Journal of Data Science **10**(2012), 321-343

Variable Selection by sNML Criterion in Logistic Regression with an Application to a Risk-Adjustment Model for Hip Fracture Mortality

Antti Liski[1*], Ioan Tăbuş[1], Reijo Sund[2] and Unto Häkkinen[2]
[1] *Tampere University of Technology and*
[2] *National Institute for Health and Welfare*

Abstract: When comparing the performance of health care providers, it is important that the effect of such factors that have an unwanted effect on the performance indicator (eg. mortality) is ruled out. In register based studies randomization is out of question. We develop a risk adjustment model for hip fracture mortality in Finland by using logistic regression. The model is used to study the impact of the length of the register follow-up period on adjusting the performance indicator for a set of comorbidities. The comorbidities are congestive heart failure, cancer and diabetes. We also introduce an implementation of the minimum description length (MDL) principle for model selection in logistic regression. This is done by using the normalized maximum likelihood (NML) technique. The computational burden becomes too heavy to apply the usual NML criterion and therefore a technique based on the idea of sequentially normalized maximum likelihood (sNML) is introduced. The sNML criterion can be evaluated efficiently also for large models with large amounts of data. The results given by sNML are then compared to the corresponding results given by the traditional AIC and BIC model selection criteria. All three comorbidities have clearly an effect on hip fracture mortality. The results indicate that for congestive heart failure all available medical history should be used, while for cancer it is enough to use only records from half a year before the fracture. For diabetes the choice of time period is not as clear, but using records from three years before the fracture seems to be a reasonable choice.

Key words: Code length, hip fracture, logistic regression, maximum likelihood.

1. Introduction

*Corresponding author.

Profiling medical care providers on the basis of quality of care and utilization of resources has become a widely used analysis in health care policy and research. A major initiative to evaluate hospital performance in the United States was launched by the Health Care Financing Administration (HCFA) in 1987 with the annual release of hospital-specific data comprising observed and expected mortality rates for Medicare patients. Hospitals observed to have higher-than-expected mortality rates were flagged as institutions with potential quality problems. HCFA derived mortality rates by estimating a patient-level model of mortality for disease-based cohorts using administrative data (Normand, Glickman and Gatsonis, 1997).

Risk-adjustment is desirable when comparing hospitals or hospital districts with respect to a performance indicator such as mortality. Adjustment is intended to account for possible differences in patient case mix (Iezzoni, 1994; Landon, Iezzoni, Ash, Shwartz, Daley, Hughes and Mackiernan, 1996; Salem-Schatz, Moore, Rucker and Pearson, 1994). The methodologic aspects of risk-adjustment have been extensively discussed in the literature on observational studies (see Rosenbaum, 2002 and references therein).

While using administrative register-based data, the comorbidities to be adjusted are typically identified from the data using the disease grouping rules defined in Charlson or Elixhauser indices (Quan, Sundararajan, Halfon, Fong, Burnand, Luthi, Saunders, Beck, Feasby and Ghali, 2005). A salient issue in adjusting performance indicators for patients' comorbidities using administrative data is to decide the length of comorbidity lookup period, i.e. to decide how far we have to go back in patient's history (recorded in the registers) in order to effectively identify comorbidities to be adjusted (Preen, Holman, Spilsbury, Semmens and Brameld, 2006). This is an important question, because all conditions might not affect the patient anymore after a certain amount of time has passed. Therefore looking back too far for a certain condition, might even make the adjustment worse. Another reason is the fact that we might have only a few years historical data available or that it is very costly to collect additional historical data. It is not desirable to collect expensive extra data if we get the same results with less information.

Often the evaluation of a risk-adjustment model for a binary response is done using the c-statistic (Iezzoni, 2003). In this approach, the probabilities estimated (typically) with logistic regression are used to predict a patient's status and the c-statistic measuring the concordance of predictions with the true events is calculated. However, accurate or inaccurate classification by c-statistic does not address the goodness of fit or the complexity of a (risk-adjustment) model (Hosmer and Lemeshow, 2000, Chapter 5). Even if the model is the correct one and thus fits very well, its classification performance may be poor. On the other hand,

the correct model may have bad fit (distances between certain observed and expected values are large) but the model still yields good classification. Clearly the aim in deciding the length of lookup period is not to find the best prediction for a single performance indicator in one data set, but to find good risk-adjustment models for further analysis. In this sense, the real model selection criteria provided should be used instead of c-statistics.

There are several traditional model selection criteria available, such as the Akaike (AIC) and the Bayesian (BIC) information criteria. Rissanen (1996) has proposed the so called minimum description length (MDL) principle which can be implemented through the normalized maximum likelihood (NML) framework. The NML distributions offer a philosophically superior approach for the model selection problem. Unfortunately, the implementation of the MDL principle for the model selection problem in logistic regression using the standard normalized maximum likelihood (NML) technique is computationally infeasible with large data sets.

This paper has two purposes. First, it develops a risk adjustment model for a binary response using logistic regression and examines the impact of the length of the register follow-up period on adjusting the performance indicator for a set of comorbidities. The second purpose of this paper is to introduce a new MDL-based model selection criterion following the idea of sequentially normalized maximum likelihood (sNML) that was recently proposed by Rissanen and Roos (2007). We show that the sNML criterion can be evaluated efficiently and it is applicable also to large models with large amounts of data by applying this criterion in the case of a risk-adjustment model for hip fracture mortality in Finland. In this case study, the focus is on the determination of the optimal length of the register follow-up periods for comorbidities. We also compare the results given by the sNML criterion with the corresponding results given by the traditional AIC and BIC model selection criteria.

2. Setting

2.1 Hip Fracture

Hip fracture is a common and important cause of mortality in the elderly population (Keene, Parker and Pryor, 1993). In Finland, the number of hip fractures was about 7000 per year between the years 1998-2002 (Sund, 2007). Not only patients suffer from hip fractures, but they also cause remarkable costs to society (Hannan, Magaziner, Wang, Eastwood, Silberzweig, Gilbert, Morrison, McLaughlin, Orosz and Siu, 2001).

The main objective in the treatment of hip fracture is to help the patient regain his/her pre-fracture health status and level of functional ability. Because a

successful treatment should make it possible that patients are able to continue life in the same fashion as before the fracture, death is obviously a very unsuccessful outcome. Although hip fracture itself doesn't usually cause death, it is often such a shock to the whole body that especially for elderly people in lowered physical state it may mean the "beginning of the end" (Heithoff and Lohr, 1990). If the hip fracture triggers the dying process, we may assume that short-term mortality is in fact an indicator that the patient's health status before the hip fracture was already substantially lowered.

Quite often the mortality indicators for hip fracture are selected to measure death within three months or one year after the fracture. Mortality is a well defined and easily observable indicator in the sense that there is typically no argument if a patient is dead or not. The 90 days mortality reflects the risk connected to the hip fracture treatment and one year mortality reflects more the overall condition of a patient than risk of death directly caused by the shock effect of the hip fracture event.

