Health Care Manag Sci (2010) 13:170–181
DOI 10.1007/s10729-009-9121-z

Analysing the length of care episode after hip fracture: a nonparametric and a parametric Bayesian approach

Jaakko Riihimäki · Reijo Sund · Aki Vehtari

Received: 26 June 2009 / Accepted: 3 November 2009 / Published online: 14 November 2009

Abstract Effective utilisation of limited resources is a challenge for health care providers. Accurate and relevant information extracted from the length of stay distributions is useful for management purposes. Patient care episodes can be reconstructed from the comprehensive health registers, and in this paper we develop a Bayesian approach to analyse the length of care episode after a fractured hip. We model the large scale data with a flexible nonparametric multilayer perceptron network and with a parametric Weibull mixture model. To assess the performances of the models, we estimate expected utilities using predictive density as a utility measure. Since the model parameters cannot be directly compared, we focus on observables, and estimate the relevances of patient explanatory variables in predicting the length of stay. To demonstrate how the use of the nonparametric flexible model is advantageous for this complex health care data, we also study joint effects of variables in predictions, and visualise nonlinearities and interactions found in the data.

Keywords Length of stay · Hip fracture · Bayesian analysis · Multilayer perceptron · Weibull mixture · Covariate relevance

1 Introduction

Accurate and relevant information concerning the performance of health system is essential for health care management as it helps to maximise the positive impact of health system on the health of both individual patients and communities, at a cost that is acceptable to those who must directly or indirectly finance health services [1, 2]. Recently, the idea of improving performance by identifying good treatment practices has become very popular [3]. However, this has been challenging in practice because of the lack of adequate data and appropriate methodology [4, 5].

Concrete quantitative measurement of performance has mainly been conducted using various performance indicators [6, 7]. Length of stay (LOS) has been characterized as an easily available indicator of hospital activity [8]. As an indicator length of stay is difficult to manipulate, and directly comparable across institutions, but as a concept it is a multifaceted one that reflects organizational patterns of care as well as the severity of patients. For an individual patient too short LOS may lead to immediate adverse outcomes, but also too long LOS may be harmful due to reduced ability to cope at home after discharge, especially among the elderly. At the service provider level, a long LOS typically indicates more complex patients, but may also

Grant sponsors: Finnish Funding Agency for Technology and Innovation (project TERANA), Yrjö Jahnsson Foundation 5978, Academy of Finland 125349.

J. Riihimäki (✉) · A. Vehtari
Department of Biomedical Engineering and Computational Science, Helsinki University of Technology—TKK,
P.O. Box 2200, 02015 Helsinki, Finland
e-mail: jaakko.riihimaki@tkk.fi

A. Vehtari
e-mail: aki.vehtari@tkk.fi

R. Sund
Service Systems Research Unit,
National Institute for Health and Welfare-THL,
P.O. Box 30, 00271 Helsinki, Finland
e-mail: reijo.sund@thl.fi

be a reflection of non-optimal operational efficiency. In practice, LOS has been considered as a relatively simple proxy for resource consumption, and it is an important measure for evaluation of the success of shifting care from costly inpatient care toward less expensive outpatient treatment [9]. In order to provide information that helps to more effectively plan the use of limited resources, it is fundamental to model length of stay accurately.

A skewed distribution is common for the length of stay data [9]. Due to the skewness, parametric models with long tails are typically applied for describing length of stay. For example, possible models for patient length of stay are Gamma, Weibull and Lognormal distributions [8], Coxian phase-type models [10], or mixtures of parametric distributions [11]. In cases when there are covariates available, the model parameters can be set to depend linearly on covariates. For instance, Faddy and McClean [12] showed how to introduce covariates through log-linear functions in a phase-type model. However, there are two restrictive assumptions often made with parametric approaches; the possible shapes of distributions are predefined, and the ways how covariates affect in a model are fixed in advance. Further, one easily faces problems if the interactions between covariates are introduced explicitly in the model, especially when the number of covariates is large. To loosen such assumptions, more flexible approaches, such as nonparametric neural networks have been suggested to be used in the analyses of length of stay distributions [13, 14]. However, the actual applications seem to be rare, and we are not aware of any systematic comparisons between the traditional parametric models and more flexible nonparametric alternatives in the case of modelling the length of stay distributions.

In this study we present Bayesian methodology to the modelling of length of stay distribution data that includes covariates. The modelling is investigated in the case of patient length of stay in care episode after a fractured hip, where the whole care episodes were reconstructed by using large register-based data extracted from the comprehensive administrative health registers. The aim of the study is to develop a flexible nonparametric multilayer perceptron (MLP) network model as well as a parametric Weibull mixture model to be used with the LOS data, and to compare the models and their predictive performances. The Bayesian approach (see, e.g. [15, 16]) is adopted throughout for both models, and the integration over parameter spaces is approximated with stochastic Markov chain Monte Carlo (MCMC) methods. Since the parameters in the MLP and the Weibull mixture model are incomparable, we demonstrate how it is possible to focus on observables rather than the model parameters, and to perform the actual comparisons by using posterior predictive checking. We show how the performances of the models can be evaluated by estimating expected utilities. As a utility measure, we use a predictive density for an independent test data. We also present pragmatic tools to find out the most relevant covariates in predicting the patient length of stay. The idea is to assess the average predictive sensitivities of covariates, and to study the joint predictive sensitivities of two covariates by measuring the change in a predictive distribution with the information theoretic Kullback-Leibler divergence when the values of two covariates are simultaneously changed. We also compare the covariate relevances given by both models, and demonstrate how the ability of models to capture the characteristics in the data can be visually evaluated.

The paper is organised as follows. Section 2 presents the register data on hip fractures. In Section 3 the nonparametric and parametric approaches and MCMC techniques are explained. Section 4 shows the results on posterior predictive checking, and conclusion is provided in Section 5.

2 Register data on hip fractures

Hip fractures are common injuries among the elderly. As the incidence of the injury is increasing, treatment and rehabilitation of the patients are likely to be a major challenge for the health systems in the near future [17]. The treatment of hip fractures virtually always requires a surgical operation at hospital, and a typically rather lengthy follow-up care at local rehabilitation facilities [18]. Length of stay is a major determinant in the cost of hip fracture treatment as less than one-fourth of the costs is caused by acute care [19]. It is also likely that the most differences between service providers in LOS following the hip fracture are due to patterns of care rather than characteristics of patients [20]. In addition, hip fracture has been characterized as a tracer condition in health systems, testing how well health and social services are integrated in the provision of acute care, rehabilitation, and continuing support for a large and vulnerable group of elderly patients [21]. In this sense, LOS following the hip fracture is a very attractive and important performance indicator.

The pragmatic problem with this indicator has been that the treatment of hip fracture patients typically consists of several phases in various facilities, and it may be challenging to obtain adequate data.

In this study, Finnish register data were used. The collection of data set has been described in detail

elsewhere [22]. In brief, a total population of patients diagnosed with a fractured hip (International Classification of Diseases revision 10 diagnosis codes S72.0, S72.1 and S72.2) in Finland between 1998–2001 was identified from the Finnish Health Care register. The medical histories of these patients during the years 1987–2002, including data on hospital care, nursing home care, and causes of deaths were obtained from the Finnish Hospital Discharge Register, the Finnish Health and Social Welfare Care Register, and the National Causes of Death Statistics using the unique national personal identity codes of the patient population. Each record in the data corresponded to one care period and included variables such as provider and area codes, dates of admission, operation and discharge, as well as diagnoses and operation codes. As virtually all hip fracture patients need inpatient hospital care and as the diagnosis is relatively straightforward, the total hip fracture population can be reliably identified from the register data. The completeness of registration and the accuracy of easily measurable variables in the Finnish register-data has been found to be very good for the purposes of hip fracture follow-up studies [23].

