Insurance: Mathematics and Economics 71 (2016) 244–252

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance

Andreas Bayerstadler[a,*], Linda van Dijk[b], Fabian Winter[a]

[a] *Munich Health, Munich Re, Königinstraße 107, 80802 Munich, Germany*[1]

[b] *Damen, National Health Insurance Company, Abu Dhabi, United Arab Emirates*

ARTICLE INFO

Article history:
Received November 2015
Received in revised form
September 2016
Accepted 18 September 2016
Available online 30 September 2016

Keywords:
Fraud and abuse detection
Health insurance
Predictive model
Bayes
Latent variable

ABSTRACT

Healthcare fraud and abuse are a serious challenge to healthcare payers and to the entire society. This article presents a predictive model for fraud and abuse detection in health insurance based on a training dataset of manually reviewed claims. The goal of the analysis is to predict different fraud and abuse probabilities for new invoices. The prediction is based on a wide framework of fraud and abuse reports which examine the behavior of medical providers and insured members by measuring systematic deviation from usual patterns in medical claims data. We show that models which directly use the results of the reports as model covariates do not exploit the full potential in terms of predictive quality. Instead, we propose a multinomial Bayesian latent variable model which summarizes behavioral patterns in latent variables, and predicts different fraud and abuse probabilities. The estimation of model parameters is based on a Markov Chain Monte Carlo (MCMC) algorithm using Bayesian shrinkage techniques. The presented approach improves the identification of fraudulent and abusive claims compared to different benchmark approaches.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

1.1. Challenge and approach

Fraud, abuse and waste in healthcare strongly contribute to the increase in total healthcare expenditure. Therefore, they are serious issues for public and private payers of healthcare and, as costs are usually transferred to the collective of insured persons, also to the insured persons themselves and to the entire society. In the US, loss estimations from fraud and abuse range from 9% to 19% of total healthcare expenditure (Berwick and Hackbarth, 2012). Assuming an average of 14% this means a total loss of 369 billion US dollars in 2011, or more than 1,000 US dollars per US citizen. In Europe similar rates are assumed (Gee and Button, 2014; Ekina, 2013). It is self-evident that this additional burden leads to increased taxes and higher health insurance premiums for individuals.

The definitions of fraud and abuse are quite heterogeneous in the literature, and depend on market and regulatory environments. In most articles, the terms waste, abuse, and fraud are used (see Fig. 1).

In this article, we will focus on fraud and abuse, i.e. intentional behavior by patients and/or medical providers to create unjustified benefits for themselves or related persons. Overall, we apply three (widely distinct) behavioral categories:

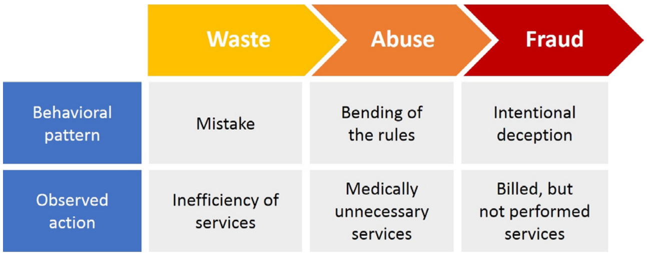
- *Unperformed services (fraud):* Medical services which are documented and charged, but not performed (e.g. insured members faking prescriptions).
- *Unjustified services (abuse):* Medical services performed without medical necessity/justification, and which deviate from medical best practice, in US literature often denoted as overutilization (e.g. prescription of antibiotics for mild respiratory diseases).
- *Other billing issues (fraud/abuse):* All other kinds of intentional misbehavior by medical providers and/or insured members (e.g. unbundling of procedure codes).

Fraud and abuse in healthcare are not only committed by medical providers. Insured patients, approvers of services and other healthcare players are also involved in fraudulent and abusive actions (Busch, 2008). Moreover, fraud and abuse are often based on the cooperation or at least complicity between different players in the health market (e.g. doctor and pharmacist, provider and insured, etc.). Nevertheless, many publications on fraud and

[*] Corresponding author.
E-mail address: abayerstadler@munichhealth.com (A. Bayerstadler).

[1] Munich Health in Cooperation with the Institute of Statistics of the Ludwig–Maximilians-University Munich.

http://dx.doi.org/10.1016/j.insmatheco.2016.09.013
0167-6687/© 2016 Elsevier B.V. All rights reserved.