2.2 Adjusting Mortality with Comorbidities

In order to compare mortality indicators between different areas or in time, the differences or changes in the patient population must be risk-adjusted (Iezzoni, 2003). In other words, we wish to find factors that explain the mortality following hip fracture, measured as a binary variable, in order to obtain a set of covariates which profile a patient's medical condition at the time of the hip fracture. Our interest is in comorbidities that a patient has had before the hip fracture and which may have effect on the outcome of the treatment. The special focus in our study is to examine how far we have to follow the patients medical history, and various lengths of the follow-up period (180 days, 1 year, 3-, 5- and 10-years) are modeled in order to find the shortest period to effectively adjust for each comorbidity. For pragmatic reasons, only three comorbidities are used in this study: congestive heart failure, cancer and diabetes. Each time period and comorbidity is analyzed separately. The analysis for other comorbidities could be done in a similar fashion. On top of the comorbidities, age, hip fracture type and sex are considered as factors to be adjusted in our risk-adjustment model.

2.3 Data

The National Institute for Health and Welfare maintains a register which contains all in hospital care periods taken place in Finland. From this register all 50 year or older first time hip fracture patients were identified from the years $1999 - 2005$. We further excluded patients who were institutionalized before the fracture. This resulted in a total of $n = 28797$ patients. For these patients (back-

wards) hospitalization history was available up to 10 years before the fracture. This information was complemented with data obtained from the register maintained by the Social Insurance Institution of Finland. From this second register, information on drug reimbursements granted for the medication of the three comorbidities stated above, was obtained. The mortality was followed using the Causes of Death register of Statistics Finland. In our final data we have combined the information obtained from these three registers. It has been shown that the quality of Finnish register data on the case of hip fractures is good (Sund, Nurmi-Lüthje, Lüthje, Tanninen, Narinen and Keskimäki, 2007). The dataset is based on the data used in the PERFECT (PERFormance, Effectiveness and Cost of Treatment episodes) project in the National Institute for Health and Welfare in Finland.

Many basic characteristics can be straightforwardly extracted from the data. These include the date of hip fracture, sex, age, the type of hip fracture (subtrochanteric, trochanteric or femoral neck fracture), and the date of death. In addition, we used ten years of medical history to construct five variables for each comorbidity which scan different time periods before hip fracture. The time intervals of interest were 180 days, 1 year, 3 years, 5 years and 10 years before the fracture. There were two ways to get an indication for a comorbidity from our data. In the first we have data on a patient's all hospitalization preceding the hip fracture until a certain (historical) time point. Now if the patient has been hospitalized because of the chosen comorbidity between this time point and the hip fracture, we get indication that the patient has had that comorbidity. The second way to get indication for a comorbidity comes through information on drug reimbursements. Now we have to check if a patient has received the right for drug reimbursements for that comorbidity and that it was still valid when the hip fracture occurred. This means that if a patient has had the right for drug reimbursements when the hip fracture occurred, then the patient will have indication for that comorbidity for all time periods.

Let us take an example where we choose the 3 year time interval. This means we jump back three years in time from the hip fracture event. We now choose one patient whose hospitalization record we start following towards the hip fracture event. Assume the patient has been hospitalized because of cancer for three weeks two years before the hip fracture. Now this patient will be identified for cancer based on the information on hospital care records. It is also checked if the patient has a right for drug reimbursements for some of the three comorbities that we are interested in at the moment of hip fracture. Say we find out that the patient has the right for drug reimbursements because of cancer but also for congestive heart failure. Therefore this patient receives indication for cancer (based on information from both registers) and congestive heart failure with a

three year lookback period.

The setting is actually quite challenging from the model selection point of view, since the number of the occurrences of a disease does not increase much when the length of inspection period increases. If we change our view for example from 180 days to one year before the fracture, the increase in the number of occurrences is typically small. Therefore it may be difficult to distinguish between models that use different time period variables. Further, if we look further back in history, more occurrences appear, but the effect of these occurrences on the dependent variable may become weaker, and we assume that this time dependence may not be same for all comorbidities.

3. Modeling Mortality with Logistic Regression

With n patients, we define $y_t = 1$ if the tth patient died within a 90 days period after the hip fracture and $y_t = 0$ otherwise (A corresponding model for the 365 days mortality is also analysed). We treat the n binary outcome variables $y_1, \cdots, y_n$ as independent. Let

$$\pi(\mathbf{x}_t; \boldsymbol{\beta}) = P(y_t = 1), \quad t = 1, \cdots, n,$$

and assume that

$$\log \frac{\pi(\mathbf{x}_t; \boldsymbol{\beta})}{1 - \pi(\mathbf{x}_t; \boldsymbol{\beta})} = \boldsymbol{\beta}^T \mathbf{x}_t, \tag{1}$$

where $\mathbf{x}_t = (x_{t1}, \cdots, x_{tk})^T$ is the vector of k covariate values of the tth patient.

The covariates (comorbidity, age, sex and hip fracture type) are such that there are many patients with the same values of covariates. For example, we may take women patients in the age group 50-69 who suffered a subtrochanteric fracture and had diabetes (inspection period one year before fracture). Let n_1 denote the number of such patients. Consequently, these patients have the same value of covariates, say $\mathbf{x}_1$, and hence the probability $\mathrm{P}(y_t = 1)$ is $\pi(\mathbf{x}_1; \boldsymbol{\beta})$ for all these n_1 patients. We say that $\mathbf{x}_1$ is the setting 1 of values of k covariates. We have only l different settings $\mathbf{x}_1, \cdots, \mathbf{x}_l$ and the number of different setting l is much smaller than n. Let n_i denote the number of the patients with the setting i, and hence we have $n = n_1 + \cdots + n_l$.

3.1 Bernoulli Likelihood

For notational convenience, we assume here that the observations are ordered such that the different settings $\mathbf{x}_1, \cdots, \mathbf{x}_l$ come first, i.e. for each $t > l$ there exists $i \leq l$ such that $\mathbf{x}_t = \mathbf{x}_i$. Since the y_t are independent and Bernoulli distributed,

the likelihood is

$$L(\boldsymbol{\beta}; \mathbf{x}_1, \cdots, \mathbf{x}_n) = \prod_{t=1}^{n} \pi(\mathbf{x}_t; \boldsymbol{\beta})^{y_t} [1 - \pi(\mathbf{x}_t; \boldsymbol{\beta})]^{1-y_t}$$
$$= \prod_{i=1}^{l} \pi(\mathbf{x}_i)^{v_i} [1 - \pi(\mathbf{x}_i)]^{n_i - v_i}, \tag{2}$$

where $v_i = \sum_{t:x_t = x_i} y_t$ is the number of deaths among the patients with the setting $\mathbf{x}_i$, $i = 1, \cdots, l$. Therefore it is sufficient to record the number of observations n_i and the number of deaths v_i corresponding to the settings $i = 1, \cdots, l$. Then v_i refers to this death count rather than to an individual binary response. We will use logistic regression (DeLong *et al.*, 1997) to assess from register data how much of medical history before fracture is needed in order to get sufficient indication of comorbidity effects.

