidden states were tried explained using inductive reasoning. The current interpretation is that the hidden states do represent an index or levels of chronicity, or sickness complexity that explains health service usage. However, this interpretation remains to be tested further in exploratory research. Though, the possibility that we may have created such a health index is quite interesting in itself, as the application of this objective measure given enough data could provide real time estimates of the health in a population. The hidden Markov model was also used to predict the events one step into the future of patient trajectories. In terms of accuracy, the model made a consistent guess 60% of the time, which is decent considering that a completely random guess would in theory only have a 25% chance of success. Such an increase in accuracy coupled with the fact that the model used is not guaranteed to be the optimal model and that the model assumptions do hold, suggests that a hidden Markov model in general could be fitted and used in practice.

9.2.3 Other remarks regarding the given Markov models

It is not the first time hidden Markov models have been used in a medical setting. For example it has been used to describe disease trajectories or progression (Guihenneuc-Jouyaux et al., 2000), since such measures can also be discretised while retaining temporal properties. One of the studies even found states that the model can be used on various sequences or strings with codes of treatment (Ohlsson et al., 2001), similar to as we have done with the patient trajectories.

In summary, it is reasonable to claim that we have been able to model, describe, predict and perform analyses on patient trajectory data. Descriptions of the patient trajectories have been these models' strongest suit and could prove to be useful on a population level in real life applications. Even though the models cannot be used in practice to predict on an individual or at a population level yet, the models presented in this thesis could inspire to further research. Research that eventually would lead to models like these to be used in practice when sufficient model verifications are complete.

Finally it could be wise to elaborate on a possible hypothetical issue if such predictive models one day were to be applied in practice to help make decisions in clinical domains. If this model's directions are followed with no regard to nuance by health personell, which could be the case, then the model will enforce bias from itself with respect to future models. In other words, the model may help construct patient trajectories that are similar to the ones the current model is built upon. This can result in suboptimal models that fail to learn or model critical information from observed patient trajectories affected by predictions. If the model providing clinical decision support is very biased, as our model in this thesis is, then the inevitable bias reinforcement should be even worse. A possible delay of the problem could be to divide the population of patients to be predicted by a model into two or more mutually representative groups, where each group train a model that is used to predict another group. Though this will only introduce indirect bias instead of a direct effect. A concrete fix to get rid of the auto inflicted bias could be to use a small proportion of the population to train a model and use it to predict trajectories for the remaining majority. The cost of such an approach is that this smaller proportion can't receive the same level of care as the majority without compromising the predicting model to some degree. Another and perhaps less unfair solution could be to change the minority for instance every second year. That way it would be possible to gather two years worth of unaffected data to train further models. There will be no ethically correct answers among the few workarounds suggested, as they could present tradeoffs related to who would live or die. In other words we would have to choose between providing suboptimal and perhaps fatal prediction for all patients or optimal prediction for a majority of patients at a time. The implications of these tradeoffs will be proportionally related to how accurate, widespread and integrated such decision support systems can become in the future.

9.3 Future work

9.3.1 Pending aspects from the regression analyses

There was one element that really caught my interest when using boosted regression to help validate standard regression. It was the fact that the boosted regression almost always found the most significant predictors in the cases that were examined. The number of cases examined were not many, thus what was observed may as well have been completely random. Therefore it would be utmost intriguing to perform many case studies, comparing the boosted regression to the standard regression with generalized linear models. By performing additional case studies it could be possible to figure out if boosted regression can perform better than standard regression and whether the boosted regression can be consistently used as an automatic mining procedure to select predictors for the known regressions. With the increasing trend with larger sets of data, boosted regression could help speed up the time it takes to perform regression and help scientists chose unorthodox predictors.

If we look away from the boosting aspect, we did also find indications of a relationship between the two measuresments of quality. This connection or possible correlation should definitely be researched further. As further insight of these two covariates can determine if patients actually are more satisfied with more personalized or general information help. Regression models can for instance be utilized again, or other methods could be preferred.

9.3.2 Next possible steps regarding patient trajectories

In the patient trajectory analysis the hidden Markov model achieved in this medical setting and thesis so far can be considered to be a conceptual prototype. Since there is definitely potential in the model to be used in practice, but not directly as the model is as of yet. The next step would be to try and validate the Markov models further in the setting of patient trajectories. Validation could for instance be to increase the success probability in prediction, test on other similar data or patient groups, increase the number of hidden states and perhaps also change the type of observable states. The reason why one should change the four states *NO*, *GP*, *OP* and *IP*, is because they may be too simplistic in practice.
If by some chance it is not possible to raise the success rate sufficiently to use a hidden Markov model in practice, then the obvious next step is to look into more advanced models than the hidden Markov model within the Markov model framework. Further, if those cannot do any better we have other models outside of the Markov model framework.

Clustering of the patient trajectories, either the hidden or observable sequences, while sustaining the time dependence is another future aspect that is definitely worth checking out. It would be easier to cluster the observable sequences, but the clustering of hidden state sequences can possibly increase the insight about what the hidden states do actually represent. We propose that the clustered groups could for instance be compared to other known factors about the patients in those clusters. Knowing more precisely or exactly what the hidden states explain would be a huge step towards being able to use the hidden state sequences, per individual, as some sort of real time population index. The population index could for example, as the interpretation is now, describe the health level of possibly groups, regions or whole countries of patients. This all relies on whether appropriate models can be constructed.

Bibliography

S. G. Baker. The multinomial-poisson transformation. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 43(4):495–504, 1994. ISSN 00390526, 14679884. doi: 10.2307/2348134.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970. ISSN 00034851. URL `http://www.jstor.org/stable/2239727`.

M. B. Buntin, M. F. Burke, M. C. Hoaglin, and D. Blumenthal. The benefits of health information technology: a review of the recent literature shows predominantly positive results. *Health affairs*, 30(3):464–471, 2011. doi: 10.1377/hlthaff.2011.0178.

A. Charnes, E. L. Frome, and P. Yu. The equivalence of generalized least squares and maximum likelihood estimates in the exponential family. *Journal of the American Statistical Association*, 71(353):169–171, 1976. doi: 10.2307/2285762.

G. G. Chowdhury. Natural language processing. *Annual Review of Information Science and Technology*, 37(1):51–89, 2003. ISSN 1550-8382. doi: 10.1002/aris.1440370103.

G. Cohen. Age and health status in a patient satisfaction survey. *Social Science & Medicine*, 42(7):1085–1093, 1996. ISSN 0277-9536. doi: 10.1016/0277-9536(95)00315-0.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246. URL `http://www.jstor.org/stable/2984875`.

J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 00905364. URL `http://www.jstor.org/stable/2699986`.

J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367 – 378, 2002. ISSN 0167-9473. doi: 10.1016/S0167-9473(01)00065-2. Nonlinear Methods and Data Mining.

C. Guihenneuc-Jouyaux, S. Richardson, and I. M. Longini. Modeling markers of disease progression by a hidden markov process: Application to characterizing cd4 cell decline. *Biometrics*, 56(3):733–741, 2000. ISSN 1541-0420. doi: 10.1111/j.0006-341X.2000.00733.x.

J. A. Hall and M. C. Dornan. Patient sociodemographic characteristics as predictors of satisfaction with medical care: A meta-analysis. *Social Science & Medicine*, 30(7):811 – 818, 1990. ISSN 0277-9536. doi: 10.1016/0277-9536(90)90205-7.