	Waste	Abuse	Fraud
Behavioral pattern	Mistake	Bending of the rules	Intentional deception
Observed action	Inefficiency of services	Medically unnecessary services	Billed, but not performed services

Fig. 1. Definition of fraud and abuse in healthcare.

abuse, and related analytical detection methods, focus on medical providers (Busch, 2008) due to a higher underlying saving potential and a broader data basis.

As major contributors of healthcare funding, private and public insurers have a strong motivation to prevent losses arising from fraud and abuse. However, there is often no systematic approach in place to deal with this issue. In many insurance companies, fraud and abuse detection is limited to opportunistic checks of known patterns within the standard claims cycle. The main focus is usually on historic provider behavior as well as diagnoses and procedures frequently related to fraud and abuse cases, whereas member behavior and the interaction of different players play a minor role. In this setup, new and more complex patterns can only be detected by coincidence. Another difficulty is the lack of data experience on fraud and abuse cases, i.e. a small sample size of reviewed cases due to a lack of resources for the topic, especially in smaller insurance companies. Other challenges are the lack of defined actions if a suspicious case is identified, and regulatory conditions that impede the recovery of money once claims are paid.

Therefore, we have developed a systematic analytical approach for the identification of fraudulent and abusive behavior which is based on

(a) a vast data basis of reviewed fraud and abuse cases in the target market to be used as quality assured response,
(b) comprehensive knowledge of fraud and abuse patterns in different markets,
(c) a "reporting factory" which quantifies the behavior of providers and members, and considers network constellations (i.e. the interaction between different players) and invoice properties (e.g. diagnoses, procedures),
(d) a predictive scoring model which classifies new invoices into the three categories described above and outputs reasons for the classification as a starting point for further investigation.

This article focuses on the methodology of the predictive scoring model (item (d)), a Bayesian multinomial latent variable model which uses the knowledge generated from items (a) to (c) (see Section 2). Due to the high number of partly highly correlated reporting results, we use latent variables as model covariates which can be interpreted as behavioral scores for providers, members and networks. In this way, we avoid overfitting to training data and stabilize the model to reach higher predictive power and transferability. For the insurance company, higher predictive power means a more efficient claims adjudication process by a more targeted investigation of invoices. In Section 3, we present the prediction results of the model based on a test dataset and compare them to alternative classification techniques.

1.2. Literature

Analytical fraud detection methods are being successfully applied in many areas of the financial industry. Bolton and Hand (2002) as well as Ngai et al. (2011) provide a comprehensive overview of applications and statistical techniques. Statistical techniques gain importance where mass data (Big Data) needs to be analyzed, such as in credit card fraud detection (Bhattacharyya et al., 2011).

In general, the approaches applied can be divided into supervised techniques, like classification and regression, where a learning dataset of identified fraud cases is available, and unsupervised techniques, like clustering and outlier detection, where the focus is on detecting abnormal patterns (Bolton and Hand, 2002; Phua et al., 2005).

Within the insurance industry, analytical fraud and abuse detection is most widespread in motor and health insurance, but there are also examples from other lines of business (Jin et al., 2005; Peng et al., 2007). An extensive overview of statistical fraud detection methods in motor insurance can be found in Viaene et al. (2002). Bayesian learning is, for instance, applied by Viaene et al. (2004a,b, 2005). Similar to health insurance, relations between different players are very important for motor insurance leading to an analysis of (social) networks (Šubelj et al., 2011).

In health insurance, various classification techniques to identify fraudulent and abusive behavior are applied. Joudaki et al. (2014) and Li et al. (2008) give a general overview of data mining techniques supporting fraud identification in health insurance, Dua and Bais (2014) focus on supervised classifications methods. For instance, Thornton et al. (2013) propose a multidimensional classification model for different kinds of healthcare fraud and abuse in the US Medicaid system. Most authors base their approaches on expert systems which are used as input (Major and Riedinger, 2002; Musal, 2010). Machine learning techniques, like Neural Networks or Feature Selection, are applied to identify new patterns and increase automation (He et al., 1997; Yang and Hwang, 2006; Aral et al., 2012). Some researchers also apply Bayesian techniques to identify fraudulent behavior in health insurance (Ekina et al., 2013).

