Distinguishing Differences in Quality of Care of Traumatic Brain Injury among Hospitals

Didier Nibbering
339536dn@student.eur.nl
Erasmus University Rotterdam
PO Box 1738, NL-3000
Rotterdam, the Netherlands

ABSTRACT

There has been an increasing demand for insights into the quality of health care. In this paper we analyse differences in quality of care among hospitals, using an unbalanced dataset containing observations of patients with traumatic brain injury. We discuss random effects and fixed effects models which are traditionally used for modeling variation in performance of hospitals. Results demonstrate that these methods leave a lot of uncertainty about the existence of individual differences. Findings also reveal that actual differences are hardly distinguishable. Hence, we propose a finite mixture approach. Our empirical results suggest that three quality groups with hospitals are sufficient to describe the variation in health care quality. Classifying hospitals in quality clusters provides a correct alternative to current shaky hospital rankings.

Keywords

Quality of Health Care, Fixed Effects, Finite Mixture

1. INTRODUCTION

Health care is nowadays one of the most challenging issues. On the one hand we see strongly increasing health care costs. In the Netherlands the expenditures on health care increased to 70.1 billion euro in 2010, which equals 11.9 percent of the Dutch gross domestic product (CBS, 2012). Because of the expected ageing of the world's population it is plausible that the expenditures on care keep rising (Lutz et al., 2008). On the other hand people demand only the best care for their health problems. So hospitals have to provide high quality care while controlling costs. Just two decades ago only physicians had a social mandate to judge the quality of care (Blumenthal, 1996). Today, patients require transparancy to compare hospitals and to assess the quality of care by themselves.

There are more stakeholders interested in health care quality. First, hospitals and physicians could improve their performance by comparing different health care providers and learn from best practices. Second, goverments try to monitor hospitals to ensure a certain level of quality. The Dutch government requires health care providers to make information about the quality of their care accessible to patients (Rijksoverheid, 2013). Finally, health insurance companies are also interested in care quality, especially in the differences between hospitals. Dutch health insurers consider the purchasing of health care on quality as main theme (ZN, 2013).

Hence, there are many institutes which require insight into the differences in quality among hospitals. A commonly used method to measure and publicly report quality of care provided by hospitals is to determine performance by ranking hospitals (Anderson et al., 2007). In the Netherlands we have several of these rankings. The Ministry of Public Health makes information publicly accessible from reports of the inspection about quality of care. Clients could easily compare hospitals on a website operated by the government (RIVM, 2013). Each year the newspaper 'Algemeen Dagblad' publishes the 'Hospital Top Hundred' (AD, 2012). Weekly magazine 'Elsevier' also compares all Dutch hospitals each year (SIRM, 2012).

It is doubtful whether these rankings represent the quality of the hospitals. Anderson et al. (2007) found that "considerable uncertainty exists in ranking of hospitals" and subsequently "calls into question the use of rank ordering as a determinant of performance". Also Jacobs et al. (2005) states that great care is warranted in interpreting the rankings of hospitals. According to Ranstam et al. (2008) it is even doubtfull if a correct ranking can be achieved, due to an insufficient number of patient observations. When patients and insurers draw conclusions based on incorrect rankings there are large consequences for the health care. Lilford et al. (2004) states that it can result in capricious sanctions, unjustified rewards and the risk of stigmatising an entire institution.

Therefore, it is important to find a more appropriate way to compare hospitals. Lingsma (2010) states that measuring care with outcome measures such as mortality poses two major methodological problems. First, because of differences between patient samples of hospitals, outcomes could differ between hospitals regardless the variation in quality of care. Therefore comparisons between hospitals need an adjustment for each patient's characteristics. Second, when sample sizes are small the variation in outcome between hospitals could easily lead to overinterpretation of differences between hospitals. Taking into account individual patient

characteristics and statistical uncertainty can improve quality of care measures.

In this paper we discuss methods to estimate individual differences in quality of health care among hospitals. We use an unbalanced dataset with traumatic brain injury patient observations in different hospitals. We examine fixed effects models and random effects models and show that differences in quality are hardly distinguisable. Therefore we propose a method to obtain clusters of hospitals in which hospitals provide similar quality, while the quality of care differs across the clusters. We estimate these clusters by means of a finite mixture model. This approach provides an alternative to current hospital rankings.

Previous studies have focused on the tradeoff between fixed effects and random effects in modeling differences in quality of care between hospitals. Lingsma (2010) calculated fixed and random hospital effects for ten hospitals from the Netherlands Stroke Survey 2002-2003. She stated that using fixed effects models to estimate effects for the relative quality of hospitals with small patient populations, could lead to exploding estimates and over-interpretation of differences between hospitals. She found that random effects are more conservative in estimating differences between hospitals. Austin et al. (2003) used Monte Carlo simulations to examine the ability of random effects and fixed-effects models to correctly classify hospitals according to their performance. They showed that "the sensitivity of the random effects method was inferior to that of the fixed-effects model, whereas under most scenarios examined, the specificity of the random-effects model was greater than that of the fixed-effects model". Glance et al. (2006) investigated how robust health outcomes report cards are to changes in the underlying statistical methodology. They discussed random effects and fixed effects models for estimating differences between hospitals and concluded that findings vary using different statistical methods, due to the fact that models always rest on assumptions.

Finite mixture models are used in a variety of scientific disciplines. Vermunt (2008) described finite mixture models for the analysis of hierarchical data sets and provided several applications. For instance, he considered the antibiotics prescriptions of doctors in two Chinese counties and used finite mixture to identify clusters of doctors with similar prescription behaviour. Galbraith and Green (1990) gave the example of a random sample of grains from a population in which there are three different ages and the distribution over the ages is unknown. This is comparable to our case in the sense that we also do not know the proportions of hospitals which are assigned to different quality clusters. However, in addition to the unknown proportions, we also have an unknown number of quality clusters. Paap et al. (2005) used a finite mixture model with an application to clustering countries on growth rate. They proposed a model which "allows a data-based classification of countries into clusters such that within a cluster countries have the same average growth rate". We follow broadly the same approach, with the countries replaced by hospitals.

The outline of this paper is as follows. In Section 2 we explain the dataset we use, how the dataset has been cleaned and the variables we use in our models. In Section 3 we discuss the methods we use to determine and evaluate differences in quality of health care among hospitals. In Section 4 we discuss the results of the models and the implications on modeling and interpretation of differences in quality of health care. In the last section, Section 5, we provide a discussion on the main results, discuss some limitations of our research and give suggestions for further research.

2. DATA

In this paper we use a dataset from Erasmus Medical Center containing information about patients with traumatic brain injury (TBI), a leading cause of disability and death worldwide (Perel et al., 2008). This dataset is based on the International Mission for Prognosis And Clinical Trial (IMPACT) database. In this section we discuss the content of the dataset and how the dataset has been cleaned. Next, we discuss the variables we use in our models.

Marmarou et al. (2007b) describe the design and content of the IMPACT database of traumatic brain injury. The database contains data over 11988 individual patients with moderate and severe TBI from randomized controlled trails and observational studies. Patients with missing outcomes, missing age, younger than fourteen, and/or with missing hospital indicator are excluded from the dataset. Also patients from one single-center study are excluded. In order to estimate fixed effects models we can only include hospitals in our analysis within which patient outcomes vary, as we explain in Section 3. Hence, we drop 36 hospitals from our data because of all positive or all negative outcomes in addition to the cleaning of the Erasmus Medical Center. Ultimately, we have a dataset with 10011 individual patients enrolled at 230 different hospitals. Each hospital has one unique code when it participated in multiple studies in the database.