In order to determine the full LOS following the hip fracture, a care episode approach was utilized [24]. The care episode after the fracture was defined as time between the beginning of operative treatment at the surgical ward, and discharge to home. The care episode included all care periods at different wards in hospitals, at primary care inpatient care wards, at nursing homes and at other inpatient residential care as far as there was no discharge to home between the consecutive periods. The comprehensive Finnish register data with the possibility for deterministic record linkage made this care episode reconstruction possible. In fact, it has been demonstrated that register data outperform prospective audit data in the recording of inpatient care history [23].

If the care episode lasted over four months, the patient was classified as a long-term patient [25]. Since the main interest was on modelling successful care episodes, only short-term patients were included in this study. In order to concentrate on the geriatric rehabilitation of patients, the patients aged less than 65 years or not living at home at the time of fracture were also excluded. In modelling care episode times, we used an accuracy of one week, as it was considered sufficient precision with respect to the interests of the study.

Available background variables included age, sex, fracture type, the place from which the patient was admitted to the surgical ward, and the days of inpatient care during one year before the fracture. Fracture types were classified into intracapsular and extracapsular hip fractures, and the inpatient care days covariate was transformed by square root transformation because the count of days could be zero which prevented the use of log-transformation. Pre-existing comorbid conditions were also identified for each patient similarly as in Sund and Liski [26].

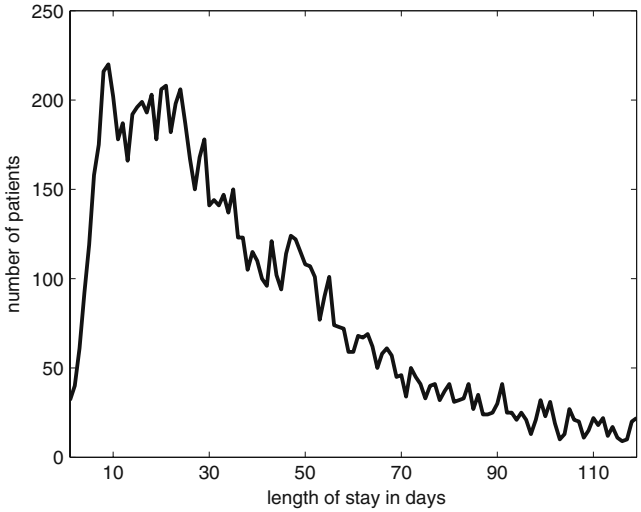

Fig. 1 Patient length of stay in care episode after a hip fracture in Finland during 1998–2001

The final study population consisted of 10 058 persons. The shape of the length of stay distribution was right-skewed (Fig. 1) that is typical for survival data. The lower quartile was 17, median 31 and upper quartile 52 days. Table 1 lists all the covariates of this study.

Table 1 Covariates for use in predicting patient length of stay

	Variable description	Range
x_1	Age	65–103
x_2	Women sex	0/1
x_3	Intracapsular fracture	0/1
x_4	Admitted from home	0/1
x_5	Days of preceding care (square root)	0–17.86
x_6	Malignancy	0/1
x_7	Chronic obstructive pulmonary disease	0/1
x_8	Ischemic heart disease	0/1
x_9	Previous myocardial infarction	0/1
x_{10}	Congestive heart failure	0/1
x_{11}	Cerebrovascular disease	0/1
x_{12}	Peripheral vascular disease	0/1
x_{13}	Diabetes without complications	0/1
x_{14}	Osteoarthritis	0/1
x_{15}	Parkinson's disease	0/1
x_{16}	Dementia	0/1
x_{17}	Alcoholism	0/1

3 Statistical models

Since health care processes are complex, and humans are a large source of variation, a flexible nonparametric Bayesian multilayer perceptron model is considered for modelling the patient length of stay. As a parametric reference model, we use a model based on a Weibull distribution, and in order to allow more flexibility in modelling, we extend our analysis to a mixture model, and use a two component Weibull mixture as the parametric approach to analyse length of stay. Bayesian modelling is employed in analysis for both models. In Bayesian analysis uncertain quantities are modelled as probability distributions, and inference is performed by computing the posterior conditional probabilities for the unobserved variables of interest, given the observed data and prior assumptions [15, 16].

3.1 Nonparametric multilayer perceptron approach

To study the distribution of patient length of stay without assuming possible shapes in advance, we apply a nonparametric approach based on a Bayesian multilayer perceptron model, and treat the modelling of length of stay as a classification problem where each class corresponds to the patient length of stay at a desired accuracy. No dependencies between the sequential classes are assumed a priori, and possible dependencies are only in a posteriori sense. Therefore, the MLP approach allows a flexible way to present the forms of distributions without assuming functional forms in advance. Another advantage with the MLP is that possible nonlinearities and implicit interactions can be automatically learned from the data.

For multilayer perceptron neural networks, the Bayesian approach is reviewed for instance by Neal [27] and Lampinen and Vehtari [28]. An MLP is a feedforward neural network model comprising of successive input, hidden, and output layers. The MLP function, corresponding to the neural network with a single hidden layer, is written as

$$f_k(\boldsymbol{x}_i, \boldsymbol{w}) = w_{k0} + \sum_{l=1}^{L} w_{kl} \tanh\left(w_{l0} + \sum_{d=1}^{D} w_{ld} x_d^{(i)}\right), \quad (1)$$

where $\boldsymbol{w}$ represents all the weight parameters w_{kl} and w_{ld} and bias parameters w_{k0} and w_{l0} of the model [27]. Indices d and l correspond to input and hidden layers, and the D-dimensional input vector is denoted by $\boldsymbol{x}_i = (x_1^{(i)}, \ldots, x_D^{(i)})^T$. In the context of neural networks, L represents the number of hidden units in the hidden layer of the network. Precisely, the MLP would become nonparametric when the number of hidden units approaches infinity, but here we approximate the infinite network with a finite one, and in a practical sense consider the finite network to be nonparametric. With the MLP function interactions between the covariates are possible, and nonlinear hyperbolic tangent (tanh) activation functions allow the hidden units to represent nonlinearities. In the classification of length of stay times, we apply the softmax model. With K possible output classes, the probability that a class target y_i has value j, is computed using softmax likelihood

$$p(y_i = j|\boldsymbol{x}_i, \boldsymbol{w}) = \frac{\exp(f_j(\boldsymbol{x}_i, \boldsymbol{w}))}{\sum_{k=1}^{K} \exp(f_k(\boldsymbol{x}_i, \boldsymbol{w}))}, \quad (2)$$

as it was done for instance in Lampinen and Vehtari [28].