J. A. Hall, M. A. Milburn, and A. M. Epstein. A causal model of health status and satisfaction with medical care. *Medical Care*, 31(1):84–94, 1993. ISSN 00257079. URL `http://www.jstor.org/stable/3765767`.

M. R. Hassan and B. Nath. Stock market forecasting using hidden markov model: a new approach. In *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pages 192–196, Sept 2005. doi: 10.1109/ISDA.2005.85.

R. J. Hijmans, S. Phillips, J. Leathwick, and J. Elith. *dismo: Species Distribution Modeling*, 2016. URL `https://CRAN.R-project.org/package=dismo`. R package version 1.1-1.

D. U. Himmelstein, A. Wright, and S. Woolhandler. Hospital computing and the costs and quality of care: A national study. *The American Journal of Medicine*, 123(1):40 – 46, 2010. ISSN 0002-9343. doi: 10.1016/j.amjmed.2009.09.004.

S. Hinduja and J. W. Patchin. Bullying, cyberbullying, and suicide. *Archives of Suicide Research*, 14(3):206–221, 2010. doi: 10.1080/13811118.2010.494133. PMID: 20658375.

M. Hsieh and J. Kagle. Understanding patient satisfaction and dissatisfaction with health care. *Health & Social Work*, 16(4):281 – 290, 1991. ISSN 0360-7283. URL `http://search.ebscohost.com/login.aspx?direct=true&db=rzh&AN=107481272&site=ehost-live`.

P. B. Jensen, L. J. Jensen, and S. Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012. doi: 10.1038/nrg3208.

C. Karlof and D. Wagner. Hidden markov model cryptanalysis. In C. D. Walter, Ç. K. Koç, and C. Paar, editors, *Cryptographic Hardware and Embedded Systems - CHES 2003: 5th International Workshop, Cologne, Germany, September 8–10, 2003. Proceedings*, pages 17–34, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-45238-6. doi: 10.1007/978-3-540-45238-6_3.

Y. Li, S. L. Gorman, and N. Elhadad. Section classification in clinical notes using supervised hidden markov model. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 744–750, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0030-8. doi: 10.1145/1882992.1883105.

S. Linder-Pelz. Social psychological determinants of patient satisfaction: A test of five hypotheses. *Social Science & Medicine*, 16(5):583–589, 1982. ISSN 0277-9536. doi: 10.1016/0277-9536(82)90312-4.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943. ISSN 1522-9602. doi: 10.1007/BF02478259.

K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive computation and machine learning. MIT Press, 2012. ISBN 9780262018029. URL `https://books.google.no/books?id=NZP6AQAAQBAJ`.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 00359238. doi: 10.2307/2344614.

D. Nutbeam. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. *Health Promotion International*, 15(3):259–267, 2000. doi: 10.1093/heapro/15.3.259.

M. Ohlsson, C. Peterson, and M. Dictor. Using hidden markov models to characterize disease trajectories. In *Proceeding of the neural networks and expert systems in medicine and healthcare conference*, pages 324–326. Citeseer, 2001.

R. H. Perlis, D. V. Iosifescu, V. M. Castro, S. N. Murphy, V. S. Gainer, J. Minnier, T. Cai, S. Goryachev, Q. Zeng, P. J. Gallagher, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychological medicine*, 42(1): 41–50, 2012. doi: 10.1017/S0033291711000997.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989. ISSN 0018-9219. doi: 10.1109/5.18626.

M. Rahmqvist. Patient satisfaction in relation to age, health status and other background factors: a model for comparisons of care units. *International Journal for Quality in Health Care*, 13(5):385–390, 2001. doi: 10.1093/intqhc/13.5.385.

G. Ridgeway. The state of boosting. *Computing Science and Statistics*, pages 172–181, 1999.

G. Ridgeway et al. *gbm: Generalized Boosted Regression Models*, 2015. URL `https://CRAN.R-project.org/package=gbm`. R package version 2.1.1.

A. Rind, T. D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, B. Shneiderman, et al. Interactive information visualization to explore and query electronic health records. *Foundations and Trends® in Human–Computer Interaction*, 5(3):207–298, 2013. doi: 10.1561/1100000039.

B. D. Ripley. Pattern recognition via neural networks. *A volume of Oxford Graduate Lectures on Neural Networks, title to be decided. Oxford University Press.*, 1996.

R. Rosenheck, N. Wilson, and M. Meterko. Influence of patient and hospital factors on consumer satisfaction with inpatient mental health treatment. *Psychatric Services*, 48(12):1553–1561, 1997. doi: 10.1176/ps.48.12.1553.

S. M. Ross. Chapter 4 - markov chains. In S. M. Ross, editor, *Introduction to Probability Models (Tenth Edition)*, pages 191 – 290. Academic Press, Boston, tenth edition, 2010. ISBN 978-0-12-375686-2. doi: 10.1016/B978-0-12-375686-2.00009-1.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 71(2):319–392, 2009. doi: 10.1111/j.1467-9868.2008.00700.x.

H. Rue, A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren. Bayesian computing with inla: a review. *Annual Review of Statistics and Its Application*, 4:395–421, 2017. doi: 10.1146/annurev-statistics-060116-054045.

M. Schonlau et al. Boosted regression (boosting): An introductory tutorial and a stata plugin. *Stata Journal*, 5(3):330, 2005. URL `http://schonlau.net/publication/05stata_boosting.pdf`.

A. J. Schuit, A. J. M. van Loon, M. Tijhuis, and M. C. Ocké. Clustering of lifestyle risk factors in a general adult population. *Preventive medicine*, 35 (3):219–224, 2002. doi: 10.1006/pmed.2002.1064.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002. doi: 10.1111/1467-9868.00353.

Statistisk Sentralbyrå. Lønn, alle ansatte, 2016, februar 2017. URL `https://www.ssb.no/arbeid-og-lonn/statistikker/lonnansatt/aar/2017-02-01`.

P. C. Tang and C. J. McDonald. Electronic health record systems. In *Biomedical informatics*, pages 447–475. Springer, 2006. doi: 10.1007/0-387-36278-9_12.

P. L. N. Thi, S. Briançon, F. Empereur, and F. Guillemin. Factors determining inpatient satisfaction with care. *Social Science & Medicine*, 54(4): 493 – 504, 2002. ISSN 0277-9536. doi: 10.1016/S0277-9536(01)00045-4.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4`. ISBN 0-387-95457-0.

Chapter 10

Appendix

10.1 Other methods

This section will contain methods that is not of core importance, but still have been utilized or mentioned in the thesis.

10.1.1 K-fold cross validation

The k in the term k-fold cross validation refers to how many equally sized subsamples that are to be randomly created from the original data. These subsamples are in turn changing being the $k-1$ training data samples and the one validation data sample. This results in k different combinations of cross validation that can be used to make more accurate decisions about the problem at hand. For example when in a regression setting this could mean that the k different samples are used to make k models and equally many fitted values. The fitted values in combination with the true observed values may then be utilized to calculate the residual deviances.

10.1.2 Welch two sample t-test

Welch two sample t-test is as the name implies a t-test, but there are some differences. The assumption about normality is kept, but the variances of the two samples (lets call them sample one and two) are assumed to be unequal. The t statistic is calculated similarly (assuming equal means)

$$t = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{10.1}$$

The degree of freedom is then calculated by

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{s_1^4/(n_1^2\nu_1) + s_2^4/(n_2^2\nu_2)} \tag{10.2}$$

Finally now one only needs to use the t-statistic and the calculated degrees of freedom in a t-distribution to find a p-value.