In this paper we investigate the differences between hospitals in patient outcome after traumatic brain injury. We use the Glasgow Outcome Scale (GOS) as outcome measure. Jennett and Bond (1975) describe this five-point scale which we dichotomize as favourable versus unfavourable outcome, at six months after injury. The favourable outcome includes the categories good recovery and moderate disability. The unfavourable outcome is composed of severe disability, vegetative state or death.

We use an unfavourable outcome according to GOS as binary dependent variable in our analysis. Three main predictors of outcome in TBI are included as independent variables. These control variables for heterogeneity in patient populations per hospital contain patient characteristics measured at admission. First, we consider the age of a patient as continuous variable. Second, we consider the pupillary reactivity as categorical variable. The pupil reactivity is divided into three categories; both reacting, one reacting, and neither reacting. The Glasgow Coma Scale (GCS) motor score is our third independent categorical variable (Teasdale and Jennett, 1974). We distinguish seven different categories for the motor respones; makes no movements, extension to painful stimuli, abnormal flexion to painful stimuli, flexion or withdrawal to painful stimuli, localizes painful stimuli, obeys comands, and untestable. The last category is included to deal with patients sedated at admission. Because sedation can be caused by either severe TBI or by other injuries, we consider these patients as untestable.

Multiple studies used these three variables in prognostic analyses in TBI (Perel et al., 2008; Murray et al., 2007; Hukkelhoven et al., 2005; Lingsma et al., 2011). Moreover, Marmarou et al. (2007a) found a strong association between

the GCS motor score and pupil reactivity and 6-month GOS. Mushkudiani et al. (2007) demonstrated that increasing age is strongly related to poorer outcome assessed by the GOS after TBI.

We consider 230 unique hospitals with widely varying patient numbers. Figure 1 shows the number of patiens per hospital. The hospitals with the smallest patient populations treat only two patients and the largest population amounts to 517 patients. The average number of patients per hospital is 43 and the median equals 22. Of all the patients 48 percent have an unfavourable outcome. The median age equals 30 and ranges from 14 to 93. The distribution of patients over the categories in pupillary reactivity and GCS motor score are shown in Appendix A.

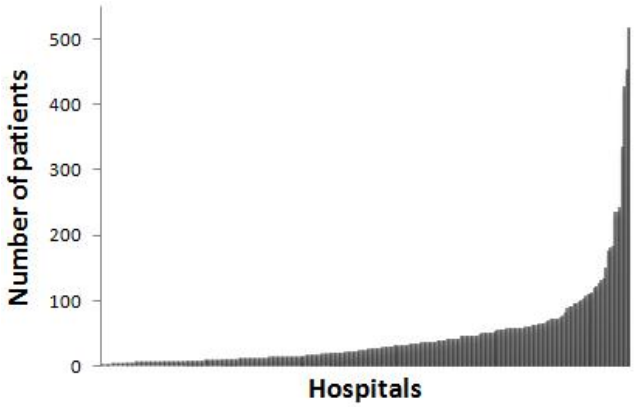

Figure 1: Observed Number of Patients per Hospital

3. METHODS

In this section we explain the methods that we use to determine and evaluate differences in quality of health care among hospitals. First, we provide the general model specification. Second, we explain three specific ways of estimating quality of care per hospital, namely with fixed effects, with random effects, or with a finite mixture approach. After specifying the models we discuss the methods which we use to evaluate and compare the different models.

3.1 Modeling Differences Between Hospitals

Patient characteristics and quality of care are the factors which influence patient outcome. We take the patient characteristics into account with independent variables which indicate the condition of the patient. With respect to the quality of care we take the differences between hospitals into consideration. These differences could be modeled with random or fixed effects. These models are often used in the analysis of panel data. We also propose a finite mixture approach. We apply these models to our case in which the data is grouped by N different hospitals and each hospital i has T_i different observations. One observation represents here one unique patient. We define the dependent variable y_{it} as the outcome of patient t in hospital i six months after injury, with $i = 1, ..., N$, $t = 1, .., T_i$, and $T = \sum_{i=1}^{N} T_i$ is the total number of patients.

$$y_{it} = \begin{cases} 1 & \text{if unfavourable outcome of patient } t \text{ in hospital } i \\ 0 & \text{otherwise} \end{cases}$$

We define x_{it} as the vector with an intercept and K explanatory variables which describe the characteristics of patient t in hospital i. Because of the binary dependent variable we use a logit model:

$$Logit(P[y_{it} = 1|x_{it}]) = \alpha_i + x'_{it}\beta \tag{1}$$

where β is the vector with parameters corresponding to x_{it} and α_i is a hospital-specific parameter. The α_i's capture all the effects peculiar to hospital i. Therefore we use the estimate of α_i as indicator for quality of care in hospital i.

3.1.1 Fixed Effects Specification

When we see the α_i's in (1) as N fixed unknown parameters we have a fixed effects logit model. We estimate the models by means of maximum likelihood. To demonstrate the log-likelihood function we first rewrite the logit model of (1) and define the cumulative distribution function $F(.)$ of the logit model:

$$P[y_{it} = 1|x_{it}] = F(\alpha_i + x'_{it}\beta) = \frac{e^{\alpha_i + x'_{it}\beta}}{1 + e^{\alpha_i + x'_{it}\beta}} \tag{2}$$

In the case of a fixed effects model we treat the α_i's as fixed unknown parameters. Therefore, we can include $N - 1$ dummy variables for the hospitals in the standard logit model. This gives us the following log-likelihood function:

$$\begin{aligned} \mathcal{L}(\beta, \alpha_1, ..., \alpha_N) = \sum_{i=1}^{N}\sum_{t=1}^{T_i}(y_{it} \ln(F(\alpha_i + x'_{it}\beta)) \\ + (1 - y_{it}) \ln(1 - F(\alpha_i + x'_{it}\beta))) \end{aligned} \tag{3}$$

Maximizing this unconditional log-likelihood function results in consistent parameter estimates provided that the number of patients goes to infinity. This implies that we can only estimate the fixed effects consistently if the number of observations for each hospital grows, which requires a large number of patients (Verbeek, 2004). Because of superior asymptotic properties we also consider conditional maximum likelihood. In this case we consider the likelihood function conditional upon the sufficient statistic $\sum_{t=1}^{T_i} y_{it}$. This means that the conditional log-likelihood function no longer depends upon α_i but still depends upon the other parameters β:

$$\mathcal{L}_c(\beta) = \sum_{i=1}^{N} \ln(f(y_{i1}, ..., y_{i,T_i} | \sum_{t=1}^{T_i} y_{it}, x_{1t}, x_{2t}, ...)) \tag{4}$$

where $f(.)$ is the probability density function of $y_{i1}, ..., y_{i,T_i}$. Chamberlain (1982) gives a further derivation of the conditional log-likelihood function. Conditional maximum likelihood makes consistent estimation possible but cannot produce any estimates of the hospital-specific parameters. Therefore, we cannot use this estimation method to compare the quality of care between hospitals. Katz (2001) investigated whether the use of unconditional maximum likelihood could be justified on the basis of finite-sample properties. He found a negligible amount of bias in the unconditional maximum likelihood estimates when the observations per unit, in our case a hospital, are greater than fifteen.

In Section 2 we noted that we discard all the observations from hospitals with only positive or negative outcomes. From the likelihood functions of the fixed effects logit model it becomes clear that when in hospital i $y_{it} = 1$ for all t then the maximum likelihood estimate of α_i is ∞ and if $y_{it} = 0$ for all t the maximum likelihood estimate of α_i is

$-\infty$. Therefore we only include the hospitals in our analysis within which y_{it} varies.