3.2 Model Selection in Logistic Regression

Let Γ be the set of all $1 \times k$ vectors of the form $\gamma = (\gamma_1, \cdots, \gamma_k)$, where $\gamma_j = 0$ or 1 for $j = 1, \cdots, k$. There are 2^k such vectors in Γ. A variable selection procedure is then equivalent to first selecting $\gamma \in \Gamma$. If $\gamma_j = 1$, the variable x_j, $1 \leq j \leq k$ is selected and the corresponding β_j is estimated, otherwise $\gamma_j = 0$ and $\beta_j = 0$, i.e. x_i is not selected. Let $\boldsymbol{\beta}_\gamma = \text{diag}[\gamma]\boldsymbol{\beta}$, where $\text{diag}[\gamma]$ is the $k \times k$ diagonal matrix with diagonal elements γ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_k)^T$ is the k-dimensional parameter vector. In our application we will consider a certain subset of models from Γ (See Section 4.1) and compare them using the model selection criteria NML, AIC and BIC.

It follows from assumption (1) and the likelihood (2) that the log likelihood function of $\boldsymbol{\beta}_\gamma$ equals

$$l(\boldsymbol{\beta}_\gamma) = \sum_{i=1}^{l} v_i \, \boldsymbol{\beta}_\gamma^T \mathbf{x}_i - \sum_{i=1}^{l} n_i \, \log[1 + \exp(\boldsymbol{\beta}_\gamma^T \mathbf{x}_i)]. \tag{3}$$

The likelihood equations result from setting $\partial l(\boldsymbol{\beta}_\gamma)/\partial \boldsymbol{\beta}_\gamma = 0$, and they may be written in the form

$$\mathbf{X}^T \mathbf{v} = \mathbf{X}^T \hat{\boldsymbol{\mu}},$$

where $\mathbf{v} = (v_1, \cdots, v_l)^T, \mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_l)^T$and $\hat{\mu}_i = n_i \pi(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_\gamma)$, $i = 1, \cdots, l$. The equations are nonlinear and require iterative solution. The likelihood equations equate the sufficient statistics to the estimate of their expected values. This is a fundamental result for generalized linear models with canonical link (see eg. McCulloch and Searle 2001, Chapter 5).

4. The MDL Principle and the NML Criterion for Logistic Regression

4.1 Normalized Maximum Likelihood

Rissanen (1996) proposed his normalized maximum likelihood (NML) distribution as a theoretical basis for statistical modeling. The NML distribution for (2) may now be written as

$$\hat{P}(\mathbf{v}|\gamma) = L[\hat{\boldsymbol{\beta}}_\gamma(\mathbf{v})|\ \mathbf{v}; \mathbf{X}]/C(\gamma), \tag{4}$$

where $L[\hat{\boldsymbol{\beta}}_\gamma(\mathbf{v})|\ \mathbf{v}; \mathbf{X}]$ is the maximum of the likelihood function and

$$C(\gamma) = \sum_{\mathbf{v}\in\Omega} L[\hat{\boldsymbol{\beta}}_\gamma(\mathbf{v})|\ \mathbf{v}; \mathbf{X}] \tag{5}$$

is the normalizing constant. In (5) Ω denotes the sample space and the sum runs over all different count vectors $(v_1, \cdots, v_l)$ such that $0 \leq v_1 + \cdots + v_l \leq n$ and $v_i \geq 0,\ i = 1, \cdots, l$. The notation $\hat{\boldsymbol{\beta}}_\gamma(\mathbf{v})$ emphasizes the obvious fact that the ML estimate $\hat{\boldsymbol{\beta}}_\gamma$ is a function of $\mathbf{v}$.

There is a correspondence between so called prefix codes and probability distributions (Rissanen, 2007, Chapter 2). Let $P(\mathbf{v}|\ \boldsymbol{\beta}_\gamma)$ be the probability of $\mathbf{v}$. Then there exists a prefix code for $\mathbf{v}$ with ideal code length $\log[1/P(\mathbf{v}|\ \boldsymbol{\beta}_\gamma)] = -\log P(\mathbf{v}|\ \boldsymbol{\beta}_\gamma)$. So, every distribution defines a prefix code. After observing $\mathbf{v}$, the shortest code length is $\log(1/P[\mathbf{v}|\ \hat{\boldsymbol{\beta}}_\gamma(\mathbf{v})])$. Clearly the maximum likelihood $P[\mathbf{v}|\ \hat{\boldsymbol{\beta}}_\gamma(\mathbf{v})]$ is not a probability distribution of $\mathbf{v}$, and therefore it does not define a prefix code for $\mathbf{v}$. However, the NML distribution (4) defines a prefix code which has important optimality properties (see eg. Barron, Rissanen and Yu, 1998).

4.2 The Minimum Description Length Principle

Rissanen (1996) considers the NML distribution in the context of coding and modeling theory and takes

$$-\log \hat{P}(\mathbf{v}|\gamma) = -l[\hat{\boldsymbol{\beta}}_\gamma(\mathbf{v})|\ \mathbf{v}; \mathbf{X}] + \log C(\gamma) \tag{6}$$

as the "shortest code length" for the data $\mathbf{v}$ that can be obtained with the model γ and calls it the stochastic complexity of $\mathbf{v}$, given γ. The first term in (6) is the minimized negative log likelihood, and the second term is called parametric complexity. In essence, $-\log \hat{P}(\mathbf{v}|\gamma)$ is the minimum of the penalized log likelihood function. The minimized negative log likelihood measures goodness of fit to the data, while $\log C(\gamma)$ penalizes the complexity of the model γ. From the coding

theoretic point of view, $-\log \hat{P}(\mathbf{v}|\gamma)$ is the length of the prefix code defined by the NML distribution.

Here we consider the class of logistic regression models defined by the 2^k subsets of covariates Γ and the logistic probabilities. The aim of model selection is to pick the optimal model γ from the set Γ. For given data $\mathbf{v}$, the NML function (4) attains its maximum and the "code length" (6) its minimum at the same value of γ. According to *the MDL* (Minimum Description Length) principle (Rissanen, 2007, Chapter 8) we select the model $\hat{\gamma}$ from Γ that minimizes the stochastic complexity (6). Since $\hat{\gamma}$ maximizes (4), we may call it the NML estimate of γ within the model class Γ.

The code length interpretation of (6) provides an illustrative yardstick to compare models. The data can be considered as a sequence of zeros and ones $0010100\cdots0010$, where 1 refers to "death". The upper limit of the code length is the length $n = 28797$ of the whole sequence. If a model will capture the regular features of data well, then the prefix code based on the NML distribution (4) can compress the data sequence. Our optimal logistic regression risk adjustment model compresses the data sequence into a sequence whose length is about half of the upper limit 28797. No actual coding is needed, of course, but the stochastic complexity of a model is computed.