10.1.3 Pearson's chi-squared test

Pearson's chi-squared test is often used on categorical data to either test for goodness of fit or for independence. It is calculated as follows

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{10.3}$$

In (10.3) O_i is the observed frequency while E_i is the expected frequency. Once χ^2 is calculated and one knows the degree of freedom it can be compared with the theoretical quantile and a p-value can be found yielding a conclusion.

10.1.4 Cramer's V

Cramer's V is used as a measure of association between nominal variables. It can produce values from 0 to 1 which could be interpreted in the same fashion one interprets the Pearson's R^2. It is calculated by

$$V = \sqrt{\frac{\chi^2/N}{(\min(r,k) - 1)}} \tag{10.4}$$

Where N is the total number of observations, while r and k is the number of categories in each respective nominal variable. Lastly χ^2 is calculated by using the Pearson's chi-squared test (section 10.1.3) and inputting the two groups.

10.1.5 Viterbi algorithm

The Viterbi algorithm is a tool that can be used with hidden Markov models. It is used to find the overall most probable sequence of hidden states of length N, $\boldsymbol{X}_N$, given a sequence of observations of equal length, $\boldsymbol{S}^N$. Also a complete hidden Markov model has to be specified to make the algorithm function. The algorithm is based on dynamic programming which does not guarantee a 100% that the most probable path is found, but it should be a decent enough estimate to consider.

$$\text{Viterbi}(\boldsymbol{S}_N) = \text{arg max}_{\boldsymbol{X}_N} P(\boldsymbol{X}_N \mid \boldsymbol{S}^N) \tag{10.5}$$

From a general perspective equation (10.5) should show how the algorithm functions in concept, but the step by step recursion and iteration often present in dynamic programming is left out.

10.2 Attachments

This section will contain a few of the many documents that have been a foundation for my work. Only the most important documents are included, i.e. those that help understand the raw interpretation of variables in the analyses from chapter 6 and 7.

10.2.1 Variable translation table

As new variable names have been constructed to ease the interpretation of the analyses, Table 10.1 will contain their corresponding translations. As such, their original names with corresponding values and descriptions in the attachments are accessible.

Original name:	New name:
PersAge2013	Age
ErMann_PID_korr	Gender
N_TOT_KONT	Services
LOS_TOT	TimeInHospital
N_TYPE	TypeServices
N_RE	Readmissions
N_DEPT	Departments
N_WARD	Wards
N_PROC	Procedures
N_DIA	DiagnosesICD
N_KAT	CategoriesICD
N_KAP	ChaptersICD
N_ICPC	DiagnosesICPC
N_ICPC_KAP	ChaptersICPC
S1	SelfRateHealth
S4A	NoLongTermSick
S11	Education
S12	Occupation
S14_1	LightExercise
S14_2	ToughExercise
S15	Smoking
S16	Alcohol
S17	NoDebased
S18	HasSupport
S19_1	Income

Table 10.1: Variable name translation table - For accessing information in the enclosed documents.

10.2.2 Part 1 of questionnaire used and available interpretations

NTNU
Kunnskap for en bedre verden

NST
Nasjonalt senter for
samhandling og telemedisin

Spørreskjema om helsetjenester du har mottatt i 2012 og 2013

Fordi vi skal spørre om helsetjenester du har mottatt, får du tilsendt to ting
1) Et gult ark med en oversikt over dine helsetjenester i 2012-2013 (tjenester for psykisk helse ol er ikke med)
2) Dette spørreskjema som vi vil be deg svare på og sende tilbake i vedlagte svarkonvolutt (portoen er betalt)

Dersom du ikke har svart innen 2 uker, vil du få en påminnelse i posten.

1. Stort sett, vil du si at din helse er?

Utmerket .... □
Meget god .... □
God .... □
Nokså god .... □
Dårlig .... □

2. Hvilke utsagn passer best på din helsetilstand i dag?

Å gå *(sett ett kryss)*
Jeg har ingen problemer med å gå omkring .... □
Jeg har litt problemer med å gå omkring .... □
Jeg er sengeliggende .... □

Å stelle meg selv *(sett ett kryss)*
Jeg har ingen problemer med personlig stell .... □
Jeg har litt problemer med å vaske meg eller kle meg □
Jeg er ute av stand til å vaske meg eller kle meg.......... □

Vanlige gjøremål *(for eksempel arbeid, studier, husarbeid, familie- eller fritidsaktiviteter) (sett ett kryss)*
Jeg har ingen problemer med å utføre mine vanlige gjøremål .... □
Jeg har litt problemer med å utføre mine vanlige gjøremål .... □
Jeg er ute av stand til å utføre mine vanlige gjøremål. □

Smerte / ubehag *(sett ett kryss)*
Jeg har verken smerter eller ubehag .... □
Jeg har moderat smerte eller ubehag .... □
Jeg har sterk smerte eller ubehag.......... □

Angst / depresjon *(sett ett kryss)*
Jeg er verken engstelig eller deprimert .... □
Jeg er noe engstelig eller deprimert .... □
Jeg er svært engstelig eller deprimert .... □

3. Alt i alt – hvordan er din evne til å ta vare på din egen helse nå?

Utmerket .... □
Meget god .... □
God .... □
Nokså god .... □
Dårlig .... □

4a. Har du noen langvarig sykdom, skade eller lidelse av fysisk eller psykisk art som gjør det vanskelig for deg å fungere i dagliglivet (Med langvarig menes at det har vart, eller vil vare i minst 1 år)? *(sett ett kryss)*

Ja -> Svar på spørsmål 4b og 4c .... □
Nei -> Gå til neste spørsmål (nr 5) .... □

4b. Hvor lenge har du levd med tilstand(er) som gjør det vanskelig for deg å fungere i ditt daglige liv? *(sett ett kryss)*

Hele livet / siden barndommen .... □
I mer enn ett år .... □
Mindre enn ett år .... □
Usikker / vet ikke .... □

4c. Hvordan tror du selv du vil fungere i dagliglivet om ett år fra nå? *(sett ett kryss)*

Frisk, tilbake til normal funksjon .... □
Bedring av funksjon, men ikke frisk.......... □
Uendret funksjon .... □
Usikker / vet ikke .... □

5. Når det skal gjøres valg om den behandlingen du skal ha, hva foretrekker du? *(sett ett kryss)*

Foretrekker å ta beslutningen selv.......... □
Foretrekker at legen tar beslutningen .... □
Foretrekker at legen og jeg tar beslutningen sammen.......... □

35478

6. Under står noen utsagn som folk av og til bruker når de snakker om helsen sin. **Marker i hvor stor grad du er enig eller uenig med hvert utsagn ved å sette et kryss ved det svaret du mener passer for deg**.