3.1.2 Random Effects Specification

When the α_i's are drawings from a normal distribution with mean zero and variance σ_α^2 the model in (1) is referred to as the random effects logit model. The log-likelihood function of the random effects logit model equals:

$$\mathcal{L}(\beta, \alpha_1, ..., \alpha_N) = \sum_{i=1}^{N} \ln(\int_{-\infty}^{\infty} \frac{\mathrm{e}^{-\alpha_i^2/2\sigma_\alpha^2}}{\sqrt{2}\sigma_\alpha} \times \{ \prod_{t=1}^{T_i} F(\alpha_i + x_{it}'\beta) \} \, \mathrm{d}\alpha_i) \tag{5}$$

Because we only have to estimate one distribution parameter instead of separate hospital effects, the random effects estimator is more efficient than the fixed effects estimator. However, the random effects model provides inconsistent estimates when the hospital-specific effects are correlated with the independent variables. Glance et al. (2006) states that there is reason to suspect that patient characteristics are not independent of provider effect. For example, when older patients are more likely to be treated by high-quality hospitals, the age of the patients are correlated with the hospital-specific effects.

3.1.3 Finite Mixture Specification

Besides distinguishing differences between individual hospitals, we can also differentiate different segments with hospitals. This is a good idea when we do not have enough power to find individual differences. We use a finite mixture logit model to group the hospitals in clusters with similar performance. We assume that within each cluster hospitals have the same quality of care, while quality of care is different across the distinct clusters. The model estimates from the data how many clusters there are and which hospitals belong to which cluster. Paap et al. (2005) describe this approach in detail.

We assume that each hospital-specific effect α_i is equal to one of the J different values γ_j with probability $p_j = P(s_i = j) = 1, ..., J$ and $\sum_{j=1}^{J} p_j = 1$. Here is s_i the cluster indicator for hospital i. So the hospitals can be classified into J clusters, where γ_j is the cluster-specific effect. This gives us the following model:

$$P[y_{it} = 1 | x_{it}] = \sum_{j=1}^{J} p_j f(\gamma_j + x_{it}'\beta) \tag{6}$$

with

$$f(\gamma_j + x_{it}'\beta) = F(\gamma_j + x_{it}'\beta)^{y_{it}} (1 - F(\gamma_j + x_{it}'\beta))^{1-y_{it}} \tag{7}$$

To estimate this finite mixture model we first define the likelihood function of (6):

$$l(y; \beta, \gamma_1, ..., \gamma_J) = \prod_{i=1}^{N} (\sum_{j=1}^{J} p_j (\prod_{t=1}^{T_i} f(\gamma_j + x_{it}'\beta))) \tag{8}$$

When we assume that the value of $s = (s_1, ..., s_N)$ is known, we can consider the complete data likelihood function:

$$l(y, s; \beta, \gamma_1, ..., \gamma_J) = \prod_{i=1}^{N} (\sum_{j=1}^{J} (p_j \prod_{t=1}^{T_i} f(\gamma_j + x_{it}'\beta))^{I[s_i = j]}) \tag{9}$$

We estimate the parameters of our model in (6) using the EM algorithm of Dempster et al. (1977), which provides a maximum of the log-likelihood function. This iterative two-step procedure consists of an expectation and a maximization step. In the first step (E-step) we take the expectation of the log complete data likelihood function with respect to $s|y$ given the current estimates of $\beta, \gamma_1, ..., \gamma_J$:

$$\begin{aligned} &E_{s|y}[\mathcal{L}(y, s; \beta, \gamma_1, ..., \gamma_J)] \\ &\quad = \sum_{i=1}^{N} (\sum_{j=1}^{J} \hat{w}_{ij} (\ln p_j + \sum_{t=1}^{T_i} \ln f(\gamma_j + x_{it}'\beta))) \end{aligned} \tag{10}$$

where

$$\begin{aligned} \hat{w}_{ij} &= P(s_i = j | y_{i1}, ..., y_{iT}, \hat{\beta}, \hat{\gamma}_1, ..., \hat{\gamma}_J) \\ &= \frac{\hat{p}_j \prod_{t=1}^{T_i} f(\hat{\gamma}_j + x_{it}'\hat{\beta})}{\sum_{j=1}^{J} \hat{p}_j \prod_{t=1}^{T_i} f(\hat{\gamma}_j + x_{it}'\hat{\beta})} \end{aligned} \tag{11}$$

In the second step (M-step) we maximize the expectation of the complete data log-likelihood function with respect to the parameters $p_j, \beta, \gamma_1, ..., \gamma_J$. This provides a value of $\hat{p}_j$:

$$\hat{p}_j = \frac{1}{N} \sum_{i=1}^{N} \hat{w}_{ij} \tag{12}$$

The first and second order conditions used in the maximization step can be found in Appendix B.

Before we start the first iteration of the EM algorithm, we initialize the weights $\hat{w}_{ij}$ and the parameters $\hat{\beta}, \hat{\gamma}_1, ..., \hat{\gamma}_J$. Subsequently we use the parameters to calculate new weights in the E-step and thereafter we update the parameters using the new weights in the maximization of the expectation of the complete data loglikelihood function in (10). The two steps of the EM algorithm are repeated until convergence. When the estimates of the parameters are converged to their final values, we can use the final weights to classify hospitals into clusters with similar quality of care. We assign each hospital i to cluster j for which $\hat{w}_{ij} = \max_j \hat{w}_{ij}$. This results in the following cluster model:

$$P[y_{it} = 1 | x_{it}, s_i = j] = F(\gamma_j + x_{it}'\beta) \tag{13}$$

To determine the appropriate number of clusters in this model, we estimate the finite mixture logit model in (6) for different values of J. We prefer the model with the lowest Bayesian Information Criterion:

$$\mathrm{BIC} = -2(\mathcal{L}(y; \beta, \gamma_1, ..., \gamma_J)) + (K + J) \ln(T) \tag{14}$$

Because we have to ensure that the parameters of our finite mixture model are identified, we cannot estimate every number of clusters. A sufficient condition for identifiability in a finite mixture of binomial models requires that (Follmann and Lambert, 1991):

$$J \leq \frac{1}{2}(\min(T_i) + 1) \tag{15}$$

This means that if we want to estimate eight clusters, we have to discard the hospitals with less than fifteen patients from our analysis.

3.2 Evaluating Modeling Methods

We discuss the different methods to estimate differences in quality of care among hospitals. First, we describe the Hausman specification test which is used to decide between

a fixed effects model and a random effects model. We also use this test to compare the fixed effects model with the finite mixture model. Second, we investigate bias in conditional and unconditional maximum likelihood estimates of the fixed effects model. Finally, we determine the power of tests on differences between quality of health care of individual hospitals.

3.2.1 Hausman Test

To decide which model is most appropriate, we perform the Hausman specification test (Hausman, 1978). First, we test the random effects model against the fixed effects model. Under the null-hypothesis of this test both estimators are consistent but random effects is the prefered model due to higher efficiency. Under the alternative the fixed effects model provides still consistent estimates whereas the random effects estimator is inconsistent. When the individual effects per hospital are uncorrelated with the independent variables, the fixed effects and random effects estimators should not be statistically different. In this case we cannot reject the null-hypothesis. The fixed effects model provides consistent estimates even when the hospital effects are correlated with the patient characteristics. So under the alternative hypothesis the fixed effects model is considered as most appropriate. The Hausman statistic is distributed as Chi-squared with the number of parameters in β as degrees of freedom. The test statistic is computed as follows:

$$H = (\widehat{\beta_F} - \widehat{\beta_R})'(\widehat{V_F} - \widehat{V_R})^{-1}(\widehat{\beta_F} - \widehat{\beta_R}) \tag{16}$$

where $\widehat{\beta_F}$ is the estimated β in the fixed effects model, $\widehat{\beta_R}$ is the estimated β in the random effects model, $\widehat{V_F}$ is the estimated covariance matrix of the fixed effects model and $\widehat{V_R}$ is the estimated covariance matrix of the random effects model. Second, we test the finite mixture model against the fixed effects model. Under the null-hypothesis both estimators are consistent. Under the alternative only the fixed effects model provides consistent estimates. Since the finite mixture model only has to estimate the cluster-specific effects instead of separate hospital effects, the finite mixture estimator is more efficient than the fixed effects estimator. The test statistic is described in (16) with the estimated β and covariance matrix of the random effects model replaced by the estimated β and covariance matrix of the finite mixture model. When we reject the null-hypothesis we have to conclude that the estimates of the finite mixture model are biased.