Unfortunately, the computational burden becomes too heavy to determine the value of $C(\gamma)$ for logistic regression models with moderate number of covariates when n is large. Let k_γ denote the number of covariates in the model γ and l_γ the number of different settings of covariate values in the data under the model γ. Then the sum in (5) runs over all different count vectors $(v_1, \cdots, v_{l_\gamma})$ such that $0 \le v_1 + \cdots + v_{l_\gamma} = v_\gamma \le n$ and $0 \le v_i \le n_i,\ i = 1, \cdots, l_\gamma$, where $n = 28797$. Let γ be a model with two covariates ($k_\gamma = 2$), say. When the covariates are dichotomous, there are 2^2 possible covariate settings. Suppose that in the data occur only the settings $(0,0), (1,0)$ and $(0,1)$, and hence $l_\gamma = 3$. Then v_γ takes the values $0, 1, \cdots, n$ and for each v_γ the the count vectors are obtained by determining all different partitionings of v_γ into (v_1, v_2, v_3) such that $v_1 + v_2 + v_3 = v_\gamma, 0 \le v_i \le n_i,\ i = 1, 2, 3$ and $n_1 + n_2 + n_3 = n$. The ML estimate has to be computed for each count vector. It is obvious that the computation of the code length for just one model is excessive not to mention the situation where we wish to compare several models.

Tabus and Rissanen (2006) presented an algorithm for the computation of the stochastic complexity (6) for logistic regression. If the number of covariates is $k = 3$, say, their algorithm is practical only in cases with a maximum of a few hundred observations. The sequentially normalized ML technique will decrease computational burden dramatically, and consequently it makes the MDL model selection practical also for models with large k and n.

4.3 Sequential NML

The sequentially normalized maximum likelihood was introduced by Roos and Rissanen (2007). This approach has the advantage that the normalizing constant is much easier to compute than in the case of the standard NML. Now we only need to normalize over the last observation, which simplifies computations substantially. On the other hand, if we don't have a strict order for the data, we have to choose one, and this ordering has naturally an effect on the results.

Roos and Rissanen (2008, equation 4) presented the sequentially normalized maximum likelihood (sNML) function. Let $X^n = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ denote the regressor matrix and $\mathbf{y}^n = (y_1, \cdots, y_n)$ a sequence of the binary outcome variables. Note that here $\mathbf{x}_i$ denotes the regressor vector of the ith patient and X^n may contain identical regressor vectors unlike $\mathbf{X}$ in the model described in (3). In the logistic regression case, the sNML function may be written as

$$\hat{P}(\mathbf{y}^n|X^n) = \hat{P}(\mathbf{y}^m|X^m) \prod_{t=m+1}^{n} \hat{P}(y_t|\mathbf{y}^{t-1}, X^t), \tag{7}$$

where $\hat{P}(\mathbf{y}^n|X^n)$ is the estimated probability to observe the string $\mathbf{y}^n$ having observed X^n.

The last term from (7) is the NML function for y_t

$$\hat{P}(y_t|\mathbf{y}^{t-1}, X^t) = \frac{P(y_t|\mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}(\mathbf{y}^t))}{K(\mathbf{y}^{t-1})}, \tag{8}$$

where

$$K(\mathbf{y}^{t-1}) = P(y_t = 0|\mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}_0) + P(y_t = 1|\mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}_1) \tag{9}$$

is the normalizing constant.

Here $\hat{\boldsymbol{\beta}}_i$, denotes the ML estimates of $\boldsymbol{\beta}$ from the binary outcome vector $(\mathbf{y}^{t-1}, i)$, $i = 0, 1$ respectively. As can be seen from (8) we only normalize over the last observation which simplifies the computation of the normalizing constant compared to the standard NML.

Because the observations are independent, we have

$$P(y_t = i|\mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}_i) = \frac{(e^{\hat{\boldsymbol{\beta}}_i^T \mathbf{x}_t})^i}{1 + e^{\hat{\boldsymbol{\beta}}_i^T \mathbf{x}_t}}, \qquad i = 0, 1,$$

and (8) becomes

$$\hat{P}(y_t = i|\mathbf{y}^{t-1}, X^t) = \left(\frac{(e^{\hat{\boldsymbol{\beta}}_i^T \mathbf{x}_t})^i}{1 + e^{\hat{\boldsymbol{\beta}}_1^T \mathbf{x}_t}} \right) / \left(\frac{1}{1 + e^{\hat{\boldsymbol{\beta}}_0^T \mathbf{x}_t}} + \frac{e^{\hat{\boldsymbol{\beta}}_1^T \mathbf{x}_t}}{1 + e^{\hat{\boldsymbol{\beta}}_1^T \mathbf{x}_t}} \right), \qquad i = 0, 1.$$

The negative logarithm of the sNML function (7) is

$$\begin{aligned} -\log \hat{P}(\mathbf{y}^n|X^n) &= -\log \hat{P}(\mathbf{y}^m|X^m) - \sum_{t=m+1}^{n} \log \hat{P}(y_t|\mathbf{y}^{t-1}, X^t) \\ &= -\log \hat{P}(\mathbf{y}^m|X^m) - \sum_{t=m+1}^{n} \log P(y_t|\mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}(\mathbf{y}^t)) \\ &\quad + \sum_{t=m+1}^{n} \log K(\mathbf{y}^{t-1}). \end{aligned} \tag{10}$$

The computational burden of $\sum_{t=m+1}^{n} \log K(\mathbf{y}^{t-1})$ in (10) is trivial contrary to the computation of $\log[C(\gamma)]$ in (4). Note that $-\log \hat{P}(\mathbf{y}^n|X^n)$ can be interpreted as the code length for data when a given model is used, as explained in Subsection 4.2.

4.4 Individual Code Lengths

Taking the negative base two logarithm of (7), yields

$$-\log_2 \hat{P}(\mathbf{y}^n|X^n) = -\log_2 \hat{P}(\mathbf{y}^m|X^m) - \sum_{t=m+1}^{n} \log_2[\hat{P}(y_t|\mathbf{y}^{t-1}, X^t)],$$

where the last term is just a sum of the code lengths of individual observations from $m+1$ to n. Thus we are able to consider the contribution of individual observations to the total code length. Let $\mathcal{S}$ denote a subsequence $s_1, s_2, \cdots, s_v$ of the sequence $m+1, m+2, \cdots, n$ of indices. Thus $m+1 \leq s_1 < s_2 < \cdots < s_v \leq n$, where $v \leq n-m$ is the number of indices in S. Next take a single index $s \in S$ and let X^s denote the sequence $(\mathbf{x}_1, \cdots, \mathbf{x}_m, x_{m+1}, \cdots, x_{s-1}, x_s)$. The sequence X^s has s elements, $m+1 \leq s \leq n$. In a similar fashion $\mathbf{y}^s$ denotes the corresponding sequence of binary outcomes $(y_1, \cdots, y_m, y_{m+1}, \cdots, y_s)$.