Svaret ditt skal være det du mener og ikke hva du tror legen eller andre ønsker at du skal svare. Hvis utsagnet ikke gjelder for deg kan du krysse av for "Ikke aktuelt" *(sett ett kryss for hver linje).*

	Helt uenig	Nokså uenig	Nokså enig	Helt enig	Ikke aktuelt
Når alt kommer til alt er jeg selv ansvarlig for å ta hånd om min egen helse	□	□	□	□	□
Det aller viktigste for min egen helse og funksjonsevne er at jeg tar aktiv del i behandlingen	□	□	□	□	□
Jeg er sikker på at jeg kan gjøre det som er nødvendig for å forebygge eller redusere symptomer eller problemer som skyldes min helsetilstand	□	□	□	□	□
Jeg vet hvordan de forskjellige medisinene jeg har fått foreskrevet skal virke	□	□	□	□	□
Jeg vet når jeg trenger medisinsk hjelp for et helseproblem og når jeg kan ta hånd om det selv	□	□	□	□	□
Jeg er trygg nok til å kunne ta opp det jeg ønsker, selv om helsepersonell ikke spør	□	□	□	□	□
Jeg er sikker på at jeg kan gjennomføre den foreskrevne medisinske behandlingen hjemme	□	□	□	□	□
Jeg forstår både hva helseproblemene mine dreier seg om og årsaken til dem	□	□	□	□	□
Jeg vet om de ulike behandlingsmuligheter for min helsetilstand	□	□	□	□	□
Jeg har opprettholdt de endringer i livsstil som jeg har gjort for helsens skyld	□	□	□	□	□
Jeg vet hvordan jeg skal forebygge forverring av min helsetilstand	□	□	□	□	□
Jeg kan finne løsninger når det oppstår nye situasjoner eller problemer med min helsetilstand	□	□	□	□	□

7. Tenk på alle helsetjenestene du har hatt kontakt med i 2012 og 2013. Hvor enig eller uenig er du i følgende utsagn?

(Sett ett kryss for hver linje)

	Helt uenig	Nokså uenig	Hverken eller	Nokså enig	Helt enig	Ikke aktuelt
Jeg er alt i alt **godt fornøyd** med de(n) helsetjenesten(e) jeg har mottatt	□	□	□	□	□	□
Helsearbeiderne fra forskjellige tjenester har **samarbeidet** godt med hverandre	□	□	□	□	□	□
Helsearbeiderne jeg har møtt har vært opptatt av å hjelpe meg med det som er **viktigst** for meg	□	□	□	□	□	□
Helsearbeiderne jeg har møtt har IKKE vært godt **informert** om min helse	□	□	□	□	□	□
Helsearbeiderne jeg har møtt har **forberedt meg** godt på det som skulle skje videre med meg	□	□	□	□	□	□
Helsearbeiderne jeg har møtt har IKKE kjent til mine **behov og verdier**	□	□	□	□	□	□
Helsearbeiderne jeg har møtt har gitt meg **motstridende råd**	□	□	□	□	□	□
Helsearbeiderne jeg har møtt har gitt meg **riktig behandling**	□	□	□	□	□	□

8. Vil du si at din funksjon er ikke, litt, middels eller mye nedsatt? *(sett ett kryss for hver linje)*.

	Ikke	Litt	Middels	Mye
Er bevegelseshemmet	□	□	□	□
Har nedsatt syn	□	□	□	□
Har nedsatt hørsel	□	□	□	□
Hemmet pga. kroppslig sykdom	□	□	□	□
Hemmet pga psykiske plager	□	□	□	□

Se det gule arket - DEL A Dine helsetjenester

9. Er framstillingen av tjenester på det gule arket riktig?
Tjenester innenfor psykisk helse og rus er ikke tatt med og mangler derfor for alle *(Sett ett eller flere kryss)*

Ja. Den ser riktig ut.. □
Delvis rett, men noen tjenester mangler.. □
Delvis rett, men noen tjenester har jeg IKKE har brukt.. □
Nei – alt er feil.. □
Vet ikke – husker ikke.. □

Skriv inn Tjenestebokstav fra det gule arket

Hvis feil, skriv inn Tjenestebokstav for de tjenestetypene du IKKE har mottatt □ □ □ □ □

....

10. Hvilke tjenester hjalp deg med følgende utfordringer?

Skriv inn bokstaven for den eller de tjenestene som best har lagt til rette for:	*Skriv inn tjeneste-bokstav fra det gule arket*					Ingen av tjenestene
Eksempel på utfylling: at du skulle forstå dine helseproblem(er)............................	□	□	□	□	□	□
at du skulle **forstå** dine helseprobleme(er)..	□	□	□	□	□	□
å **lytte** til hva som betyr mest for deg når det gjelder dine helseproblem(er)..	□	□	□	□	□	□
at du selv kunne fortelle om problemer eller bivirkninger av **medisinene** dine..	□	□	□	□	□	□
å informere deg om hva som var **planlagt** at skulle skje videre med dine helseproblem(er)..	□	□	□	□	□	□
å ta hensyn til hva som var **viktigst for deg**, når det ble **bestemt hva som skulle gjøres videre** med dine helseproblem(er)..	□	□	□	□	□	□
å gi deg **råd** for hvordan du kunne ta vare på din **egen helse**..............................	□	□	□	□	□	□
å foreslå at du skulle delta på **kurs** eller i **grupper** som er relevant for helsa di..	□	□	□	□	□	□
å sette mål sammen med deg for **kosthold** og/eller **mosjon**..............................	□	□	□	□	□	□
å hjelpe dine nærmeste **pårørende** med problemer knyttet til dine helseproblem(er)..	□	□	□	□	□	□
å hjelpe med å løse **økonomiske** følger av dine helseproblem(er)......................	□	□	□	□	□	□

NTNU
Kunnskap for en bedre verden

NST
Nasjonalt senter for
samhandling og telemedisin

11. Hvilken utdanning er den høyeste du har fullført? *(sett ett kryss)*

Grunnskole, framhaldsskole, folkehøgskole .... □

Realskole, middelskole, yrkesskole .... □

Artium, økonomisk gymnas, allmennfaglig retning i videregående skole .... □

Høgskole/universitet, mindre enn 4 år.................... □

Høgskole/universitet, 4 år eller mer .... □

12. Hvilken av følgende er mest dekkende for din hovedaktivitet? *(sett ett kryss)*

Ansatt eller selvstendig næringsdrivende .... □

Alderspensjonist .... □

Annen pensjon / trygd.................... □

Hjemmeværende .... □

Student .... □

Arbeidssøker .... □

13. Hvem bor du sammen med? *(sett ett eller flere kryss)*

Bor alene.................... □

Partner eller ektefelle .... □

Andre under 18 år .... □

Andre over 18 år.................... □

14. Hvordan har din fysiske aktivitet i FRITIDEN vært det siste året? *(Tenk deg et gjennomsnitt for året. Arbeidsvei regnes som fritid)*

Lett aktivitet (ikke svett/andpusten)? *(sett ett kryss)*

Ingen .... □

Under 1time pr uke.................... □

1–2 timer per uke.................... □

3 timer eller mer pr uke .... □

Hard fysisk aktivitet (svett/andpusten? *(sett ett kryss)*

Ingen .... □

Under 1t pr uke.................... □

1–2 timer per uke.................... □

3 timer eller mer pr uke .... □

15. Røyker du? *(sett ett kryss)*

Nei, jeg har aldri røykt.................... □

Nei, jeg har sluttet å røyke .... □

Ja, sigaretter av og til (fest/ferie, ikke daglig) .... □

Ja, sigarer/sigarillos/pipe av og til .... □

Ja, sigaretter daglig .... □

Ja, sigarer/sigarillos/pipe daglig .... □

16. Hvor ofte drikker du 5 glass eller mer av øl, vin eller brennevin ved samme anledning? *(sett ett kryss)*

Daglig .... □

Ukentlig.................... □

Månedlig .... □

Sjeldnere .... □

Aldri.................... □

17. Har du noen gang i livet opplevd at noen over lengre tid har forsøkt å kue, fornedre eller ydmyke deg? *(sett ett kryss)*