Latent variable approaches have already been used in different research areas. An overview can be found in Skrondal and Rabe-Hesketh (2004). Bayesian latent variable approaches, similar to the one we introduce in this article, have been applied in AIDS prevention (Adebayo et al., 2011), in the examination of human birth defects (Sammel et al., 1997) and in social science (Fahrmeir and Raach, 2007; Fahrmeir and Steinert, 2006). Bermúdez et al. (2008) apply a Bayesian model with latent variables in motor insurance fraud detection. They use latent variables to model the skewness of the response distribution in a dichotomous classification model. In our setting, this skewness is balanced by undersampling (see Section 1.3) and latent variables serve as behavioral scores. As far as we know, Bayesian latent variable models have not yet been applied for fraud detection in health insurance. Also, Bayesian latent variable models with multinomial response as well as the combination with Bayesian shrinkage techniques are new fields of statistical research.

1.3. Data environment

One key element of an efficient analytical fraud and abuse detection method is the availability of a sufficient number of cases with quality assured response that can be used for supervised learning. As basis of our analyses, we use a fraud and abuse dataset from an insurance company in the Middle East. This dataset includes more than 100,000 manually reviewed cases collected over a period of four years, which have explicitly been assigned to one of the three response categories "unperformed services", "unjustified services", "other billing issues", or the reference category "no irregularities". We will illustrate our approach using the example of outpatient invoices of which $n = 36{,}796$ are in the training dataset, even though the described methodology can be applied to all kind of medical invoices.

Despite this comparably large training dataset, there are some obstacles in the development of a prediction model which shall be applied to new invoices of the same company.

The $n = 36{,}796$ outpatient invoices have already been pre-selected by claims experts based on a simple rule model (focus on providers who have already been involved in fraud and abuse, as well as on certain diagnoses and procedures). This pre-selection leads to a high danger of overfitting to training data if the model will be applied to non-preselected claims. For example, the share of invoices in the training dataset with actually detected irregularities is, at 36.9%, very high and not realistic compared to the assumptions in literature (about 10% according to Gee and Button, 2014; Berwick and Hackbarth, 2012; Ekina, 2013). Moreover, the distribution of the response categories is very unbalanced (33.3% "unperformed services", 1.1% "unjustified services" and 2.4% "other billing issues").

The reason for the high percentage of detected cases of unperformed services is that the reviewers focused on this kind of misbehavior because it usually leads to a direct recovery of payments. Especially "unjustified services" are usually harder to prove and need a more extensive investigation process. However, many authors, like Ezekiel and Fuchs (2008), assume that in most markets the major part of the loss is related to abusive behavior. Therefore, a weighting approach is needed which enables the prediction model to also detect cases from the categories "unjustified services" and "other billing issues" based on the existing experience.

Consequently, we apply a sampling strategy that helps to control overfitting and to increase the transferability of the model (see Fig. 2).

First, we randomly branch off a 10% test sample from the original dataset of outpatient invoices ($n_{\text{test}} = 3{,}680$) which is used to measure the predictive performance of the models. The remaining 90% of the training sample ($n_{\text{train}} = 33{,}116$) is used for model building. To avoid overfitting to and increase the transferability of the model, we apply a subsampling approach (Breiman, 1996) based on repeated undersampling of the training sample. More precisely, we draw several stratified subsamples from the training sample so that the relative frequencies for the response categories "unperformed services" ($f_{\text{UP,strain}}$), "unjustified services" ($f_{\text{UJ,strain}}$), "other billing issues" ($f_{\text{BI,strain}}$) and "no irregularities" ($f_{\text{NI,strain}}$) are approximately balanced (exact sampling ratios are given in Fig. 2). This balancing approach is proposed by several authors, e.g. Wallace et al. (2011), and has already been applied in (credit card) fraud detection based on highly unbalanced samples (Sahin and Duman, 2011a,b). In this way, the predictive performance for less frequently observed response categories is improved and abusive behavior gets more attention.

The number of subsamples B is chosen so that sufficient data information is preserved. With $B = 50$ approximately 75% of all fraud and abuse cases in the training sample occur in at least one of the subsamples. In order to monitor performance, we use the non-selected invoices from each subsample as a validation sample.

In our sampling process, we only apply the balancing of response categories to the training and validation subsamples, but not to the test samples. We are aware that this proceeding may lead to a model calibration bias and a lower predictive performance on the defined test sample. We accept this decrease in performance because we assume that the balanced subsampling approach reduces overfitting to the training data biased through the pre-selection of invoices and increases the transferability of the model.