We may now compare how changing explanatory variables affects the code length. Let X_1 and X_2 be two different sets of explanatory variables. The change in code length y_s (or description for y_s) is obtained by computing

$$\log_2[\hat{P}(y_s|\mathbf{y}^{s-1}, X_2^s)] - \log_2[\hat{P}(y_s|\mathbf{y}^{s-1}, X_1^s)].$$

By summing up these individual differences over $\mathcal{S}$ we obtain

$$d_S(X_1, X_2) = \sum_{s \in S} \{\log_2[\hat{P}(y_s|\mathbf{y}^{s-1}, X_2^s)] - \log_2[\hat{P}(y_s|\mathbf{y}^{s-1}, X_1^s)]\}, \tag{11}$$

which tells us how much the observations belonging to $\mathcal{S}$ affect the total code length as we switch our set of explanatory variables from X_2 to X_1. Let X_1^C be the comorbidity variable "cancer" using one year of a patient's medical history, and $X_{1/2}^C$ the corresponding variable using a half year medical history. Then $d_S(X_1^C, X_{1/2}^C)$ gives the change of the code length when the patients belong to the set S and one year of the medical history is used instead of half a year when cancer is the comorbidity variable. Here we may understand as well that there are also other variables in the model but only the variable X_1^C is changed to $X_{1/2}^C$.

We note that (11) is generally not the same as

$$\log_2 \hat{P}(\mathbf{y}^n|X_2^n) - \log_2 \hat{P}(\mathbf{y}^n|X_1^n). \tag{12}$$

In the case where $\mathcal{S}$ is the full sequence of $n-m$ indices $m+1, m+2, \cdots, n$, we have equality between (11) and (12).

4.5 Nonconstant Covariate Effects

If we assume that the effects of covariates may change over time, the calculation for the code length of each observation should be done by using an appropriately selected subdata from the near past. One choice is to slide a window over the data. The sNML approach is suitable for this purpose, although not without problems. Let us consider a window of w observations. Now to encode the whole data, we need to calculate first the 'regular' NML code length for the first w observations (term $\hat{P}(\mathbf{y}^m|X^m)$ in (7)). If now w is large, we face the same problems as before in the calculation of the normalizing constant (5). In the case study, we have circumvented this problem by just focusing on the comparison of the code length calculated for the patients with the indices $[501, 28797]$. This way we are using the information from the first 500 observations in encoding, but we do not include the cost of coding of the first 500 observations in the total code length.

5. Statistical Analysis

We analyse Finnish register data on hip fracture patients from the years 1999-2005. The data was described in Section 2.3. We have two binary outcome variables, the 90 days mortality and the 365 days mortality, with possible outcomes 1 (died) and 0 (alive). These mortality variables indicate if the patient has died within the 90 days (or the 365 days, respectively) period after the hip fracture.

The basic set of explanatory variables consists of five constructed dichotomous variables. From the national registries we have information on the hip fracture

type categorized in three classes, trochanteric, subtrochanteric and femoral neck fractures. Patients are classified into three age groups (50-69, 70-89, 90-). We use five dummy variables: two dummies for the hip fracture type, two dummies for age and one dummy for sex.

The outcome variables 90 days mortality and one year mortality will be modeled separately. The comorbidities of interest, congestive heart failure, cancer and diabetes are measured in five time intervals. Therefore we have fifteen comorbidity variables. The five basic explanatory variables are included in all models. In addition to them, one comorbidity variable is selected from the set of 15 comorbidity variables, giving 15 alternative models with six explanatory variables and a constant in each model.

We will do the analysis under two different assumptions: (1) the covariate effects change over time and (2) the covariate effects stay constant. Under the second assumption, we utilize the full medical history at each point in the computation of the sNML criterion. Under the first assumption, a sliding window technique is used.

We compute sNML with $m = 25$ (see (7)). The value $m = 25$ was chosen to make sure we have enough dead and alive patients in the initial calculation of sNML (done with the regular NML) (see Albert and Anderson, 1984). We cannot estimate a model if we only have for example dead patients in our data.

When using the sliding window, we use only a limited number of past observations to calculate the code length for an observation y_t. Now (7) becomes

$$\hat{P}(\mathbf{y}^n|X^n) = \hat{P}(\mathbf{y}^m|X^m) \prod_{t=m+1}^{n} \hat{P}(y_t|\mathbf{y}^{t-w-1,\cdots,t-1}, X^{t-w-1,\cdots,t}), \qquad (13)$$

where w is the window length and $\mathbf{y}^{t-w-1,\cdots,t-1} = (y_{t-w-1}, \cdots, y_{t-1})$ and $X^{t-w-1,\cdots,t-1}$ is the corresponding regressor matrix. In our calculations with the sliding window we take $m = 500$ and drop the term $\hat{P}(\mathbf{y}^m|X^m)$ from our code lengths because it is not possible to calculate the regular NML with 500 observations. In our setting $\hat{P}(\mathbf{y}^m|X^m)$ with $m = 500$ is always constant (or very close to constant) between the different models so omitting it doesn't really make a difference to our comparisons as they are done between different time periods of a comorbidity.

As we increase the time backwards from the hip fracture event, we get more occurrences for each comorbidity. This means that in the data some 0's of explanatory variables turn into 1's, but otherwise the data stays exactly the same when increasing the time period.

Let $\mathcal{A}$ be the set of indices of the observations that change as we switch from the comorbidity variable X_{t_1} to X_{t_2}. The subindices t_1 and t_2 indicate the length of the time periods that we look backwards from the hip fracture

event. Now the length of time period 1 is less than time period 2. By (11), we compute $d_{\mathcal{A}}(X_{t_1}, X_{t_2})$ to obtain the change in total code length due to changing observations.

5.1 Results

We observe that in Table 1 all model selection criteria give results consistent with each other. The AIC and BIC values were calculated from the whole data (not sequentially) to show that in this case the sequential and non-sequential approaches give similar results. The formulas for AIC and BIC are

$$\text{AIC} = -2 \log P(\mathbf{y}^n | X^n; \hat{\boldsymbol{\beta}}) + 2k$$

and

$$\text{BIC} = -2 \log P(\mathbf{y}^n | X^n; \hat{\boldsymbol{\beta}}) + k \log n,$$

where k is the number of estimated parameters in the model (see eg. Burnham & Anderson, 2002, Chapter 6). These criteria can be easily calculated in the case of logistic regression model.

C-statistic values (calculated from the whole data) are reported because they are often used in this kind of analysis. Also notice that the comorbidities seem to behave quite different from each other. Congestive heart failure works best as an explanatory variable if we use all of the data available to us. Cancer is a good explanatory variable for mortality with just information from 180 days preceding the fracture. In the case of diabetes, it is difficult to distinguish between time periods. There is very little variation in the values of the model selection criteria and the time periods from three to ten years give virtually the same values. However, all the model selection criteria except c-statistic seem to make the same choice of time period also for diabetes. We notice also that our models fit better the 90 days mortality than one year mortality.