Ja.................... □

Nei .... □

18. Har du nok familie/ venner som kan gi deg hjelp når du trenger det? *(sett ett kryss)*

Ja.................... □

Nei .... □

19. Hva er husstandens samlede brutto årsinntekt? *(sett ett kryss)*

Under 250.000,- .... □

250.000,- til 500.000,- .... □

500.000,- til 750.000,- .... □

750.000,- til 1 million .... □

Over 1 million.................... □

Hvem fyller ut spørreskjemaet? *(Sett ett kryss)*

Pasienten selv.................... □

Noen andre på vegne av pasienten (Pårørende, verge el) .... □

Hoved (første spørreskjema) - DEL 1

Variabelnavn	Spørsmål	Verdier / svaralternativ
NR	Stedskode	
UTFYLT	Hvem fyller ut spørreskjemaet?	1= Pasienten selv 2= Noen andre på vegne av pasienten (Pårørende, verge el)
S3	Alt i alt – hvordan er din evne til å ta vare på din egen helse nå?	1= Utmerket 2= Meget god 3= God 4= Nokså god 5= Dårlig
S1	Stort sett, vil du si at din helse er?	1= Utmerket 2= Meget god 3= God 4= Nokså god 5= Dårlig
S2_1	Hvilke utsagn passer best på din helsetilstand i dag? Å gå	1= Jeg har ingen problemer med å gå omkring 2= Jeg har litt problemer med å gå omkring 3= Jeg er sengeliggende
S4A	Har du noen langvarig sykdom, skade eller lidelse av fysisk eller psykisk art som gjør det vanskelig for deg å fungere i dagliglivet (Med langvarig menes at det har vart, eller vil vare i minst 1 år)?	1= Ja -> Svar på spørsmål 4b og 4c 2= Nei -> Gå til neste spørsmål (nr 5)
S2_2	Hvilke utsagn passer best på din helsetilstand i dag? Å stelle meg selv	1= Jeg har ingen problemer med personlig stell 2= Jeg har litt problemer med å vaske meg eller kle meg 3= Jeg er ute av stand til å vaske meg eller kle meg
S4B	Hvor lenge har du levd med tilstand(er) som gjør det vanskelig for deg å fungere i ditt daglige liv?	1= Hele livet / siden barndommen 2= I mer enn ett år 3= Mindre enn ett år 4= Usikker / vet ikke
S2_3	Hvilke utsagn passer best på din helsetilstand i dag? Vanlige gjøremål (for eksempel arbeid, studier, husarbeid, familie- eller fritidsaktiviteter)	1= Jeg har ingen problemer med å utføre mine vanlige gjøremål 2= Jeg har litt problemer med å utføre mine vanlige gjøremål 3= Jeg er ute av stand til å utføre mine vanlige gjøremål
S4C	Hvordan tror du selv du vil fungere i dagliglivet om ett år fra nå?	1= Frisk, tilbake til normal funksjon 2= Bedring av funksjon, men ikke frisk

		3= Uendret funksjon 4= Dårligere funksjon 5= Usikker / vet ikke
S2_4	Hvilke utsagn passer best på din helsetilstand i dag? Smerte / ubehag	1= Jeg har verken smerter eller ubehag 2= Jeg har moderat smerte eller ubehag 3= Jeg har sterk smerte eller ubehag
S5	Når det skal gjøres valg om den behandlingen du skal ha, hva foretrekker du?	1= Foretrekker å ta beslutningen selv 2= Foretrekker at legen tar beslutningen 3= Foretrekker at legen og jeg tar beslutningen sammen
S2_5	Hvilke utsagn passer best på din helsetilstand i dag? Angst / depresjon	1= Jeg er verken engstelig eller deprimert 2= Jeg er noe engstelig eller deprimert 3= Jeg er svært engstelig eller deprimert
S6_1	Når alt kommer til alt er jeg selv ansvarlig for å ta hånd om min egen helse	1= Helt uenig 2= Nokså uenig 3= Nokså enig 4= Helt enig 5= Ikke aktuelt
S6_2	Det aller viktigste for min egen helse og funksjonsevne er at jeg tar aktiv del i behandlingen	1= Helt uenig 2= Nokså uenig 3= Nokså enig 4= Helt enig 5= Ikke aktuelt
S6_3	Jeg er sikker på at jeg kan gjøre det som er nødvendig for å forebygge eller redusere symptomer eller problemer som skyldes min helsetilstand	1= Helt uenig 2= Nokså uenig 3= Nokså enig 4= Helt enig 5= Ikke aktuelt
S6_4	Jeg vet hvordan de forskjellige medisinene jeg har fått foreskrevet skal virke	1= Helt uenig 2= Nokså uenig 3= Nokså enig 4= Helt enig 5= Ikke aktuelt
S6_5	Jeg vet når jeg trenger medisinsk hjelp for et helseproblem og når jeg kan ta hånd om det selv	1= Helt uenig 2= Nokså uenig 3= Nokså enig 4= Helt enig 5= Ikke aktuelt
S6_6	Jeg er trygg nok til å kunne ta opp det jeg ønsker, selv om helsepersonell ikke spør	1= Helt uenig 2= Nokså uenig 3= Nokså enig 4= Helt enig 5= Ikke aktuelt

S6_7	Jeg er sikker på at jeg kan gjennomføre den foreskrevne medisinske behandlingen hjemme	1= Helt uenig 2= Nokså uenig 3= Nokså enig 4= Helt enig 5= Ikke aktuelt
S6_8	Jeg forstår både hva helseproblemene mine dreier seg om og årsaken til dem	1= Helt uenig 2= Nokså uenig 3= Nokså enig 4= Helt enig 5= Ikke aktuelt
S6_9	Jeg vet om de ulike behandlingsmuligheter for min helsetilstand	1= Helt uenig 2= Nokså uenig 3= Nokså enig 4= Helt enig 5= Ikke aktuelt
S6_10	Jeg har opprettholdt de endringer i livsstil som jeg har gjort for helsens skyld	1= Helt uenig 2= Nokså uenig 3= Nokså enig 4= Helt enig 5= Ikke aktuelt
S6_11	Jeg vet hvordan jeg skal forebygge forverring av min helsetilstand	1= Helt uenig 2= Nokså uenig 3= Nokså enig 4= Helt enig 5= Ikke aktuelt
S6_12	Jeg kan finne løsninger når det oppstår nye situasjoner eller problemer med min helsetilstand	1= Helt uenig 2= Nokså uenig 3= Nokså enig 4= Helt enig 5= Ikke aktuelt
S7_1	Jeg er alt i alt godt fornøyd med de(n) helsetjenesten(e) jeg har mottatt	1= Helt uenig 2= Nokså uenig 3= Hverken eller 4= Nokså enig 5= Helt enig 6= Ikke aktuelt
S7_2	Helsearbeiderne fra forskjellige tjenester har samarbeidet godt med hverandre	1= Helt uenig 2= Nokså uenig 3= Hverken eller 4= Nokså enig 5= Helt enig 6= Ikke aktuelt
S7_3	Helsearbeiderne jeg har møtt har vært opptatt av å hjelpe meg med det som er viktigst for meg	1= Helt uenig 2= Nokså uenig 3= Hverken eller 4= Nokså enig 5= Helt enig 6= Ikke aktuelt
S7_4	Helsearbeiderne jeg har møtt har IKKE vært godt informert om min helse	1= Helt uenig 2= Nokså uenig 3= Hverken eller