The choice of the described sampling approach is based on different tests of the predictive quality of resulting models and their stability. First, the number of subsamples ($B = 50$) and the non-category-specific sampling ratio of 80% have been varied which has not lead to a major change of model outcomes. Second, the category-specific balancing weights (1/60, 1/30, 1, 1/2) have been increased (anti-proportionally) to decrease the degree of undersampling. In the extreme case of no undersampling (i.e. weights of 1, 1, 1, 1) the overall predictive results slightly improved, due to a better model fit in the over-represented categories. As the model fit in the under-represented categories, however, dropped drastically, we kept the original balancing weights to get a more transferable model with equal attention to all response categories. Third, we applied an oversampling approach of under-represented response categories, i.e. we allowed drawing with replacement for these categories. While increasing the degree of oversampling, we observed a decrease in predictive quality due to an over-adaptation to the training data.

2. Methodology

2.1. Model structure

The basic idea of the modeling approach is to predict the probabilities π_{UP}, π_{UJ}, π_{BI} and π_{NI} that a new invoice belongs to one of the response categories "unperformed services", "unjustified services", "other billing issues" or "no irregularities", respectively. According to the clear definition in Section 1.1, we assume that response categories are disjoint and the response probabilities π_{UP}, π_{UJ}, π_{BI} and π_{NI} add up to one. This implies that the class affiliation y_i of invoice i follows a multinomial distribution:

$$y_i \sim \text{Mult}((\pi_{\text{UP}i}, \pi_{\text{UJ}i}, \pi_{\text{BI}i}, \pi_{\text{NI}i})) \quad \text{with } \pi_{\text{NI}i} = 1 - (\pi_{\text{UP}i} + \pi_{\text{UJ}i} + \pi_{\text{BI}i}). \tag{1}$$

Based on this assumption, a multinomial logit model (McCullagh and Nelder, 1989) can be applied to estimate and predict the response probabilities. On the covariate side, a lot of potential influence parameters are available:

- reporting results for provider behavior $p_1, \ldots, p_{r_P}$ ($r_P = 50$), e.g. the number of prescriptions per provider
- reporting results for member behavior $m_1, \ldots, m_{r_M}$ ($r_M = 30$), e.g. the number of different doctors visited
- reporting results for network behavior $n_1, \ldots, n_{r_N}$ ($r_N = 10$), e.g. the distance between member and provider
- different invoice parameters $c_1, \ldots, c_{r_C}$ ($r_C = 10$), e.g. the number of invoicelines.

Due to the high amount of potential influence factors and the fact that they are partly highly correlated, a full model with all of these factors as covariates would be highly unstable and have little predictive power. Instead, we summarize all behavioral aspects of provider, member and network behavior in latent variables υ_P, υ_M and υ_N which can be interpreted as behavioral scores and used as model covariates. In this way, we reduce the dimension of the covariate space and increase the stability and transferability of the model for prediction. The model equation then has the following form:

$$\log\left(\frac{\pi_{ij}}{\pi_{\text{NI}i}}\right) = \alpha_{j0} + \alpha_{j1}c_{i1} + \cdots + \alpha_{jr_C}c_{ir_C} + \sum_j \upsilon_{Pij}\beta_{Pj} + \sum_j \upsilon_{Mij}\beta_{Mj} + \sum_j \upsilon_{Nij}\beta_{Nj} \quad \text{with } j \in \{UP, UJ, BI\}\,. \tag{2}$$

This means that we are modeling the (logarithmized) odds that an invoice belongs to one of the fraud and abuse categories UP, UJ and BI instead of to the reference category NI. In this first stage model, the αs are the regression coefficients belonging to the invoice parameters, and the βs the regression coefficients belonging to the behavioral scores related to invoice i.