Table 2 reports the number of occurrences of congestive heart failure (CHF), cancer and diabetes in each five time periods. The increase in occurrences is not very big compared to the size of the whole data (n=28797). This might be the reason why the model selection criteria do not clearly prefer any model over the others. Especially this is the case with diabetes. The maximum increase of occurrences is in congestive heart failure as we extend the period from 180 days to ten years (1376 occurrences). If we don't include any of the comorbidities in the model, we get the code length (sNML) of 15113 bits for the 90 days mortality and 21339 bits for one year mortality. Even though the time periods within diabetes do not differ from each other, they all clearly improve the models compared to the models without any comorbodities.

Table 1: Code lengths (sNML), AIC, BIC and c-statistic values for 30 mortality models for each comorbidity are given. The basic variables fracture type, age and sex are included in all models and exactly one of the 15 comorbidity variables is selected for each alternative model. Models for 90 day and one year (values in brackets) mortality are given

	CHF			
	sNML	AIC	BIC	c-statistic
180 days	14898 (21010)	20611 (29086)	20669 (29144)	0.6746 (0.6585)
1 year	14869 (20971)	20572 (29033)	20629 (29091)	0.6767 (0.6609)
3 years	14837 (20924)	20526 (28968)	20584 (29026)	0.6799 (0.6635)
5 years	14833 (20902)	20520 (28936)	20578 (28993)	0.6804 (0.6649)
10 years	**14825** (**20879**)	**20509** (**28904**)	**20567** (**28962**)	**0.6812** (**0.6658**)
	CANCER			
	sNML	AIC	BIC	c-statistic
180 days	14980 (**21011**)	**20719** (**29079**)	**20777** (**29137**)	0.6644 (**0.6599**)
1 year	**14979** (21013)	20721 (29085)	20779 (29143)	0.6647 (0.6560)
3 years	14992 (21026)	20738 (29103)	20796 (29160)	**0.6653** (0.6550)
5 years	14991 (21062)	20765 (29152)	20823 (29210)	0.6640 (0.6538)
10 years	15005 (21067)	20757 (29160)	20815 (29218)	0.6646 (0.6540)
	DIABETES			
	sNML	AIC	BIC	c-statistic
180 days	15093 (21278)	20882 (29457)	20940 (29515)	0.6549 (0.6409)
1 year	15091 (21273)	20879 (29449)	20937 (29507)	0.6553 (0.6414)
3 years	**15089** (**21272**)	**20877** (**29448**)	**20934** (**29506**)	0.6557 (0.6420)
5 years	15090 (21273)	20878 (29450)	20936 (29508)	0.6560 (0.6422)
10 years	15090 (21274)	**20877** (29451)	20935 (29509)	**0.6561** (**0.6423**)

Table 2: Number of occurences of the comorbidities within different time periods

time period	CHF	CANCER	DIABETES
180days	4654	2205	4064
1 year	4947	2470	4152
3 years	5570	2926	4305
5 years	5820	3237	4374
10 years	6030	3548	4420

In Table 3 we have reported the code lengths computed for the three comorbidities by using sliding windows of different lengths. The performance of sNML with various window lengths is close to that presented in Table 1, except for

90 days mortality with a window length of 25 observations and congestive heart failure as comorbidity. For diabetes the models with various time periods are still quite close to each other. Note, however, that Table 3 and Table 1 are not directly comparable because in the calculations for Table 3 we have omitted the code length for the first 500 observations.

Table 3: Code lengths (sNML) for 30 mortality models for each comorbidity (as in Table 1) using sliding windows of different lengths (25, 50, 100 and 500 observations). Code lengths for one year mortality are in brackets

	CHF			
	25 obs	50 obs	100 obs	500 obs
180 days	**18883** (23710)	17059 (22760)	15816 (21814)	14838 (20836)
1 year	**18883** (23689)	17044 (22728)	15793 (21785)	14811 (20801)
3 years	18935 (23669)	17033 (22685)	15761 (21736)	14783 (20755)
5 years	18950 (23673)	**17022** (22663)	15751 (21704)	14783 (20733)
10 years	18954 (**23658**)	17025 (**22646**)	**15744** (**21692**)	**14777** (**20712**)
	CANCER			
	25 obs	50 obs	100 obs	500 obs
180 days	**18676** (**23571**)	**17098** (**22640**)	**15901** (21785)	**14913** (20818)
1 year	18760 (23625)	17111 (22644)	15907 (**21782**)	14915 (**20817**)
3 years	18850 (23680)	17147 (22671)	15914 (21801)	14927 (20835)
5 years	18916 (23717)	17187 (22729)	15932 (21832)	14945 (20869)
10 years	18967 (23733)	17192 (22741)	15934 (21854)	14945 (20877)
	DIABETES			
	25 obs	50 obs	100 obs	500 obs
180 days	**19169** (24047)	17335 (23006)	16028 (22078)	15030(21110)
1 year	19180 (24042)	17326 (23005)	16021 (22068)	15027(21104)
3 years	19186 (**24036**)	17327 (22997)	**16017** (**22053**)	15023(21099)
5 years	19190 (24041)	**17323** (**22996**)	16021 (22056)	15023(21097)
10 years	19193 (24042)	17329 (22999)	16027 (22062)	**15022**(**21095**)

In Table 4 we have the change in code length within the subset of added occurrences and the whole data. With added occurrences we mean the observations that will become new occurrences of a comorbidity as we extend the time period. Let $\mathcal{A}$ denote the set of added occurrences as in Section 5. Note that for all different pairs of time periods (in connection of a given comorbidity) we have a different set of added occurrences. For example, let $X_{t_i}^{\mathrm{CHF}}$ be the comorbidity variable "congestive heart failure", when the period t_i of patients medical history before hip fracture is used. If $t_1 = 1/2$ year and $t_2 = 1$ year, then by (11) $d_{\mathcal{A}}(X_{t_1}^{\mathrm{CHF}}, X_{t_2}^{\mathrm{CHF}})$ is the first figure (24.3097) in the first row of Table 4.