		4= Nokså enig 5= Helt enig 6= Ikke aktuelt
S7_5	Helsearbeiderne jeg har møtt har forberedt meg godt på det som skulle skje videre med meg	1= Helt uenig 2= Nokså uenig 3= Hverken eller 4= Nokså enig 5= Helt enig 6= Ikke aktuelt
S7_6	Helsearbeiderne jeg har møtt har IKKE kjent til mine behov og verdier	1= Helt uenig 2= Nokså uenig 3= Hverken eller 4= Nokså enig 5= Helt enig 6= Ikke aktuelt
S7_7	Helsearbeiderne jeg har møtt har gitt meg motstridende råd	1= Helt uenig 2= Nokså uenig 3= Hverken eller 4= Nokså enig 5= Helt enig 6= Ikke aktuelt
S7_8	Helsearbeiderne jeg har møtt har gitt meg riktig behandling	1= Helt uenig 2= Nokså uenig 3= Hverken eller 4= Nokså enig 5= Helt enig 6= Ikke aktuelt
S8_1	Er bevegelseshemmet	1= Ikke 2= Litt 3= Middels 4= Mye
S8_2	Har nedsatt syn	1= Ikke 2= Litt 3= Middels 4= Mye
S8_3	Har nedsatt hørsel	1= Ikke 2= Litt 3= Middels 4= Mye
S8_4	Hemmet pga. kroppslig sykdom	1= Ikke 2= Litt 3= Middels 4= Mye
S8_5	Hemmet pga psykiske plager	1= Ikke 2= Litt 3= Middels 4= Mye
S15	Røyker du?	1= Nei, jeg har aldri røykt 2= Nei, jeg har sluttet å røyke 3= Ja, sigaretter av og til (fest/ferie, ikke daglig)

		4= Ja, sigarer/sigarillos/pipe av og til 5= Ja, sigaretter daglig 6= Ja, sigarer/sigarillos/pipe daglig
S11	Hvilken utdanning er den høyeste du har fullført?	1= Grunnskole, framhaldsskole, folkehøgskole 2= Realskole, middelskole, yrkesskole 3= Artium, økonomisk gymnas, allmennfaglig retning i videregående skole 4= Høgskole/universitet, mindre enn 4 år 5= Høgskole/universitet, 4 år eller mer
S16	Hvor ofte drikker du 5 glass eller mer av øl, vin eller brennevin ved samme anledning?	1= Daglig 2= Ukentlig 3= Månedlig 4= Sjeldnere 5= Aldri
S12	Hvilken av følgende er mest dekkende for din hovedaktivitet?	1= Ansatt eller selvstendig næringsdrivende 2= Alderspensjonist 3= Annen pensjon / trygd 4= Hjemmeværende 5= Student 6= Arbeidssøker
S17	Har du noen gang i livet opplevd at noen over lengre tid har forsøkt å kue, fornedre eller ydmyke deg?	1= Ja 2= Nei

S13	Hvem bor du sammen med?	alene= Bor alene
		partner= Partner eller ektefelle
		under18= Andre under 18
		over18= Andre over 18

ALENE	Hvem bor du sammen med?	0= ikke avkrysset 1= Bor alene
PARTNER	Hvem bor du sammen med?	0= ikke avkrysset 1= Partner eller ektefelle
UNDER18	Hvem bor du sammen med?	0= ikke avkrysset 1= Andre under 18
OVER18	Hvem bor du sammen med?	0= ikke avkrysset 1= Andre over 18
S18	Har du nok familie/ venner som kan gi deg hjelp når du trenger det?	1= Ja 2= Nei
S19_1	Hva er husstandens samlede brutto årsinntekt?	1= Under 250.000,- 2= 250.000,- til 500.000,- 3= 500.000,- til 750.000,- 4= 750.000,- til 1 million

		5= Over 1 million
S14_1	Hvordan har din fysiske aktivitet i FRITIDEN vært det siste året? (Tenk deg et gjennomsnitt for året. Arbeidsvei regnes som fritid) Lett aktivitet (ikke svett/andpusten)?	1= Ingen 2= Under 1time pr uke 3= 1–2 timer per uke 4= 3 timer eller mer pr uke
S14_2	Hvordan har din fysiske aktivitet i FRITIDEN vært det siste året? Hard fysisk aktivitet (svett/andpusten?	1= Ingen 2= Under 1t pr uke 3= 1–2 timer per uke 4= 3 timer eller mer pr uke

10.2.3 Raw aggregated data variable descriptions

COUNT.11	Antall øhj innleggelser	Hastegrad=1(øhj) Omsorgsnivå=1(innl), Antall basert på telling av aggregerte sykehusopphold.
		Sykehusopphold består av overlappende episoder hvor minst én av episodene er en innleggelse.
		Siden rapporteringsformat for data baserer seg på utskrivelsesdato i 2012/13 vil det telles med noen innleggelser fra 2011 som flyter over i 2012 og man mister innleggelser fra 2013 hvor utskrivelsesdato var i 2014. Disse antas å veie opp for hverandre.
			SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv)
				Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol)
				Inn=IPAdmitDateTime
				Ut=IPDischDateTime

COUNT.12	Antall øhj dagbehandling	Hastegrad=1(øhj) Omsorgsnivå=2(dag), Antall basert på telling av aggregerte sykehusopphold.
		Sykehusopphold består av overlappende episoder hvor minst én av episodene er en innleggelse.
		Siden rapporteringsformat for data baserer seg på utskrivelsesdato i 2012/13 vil det telles med noen innleggelser fra 2011 som flyter over i 2012 og man mister innleggelser fra 2013 hvor utskrivelsesdato var i 2014. Disse antas å veie opp for hverandre.
			SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv)
				Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol)
				Inn=IPAdmitDateTime
				Ut=IPDischDateTime

COUNT.13	Antall øhj poliklinikk	Hastegrad=1(øhj) Omsorgsnivå=3(pol), Antall basert på telling av aggregerte sykehusopphold.
		Sykehusopphold består av overlappende episoder hvor minst én av episodene er en innleggelse.
		Siden rapporteringsformat for data baserer seg på utskrivelsesdato i 2012/13 vil det telles med noen innleggelser fra 2011 som flyter over i 2012 og man mister innleggelser fra 2013 hvor utskrivelsesdato var i 2014. Disse antas å veie opp for hverandre.
			SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv)
				Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol)
				Inn=IPAdmitDateTime
				Ut=IPDischDateTime

COUNT.21	Antall elektive innleggelser	Hastegrad=2(el) Omsorgsnivå=1(innl), Antall basert på telling av aggregerte sykehusopphold. Sykehusopphold består av overlappende episoder hvor minst én av episodene er en innleggelse. Siden rapporteringsformat for data baserer seg på utskrivelsesdato i 2012/13 vil det telles med noen	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv) Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh.,

		innleggelser fra 2011 som flyter over i 2012 og man mister innleggelser fra 2013 hvor utskrivelsesdato var i 2014. Disse antas å veie opp for hverandre.		2=pol)
				Inn=IPAdmitDateTime
				Ut=IPDischDateTime