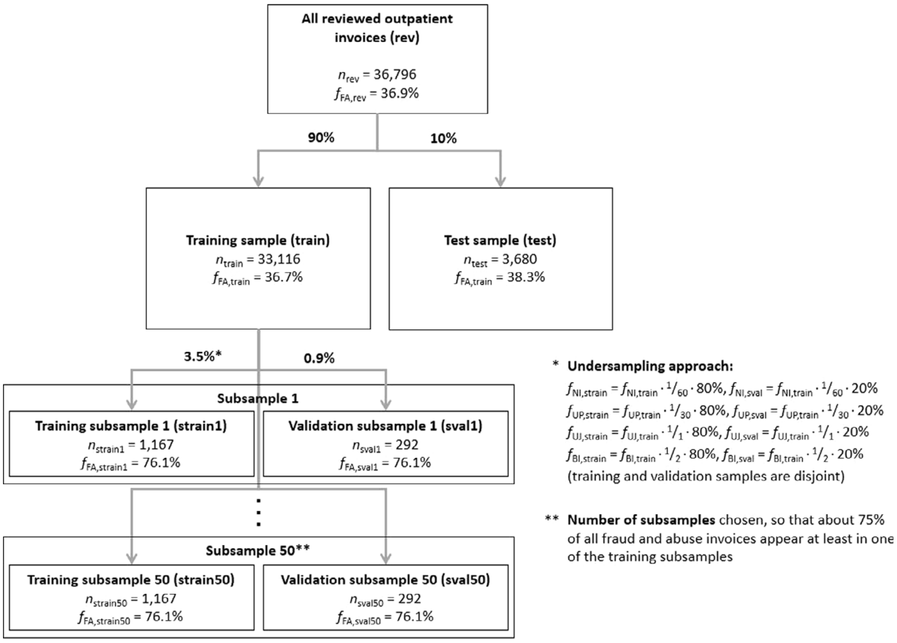

All reviewed outpatient invoices (rev)
$n_{\text{rev}} = 36{,}796$
$f_{\text{FA,rev}} = 36.9\%$
90%
10%
Training sample (train)
$n_{\text{train}} = 33{,}116$
$f_{\text{FA,train}} = 36.7\%$
Test sample (test)
$n_{\text{test}} = 3{,}680$
$f_{\text{FA,train}} = 38.3\%$
3.5%*
0.9%
Subsample 1
Training subsample 1 (strain1)
$n_{\text{strain1}} = 1{,}167$
$f_{\text{FA,strain1}} = 76.1\%$
Validation subsample 1 (sval1)
$n_{\text{sval1}} = 292$
$f_{\text{FA,sval1}} = 76.1\%$
Subsample 50**
Training subsample 50 (strain50)
$n_{\text{strain50}} = 1{,}167$
$f_{\text{FA,strain50}} = 76.1\%$
Validation subsample 50 (sval50)
$n_{\text{sval50}} = 292$
$f_{\text{FA,sval50}} = 76.1\%$
* Undersampling approach:
$f_{\text{NI,strain}} = f_{\text{NI,train}} \cdot {}^1/_{60} \cdot 80\%, f_{\text{NI,sval}} = f_{\text{NI,train}} \cdot {}^1/_{60} \cdot 20\%$
$f_{\text{UP,strain}} = f_{\text{UP,train}} \cdot {}^1/_{30} \cdot 80\%, f_{\text{UP,sval}} = f_{\text{UP,train}} \cdot {}^1/_{30} \cdot 20\%$
$f_{\text{UJ,strain}} = f_{\text{UJ,train}} \cdot {}^1/_{1} \cdot 80\%, f_{\text{UJ,sval}} = f_{\text{UJ,train}} \cdot {}^1/_{1} \cdot 20\%$
$f_{\text{BI,strain}} = f_{\text{BI,train}} \cdot {}^1/_{2} \cdot 80\%, f_{\text{BI,sval}} = f_{\text{BI,train}} \cdot {}^1/_{2} \cdot 20\%$
(training and validation samples are disjoint)
** Number of subsamples chosen, so that about 75% of all fraud and abuse invoices appear at least in one of the training subsamples

Fig. 2. Sampling strategy to control overfitting and to increase model transferability.

The behavioral scores/latent variables are the result of aggregating provider, member and network reports. Similar to the idea of structural equation models (Goldberger, 1972), this aggregation process is performed by fitting linear second stage models of the form

$$\begin{aligned} v_{Pij} &= \gamma_{Pj0} + \gamma_{Pj1} p_{i1} + \cdots + \gamma_{Pjr_P} p_{ir_P} + \varepsilon_{Pij} \\ v_{Mij} &= \gamma_{Mj0} + \gamma_{Mj1} m_{i1} + \cdots + \gamma_{Mjr_M} m_{ir_M} + \varepsilon_{Mij} \\ v_{Nij} &= \gamma_{Nj0} + \gamma_{Nj1} n_{i1} + \cdots + \gamma_{Njr_N} n_{ir_N} + \varepsilon_{Nij} \\ &\text{with } j \in \{\text{UP,UJ,BI}\}. \end{aligned} \tag{3}$$