Table 4: Differences of code lengths (sNML) for changing observations and for the whole data. In the table we have $d_{\mathcal{A}}(X_{t_1}, X_{t_2})$ and $d(X_{t_1}, X_{t_2})$ (see (11) and Section 5) values with different time periods for t_1 and t_2. If for example $t_1 = 1$ year and $t_2 = 10$ years, take congestive heart failure as comorbidity and 90 days mortality as outcome, then $d_{\mathcal{A}}(X_{t_1}, X_{t_2}) = 30.6144$. For the whole data $d(X_{t_1}, X_{t_2}) = 44.0480$. Values for one year mortality are in brackets

		CHF	
		ch obs	all
180 days	1 year	24.3097 (34.2059)	28.2030 (38.2309)
	3 years	46.8802 (70.0761)	60.9309 (85.1012)
	5 years	48.1942 (88.2988)	64.5818 (107.7948)
	10 years	52.6179 (104.8533)	72.2511 (130.1355)
1 year	3 years	23.9742 (37.7484)	32.7278 (46.8703)
	5 years	25.5732 (56.4925)	36.3788 (69.5639)
	10 years	30.6144 (74.1631)	44.0480 (91.9046)
3 years	5 years	1.6602 (19.9655)	3.6510 (22.6937)
	10 years	6.9804 (38.7493)	11.3202 (45.0343)
5 years	10 years	5.2479 (19.2118)	7.6692 (22.3407)
		CANCER	
		ch obs	all
180 days	1 year	-5.0550 (-9.0281)	0.3741 (-1.9093)
	3 years	-18.9243 (-25.5112)	-11.8064 (-14.5274)
	5 years	-35.3211 (-60.0382)	-31.4578 (-50.5906)
	10 years	-32.8788 (-65.0620)	-25.2902 (-55.9012)
1 year	3 years	-15.9025 (-20.8443)	-12.1805 (-12.6182)
	5 years	-33.0806 (-57.6566)	-31.8319 (-48.6814)
	10 years	-30.6938 (-63.7291)	-25.6643 (-53.9920)
3 years	5 years	-20.5161 (-41.9101)	-19.6514 (-36.0632)
	10 years	-18.1913 (-50.8998)	-13.4838 (-41.3738)
5 years	10 years	2.1115 (-12.5200)	6.1676 (-5.3106)
		DIABETES	
		ch obs	all
180 days	1 year	2.5761 (5.4076)	2.5049 (5.3779)
	3 years	4.3044 (5.6134)	4.1464 (6.3606)
	5 years	2.6456 (3.5228)	3.3646 (4.8051)
	10 years	2.9066 (2.6675)	3.7320 (4.1115)
1 year	3 years	1.7909 (0.2711)	1.6415 (0.9828)
	5 years	0.1771 (-1.8013)	0.8597 (-0.5727)
	10 years	0.4110 (-2.6583)	1.2271 (-1.2664)
3 years	5 years	-1.6343 (-2.1389)	-0.7818 (-1.5555)
	10 years	-1.3560 (-2.9672)	-0.4144 (-2.2492)
5 years	10 years	0.3228 (-0.8389)	0.3674 (-0.6937)

If $t_1 = 5$ years and $t_1 = 10$ years, then $d_{\mathcal{A}}(X_{t_1}^{\mathrm{CHF}}, X_{t_2}^{\mathrm{CHF}}) = 5.2479$ is the first figure in the tenth row of Table 4. It is understood here that the basic explanatory variables (hip fracture type, age and sex) and a constant are included in all models.

Table 4 shows the same tendency as the results in Table 1. We observe how much the code length changes within the subset of added occurrences and the whole data. In the case of congestive heart failure, increasing the time period shortens the code length among the added occurrences and also within the rest of the data. This means the added occurrences fit the data better with outcome value 1 than with value 0 and also improve the fit for observations coming after them.

Cancer behaves differently. There we can see that the difference in code length is larger for the subset of added occurrences than for the whole data. As we increase the time period, new occurrences worsen the overall model. As seen in Table 4, the increase in code length is largely due to the new occurrences.

For diabetes there are no big differences in code lengths between the time periods. Pairwise comparison in Table 4 shows that the improvement in code length is largest as we increase the time period from 180 days to 3 years. The comparison of three years to longer time periods indicates that we will not improve our model if we extend the time period. Again the differences between models were very small. The three year time period seems to be a reasonable choice also on basis of Table 4.

The worst code length for both of our mortality sequences is 28797, which would mean that we are not able to compress our original data at all. With the models used in this paper we obtain a code length which is approximately half of the worst code length. If we compress both of the mortalities with the Lempel-Ziv algorithm (Ziv and Lempel, 1978), we can get an idea of the size of the entropy for the sequences. With Lempel-Ziv the code length obtained for 90 days mortality is 3321 bits and 4219 bits for 365 days mortality. This means that both of the sequences could still be compressed much more than we were able to do with the logistic regression model. On the other hand, our compression with the logistic regression model uses information from the explanatory variables while Lempel-Ziv uses the sequence itself. Therefore comparison based solely on compression capability is not fair for the method presented in this paper.

6. Discussion

In this paper we have presented a sNML model selection criterion for logistic regression. The sequential approach enables us to compute the normalized maximum likelihood criterion also for large datasets. This was previously not possible for logistic regression models because of computational difficulties in the

normalizing constant of the NML criterion.

If the data doesn't have a natural ordering, we have to find one. This ordering should make sense from the applications perspective, which may be difficult in some cases. With the hip fracture data a natural (although not unique) ordering was obtained by using the date of arrival to hospital. The sequential approach also enables us to assume that the covariate effects develop over time. By using a sliding window in the calculation of the code length we are able to take this development in covariate effects into account and if necessary try to find the most suitable window length. If we approach the problem non-sequentially, we have to assume that covariate effects stood constant over the data.

Our objective was to find how far back in time we should look for three different comorbidities to get a good model for the mortality of hip fracture patients. We viewed each comorbidity separately from the other comorbidities. In our analysis we found out that we should use a different time period for each comorbidity. The results from sNML, AIC, BIC and c-statistic all pointed to the same direction for the choice of time period. This is a good result because the agreement of the different methods gives us stronger confirmation on the behavior of the comorbidities as explanatory variables. It also seems that in this case the sequential approach gives results which are in line to non-sequential approaches.

Our results indicate that for congestive heart failure we should use all medical history available to us, while for cancer it is enough to use only records from half a year before the fracture. For diabetes the message is not clear, but using records from three years before the fracture seems to be a reasonable choice. The results obtained by using a sliding window do not change our previous conclusions on the effect of different comorbidities. This suggests that there has not been any remarkable changes in covariate effects within the time period under consideration.

We were also able to distinguish how much of the change in codelength is due to the observations that become new indications of a comorbidity as we increase the time period that we look back in time. In congestive heart failure the fit of the whole data improves as we get new indications of that comorbidity. On the other hand, with cancer the model fits worse especially among the new cancer indications. Also this suggests that these two comorbidities behave in a quite different manner from each other.

All of the comorbidities improved the model. If we use the codelength obtained with Lempel-Ziv algorithm as a yardstick how far we are from the entropy, we can see that there is still a lot to improve. However, we do not want to lose interpretations about the explanatory variables effects on the outcome. Therefore we cannot construct a method aiming purely for maximum compression of data.

Acknowledgements

The work of Reijo Sund was financially supported by the Yrjö Jahnsson Foundation (grant, 5978).