3.2.2 Consistency Fixed Effects Estimators

The conditional maximum likelihood estimator of the fixed effects model provides consistent estimates but cannot estimate hospital-specific effects. Therefore, we have to estimate the fixed effects model by means of unconditional maximum likelihood, which is only consistent if the number of patients in each hospital is large enough. To examine whether we have a significant amount of bias in the unconditional maximum likelihood estimates of the fixed effects logit model, we perform a Monte Carlo study on bias in both estimators. When we find significantly biased estimators, we assess whether the bias of the estimators is acceptable.

In the Monte Carlo simulations, the data generating process is the logit model in terms of latent variables (Cameron and Trivedi, 2005):

$$y^*_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it} \tag{17}$$

where ϵ_{it} is the unobserved individual-specific effect, which has a standard logistic distribution. The latent variable y^*_{it} is related to the binary dependent variable y_{it} as follows:

$$y_{it} = \begin{cases} 1 & \text{if } y^*_{it} \geq 0 \\ 0 & \text{if } y^*_{it} < 0 \end{cases}$$

Before running the simulations we choose the parameter values in α_i and β. We draw α_i from a normal distribution with a mean of zero and standard deviation of 0.5 and take fixed values for β. Thereafter we run M simulations. First we randomly divide T patients over N hospitals. The number of patients is distributed to hospitals in the same way as in the real dataset. So each hospital has the same number of patients in every simulation, but the characteristics of the patients per hospital differ in each simulation. Second, we randomly generate ϵ_{it} from a standard logistic distribution and compute y_{it} using the data generating process in (17). In the last step of the simulation we estimate three models with the dependent variable y_{it} and the three indepent variables which describe age, pupil reactivity and motor score of the patients; a fixed effects logit model with conditional maximum likelihood estimators, a fixed effects logit model with unconditional maximum likelihood estimators, and a simple logit model which does not take the fixed effects into consideration.

After performing the simulations, we take for each model the average of the estimates for β in each simulation $\bar{\hat{\beta}}$. We refer to this statistic as the Monte Carlo mean which is approximately normally distributed:

$$\bar{\hat{\beta}}_k = \sum_{i=1}^{M} \hat{\beta}_{k,i} \approx \mathcal{N}(\beta_k, \frac{s_k^2}{M})$$

where

$$s_k^2 = \frac{1}{M-1}\sum_{i=1}^{M}(\hat{\beta}_{k,i} - \bar{\hat{\beta}}_k)^2 \tag{18}$$

and $k = 1, ..., K$ with K the number of explanatory variables. When we have to reject the null-hypothesis of no difference between the value of parameter β_k and the corresponding estimate $\hat{\beta}_k$ we conclude that this estimate is biased.

When we find significantly biased estimators, we use four criteria to assess whether the bias of the estimators $\hat{\beta}_k$ with $k = 1, ..., K$ is acceptable. These criteria are described by Boomsma and Hoogland (2001). First, we define the relative bias of estimator $\hat{\beta}_k$ for parameter β_k as follows:

$$B(\hat{\beta}_k) = \frac{\bar{\hat{\beta}}_k - \beta_k}{\beta_k} \tag{19}$$

Second, we use the mean absolute relative bias (MARB) to compare the bias of the parameter estimators:

$$\text{MARB}(\hat{\beta}_k) = \frac{1}{k}\sum_{k=1}^{K}|B(\hat{\beta}_k)| \tag{20}$$

We also apply the above described criteria for the parameter

estimators to the standard error estimators:

$$B[\hat{se}(\hat{\beta}_k)] = \frac{\hat{se}(\hat{\beta}_k) - SD(\hat{\beta}_k)}{SD(\hat{\beta}_k)} \tag{21}$$

where $\hat{se}(\hat{\beta}_k)$ is the average of the estimated standard errors for the parameter estimate in each simulation and $SD(\hat{\beta}_k)$ is the standard deviation of the Monte Carlo mean which equals s_k in (18). The MARB of standard error estimators is defined as follows:

$$\text{MARB}(\hat{se}(\hat{\beta}_k)) = \frac{1}{k}\sum_{t=1}^{K} |B[\hat{se}(\hat{\beta}_k)]| \tag{22}$$

3.2.3 Determining Power and Sample Size

In the fixed and random effects models we estimate the effects α_i to draw conclusions about differences in quality of care between individual hospitals. For this aim we have to test the significance of the differences between the hospital-specific effects. Moreover, we are interested whether we can find actual differences. Therefore, we determine the statistical power of this test between two hospitals at specific sample sizes by means of a Monte Carlo simulation.

We perform the Monte Carlo simulation for several sample sizes. For each sample size T we draw for each hospital 1000 times T random patients, with corresponding patient characteristics, from our dataset. After generating the patient populations both hospitals contain T patients. Thereafter we calculate the patient outcome using the logit data generating process as described in (17), with $i = 1, 2$, $\alpha_1 = 0$ and α_2 equals the magnitude of the difference between quality of care of the two hospitals we would like to test. Then we perform the likelihood-ratio test with under the null-hypothesis a model without hospital-specific effects included and under the alternative a model with hospital-specific effects:

$$LR = 2\mathcal{L}(\hat{\beta}, \hat{\alpha_1}, \hat{\alpha_2}) - 2\mathcal{L}(\hat{\beta}) \tag{23}$$

We have to reject the null-hypothesis when LR is sufficiently large. In this case the test finds the actual difference in quality of care between the two hospitals. To find the power of the test, we sum the number of simulations in which the likelihood-ratio test rejects the null-hypothesis and divide it by the total number of simulations.

4. RESULTS

In this section we discuss the results of the models and implications on modeling and interpretation of differences in quality of health care. First, we show that the fixed effects model is more appropriate in modeling the patient outcomes than the random effects model. Second, we provide the results of the Monte Carlo simulations with regard to inconsistent estimators which indicate that the unconditional estimation method of the fixed effects model generates only a small amount of bias. Third we discuss whether there are genuine individual differences between hospitals and whether we can distinguish those differences. Our findings call the validity of ranking hospitals on individual differences in quality in question. Finally, we discuss the results of the finite mixture logit model. We find that three clusters are sufficient to describe the differences in quality of care among the hospitals.

4.1 Random Effects or Fixed Effects?

We estimate the hospital-specific effects with fixed and random effects. Table 1 shows the conditional estimates of the fixed effects model and the estimates of the random effects model. The signs of the estimated parameters are in line with our expectations. The older the patient, the higher the probability of an unfavourable outcome after traumatic brain injury. The estimated parameters of the categorical variable for the motor scale show that a lower motor response corresponds to a higher likelihood of an unfavourable outcome. The seventh category represents patients which are untestable on motor score because of sedation at admission in the hospital. We find that the effect of this category is around the average of the effects of the other (testable) motor scales. Finally, we can derive that the chances of a patient rise significantly as the pupillary reactivity increases.