In the MLP model the prior is set indirectly to a function space via network and weight priors. We assume a hierarchical prior, where the uncertainty of parameter values can be transferred into higher levels, and fix the hyperprior values similar to those in Neal [27] and Lampinen and Vehtari [28]. For the input to hidden weights, we use the following Automatic Relevance Determination (ARD) hierarchical structure

$$w_{ld} \sim N(0, \gamma_d^2) \quad (3)$$

$$\gamma_d^2 \sim \text{Inv-gamma}(\gamma_{\text{ave}}^2, \nu_\gamma) \quad (4)$$

$$\gamma_{\text{ave}}^2 \sim \text{Inv-gamma}(\gamma_0^2, \nu_{\gamma,\text{ave}}). \quad (5)$$

In the ARD prior, the weights connected to the same input have a common variance term γ_d^2. The variance γ_d^2 is controlled by the next level hyperparameters γ_{ave}^2 and ν_γ, which further are controlled by the third level hyperparameters γ_0^2 and $\nu_{\gamma,\text{ave}}$. We fixed the following values in the prior: $\nu_\gamma = 1$, $\gamma_0^2 = (0.05/D^{1/\nu_{\gamma,\text{ave}}})^2$ and $\nu_{\gamma,\text{ave}} = 2$. The scaling of prior depends on the number of inputs D, thus the more input units there are in the network, the smaller the average weights are assumed to be. For the biases in the hidden layer of the network, we set the prior

$$w_{l0} \sim N(0, \gamma_b^2) \quad (6)$$

$$\gamma_b^2 \sim \text{Inv-gamma}(0.05^2, 1). \quad (7)$$

Further, the weights between hidden and output layers were given the prior

$$w_{kl} \sim N(0, \gamma_l^2) \quad (8)$$

$$\gamma_l^2 \sim \text{Inv-gamma}((0.05/D)^2, 1), \quad (9)$$

and the output biases w_{k0} were given $N(0, 1)$ prior. The assumed prior for the weights and biases of the

network favours smooth solutions since small weights produce smooth functions. Therefore, if there is not enough information in the data, the model automatically reverts to results similar to simpler models. For further details and justifications of the prior, see Neal [27] and Lampinen and Vehtari [28].

We used a 16-hidden-unit MLP network with 17 output classes corresponding to the length of stay times at an accuracy of one week. The integration over posterior distribution was approximated as described in Neal [27]: the weight and bias parameters were sampled with a hybrid Monte Carlo (HMC) algorithm [29], and hyperparameters with Gibbs sampling. The computations were implemented using MATLAB software with the MCMC Methods for MLP and GP and Stuff -toolbox [30]. We computed an MCMC chain of 25 000 iterations of which, after burn-in stage and thinning, 100 samples were used in the posterior analysis. There were some convergence difficulties due to the large number of parameters in the model, and the multimodality of the posterior distribution. The convergence was assessed by multiple MCMC chains, visual inspection and the potential scale reduction factor [15]. The results given by the independent MCMC chains were similar.

3.2 Parametric Weibull mixture approach

As the parametric reference approach for modelling patient length of stay, we use a Weibull mixture model where explanatory variables are introduced through log-linear functions, and a weakly informative prior is set for the model parameters. As a starting point, we consider the Weibull distribution due to its wide use in survival analysis, and because of the distribution shape of the length of stay data in Fig. 1. The two parameter Weibull distribution is parameterised as

$$f_m(y_i|\alpha_m,\lambda_m) = \alpha_m y_i^{\alpha_m-1}\exp\left(\lambda_m - \exp(\lambda_m)y_i^{\alpha_m}\right) \quad (10)$$

[31], where y_i is the ith observation and α_m and λ_m are the distribution parameters. To model possible latent classes more accurately, we proceed to apply a mixture model of Weibull distributions. The mixture approach gives flexibility in modelling length of stay by allowing subpopulations to have specific length of stay patterns. The Weibull mixture, with the known number of components M, is written as

$$f(y_i|\boldsymbol{\alpha},\boldsymbol{\lambda},\boldsymbol{\mu}) = \sum_{m=1}^{M}\mu_m\alpha_m y_i^{\alpha_m-1}\exp(\lambda_m - \exp\left(\lambda_m\right)y_i^{\alpha_m}), \quad (11)$$

where $\boldsymbol{\alpha} = (\alpha_1\ldots,\alpha_M)^T$ and $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_M)^T$. The parameters $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_M)^T$ are the mixture proportions with the constraints $0<\mu_m<1$ and $\sum_{m=1}^{M}\mu_m=1$. In modelling the length of care episode stay for the hip fracture patients, the interest in the study was on modelling the short-term patients only, leading to truncated data with a known truncation point at four months. Therefore we need to truncate the parametric model. The truncated mixture component is given by

$$f_{\mathrm{T}m}(y_i|\alpha_m,\lambda_m) = \frac{\alpha_m y_i^{\alpha_m-1}\exp(\lambda_m - \exp(\lambda_m)y_i^{\alpha_m})}{1-\exp(-\exp(\lambda_m)y_{\mathrm{tp}}^{\alpha_m})}, \quad (12)$$

where y_{tp} denotes the truncation point, and the truncated Weibull mixture is

$$f_{\mathrm{T}}(y_i|\boldsymbol{\alpha},\boldsymbol{\lambda},\boldsymbol{\mu}) = \sum_{m=1}^{M}\mu_m\frac{\alpha_m y_i^{\alpha_m-1}\exp(\lambda_m - \exp(\lambda_m)y_i^{\alpha_m})}{1-\exp(-\exp(\lambda_m)y_{\mathrm{tp}}^{\alpha_m})}. \quad (13)$$

We introduce the covariates in the Weibull distribution through λ_m parameters [31], such that $\lambda_{mi} = \tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}_m$, where the $(1+D)$-dimensional covariate vector corresponding the ith observation is $\tilde{\boldsymbol{x}}_i = (1, x_1^{(i)},\ldots,x_D^{(i)})^T$, and $\boldsymbol{\beta}_m = (\beta_0^{(m)},\beta_1^{(m)},\ldots,\beta_D^{(m)})^T$ are the regression coefficients. The likelihood as a function of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ is written as

$$p(\boldsymbol{y}|X,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu}) = \prod_{i=1}^{N}\left(\sum_{m=1}^{M}\mu_m\frac{\alpha_m y_i^{\alpha_m-1}\exp(\tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}_m - \exp(\tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}_m)y_i^{\alpha_m})}{1-\exp(-\exp(\tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}_m)y_{\mathrm{tp}}^{\alpha_m})}\right), \quad (14)$$

where $\boldsymbol{y} = (y_1,\ldots,y_N)^T$ denotes the entire set of observations, X denotes $N\times D$ matrix of covariates, and $\boldsymbol{\beta}$ contains all the regression coefficient vectors.

For α_m parameters a weakly informative Gamma distribution prior $G(\alpha_m|\alpha_0,\kappa_0)\propto\alpha_m^{\alpha_0-1}\exp(-\kappa_0\alpha_m)$ is assumed, in which the parameters are given values $\alpha_0=1$ and $\kappa_0=0.001$. For the regression coefficients, we set a zero mean normal prior $N(\boldsymbol{\beta_m}|\boldsymbol{0},\sigma^2\boldsymbol{I})$, where $\sigma^2=10^4$ and $\boldsymbol{I}$ is the identity matrix. The prior is chosen to be weakly informative, as there is no a priori information about the values of regression coefficients. A Dirichlet prior distribution is set for the mixing coefficient $\boldsymbol{\mu}\sim\mathrm{Dirichlet}(\phi_1,\ldots,\phi_M)$, where $\phi_1=\ldots=$