Appendix

We give the algorithm for the computation of sNML (see (7)) in logistic regression. The mortality sequence $\{y_1, y_2, \cdots, y_t\}$ is denoted as $\mathbf{y}^t$ and the matrix $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_t\}$ of explanatory variables as X^t. Let n be the number of observations in the whole dataset and $y^{t|a}$ is the sequence of length t where the last observation $y_t = a$. First $m = r$ must be chosen in such a way that the sequentially maximized likelihood is finite (see e.g. Albert and Anderson, 1984). Then calculate the regular NML from the r observations. Denote this by $\hat{p}(\mathbf{y}^r|X^r)$. Compute the sequential part as:

0. Initialize $\Delta = -\log_2[\hat{p}(\mathbf{y}^r|X^r)]$
    1. For $i = (r+1) : n$

        1.1 Solve the ML-estimate $\hat{\boldsymbol{\beta}}_0$ by using $\mathbf{y}^{i|0}$ and X^i (see (3))

        1.2 Solve the ML-estimate $\hat{\boldsymbol{\beta}}_1$ by using $\mathbf{y}^{i|1}$ and X^i

        1.3 Compute

        $P(y_t = 0|\mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}_0(\mathbf{y}^t)) = 1/(1 + e^{\hat{\boldsymbol{\beta}}_0^T \mathbf{x}_t})$,

        $P(y_t = 1|\mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}_1(\mathbf{y}^t)) = (e^{\hat{\boldsymbol{\beta}}_1^T \mathbf{x}_t})/(1 + e^{\hat{\boldsymbol{\beta}}_1^T \mathbf{x}_t})$ and

        $K(\mathbf{y}^{t-1}) = P(y_t = 0|\mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}_0(\mathbf{y}^t)) + P(y_t = 1|\mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}_1(\mathbf{y}^t))$

        1.4 If $y_t = 0$ then

        $\Delta_i = -\log_2 P(y_t = 0|\mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}_0(\mathbf{y}^t)) + \log_2 K(\mathbf{y}^{t-1})$

        else

        $\Delta_i = -\log_2 P(y_t = 1|\mathbf{y}^{t-1}, X^t, \hat{\boldsymbol{\beta}}_0(\mathbf{y}^t)) + \log_2 K(\mathbf{y}^{t-1})$

        1.5 Set $\Delta = \Delta + \Delta_i$

The codelength for the data is Δ.

References

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1-10.

Barron, A., Rissanen J. and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory* **44**, 2743-2760.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference*: *A Practical Information-Theoretic Approach.* Springer, New York.

DeLong, E. R., Peterson, E. D., DeLong, D. M., Muhlbaier, L. H., Hacket, S. and Mark, D. B. (1997). Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine* **16**, 2645-2664.

Hannan, E. L., Magaziner J., Wang, J. J., Eastwood, E. A., Silberzweig, S. B., Gilbert, M., Morrison, R. S., McLaughlin, M. A., Orosz, G. M. and Siu, A. L. (2001). Mortality and locomotion 6 months after hospitalization for hip fracture: risk factors and risk-adjusted hospital outcomes. *Journal of the American Medical Association* **285**, 2736-2742.

Heithoff, H. A. and Lohr, K. N. (1990). Hip fracture: setting priorities for effective research. Report of a study by the Institute of Medicine, Division of Health Care Services, National Academy of Sciences, National Academy press, Washington, District of Columbia, 61-64.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd edition. Wiley, New York.

Iezzoni, L. I. (1994). Using risk-adjusted outcomes to assess clinical practice: an overview of issues pertaining to risk adjustment. *Annals of Thoracic Surgery* **58**, 1822-1826.

Iezzoni, L. I. (2003). Range of risk factors. In *Risk Adjustment for Measuring Health Care Outcomes* (Edited by L. I. Iezzoni), 3rd edition. Health Administration Press, Chicago.

Keene, G. S., Parker, M. J. and Pryor, G. A. (1993). Mortality and morbidity after hip fractures. *British Medical Journal* **307**, 1248-1250.

Landon, B., Iezzoni, L. I., Ash, A. S., Shwartz, M., Daley, J., Hughes, J. S. and Mackiernan, Y. D. (1996). Judging hospitals by severity-adjusted mortality rates: the case of CABG surgery. *Inquiry* **33**, 155-166.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear and Mixed Models.* Wiley, New York.

Normand, S-L. T., Glickman, M. E. and Gatsonis, C. A. (1997). Statistical methods for profiling providers of medical care: issues and applications. *Journal of the American Statistical Association* **92**, 803-814.

Preen, D. B., Holman, C. D. J., Spilsbury, K., Semmens, J. B. and Brameld, K. J. (2006). Length of comorbidity lookback period affected regression model performance of administrative health data. *Journal of Clinical Epidemiology* **59**, 940-946.

Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi, J. C., Saunders, L. D., Beck, C. A., Feasby, T. E. and Ghali, W. A. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care* **43**, 1130-1139.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory* **42**, 40-47.

Rissanen, J. (2007). *Information and Complexity in Statistical Modeling.* Springer, New York.

Rissanen J. and Roos T. (2007). Conditional NML universal models. In *Proceedings 2007 Information Theory and Applications Workshop, IEEE press*, 337-341.

Roos T. and Rissanen J. (2008). On sequentially normalized maximum likelihood models. *Workshop on Information Theoretic Methods in Science and Engineering* (*WITMSE-08*). Tampere, Finland, August 18-20.

Rosenbaum, P.D. (2002). *Observational Studies.* Springer, New York.

Salem-Schatz, S., Moore, G., Rucker, M. and Pearson, S. D. (1994). The case for case-mix adjustment in practice profiling: when good apples look bad. *Journal of the American Medical Association* **272**, 871-874.

Sund, R. (2007). Utilization of routinely collected administrative data in monitoring the incidence of aging dependent hip fracture. *Epidemiologic Perspectives & Innovations*, **4**:2.
http://www.epi-perspectives.com/content/4/1/2

Sund, R., Nurmi-Lüthje, I., Lüthje, P., Tanninen, S., Narinen, A. and Keskimäki, I. (2007). Comparing properties of audit data and routinely collected register data in case of performance assessment of hip fracture treatment in Finland. *Methods of Information in Medicine* **46**, 558-566.

Tabus, I. and Rissanen, J. (2006). Normalized maximum likelihood models for logit regression. In *Festschrift for Tarmo Pukkila on his 60th Birthday* (Edited by Liski, E. P., Isotalo, J., Niemelä, J., Puntanen, S., and Styan, G. P. H.), 159-172. University of Tampere, Department of Mathematics, Statistics and Philosophy, Report A 368.

Ziv, J. and Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory* **24**, 530-536.

Received September 22, 2010; accepted January 12, 2011.

Antti Liski
Department of Signal Processing
Tampere University of Technology
PO Box 553, 33101 Tampere, Finland
antti.liski@tut.fi

Ioan Tăbuş
Department of Signal Processing
Tampere University of Technology
PO Box 553, 33101 Tampere, Finland
ioan.tabus@tut.fi

Reijo Sund
Service Systems Research Unit
National Institute for Health and Welfare
THL, PO Box 30, FI-00271 Helsinki, Finland
reijo.sund@thl.fi

Unto Häkkinen
Center for Health Economics
National Institute for Health and Welfare
THL, PO Box 30, FI-00271 Helsinki, Finland
unto.hakkinen@thl.fi