Table 1: FE and RE Model Estimates

	Fixed Effects		Random Effects	
	Coef.	s.e.	Coef.	s.e.
age	0.038	0.002	0.038	0.002
motor 2	0.615	0.098	0.655	0.095
motor 3	-0.034	0.091	-0.024	0.088
motor 4	-0.687	0.081	-0.665	0.078
motor 5	-1.322	0.083	-1.276	0.080
motor 6	-1.502	0.164	-1.466	0.162
motor 7	-0.354	0.121	-0.303	0.116
pupil 2	0.799	0.066	0.811	0.065
pupil 3	1.447	0.069	1.448	0.068
constant			-1.361	0.091
σ_α			0.407	0.038

We use the estimates in Table 1 to perform the Hausman test. The Hausman test statistic $H \sim \chi^2(9)$ equals 26.18 corresponding to a p-value of 0.002. This means that the explanatory variables are correlated with the hospital-specific effects, which results in biased estimates of the random effects model. Therefore we have to reject the random effects model, irrespective of the hospital-specific effects are stochastic or not (Greene, 2008).

Any correlation between hospital-specific effects and the explanatory variables can imply an omitted variable. The confounding effect of an omitted variable on estimates of effects of explanatory variables in the fixed effect model are removed by estimating separate unit effects. Since the random effects model does not estimate separate unit effects but models a probability distribution for the hospital-specific effects, any correlation between the explanatory variables and the unit effects produces biased estimates. Because our aim is to model and examine the differences in quality of care between hospitals, we have to avoid that we estimate differences systematically too large or too small. Hence, the consistency of estimators is of great importance. Therefore we use the fixed effects model in further analysis and we reject the random effects model for modeling differences between hospitals.

4.2 Bias in Maximum Likelihood Estimators

Since we have decided to use fixed effects, we have to determine whether the unconditional estimation method of this model is appropriate. Conditional maximum likelihood

estimation provides consistent estimates of the parameters of the explanatory variables concerning patient characteristics, but cannot estimate the hospital-specific effects. Unconditional maximum likelihood estimation does provide these effects, but is only consistent if the number of observations for each hospital is large enough. The results in Table 2 and Table 3 indicate whether this is the case.

Table 2: Results Simulations Conditional MLE

	β	$\bar{\hat{\beta}}$	$SD(\hat{\beta})$	$\hat{se}(\hat{\beta})$	p-value
age	0.040	0.040	0.002	0.002	0.460
motor 2	0.600	0.601	0.091	0.092	0.630
motor 3	-0.030	-0.029	0.088	0.086	0.617
motor 4	-0.700	-0.698	0.076	0.077	0.436
motor 5	-1.400	-1.396	0.078	0.080	0.089
motor 6	-1.500	-1.501	0.164	0.167	0.782
motor 7	-0.400	-0.394	0.114	0.112	0.071
pupil 2	0.800	0.802	0.066	0.066	0.433
pupil 3	1.500	1.504	0.066	0.066	0.050

These tables contain the summarized results of the Monte Carlo study of bias in the conditional and unconditional maximum likelihood estimators of the fixed effects model. The table with simulation results of the maximum likelihood estimates of a simple logit model without fixed effects are shown in Appendix C. The second column of the tables shows the actual value of the parameters. The third column shows the Monte Carlo mean of the estimated parameter values of the variables for the patient characteristics. The numbers in the remaining columns represent the standard deviations of the Monte Carlo means, the average of the estimated standard errors for the parameter estimates in each simulation, and the p-value of the test on equality between the Monte Carlo mean and the actual value of the parameters, respectively.

Table 3: Results Simulations Unconditional MLE

	β	$\bar{\hat{\beta}}$	$SD(\hat{\beta})$	$\hat{se}(\hat{\beta})$	p-value
age	0.040	0.041	0.002	0.002	0.000
motor 2	0.600	0.618	0.093	0.094	0.000
motor 3	-0.030	-0.029	0.090	0.087	0.832
motor 4	-0.700	-0.716	0.078	0.078	0.000
motor 5	-1.400	-1.432	0.080	0.081	0.000
motor 6	-1.500	-1.540	0.168	0.169	0.000
motor 7	-0.400	-0.404	0.117	0.113	0.288
pupil 2	0.800	0.823	0.067	0.067	0.000
pupil 3	1.500	1.544	0.068	0.067	0.000

We find, according to the p-values, that no Monte Carlo mean of the conditional maximum likelihood estimates is significant different from the actual values on a significance level of five percent. This is in accordance with the fact that conditional maximum likelihood of the fixed effects logit model gives consistent estimators. In contrast to the conditional estimators almost all the unconditional estimates are significant different from the real values. As expected, the estimation method in which we can retrieve the values for the hospital-specific effects provides biased parameter estimates. Since we would like to compare differences in quality of care among hospitals on the basis of these hospital-specific effects, we have to examine whether the bias is acceptable.

The mean absolute relative bias is used to compare the bias of parameter estimators. It is also a criterion for an acceptable bias. Boomsma and Hoogland (2001) stated that the estimators have to satisfy the following conditions to be regarded as acceptable:

$$\text{MARB}(\hat{\beta}) < 0.025$$
$$\text{MARB}(\hat{se}(\hat{\beta})) < 0.050$$

Table 4 shows that both the conditional as the unconditional maximum likelihood estimators meet this condition. Although the unbiased conditional estimator has a much lower value for the MARB($\hat{\beta}$), the unconditional estimator is also comfortably within the acceptable boundaries. In the Tables 2 and 3, both for the unconditional as the conditional estimates the standard errors and the standard deviations seem very similar. This is consistent with the low MARB($\hat{se}(\hat{\beta})$)'s, which are much lower than the permitted values. The values in Table 4 which represent the performance of the estimators of the simple logit model confirm the importance of including hospital-specific effects in an accurate model.

Table 4: Mean Absolute Relative Bias MLE's

	MARB($\hat{\beta}$)	MARB($\hat{se}(\hat{\beta})$)
Conditional MLE	0.009	0.015
Unconditional MLE	0.024	0.014
No FE MLE	0.060	0.017

We find in our results a confirmation of the theory of consistent estimates of conditional estimators and inconsistency in unconditional estimators of the fixed effects model. This implies that we can only estimate the hospital-specific effects in a proper way if the number of patients for each hospital grows. Since the inconsistency remains between the limits of acceptable bias, we argue that we satisfy this condition. Therefore we consider the unconditional estimates of the fixed effects in our research into differences between hospitals' quality of care.

4.3 Individual Differences between Hospitals

Estimation of the fixed effects model produces estimates of effects for each hospital. These effects describe the deviation of hospitals in the probability of an unfavourable patient outcome from other hospitals, cleaned from different patient characteristics. Because each hospital-specific effect is peculiar to one hospital, these deviations are caused by the differences among care providers. Therefore we attribute the variation in the fixed effects to differences between the quality of care in hospitals. When we try to draw conclusions based on this variation we have to investigate two things: First, we have to determine whether there are genuine differences. Second, we should examine whether we can distinguish those differences. Figure 2 shows the histogram of the distribution of the hospital-specific effects. This histogram shows that 59 percent of the fixed effects is located between -0.5 and 0.5. When we rank the hospitals on their fixed effects, the average difference between two consecutive hospitals equals 0.020.