$\phi_M = 1$. The product of likelihood and prior leads to an unnormalised posterior distribution

$$
\begin{aligned}
&p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}|\boldsymbol{y}, X) \\
&\propto \prod_{i=1}^{N}\left(\sum_{m=1}^{M}\mu_m \frac{\alpha_m y_i^{\alpha_m-1}\exp(\tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}_m - \exp(\tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}_m)y_i^{\alpha_m})}{1-\exp(-\exp(\tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}_m)y_{\mathrm{tp}}^{\alpha_m})}\right) \\
&\quad \times \prod_{m=1}^{M} N(\boldsymbol{\beta}_m|0, \sigma^2\boldsymbol{I}) \prod_{m=1}^{M} G(\alpha_m|\alpha_0, \kappa_0) \\
&\quad \times \mathrm{Dirichlet}(\boldsymbol{\mu}|\phi_1, \ldots, \phi_M). \qquad (15)
\end{aligned}
$$

We approximate the integration over the parameter space of the model using Gibbs sampling with data augmentation by introducing an unobserved indicator variable ζ_{im} in the model [15, 32]. The indicator variable is

$$
\zeta_{im} = \begin{cases} 1 & \text{if the } i\text{th observation is from the } m\text{th} \\ & \quad \text{mixture component, and} \\ 0 & \text{otherwise.} \end{cases}
$$

The following simulation steps are then repeatedly used to draw samples from the posterior:

1. Given the mixture proportions $\boldsymbol{\mu}$ and parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the latent indicators ζ_{im} are generated from a multinomial distribution with the following probabilities

$$
z_{im} = \frac{\mu_m f_{\mathrm{T}m}(y_i|\alpha_m, \lambda_m)}{\sum_{k=1}^{M}\mu_k f_{\mathrm{T}k}(y_i|\alpha_k, \lambda_k)}.
$$

2. Given the latent indicators ζ_{im} and parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the mixture proportions are sampled from the $\boldsymbol{\mu} \sim \mathrm{Dirichlet}(\phi_1 + \sum_{i=1}^{N}\zeta_{i1}, \ldots, \phi_M + \sum_{i=1}^{N}\zeta_{iM})$.
3. Given the latent indicators ζ_{im} and mixture proportions $\boldsymbol{\mu}$, the parameters α_m and $\boldsymbol{\beta}_m$ of mixture components are sampled from

$$
\begin{aligned}
&p(\alpha_m, \boldsymbol{\beta}_m|\boldsymbol{y}, X, \boldsymbol{\zeta}_m, \boldsymbol{\mu}) \\
&\propto \prod_{i=1}^{N}\left(\frac{\alpha_m y_i^{\alpha_m-1}\exp(\tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}_m - \exp(\tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}_m)y_i^{\alpha_m})}{1-\exp(-\exp(\tilde{\boldsymbol{x}}_i^T\boldsymbol{\beta}_m)y_{\mathrm{tp}}^{\alpha_m})}\right)^{\zeta_{im}} \\
&\quad \times N(\boldsymbol{\beta}_m|0, \sigma^2\boldsymbol{I}) G(\alpha_m|\alpha_0, \kappa_0)
\end{aligned}
$$

with slice sampling [33], where $\boldsymbol{\zeta}_m = (\zeta_{1m}, \ldots, \zeta_{Nm})^T$.

We modelled the patient length of stay with a two component Weibull mixture. By increasing the number of mixture components, the model would become more accurate, but also the estimation of parameters becomes more difficult. To approximate the integration over the posterior distribution, we ran several independent MCMC chains of 100 000 iterations. The convergence was checked similarly as with the MLP model, and after burn-in stage and thinning, 100 samples were used in the posterior analysis.

4 Predictive comparisons

A natural way to assess the goodness of a model is to evaluate its predictive ability for future observations by estimating expected utilities [16, 34]. The posterior predictive distribution for the test input $\boldsymbol{x}_{(N+1)}$ is given by

$$
p(y|\boldsymbol{x}_{(N+1)}, \boldsymbol{y}, X) = \int p(y|\boldsymbol{x}_{(N+1)}, \boldsymbol{y}, X, \theta)p(\theta|\boldsymbol{y}, X)d\theta, \qquad (16)
$$

where θ denotes all the model parameters, and $\boldsymbol{y}$ and X are the training data. The logarithm of predictive density is chosen as a utility measure since it measures how good the model is in modelling the whole predictive distribution. A utility u_h for a future observation $(\boldsymbol{x}_{(N+h)}, y_{(N+h)})$, is given by

$$
u_h = \log p(y_{(N+h)}|\boldsymbol{x}_{(N+h)}, \boldsymbol{y}, X), \qquad (17)
$$

where $h = 1, 2, \ldots$ indexes all the observations in the test data set. We use the mean

$$
\bar{u} = E_h[u_h] \qquad (18)
$$

as a summary quantity for the predictive ability of the models. Since the observations $\boldsymbol{y}$ are discrete, and in the parametric approach their distribution is approximated by the continuous Weibull mixture model, we do a continuity correction. The correction is done by using the integration limits $y \pm 1/2$ in the Weibull mixture when computing the predictive density Eq. 16 for the observation y.

We divided the data into two equally sized parts, and estimated the expected utility for both models with the test part of the data. The expected utilities with 95% credible intervals were -2.46 ± 0.02 for the MLP, and -2.48 ± 0.02 for the Weibull mixture (-2.51 ± 0.02 with a single truncated Weibull distribution). In a pairwise test the MLP was better than the two component Weibull mixture with a probability greater than 0.999, indicating more accurate predictions with the MLP model. Further, we computed the expected utilities separately for the patients diagnosed with either intra- or extracapsular fracture. The MLP gave the expected utility -2.37 for intracapsular, and -2.64 for extracapsular patients. With the Weibull mixture the utilities

were −2.38 for intracapsular, and −2.66 for extracapsular patients (−2.42 and −2.68 with a single truncated Weibull distribution). The results suggest that the MLP gives more accurate predictions, and that the length of care episode stay is predicted more accurately for intracapsular than extracapsular patients with both models.

Figure 2 illustrates the mean distributions of length of stay for both fracture types using the MLP and the Weibull mixture. The distributions for intracapsular patients, with unimodal and right-skewed shapes, are similar with both models. Noticeable difference is in the distributions for extracapsular patients. The Weibull mixture gives a simple unimodal distribution, whereas the MLP suggests that the length of stay distribution is multimodal with no systematic decrease in probabilities until the mode at 7 weeks. The second mode is missed with the Weibull mixture, and may be due to the fixed forms of parametric distributions. The length of stay distributions for extracapsular patients are wider than for intracapsular patients with both models, explaining the worse expected utilities for extracapsular patients.