COUNT.22	Antall elektiv dagbehandling	Hastegrad=2(el) Omsorgsnivå=2(dag), Antall basert på telling av aggregerte sykehusopphold. Sykehusopphold består av overlappende episoder hvor minst én av episodene er en innleggelse. Siden rapporteringsformat for data baserer seg på utskrivelsesdato i 2012/13 vil det telles med noen innleggelser fra 2011 som flyter over i 2012 og man mister innleggelser fra 2013 hvor utskrivelsesdato var i 2014. Disse antas å veie opp for hverandre.	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv) Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol) Inn=IPAdmitDateTime Ut=IPDischDateTime
COUNT.23	Antall elektiv poliklinikk	Hastegrad=2(el) Omsorgsnivå=3(pol), Antall basert på telling av aggregerte sykehusopphold. Sykehusopphold består av overlappende episoder hvor minst én av episodene er en innleggelse. Siden rapporteringsformat for data baserer seg på utskrivelsesdato i 2012/13 vil det telles med noen innleggelser fra 2011 som flyter over i 2012 og man mister innleggelser fra 2013 hvor utskrivelsesdato var i 2014. Disse antas å veie opp for hverandre.	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv) Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol) Inn=IPAdmitDateTime Ut=IPDischDateTime
N_ØHJ_KONT	Antall ø-hjelpskontakter	Sum hastegrad=1, Antall basert på telling av aggregerte sykehusopphold. Sykehusopphold består av overlappende episoder hvor minst én av episodene er en innleggelse. Siden rapporteringsformat for data baserer seg på utskrivelsesdato i 2012/13 vil det telles med noen innleggelser fra 2011 som flyter over i 2012 og man mister innleggelser fra 2013 hvor utskrivelsesdato var i 2014. Disse antas å veie opp for hverandre.	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv) Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol) Inn=IPAdmitDateTime Ut=IPDischDateTime
N_EL_KONT	Antall elektive kontakter	Sum hastegrad=2, Antall basert på telling av aggregerte sykehusopphold. Sykehusopphold består av overlappende episoder hvor	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv)

		minst én av episodene er en innleggelse. Siden rapporteringsformat for data baserer seg på utskrivelsesdato i 2012/13 vil det telles med noen innleggelser fra 2011 som flyter over i 2012 og man mister innleggelser fra 2013 hvor utskrivelsesdato var i 2014. Disse antas å veie opp for hverandre.		Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol) Inn=IPAdmitDateTime Ut=IPDischDateTime
N_TOT_KONT	Totalt antall kontakter	Sum alle kontakter, Antall basert på telling av aggregerte sykehusopphold. Sykehusopphold består av overlappende episoder hvor minst én av episodene er en innleggelse. Siden rapporteringsformat for data baserer seg på utskrivelsesdato i 2012/13 vil det telles med noen innleggelser fra 2011 som flyter over i 2012 og man mister innleggelser fra 2013 hvor utskrivelsesdato var i 2014. Disse antas å veie opp for hverandre.	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv) Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol) Inn=IPAdmitDateTime Ut=IPDischDateTime
N_TYPE	Antall ulike kontakttyper	Antallet ulike kontakttyper (COUNT.11/COUNT.12/…) registrert per PID. Maks 6 forskjellige kontakttyper. Antall basert på telling av aggregerte sykehusopphold. Sykehusopphold består av overlappende episoder hvor minst én av episodene er en innleggelse. Siden rapporteringsformat for data baserer seg på utskrivelsesdato i 2012/13 vil det telles med noen innleggelser fra 2011 som flyter over i 2012 og man mister innleggelser fra 2013 hvor utskrivelsesdato var i 2014. Disse antas å veie opp for hverandre.	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv) Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol) Inn=IPAdmitDateTime Ut=IPDischDateTime
N_RE	Antall re-innleggelser	Øhj innleggelse innen 30 dager etter foregående innleggelse. Fordi data kun gir informasjon om pasienter skrevet ut i 2012/2013 vil øhj innleggelser før 31. januar 2012, som ikke har en tidligere innleggelse med utskrivelse i 2012 ikke med sikkerhet kunne avskrives som re-innleggelser, selv om de ikke blir flagget i eksisterende data. Øhj innleggelser som skulle vært re-innleggelser basert på tidligere innleggelser med utskrivelsesdato i desember 2011 blir ikke funnet.	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv) Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol)

		F. eks. hvis utskrevet 5/1-12 og akuttinnlagt igjen 20/1-12 blir det telt som reinnleggelse. Men hvis var utskrevet 28/12-11 og akuttinnlagt 20/1-12 telles det ikke med. Betyr at antall reinnleggelser i januar 2012 blir lavere.		Inn=IPAdmitDateTime
				Ut=IPDischDateTime

LOS_SUM.11	Sum liggetid øhj innleggelser	Liggetid i desimaldøgn. Dvs at 4,2 døgn = 4 døgn og 4 timer og 48 minutter (0,2x24 timer= 4,8 timer) Ett døgn=86400 sekunder, som er grunnlag for alle dato-/tidsberegninger i datasettet. For innleggelser med startdato i 2011 er liggetid telt f.o.m. 01.01.2012. Innleggelser med startdato i 2013 og sluttdato i 2014 er ikke telt med da disse telles i 2014 rapport (basert på utskrivelsesdato). Siden liggetid for innleggelser som overlapper fra 2011 telles fra 1/1-12 vil det potensielt kunne være noe underestimering av total liggetid.	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv) Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol) Inn=IPAdmitDateTime Ut=IPDischDateTime
LOS_SUM.21	Sum liggetid elektive innleggelser	Liggetid i desimaldøgn. Dvs at 4,2 døgn = 4 døgn og 4 timer og 48 minutter (0,2x24 timer= 4,8 timer) Ett døgn=86400 sekunder, som er grunnlag for alle dato-/tidsberegninger i datasettet. For innleggelser med startdato i 2011 er liggetid telt f.o.m. 01.01.2012. Innleggelser med startdato i 2013 og sluttdato i 2014 er ikke telt med da disse telles i 2014 rapport (basert på utskrivelsesdato). Siden liggetid for innleggelser som overlapper fra 2011 telles fra 1/1-12 vil det potensielt kunne være noe underestimering av total liggetid.	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv) Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol) Inn=IPAdmitDateTime Ut=IPDischDateTime
LOS_TOT	Total liggetid innleggelser	Liggetid i desimaldøgn. Dvs at 4,2 døgn = 4 døgn og 4 timer og 48 minutter (0,2x24 timer= 4,8 timer) Ett døgn=86400 sekunder, som er grunnlag for alle dato-/tidsberegninger i datasettet. For innleggelser med startdato i 2011 er liggetid telt f.o.m. 01.01.2012. Innleggelser med startdato i 2013 og sluttdato i 2014 er ikke	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv) Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol)

		telt med da disse telles i 2014 rapport (basert på utskrivelsesdato) Siden liggetid for innleggelser som overlapper fra 2011 telles fra 1/1-12 vil det potensielt kunne være noe underestimering av total liggetid.		Inn=IPAdmitDateTime
				Ut=IPDischDateTime