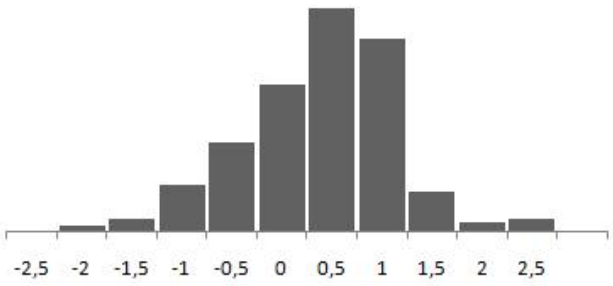

Figure 2: Distribution of the Fixed Effects

So the differences between individual hospitals in a ranking are very small. Only the best and worst hospitals diverged significantly from each other. Moreover, hardly any fixed effect is significant on a five percent level. The results of the Monte Carlo simulations of the unconditional maximum likelihood estimation of the fixed effects model provide a mean absolute relative bias of 2.369 for the fixed effects parameters. This means that there is no less than 200 percent bias in the fixed effect estimates. So we can conclude that when we rank hospitals we are far from certain that one hospital is better or worse than the other.

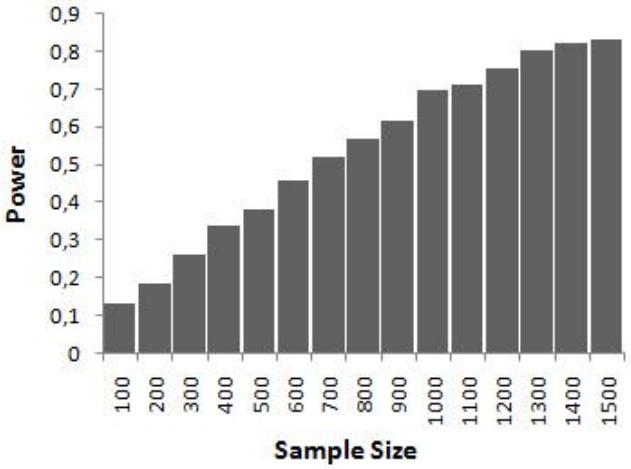

Figure 3: Power of the Test on Fixed Effect of 0.25

Furthermore, we consider the case in which there is an actual difference between the quality of two providers. We perform the likelihood-ratio test to find these differences. As hospitals actually provide another quality of care, we also want to find this distinction with our test. By way of explanation, we have to reject the null-hypothesis when there is an actual difference in quality of care. The statistical power of the test represents the probability that the test will reject the null-hypothesis when there is an actual difference. Figure 3 shows the power of the test on differences between two hospitals for different sizes of patient populations per hospital. The actual difference between the hospital-specific effects is set at 0.250. Only at a sample size of 1300 patients in each hospital, statistical power rises above 80 percent, which is commonly used as lower bound of adequate power. We recall that in our dataset the hospital with the largest patient population provides care to 517 patients, where the median of patients per hospital equals 22. The statistical power for a sample size of 500 patients is only 0.382. Moreover, most hospitals have much smaller differences in fixed effects than 0.250.

When we combine our findings regarding the existence of genuine differences and the ability to distinguish these differences, we conclude that even assuming that there are genuine differences between consecutive hospitals in rankings, we do not have enough power to identify them. Therefore we state that ranking hospitals on quality of care is an inappropriate method for distinguishing differences between hospitals.

4.4 Clustering Hospitals on Quality of Care

Since we cannot prove individual differences, rank ordering is not a good idea. Because of the considerable uncertainty in individual differences a ranking explains little about the genuine quality relationships. Probably, there is in fact only a rough separation in quality delivered by hospitals. Clustering of hospitals allows us to investigate this. In particular, it provides insight into how many groups of hospitals there are with different health care quality. For example, when the finite mixture logit model estimates from the data that we have two clusters within hospitals have the same quality of care, we state that we can only distinguish two distinct levels of quality. So we do not draw any conclusion about individual differences among hospitals. However, we can conlude that the hospitals in one cluster differ from the hospitals in other clusters. When we estimate as many clusters as there are hospitals we end up with the same ranking with which we started. Since the model estimates the number of groups between which there is a difference, there is a slight chance of this case.

We estimate the finite mixture logit model with a different number of clusters. To ensure identifiability of these models, we discard hospitals with fifteen patients or less. This results in a dataset with 9275 patients across 139 hospitals. Figure 4 shows the Bayesian Information Criteria of a model with only one cluster to a model with eight clusters. From the development of the information criteria we can conclude that three clusters are sufficient to describe the differences in quality of care. So we distinguish three levels of health care quality among the hospitals in the dataset.

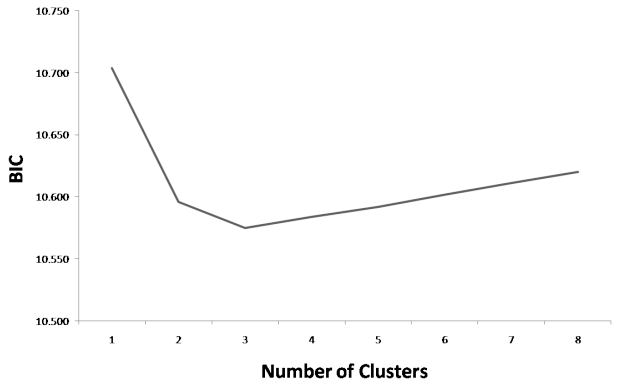

Figure 4: Development BIC for Different Number of Clusters

The finite mixture model with three clusters is shown in Table 5. We find all coefficients significant, except the coefficient of motor3. Hence, whether a patient makes no movements or whether a patient has an abnormal flexion to painful stimuli, has no significant different effect on the outcome. We notice that the signs and the order of magnitude of the coefficients of the patient characteristics are similar to the estimates in Table 1. For interpretation of these coefficients we refer to the explanation about this table.

Table 5: Finite Mixture Model Estimates

	Coef.	s.e.	p-value
age	0.038	0.002	0.000
motor 2	0.652	0.097	0.000
motor 3	-0.032	0.089	0.722
motor 4	-0.675	0.079	0.000
motor 5	-1.288	0.080	0.000
motor 6	-1.527	0.165	0.000
motor 7	-0.308	0.117	0.009
pupil 2	0.818	0.068	0.000
pupil 3	1.454	0.067	0.000
cluster low	-0.885	0.090	0.000
cluster middle	-1.434	0.090	0.000
cluster high	-2.186	0.116	0.000

Besides estimates of the coefficients of patient characteristics Table 5 shows the estimates of the cluster parameters. These coefficients are significant in contrast to the estimates for the fixed effects parameters. So in the case of clustering hospitals we have much more certainty about the existence of the estimated effects. By means of these cluster effects, we can establish a ranking of hospitals based on three groups. The low quality group has an intercept equal to -0.885. The constant of hospitals in the middle group adds up to -1.434 and the high quality hospitals posses an intercept of -2.186. Because the intercept has a positive effect on the probability of an unfavourable patient outcome, patients in last-mentioned hospitals have the lowest probability on an unfavourable outcome.

We mentioned that consistent estimates are of great importance in modeling differences in quality of care between hospitals. To test whether the finite mixture estimates in the cluster model are biased, we perform the Hausman test on the fixed effects model estimates in Table 1 against the finite mixture model estimates in Table 5. The Hausman test statistic $H \sim \chi^2(9)$ equals 13.69 corresponding to a p-value of 0.134. So we cannot reject the null-hypothesis of both consistent estimates of the fixed effects model and the finite mixture model on a significance level of five percent. Hence, the finite mixture approach is not only efficient, but also consistent.