Since the parameters of the models cannot be compared, we study the relevances of covariates in predicting the length of care episode stay. To find out the most relevant variables in the prediction, we use the average predictive comparison method proposed by Gelman and Pardoe [35]. The method estimates the expected difference in the outcome associated with a unit difference in one of the covariate. The method takes into account the uncertainty of model parameters, and averages over the population distribution of the covariates, which is useful with nonlinear models. To compute the average predictive comparison, we define the probability of early discharge from care episode as the outcome of interest. The patient is classified as early discharge patient if the discharge happens during weeks 1–5. Both models give somewhat similar relevance estimates for the covariates (Fig. 3). The estimates away from zero show more relevance, and thereby the most relevant covariates in predicting length of stay are age and fracture type. In addition to the relevance estimates of single covariates, we study further the joint relevances of two covariates since there may be interactions between the covariates. To find out the joint relevances, we measure the change in the predictive distribution Eq. 16, when the values of two covariates are simultaneously changed. The change is measured between the predictive distributions with the information theoretic Kullback-Leibler (KL) divergence

$$\delta_i^{\text{joint}} = \int p(y|\boldsymbol{x}_i, \boldsymbol{y}, X) \log\left(\frac{p(y|\boldsymbol{x}_i, \boldsymbol{y}, X)}{p(y|\boldsymbol{x}_i^{\Delta}, \boldsymbol{y}, X)}\right) dy, \quad (19)$$

where $\boldsymbol{x}_i$ denotes the observed vector of covariate values, and $\boldsymbol{x}_i^{\Delta}$ is the same vector but with two covariate values changed with a unit difference. We normalise the divergence with a euclidean distance from $\boldsymbol{x}_i$ to $\boldsymbol{x}_i^{\Delta}$, and average over the observed population as

$$\bar{v} = E_i\left[\frac{\delta_i^{\text{joint}}}{\|\boldsymbol{x}_i - \boldsymbol{x}_i^{\Delta}\|}\right], \quad (20)$$

where $\bar{v}$ is the estimate for the joint relevance of two covariates in predicting the length of care episode stay.

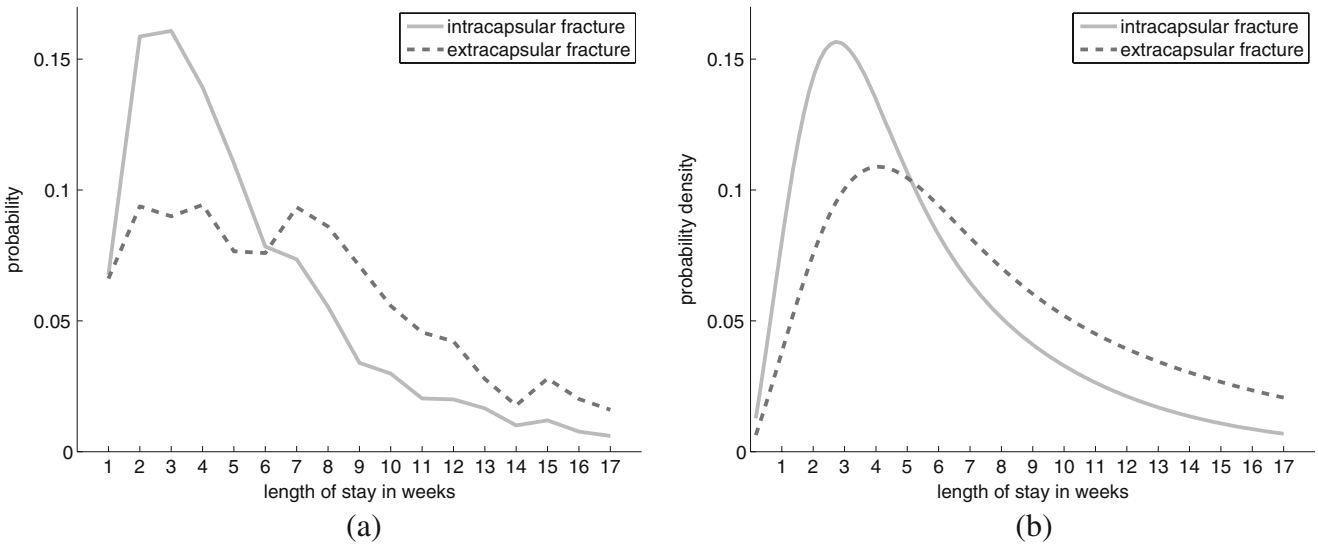

Fig. 2 Estimated mean distributions of length of stay in care episode for two fracture types with the multilayer perceptron (**a**) and the Weibull mixture model (**b**)

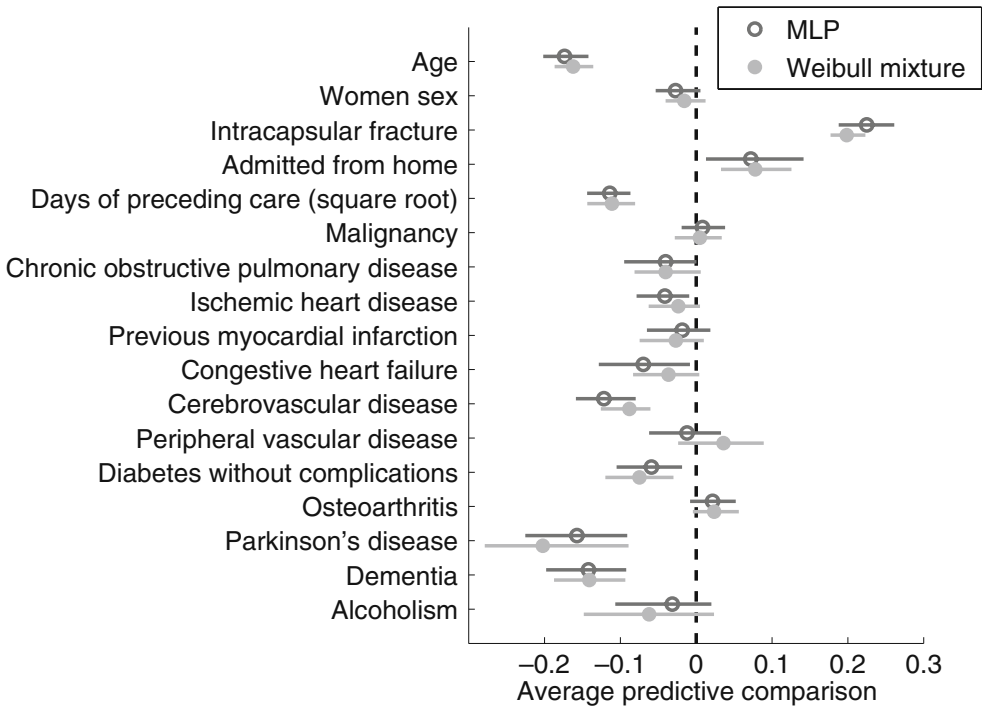

Fig. 3 Estimated average predictive comparisons for the probability of early discharge from care episode. The estimated mean values and 95% credible intervals are shown for the multilayer perceptron (MLP) and the Weibull mixture model

Figure 4 demonstrates the joint relevance estimates of other covariates with the age covariate. Fracture type and age has the largest estimate for the joint relevance with both models. Further, the Weibull mixture gives almost throughout greater estimates for the joint relevance than the MLP. The difference in the estimates is the largest at the joint estimate of Parkinson's disease and age, where the Weibull mixture gives significantly greater estimate for the joint relevance than the MLP. To examine in more detail the capabilities of the