N_DS.11	Antall døgnskiller øhj	Antall inneliggende døgnskiller -> at en innleggelse går over kl 24.00 (kl 23.59-00.01). Sammenlignet med length of stay kan det bli noen "rare" tall. Ti unike innleggelser fra 23:59-00:01 innen to døgn vil telle som 10 døgnskiller, men kun 0,14 liggedøgn. Motsatt vil ti unike innleggelser fra 00:01-23:59 samme døgn telle som tilnærmet 10 liggedøgn, men 0 døgnskiller For innleggelser med startdato i 2011 er døgnskiller telt f.o.m. 01.01.2012	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv) Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol) Inn=IPAdmitDateTime Ut=IPDischDateTime
N_DS.21	Antall døgnskiller elektiv	Antall inneliggende døgnskiller -> at en innleggelse går over kl 24.00 (kl 23.59-00.01). Sammenlignet med length of stay kan det bli noen "rare" tall. Ti unike innleggelser fra 23:59-00:01 innen to døgn vil telle som 10 døgnskiller, men kun 0,14 liggedøgn. Motsatt vil ti unike innleggelser fra 00:01-23:59 samme døgn telle som tilnærmet 10 liggedøgn, men 0 døgnskiller For innleggelser med startdato i 2011 er døgnskiller telt f.o.m. 01.01.2012	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv) Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol) Inn=IPAdmitDateTime Ut=IPDischDateTime
DS_TOT	Antall døgnskiller totalt	Antall inneliggende døgnskiller -> at en innleggelse går over kl 24.00 (kl 23.59-00.01). Sammenlignet med length of stay kan det bli noen "rare" tall. Ti unike innleggelser fra 23:59-00:01 innen to døgn vil telle som 10 døgnskiller, men kun 0,14 liggedøgn. Motsatt vil ti unike innleggelser fra 00:01-23:59 samme døgn telle som tilnærmet 10 liggedøgn, men 0 døgnskiller	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hastegrad: NprEmergencyLevel (1-3=øhj,4=elektiv) Omsorgsnivå: NprCarecat (1=innl, 2=dagbeh., 2=pol) Inn=IPAdmitDateTime Ut=IPDischDateTime

		For innleggelser med startdato i 2011 er døgnskiller telt f.o.m. 01.01.2012		
N_DEPT	Antall ulike departments/klinikker	Antall forskjellige avdelinger (overordnet) pasient har hatt kontakt med. En avdeling består av flere enheter/units Values tilsvarer "Avd.nr." i St. Olavs "places" (pastasnavn) GYLDIGE ENHETER I PAS" fra St. Olavs "	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	ContactDept
N_WARD	Antall ulike wards/avdeling/post	Antall forskjellige enheter/units (underordnet) pasient har hatt kontakt med. Dette er laveste organisasjonsnivå vi har i data. (underordnet) Values tilsvarer enhetsnr/unitnr i PAsTAs St. Olavs places og GYLDIGE ENHETER I PAS fra St. Olavs	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	StdDeptId
N_DIA	Antall unike diagnosekoder (ICD-10)	Antall unike ICD-10 diagnosekoder (dvs hvis fått samme diagnose flere ganger ved ulike besøk teller det kun som en). Både hoved- og bidiagnoser på fullt format (finnes rundt 19000 koder i ICD-10 kodeverket)	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hoveddiagnose:PDX Bidiagnoser: SDX1-SDX38
N_DIA3T	Antall unike diagnosekoder forkortet til 3 tegn (ICD-10)	Antall unike ICD-10 diagnosekoder. Både hoved- og bidiagnoser forkortet til 3 tegn (I25.0/I25.1/… => I25)	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hoveddiagnose:PDX Bidiagnoser: SDX1-SDX38
N_PROC	Antall unike prosedyrekoder (NCMP/NCSP)	Antall unike/forskjellige prosedyrekoder og slått sammen NCMP og NCSP, begge på fullt format	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	NCSP: Proc1-Proc30 NCMP: Mproc1-Mproc10
N_KAP	Antall unike diagnosekapitler (ICD-10)	Antall unike ICD-10 diagnosekapitler pasient har. Diagnosekoder fra totalt 22 kapitler I-XXII	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hoveddiagnose:PDX Bidiagnoser: SDX1-SDX38
N_KAT	Antall unike diagnosekategorier (ICD-10)	Antall unike ICD-10 diagnosekategorier pasient har. Diagnosekoder fra totalt 236 kategorier	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	Hoveddiagnose:PDX Bidiagnoser: SDX1-SDX38
SUM_DRG_KORRIGERT	Sum av DRG korrigert vekt per PID for alle episoder med utdato 2012/13	Summert korrigert DRG-vekt for alle hendelser på sykehus (StO) uavhengig av hastegrad og omsorgsnivå	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	NirvacoWeight

SUM_DRG_GRUNNVEKT	Sum av DRG grunnvekt per PID for alle episoder med utdato 2012/13	Summert DRG-grunnvekt for alle hendelser på sykehus (StO) uavhengig av hastegrad og omsorgsnivå	SAS26jun15/stolav_ip_op_avdopph.sas7bdat	NirvacoWeightBasic
N_ICPC	Antall unike diagnosekoder (ICPC)	Antall unike ICPC diagnosekoder. Prosesskoder x30-x69 er ekskludert, kun koder på format x00-x29 og x70-x99 er tatt med i beregningen.	SAS26jun15/kuhr_helfo.sas7bdat	Diagn_ICPC
N_ICPC_KAP	Antall unike diagnosekapitler (ICPC)	Antall unike ICPC diagnosekapitler pasient har diagnosekoder fra. Prosesskoder x30-x69 er ekskludert, kun koder på format x00-x29 og x70-x99 er tatt med i beregningen. Totalt 17 kapitler	SAS26jun15/kuhr_helfo.sas7bdat	Diagn_ICPC
N_KUHR_ICD	Antall unike diagnosekoder (ICD-10) fra KUHR/HELFO (avtalespesialister)	Antall unike ICD-10 diagnosekoder på fullformat fra KUHR/HELFO-data (avtalespesialister)	SAS26jun15/kuhr_helfo.sas7bdat	Diagn_ICD
MeanSpm7	Måler selvopplevd helsetjenestekvalitet	Måler gjennomsnittlig score på selvopplevd helsetjenestekvalitet på en skala fra 1-5 der 1=lite fornøyd og 5=svært fornøyd. Variabelen er satt sammen av de 8 underspm til spm 7 i spørreskjemaet: «tenk på alle helsetjenestene du har hatt kontakt med i 2012 og 2013. Hvor enig eller uenig er du i følgende utsagn?» Tre negative underspm (7.4, 7.6 og 7.7) ble reversert før de ble inkludert i skalaen. Svaralternativet «ikke aktuelt» ble kodet til sysmiss. Chronbach's Alpha: 0.88	D:\SAS26jun15/survey_inkludert.sas7bdat	S7_1, S7_2, S7_3, S7_4, S7_5, S7_6, S7_7, S7_8

MeanSpm7a	Måler selvopplevd helsetjenestekvalitet på behandling/samarbeid	Måler gjennomsnittlig score på selvopplevd helsetjenestekvalitet på behandling/samarbeid på en skala fra 1-5 der 1=lite fornøyd og 5=svært fornøyd. Utledet fra en faktoranalyse (PCA) av alle de 8 underspm til spm 7. Inkluderer spm 7.1, 7.2, 7.3, 7.5 og 7.8. svaralternativet «ikke aktuelt» ble kodet til sysmiss. Chronbach's Alpha: 0.89	D:\SAS26jun15/ survey_inkludert.sas7bdat	S7_1, S7_2, S7_3, S7_4, S7_5, S7_6, S7_7, S7_8
MeanSpm7b	Måler selvopplevd helsetjenestekvalitet på informasjonsflyt	Måler gjennomsnittlig score på selvopplevd helsetjenestekvalitet på informasjonsflyt på en skala fra 1-5 der 1=lite fornøyd og 5=svært fornøyd. Utledet fra en faktoranalyse (PCA) av alle de 8 underspm til spm 7. Inkluderer spm 7.4, 7.6 og 7.7. svaralternativet «ikke aktuelt» ble kodet til sysmiss. Variablene er reversert for å ha lik scoring som MeanSpm7a Chronbach's Alpha: 0.77	D:\SAS26jun15/ survey_inkludert.sas7bdat	S7_1, S7_2, S7_3, S7_4, S7_5, S7_6, S7_7, S7_8