Table 6: Distribution over the Clusters

	Low	Middle	High
$\hat{p}_j$	42.52	47.34	10.14
# hospitals	47	78	14
# patients	3673	4760	842

Table 6 provides an overview of the distribution of patients and hospitals over the three clusters. The estimated values $\hat{p}_j$ show that each hospital has a probability of 10.1 percent to be in the cluster with the highest quality of care. These probabilities are 47.3 and 42.5 percent for the clusters with medium and low quality, respectively. According to the estimated weights in the finite mixture logit model we assign each hospital to a cluster. We find two large groups of reasonable and poorly performing hospitals and a small top class of hospitals which provides excellent services. The number of patients is approximately in the same way divided over the clusters as the number of hospitals. A random patient has a probability of 9.1 percent to be taken care of in an excellent hospital after traumatic brain injury.

We can also look at the distribution of the patient characteristics over the three clusters. Table 7 shows the average value per characteristic in each cluster, weighted according to the weights $\hat{w}_{ij}$. We find the weighted average age per cluster, which is approximately equally distributed. The values attributed to motor score and pupil reactivity can be interpreted as percentages. For example, 13.1 percent of the patients of hospitals in the low quality cluster have motor score one. We find that also motor score and pupil reactivity are approximately equally distributed. This means that there are no large differences between the patient populations in each quality cluster. Hence, high quality hospitals have no other population composition than less performing hospitals.

Table 7: Weighted Averages Patient Characteristics

	Low	Middle	High
age	35.299	34.055	34.861
motor 1	0.131	0.150	0.179
motor 2	0.128	0.130	0.109
motor 3	0.110	0.139	0.153
motor 4	0.215	0.250	0.263
motor 5	0.291	0.273	0.213
motor 6	0.031	0.030	0.013
motor 7	0.093	0.028	0.070
pupil 1	0.631	0.669	0.631
pupil 2	0.150	0.135	0.192
pupil 3	0.220	0.196	0.177

5. DISCUSSION

In this paper we discussed methods to estimate differences in quality of health care among hospitals. We specified a random effects model and a fixed effects model for modeling the hospital-specific effects. Although random effects provide more efficient estimates, the Hausman specification test revealed the fixed effects model as most appropriate. To retrieve the hospital-specific effects in this model we had to use the unconditional maximum likelihood estimator, which is inconsistent when the number of patients per hospital is small. A Monte Carlo simulation on bias in the estimates proved the bias to be acceptable. We evaluated statistical power in differentiating individual differences in quality of care between hospitals for different sample sizes in a second Monte Carlo study. Finally, we proposed a finite mixture logit model as alternative method to current hospital rankings.

Our research shows that the estimates for hospital-specific effects are highly biased and hardly significant. Ranking hospitals according to these effects causes the problem that we are far from certain that one hospital is better or worse than the other, while in a ranking one hospital has to be first and one has to be last. It is very likely that there are no significant differences between consecutive hospitals on rankings. However, even if there are genuine differences in performance between individual hospitals, we find that we do not have the power to identify them. Therefore we argue that ranking hospitals on quality of care is an inadequate method to distinguish individual differences. An important conclusion from our empirical analysis is that we have to move our attention from individual differences between hospitals to differences between quality groups of hospitals.

This has resulted in a method to rank hospitals in a correct manner. First, we use a finite mixture logit model to group hospitals on similar quality of care. Second, we avoid problems that arise on individual differences between hospitals by only ranking the clusters with hospitals. This approach to the assessment of relative differences in quality of care offers opportunities for a more efficient policy based on less uncertainty. For instance, connecting implications to the listing on a ranking for all 348 hospitals in the Netherlands is very expensive and complicated. Policy making on three quality groups is more effictive rather than focussing on quality differences between individual hospitals. Furthermore, the risk of over-interpretation of differences in quality of care between hospitals is strongly reduced. In summary, determing performance of hospitals by rankings should be past tense.

Several potential limitations of this study should be noted. First, it is only possible to include hospitals with variation in the patient outcome in our fixed effects analysis. We have to disregard even more hospitals from our finite mixture analysis to ensure the identifiability of the models. These limitations are related to the known statistical problems around (luckily) small patient populations in hospitals.

Based on our analysis we propose recommendations for further research. To ensure that differences between patient samples of hospitals are not ascribed to differences in quality, more patient characteristics should be included in the models. The addition of extra patient characteristics may result in disappearance of the correlation between random effects and explanatory variables, which means that random effects become an appropriate method for modeling hospital-specific effects. Furthermore, we recommend to investigate whether the findings in this paper are reproducable in other datasets.

References

AD. Algemeen Dagblad. AD Ziekenhuis Top 100 - 2012. 2012. URL `http://ziekenhuis.i-serve.nl/`.

J. Anderson, M. Hackman, J. Burnich, and T. R. Gurgiolo. Determining hospital performance based on rank ordering: is it appropriate? *American Journal of Medical Quality*, 22(3):177–185, 2007.

P. C. Austin, D. A. Alter, and J. V. Tu. The use of fixed-and random-effects models for classifying hospitals as mortality outliers: a monte carlo assessment. *Medical decision making*, 23(6):526–539, 2003.

D. Blumenthal. Part i: Quality of care: What is it? *The New England Journal of Medicine*, 335(12):891–894, 1996.

A. Boomsma and J. J. Hoogland. The robustness of lisrel modeling revisited. *Structural equation models: Present and future. A Festschrift in honor of Karl Jöreskog*, pages 139–168, 2001.

A. C. Cameron and P. K. Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005.

CBS. *Centraal Bureau voor de Statistiek. Gezondheid en zorg in cijfers 2012*. CBS, 2012.

G. Chamberlain. Analysis of covariance with qualitative data, 1982.

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)*, pages 1–38, 1977.

D. A. Follmann and D. Lambert. Identifiability of finite mixtures of logistic regression models. *Journal of Statistical Planning and Inference*, 27(3):375–381, 1991.

R. Galbraith and P. Green. Estimating the component ages in a finite mixture. *International Journal of Radiation Applications and Instrumentation. Part D. Nuclear Tracks and Radiation Measurements*, 17(3):197–206, 1990.

L. G. Glance, A. Dick, T. M. Osler, Y. Li, and D. B. Mukamel. Impact of changing the statistical methodology on hospital and surgeon ranking: the case of the new york state cardiac surgery report card. *Medical care*, 44 (4):311–319, 2006.

W. H. Greene. *Econometric Analysis*. Prentice Hall, 2008.

J. A. Hausman. Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, pages 1251–1271, 1978.

C. W. Hukkelhoven, E. W. Steyerberg, J. D. F. Habbema, E. Farace, A. Marmarou, G. D. Murray, L. F. Marshall, and A. I. Maas. Predicting outcome after traumatic brain injury: development and validation of a prognostic score based on admission characteristics. *Journal of neurotrauma*, 22(10):1025–1039, 2005.

R. Jacobs, M. Goddard, and P. C. Smith. How robust are hospital ranks based on composite performance measures? *Medical care*, 43(12):1177–1184, 2005.

B. Jennett and M. Bond. Assessment of outcome after severe brain damage: a practical scale. *The Lancet*, 305(7905): 480–484, 1975.

E. Katz. Bias in conditional and unconditional fixed effects logit estimation. *Political Analysis*, 9(4):379–384, 2001.

R. Lilford, M. A. Mohammed, D. Spiegelhalter, and R. Thomson. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *The Lancet*, 363(9415):1147–1154, 2004.

H. F. Lingsma. *Measuring quality of care: methods and applications to acute neurological diseases*. Erasmus University Rotterdam, 2010.

H. F. Lingsma, B. Roozenbeek, B. Li, J. Lu, J. Weir, I. Butcher, A. Marmarou, G. D. Murray, A. I. Maas, and E. W. Steyerberg. Large between-center differences in outcome after moderate and severe traumatic brain injury in the international mission on prognosis and clinical trial design in traumatic brain injury (impact) study. *Neurosurgery*, 68(3):601–608, 2011.