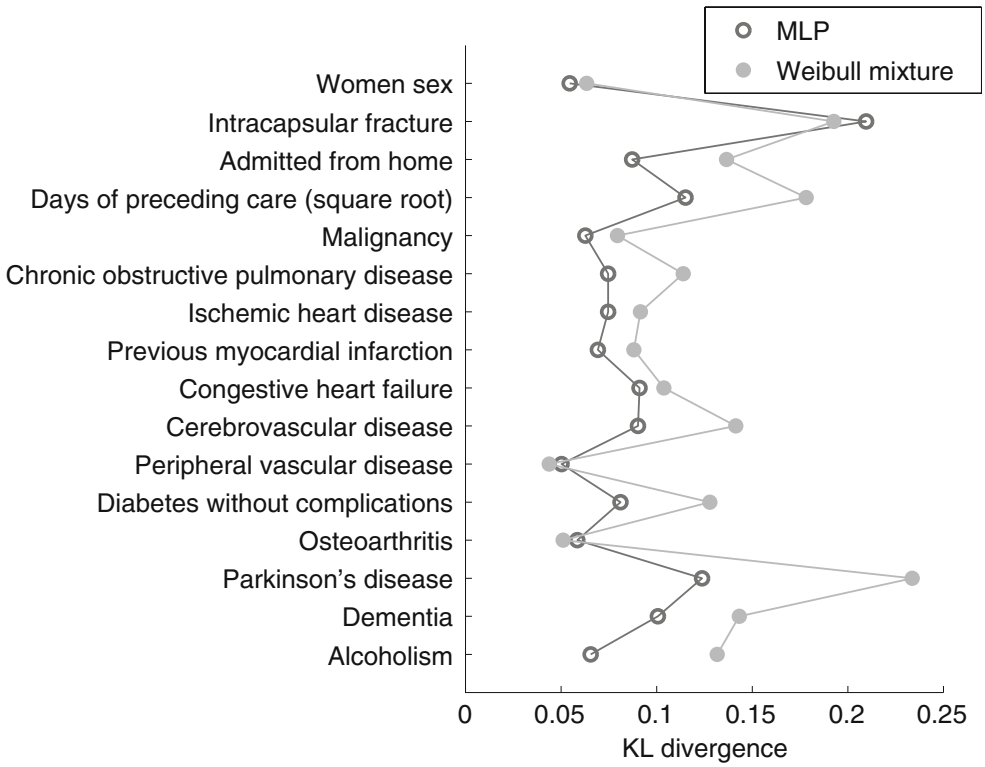

Fig. 4 Estimated joint predictive relevances of other covariates with the age covariate for the multilayer perceptron (MLP) and the Weibull mixture model

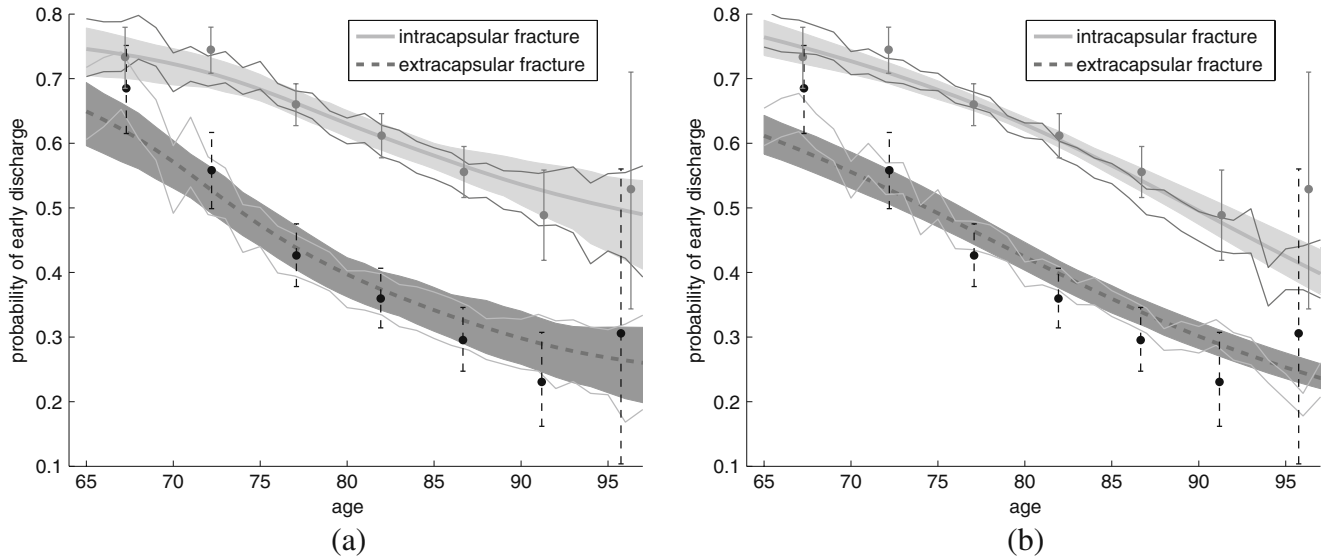

Fig. 5 Estimated mean and 95% credible intervals for the probability of early discharge as a function of age and fracture type described as shaded plots with the multilayer perceptron (**a**) and the Weibull mixture model (**b**). *Thin lines* describe the individual predictions grouped by age and fracture type. The crude estimates from the data are also illustrated as vertical segments of lines (mean and 95% credible intervals)

models to model joint effects, we use posterior predictive simulation and comparison to the data. We study the probability of early discharge from care episode (defined as earlier) as a function of fracture type and age, whose joint relevance estimate was large and similar with both models (Fig. 4). We simulate the effects of two covariates with the models by changing the values of the two covariates for a randomly chosen simulation population. The simulated mean values and 95% credible interval predictions are shown in Fig. 5. As a reference for the simulated predictions, we compute data estimates using a binomial model with a weakly informative Beta prior distribution whose mean value was set to the mean value observed in the data set. The mean values and 95% credible intervals of the data estimates are illustrated as dots and vertical segments of

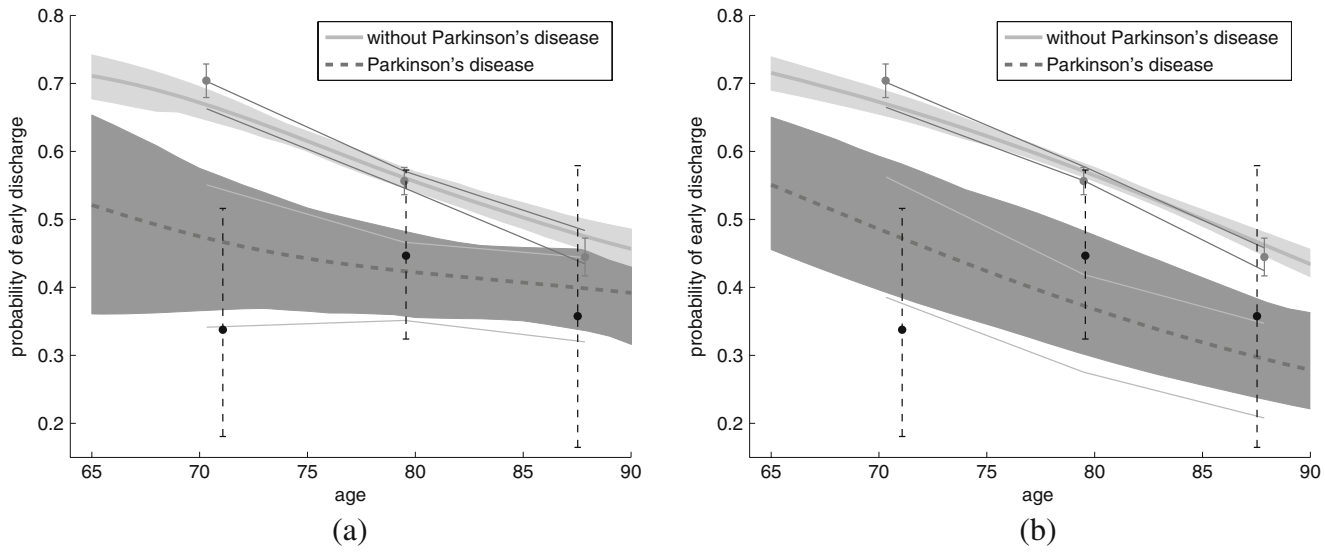

Fig. 6 Estimated mean and 95% credible intervals for the probability of early discharge as a function of age and Parkinson's disease described as shaded plots with the multilayer perceptron (**a**) and the Weibull mixture model (**b**). *Thin lines* describe the individual predictions grouped by age and whether the patient has Parkinson's disease or not. The crude estimates from the data are also illustrated as vertical segments of lines (mean and 95% credible intervals)

lines in Fig. 5. The results indicate that the MLP model gives more accurate results than the Weibull mixture when simulated values are compared to the data estimates. There is a clear interaction between fracture type and age, seen in how differently age affects the probability of early discharge depending on which of the fractures a patient has. The Weibull mixture fails to model accurately the nonlinearities and interactions since the covariates affect log-linearly in the model. Further, since the Weibull mixture has a rigid model structure, the uncertainty in the predictions is estimated to be much lower than with the MLP, giving probably overoptimistic uncertainty estimates for the predictions. This can be seen especially for patients over 90.