W. Lutz, W. Sanderson, and S. Scherbov. The coming acceleration of global population ageing. *Nature*, 451(7179): 716–719, 2008.

A. Marmarou, J. Lu, I. Butcher, G. S. McHugh, G. D. Murray, E. W. Steyerberg, N. A. Mushkudiani, S. Choi, and A. I. Maas. Prognostic value of the glasgow coma scale and pupil reactivity in traumatic brain injury assessed prehospital and on enrollment: an impact analysis. *Journal of neurotrauma*, 24(2):270–280, 2007a.

A. Marmarou, J. Lu, I. Butcher, G. S. McHugh, N. A. Mushkudiani, G. D. Murray, E. W. Steyerberg, and A. I. Maas. Impact database of traumatic brain injury: design and description. *Journal of neurotrauma*, 24(2):239–250, 2007b.

G. D. Murray, I. Butcher, G. S. McHugh, J. Lu, N. A. Mushkudiani, A. I. Maas, A. Marmarou, and E. W. Steyerberg. Multivariable prognostic analysis in traumatic brain injury: results from the impact study. *Journal of neurotrauma*, 24(2):329–337, 2007.

N. A. Mushkudiani, D. C. Engel, E. W. Steyerberg, I. Butcher, J. Lu, A. Marmarou, F. Slieker, G. S. McHugh, G. D. Murray, and A. I. Maas. Prognostic value of demographic characteristics in traumatic brain injury: results from the impact study. *Journal of neurotrauma*, 24(2): 259–269, 2007.

R. Paap, P. H. Franses, and D. Van Dijk. Does africa grow slower than asia, latin america and the middle east? evidence from a new data-based classification method. *Journal of Development Economics*, 77(2):553–570, 2005.

P. Perel, M. Arango, T. Clayton, P. Edwards, E. Komolafe, S. Poccock, I. Roberts, H. Shakur, E. Steyerberg, et al. Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. *bmj*, 336(7641):425–9, 2008.

J. Ranstam, P. Wagner, O. Robertsson, and L. Lidgren. Health-care quality registers outcome-orientated ranking of hospitals is unreliable. *Journal of Bone & Joint Surgery, British Volume*, 90(12):1558–1561, 2008.

Rijksoverheid. Rijksoverheid. Kwaliteit van de zorg. 2013. URL `http://www.rijksoverheid.nl/onderwerpen/kwaliteit-van-de-zorg/kiezen-in-de-zorg`.

RIVM. kiesBeter. Wat is kiesBeter.nl? 2013. URL `http://www.kiesbeter.nl/algemeen/overkiesbeter/`.

SIRM. *Elsevier's De beste ziekenhuizen 2012*. Elsevier, 2012.

G. Teasdale and B. Jennett. Assessment of coma and impaired consciousness: a practical scale. *The Lancet*, 304 (7872):81–84, 1974.

M. Verbeek. *A guide to modern econometrics*, volume 2. John Wiley & Sons New York, 2004.

J. K. Vermunt. Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, 17(1):33–51, 2008.

ZN. Zorgverzekeraars Nederland. Organisatie. 2013. URL `https://www.zn.nl/over-zn/organisatie/`.

APPENDIX

A. DESCRIPTIVE STATISTICS PUPILLARY REACTIVITY AND GCS MOTOR SCORE

Motor	GSC	% patients
1	makes no movements	15
2	extension to painful stimuli	12
3	abnormal flexion to painful stimuli	13
4	flexion/withdrawal to painful stimuli	24
5	localizes painful stimuli	27
6	obeys comands	3
7	untestable	6
Pupil	**Reactivity**	
1	both reacting	65
2	one reacting	15
3	neither reacting	20

This table shows the distribution of patients over the categories in pupillary reactivity and GCS motor score

B. FIRST AND SECOND ORDER CONDITIONS USED IN THE EM ALGORITHM

We use the EM algorithm to estimate the parameters in our finite mixture model for clustering hospitals. In the maximization step of this algorithm we use the Newton-Rhapson method to maximize the expectation of the complete data log-likelihood. This method requires the gradient and hessian which contain the first order conditions and the second order conditions, respectively.

The first order conditions are given by:

$$\frac{\partial E_{s|y}[\mathcal{L}(y,s;\beta,\gamma_1,...,\gamma_J)]}{\partial \gamma_j} = \sum_{i=1}^{N}\sum_{t=1}^{T_i} \hat{w}_{ij}(y_{it} - F(\gamma_j + x_{it}'\beta)) = 0 \text{ for } j = 1,...,J$$

$$\frac{\partial E_{s|y}[\mathcal{L}(y,s;\beta,\gamma_1,...,\gamma_J)]}{\partial \beta} = \sum_{i=1}^{N}\sum_{t=1}^{T_i} x_{it}\sum_{j=1}^{J} \hat{w}_{ij}(y_{it} - F(\gamma_j + x_{it}'\beta)) = 0$$

The second order conditions are given by:

$$\frac{\partial^2 E_{s|y}[\mathcal{L}(y,s;\beta,\gamma_1,...,\gamma_J)]}{\partial \gamma_j^2} = -\sum_{i=1}^{N}\sum_{t=1}^{T_i} \hat{w}_{ij}\frac{e^{\gamma_j + x_{it}'\beta}}{(1+e^{\gamma_j + x_{it}'\beta})^2} \text{ for } j = 1,...,J$$

$$\frac{\partial^2 E_{s|y}[\mathcal{L}(y,s;\beta,\gamma_1,...,\gamma_J)]}{\partial \gamma_j \partial \gamma_q} = 0 \text{ for } j = 1,...,J, q = 1,...,J \text{ and } j \neq q,$$

$$\frac{\partial^2 E_{s|y}[\mathcal{L}(y,s;\beta,\gamma_1,...,\gamma_J)]}{\partial \gamma_j \partial \beta} = -\sum_{i=1}^{N}\sum_{t=1}^{T_i} \hat{w}_{ij}x_{it}\frac{e^{\gamma_j + x_{it}'\beta}}{(1+e^{\gamma_j + x_{it}'\beta})^2} \text{ for } j = 1,...,J$$

$$\frac{\partial^2 E_{s|y}[\mathcal{L}(y,s;\beta,\gamma_1,...,\gamma_J)]}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{N}\sum_{t=1}^{T_i} x_{it}x_{it}^T\sum_{j=1}^{J} \hat{w}_{ij}\frac{e^{\gamma_j + x_{it}'\beta}}{(1+e^{\gamma_j + x_{it}'\beta})^2}$$

C. RESULTS SIMULATIONS SIMPLE MAXIMUM LIKELIHOOD ESTIMATES

	β	$\bar{\hat{\beta}}$	$SD(\hat{\beta})$	$\hat{se}(\hat{\beta})$	p-value
age	0.040	0.038	0.002	0.002	0.000
motor 2	0.600	0.569	0.088	0.089	0.000
motor 3	-0.030	-0.027	0.084	0.083	0.280
motor 4	-0.700	-0.659	0.073	0.073	0.000
motor 5	-1.400	-1.322	0.074	0.076	0.000
motor 6	-1.500	-1.424	0.158	0.161	0.000
motor 7	-0.400	-0.372	0.109	0.107	0.000
pupil 2	0.800	0.757	0.062	0.064	0.000
pupil 3	1.500	1.422	0.064	0.063	0.000

This table shows the actual parameter value, the Monte Carlo mean, the Monte Carlo standard deviation, the average standard errors, and the p-value of the test on equality between the estimated and real parameter values of the maximum likelihood estimates of the logit model without fixed effects.