It must be noticed that other covariates, which possibly correlate with fracture type and age, may have an effect on the simulation results in Fig. 5. To find out if there is some other correlating covariate, we also computed the predictions for the individuals in the simulation population without changing the covariate values. By doing age grouping for both fractures, we present 95% credible intervals for the individual predictions as thin lines (Fig. 5). The individual predictions coincide well with the simulated predictions, making evident that merely fracture type and age explain these changes in the predictions.

Since the joint relevance estimates differed mostly between the MLP and Weibull mixture models for the Parkinson's disease and age covariate pair (Fig. 4), we also visualise the early discharge probability as a function of these two variables (Fig. 6). There is only a small number of Parkinson's disease observations in the data, causing large uncertainty intervals. Both models fail to capture the lower probability between ages 65 and 70 for the patients diagnosed with Parkinson's disease. The large difference in the joint relevance estimates between the models is explained by comparing the results in Fig. 6; the age covariate affects the early discharge probability less in the MLP when the patient is diagnosed with Parkinson's disease, whereas the Weibull mixture gives a clearly decreasing trend. It seems that age is an overall indicator for health, but if patient is diagnosed with Parkinson's disease, this illness is sufficient merely to change the predictive distribution for the length of care episode stay using the MLP model.

5 Conclusions

In this paper we have studied the modelling of the length of care episode stay with the nonparametric and parametric Bayesian models. The case study was about LOS following the hip fracture. We used comprehensive Finnish register data in the reconstruction of the care episodes that crossed the boundaries of service providers. The idea of care episode approach is not new [36], but the lack of adequate data has made it difficult to utilize it in practice [37]. Fortunately, the Finnish register system with common personal identity codes in all registers and good data quality [38] made the reconstruction of inpatient care episodes for the total population of hip fracture patients in Finland feasible [22]. It has also been shown that a register-based reconstruction of care episodes performs better than with a separate prospective audit data collection [23]. The main drawback of the register-based data is the lack of clinical background variables [39].

One aim of this study was the accurate modelling of patient length of stay. For that purpose, we wanted to compare a traditional parametric model and a nonparametric multilayer perceptron network model.

In the literature, commonly utilized parametric LOS models include exponential, Weibull, lognormal, and gamma models [8, 40]. Based on the observed shape of the LOS distribution, we considered Weibull model to be a suitable parametric candidate. As the use of mixture models allows more flexibility [41], we decided to use a Weibull mixture model as a parametric reference model. It is possible that a mixture model with some other component distributions such as gamma [42], a mixture of different parametric distributions as its mixture components [11], or a Coxian phase-type distribution [43] may give a better performance for this particular length of stay data, but the testing of a large number of rather similar parametric models was out of the scope of this study.

The nonparametric multilayer perceptron network model is a flexible approach that also suits to the modelling of LOS distribution with covariates. So far the attempts to model LOS with neural networks seem to be rare [13, 14]. This may be due to the fact that the interpretation and explainability of the results become more challenging. On the other hand, the modelling of complex systems may be even more useful with complex nonparametric methods than with parametric ones, especially if the predefined parametric model cannot represent the phenomenon under study sufficiently well.

The Bayesian MLP model results in comprehensive information about the systematic patterns occurring in the data without the need to fix the relationships between the variables, as it is needed in Bayesian belief networks that have been used to predict patient length of stay [44]. Therefore, the MLP model can be advantageous also for exploratory analysis, especially

in cases when there is no a priori expertise information about causal relationships between variables. The MLP model is also suitable for modelling continuous explanatory variables, which are difficult to handle with classification or regression tree models, sometimes applied for length of stay analysis [45]. There are also other miscellaneous classifiers (see, e.g. [46]), but testing of several classifiers was out of the scope of this study. The Bayesian MLP model has performed well in public classification comparisons [47].

We adopted the Bayesian approach for modelling, and formulated a nonparametric multilayer perceptron network model as well as a parametric Weibull mixture model for the modelling of LOS data with covariates. We also wanted to compare the models and their predictive performances. Since the parameters in the MLP and the Weibull mixture model were incomparable, we focused on observables rather than the model parameters, and studied the performances of the multilayer perceptron and the Weibull mixture model by doing posterior predictive checking, and by comparing the predictive distributions visually.

There were noticeable differences in the results. The parametric Weibull mixture was insufficient to capture the characteristics in the data accurately due to the fixed forms of distributions, and log-linearly introduced covariates in the model. For instance, the anomaly in the shapes of the length of stay distribution between intra- and extracapsular fractures, were missed with the Weibull mixture. Actually the anomaly detected by the MLP model was clinically interesting, and a more careful investigation on the issue has been reported elsewhere [48]. In brief, the rehabilitation of patients with extracapsular fractures took longer than in the intracapsular group. The difference was due to the different surgical methods and especially to the different rehabilitation practices. This means that service providers may be using outdated practices, such as instructing most of the patients with extracapsular fractures to start rehabilitation with partial weight-bearing although that is in contrast with the clinical guidelines.

In order to identify the most relevant variables in predicting the care episode times, the relevances of covariates were estimated by doing average predictive comparison. To provide a better understanding of how patient covariates affect the predictions of the length of stay, the joint effects of the covariates were also studied. The average predictive comparison for both models gave similar results, but when illustrating the interactions between two covariates, the predictions differed noticeably. This was studied further by observing the probability of early discharge from care episode as a function of two covariates. The predictions given by the MLP model were closer to the data estimates than with the Weibull mixture. In general, the Weibull mixture gave over-optimistic uncertainty estimates, whereas the MLP model gave more accurate results with the larger uncertainty intervals suggesting more realistic estimates.

In conclusion, the Bayesian MLP model is a viable alternative to model the large scale health care data featuring nonlinear effects and interactions. We have also demonstrated how such a complex Bayesian approach can be used for practical performance evaluation, and showed implementations of several methodological ideas that make the concrete analyses more feasible, and are therefore likely to be useful also in more general context. Further research involves including costs in the models to study cost-effectiveness, and also taking regional variables into account. For instance, in the MLP model regional effects can be modelled without need to model the regional hierarchy explicitly in the model structure.

References

1. OECD (2002) Measuring up. Improving health system performance in OECD countries. Organisation for Economic Co-operation and Development, Paris
2. Aday LA, Begley CE, Lairson DR, Slater CH (1998) Evaluating the healthcare system: effectiveness, efficiency, and equity, 2nd edn. Health Administration Press, Chigaco
3. Perleth M, Jakubowski E, Busse R (2001) What is 'best practice' in health care? State of the art and perspectives in improving the effectiveness and efficiency of the European health care systems. Health Policy 56:235–250
4. Eddy DM (1998) Performance measurement: problems and solutions (with discussion). Health Aff 17:7–41
5. Loeb JM (2004) The current state of performance measurement in health care. Int J Qual Health Care 16(Suppl 1):i5–9
6. Freeman T (2002) Using performance indicators to improve health care quality in the public sector: a review of the literature. Health Serv Manag Res 15:126–137
7. Bird SM, Cox D, Farewell VT, Goldstein H, Holt T, Smith PC (2005) Performance indicators: good, bad, and ugly. J R Stat Soc, A 168(1):1–27
8. Marazzi A, Paccaud F, Ruffieux C, Beguin, C (1998) Fitting the distributions of length of stay by parametric models. Med Care 36(6):915–927
9. Fisher WH, Altaffer FB (1992) Inpatient length of stay measures: statistical and conceptual issues. Adm Policy Ment Health 19:311–320
10. Marshall AH, McClean SI (2004) Using Coxian phase-type distributions to identify patient characteristics for duration of stay in hospital. Health Care Manag Sci 7:285–289
11. Atienza N, García-Heras J, Muñoz-Pichardo JM, Villa R (2008) An application of mixture distributions in modelization of length of hospital stay. Stat Med 27(9):1403–1420
12. Faddy MJ, McClean SI (1999) Analysing data on lengths of stay of hospital patients using phase-type distributions. Appl Stoch Models Bus Ind 15:311–317

13. Lowell W, Davis G, Lajousky W, Stieffel S, Davis G, Breau M, Harvey S, Shirazi H (1997) A field trial using artificial neural networks to predict psychiatric in-patient length-of-stay. Neural Comput Appl 5:184–193
14. Walczak S, Scorpio RJ, Pofahl WE (1998) Predicting hospital length of stay with neural networks. In: Proceedings of the eleventh international FLAIRS conference, pp 333–337
15. Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian data analysis, 2nd edn. Chapman & Hall/CRC, Boca Raton
16. Bernardo JM, Smith AFM (1994) Bayesian theory. Wiley, New York
17. Marks R, Allegrante JP, MacKenzie CR, Lane JM (2003) Hip fractures among the elderly: causes, consequences and control. Ageing Research Reviews 2(1):57–93
18. Zuckerman JD (1996) Hip fracture. N Engl J Med 334(23):1519–1525
19. Nurmi I, Narinen A, Lüthje P, Tanninen S (2003) Cost analysis of hip fracture treatment among the elderly for the public health services: a 1-year prospective study in 106 consecutive patients. Arch Orthop Trauma Surg 123:551–554
20. Parker MJ, Todd CJ, Palmer CR, Camilleri-Ferrante C, Freeman CJ, Laxton CE, Payne BV, Rushton N (1998) Inter-hospital variations in length of hospital stay following hip fracture. Age Ageing 27:333–337
21. SIGN (2002) Prevention and management of hip fracture in older people. Scottish Intercollegiate Guidelines Network, Edinburgh
22. Sund R (2008) Methodological perspectives for register-based health system performance assessment: developing a hip fracture monitoring system in Finland. Research report 174, National Research and Development Centre for Welfare and Health. http://urn.fi/URN:ISBN:978-951-33-2132-1
23. Sund R, Nurmi-Lüthje I, Lüthje P, Tanninen S, Narinen A, 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 Inf Med 46(5):558–566
24. Hornbrook MC, Hurtado AV, Johnson RE (1985) Health care episodes: definition, measurement and use. Med Care Res Rev 42:163–218
25. Heikkinen T, Jalovaara P (2005) Four or twelve months' follow-up in the evaluation of functional outcome after hip fracture surgery? Scand J Surg 94:59–66
26. Sund R, Liski A (2005) Quality effects of operative delay on mortality in hip fracture treatment. Qual Saf Health Care 14:371–377
27. Neal RM (1996) Bayesian learning for neural networks. Springer, New York
28. Lampinen J, Vehtari A (2001) Bayesian approach for neural networks—review and case studies. Neural Netw 14(3): 7–24
29. Duane S, Kennedy AD, Pendleton BJ, Roweth D (1987) Hybrid Monte Carlo. Phys Lett, B 195(2):216–222
30. COSY (2008) MCMC Methods for MLP and GP and Stuff (for Matlab) V2.1. http://www.lce.hut.fi/research/mm/mcmcstuff/
31. Ibrahim JG, Chen MH, Sinha D (2001) Bayesian survival analysis. Springer, New York
32. Diebolt J, Robert CP (1994) Estimation of finite mixture distributions through Bayesian sampling. J R Stat Soc, B 56(2):363–375
33. Neal RM (2003) Slice sampling. Ann Stat 31:705–767
34. Vehtari A, Lampinen J (2002) Bayesian model assessment and comparison using cross-validation predictive densities. Neural Comput 14(10):2439–2468
35. Gelman A, Pardoe I (2007) Average predictive comparisons for models with nonlinearity, interactions, and variance components. Sociol Method 37(1):23–51
36. Solon JA, Feeney JJ, Jones SH, Rigg RD, Sheps CG (1969) Delineating episodes of medical care. Am J Public Health 57:401–408
37. Rosen AK, Mayer-Oakes A (1999) Episodes of care: theoretical frameworks versus current operational realities. Joint Comm J Qual Improve 25:111–128
38. Gissler M, Haukka J (2004) Finnish health and social welfare registers in epidemiological research. Norsk Epidemiol 14:113–120
39. Powell AE, Davies HTO, Thomson RG (2003) Using routine comparative data to assess the quality of health care: understanding and avoiding common pitfalls. Qual Saf Health Care 12:122–128
40. Ruffieux C, Marazzi A, Paccaud F (1993) Exploring models for the length of stay distribution. Soz Praventivmed 38:77–82
41. Quantin C, Sauleau E, Bolard P, Mousson C, Kerkri M, Brunet Lecomte P, Moreau T, Dusserre L (1999) Modeling of high-cost patient distribution within renal failure diagnostic related group. J Clin Epidemiol 52:251–258
42. Lee AH, Ng ASK, Yau KKW (2001) Determinants of maternity length of stay: a gamma mixture risk-adjusted model. Health Care Manag Sci 4:249–255
43. Marshall A, Shaw B (2008) Modeling survival of hip fracture patients using a conditional phase-type distribution. In: 21st IEEE international symposium on computer-based medical systems, vol 21, pp 518–523
44. Marshall AH, McClean SI, Shapcott CM, Hastie IR, Millard PH (2001) Developing a Bayesian belief network for the management of geriatric hospital care. Health Care Manag Sci 4:25–30
45. Harper PR (2002) A framework for operational modelling of hospital resources. Health Care Manag Sci 5:165–173
46. Bishop CM (2006) Pattern recognition and machine learning. Springer, New York
47. Neal RM (2006) Classification with Bayesian neural networks. Machine learning challenges. Evaluating predictive uncertainty, visual object classification, and recognising textual entailment. Lecture notes in computer science no 3944. Springer, New York
48. Sund R, Riihimäki J, Mäkelä M, Vehtari A, Lüthje P, Huusko T, Häkkinen U (2009) Modeling the length of the care episode after hip fracture: does the type of fracture matter? Scand J Surg 98:169–174