The γs denote the regression coefficients of the second stage models and the εs the error terms of the models. As target variables of these models, we use binary indicators which determine whether invoice i falls into the category j or not. As we use linear second stage models the estimates of the behavioral scores/latent variables $\hat{\boldsymbol{v}}_P$, $\hat{\boldsymbol{v}}_M$ and $\hat{\boldsymbol{v}}_N$ range between $-\infty$ and ∞. These estimates are interpreted as "observed behavioral scores" and serve as covariates in the first stage model. They can also be used to rank provider and members by their fraud and abuse potential, which is a beneficial side effect.

As both first and second stage models depend on the class affiliation y_i, the model fitting process is based on an iterative update of first and second stage model parameters. In the following Section 2.2 we describe a Bayesian fitting algorithm which follows this principle and unites first and second stage models in a common posterior representation. The developed model can be seen as a onedimensional special case of the multidimensional latent variable model proposed by Sammel et al. (1997). Compared to other authors who have addressed this class of models, we apply it to model a multinomial outcome instead of continuous, binary, ordinal or Poisson distributed outcomes (Fahrmeir and Steinert, 2006; Fahrmeir and Raach, 2007; Adebayo et al., 2011).

2.2. Model fitting

In order to fit the model parameters $\boldsymbol{\alpha} = (\alpha_{j0}, \ldots, \alpha_{jr_C})$, $\boldsymbol{\beta} = (\beta_{\text{PUP}}, \ldots, \beta_{\text{NBI}})$ and $\boldsymbol{\gamma} = (\gamma_{\text{PUP0}}, \ldots, \gamma_{\text{NBIr}_N})$, we apply a joint fitting algorithm for both stages based on Metropolis–Hastings sampling. Final estimates are derived from the subsampling approach already mentioned in Section 1.3 by calculating median values over all subsamples. The MCMC sampling approach is a straightforward choice for parameter estimation because it intuitively allows an alternating update of first and second stage model parameters and is computationally efficient. The algorithm uses the following (simplified) reformulation of the posterior distribution of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$:

$$\begin{aligned} \underbrace{p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} | y, \boldsymbol{v})}_{\text{posterior}} &\propto \underbrace{p(y, \boldsymbol{v} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})}_{\text{likelihood}} \cdot \underbrace{p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})}_{\text{prior}} \\ &= p(y|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \cdot p(\boldsymbol{v}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \cdot p(\boldsymbol{\alpha}, \boldsymbol{\beta}) \cdot p(\boldsymbol{\gamma}) \\ &= \underbrace{p(y|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \cdot p(\boldsymbol{\alpha}, \boldsymbol{\beta})}_{\text{first stage model}} \cdot \underbrace{p(\boldsymbol{v}|\boldsymbol{\gamma}) \cdot p(\boldsymbol{\gamma})}_{\text{second stage model}}. \end{aligned} \tag{4}$$

The likelihood in the first line is split into two parts based on the assumption of conditional independence between $\boldsymbol{v} = (v_{\text{PUP}}, \ldots, v_{\text{NBI}})$ and y given $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

The fitting algorithm starts by drawing the second stage parameters $\boldsymbol{\gamma}$ using the Metropolis–Hastings algorithm (Robert and Casella, 2004). For the linear models, we apply a hierarchical decomposition model to draw from the posterior distribution of $\boldsymbol{\gamma}$ and the error variance σ^2:

$$\underbrace{p(\boldsymbol{\gamma}, \sigma^2 | v)}_{\text{posterior}} \propto \underbrace{p(v|\boldsymbol{\gamma}, \sigma^2)}_{\text{likelihood}} \cdot \underbrace{p(\boldsymbol{\gamma}|\sigma^2) \cdot p(\sigma^2)}_{\textit{prior}}. \tag{5}$$

As an additional instrument to control overfitting to the training data, a parameter shrinkage option is included in the algorithm.

For the version without parameter shrinkage, we use Gibbs sampling as a special case of the Metropolis–Hastings algorithm (acceptance probability of proposed parameters always equal to one) implemented in the function `MCMCregress` in the R package `MCMCpack` (Martin et al., 2011). Here, a semi-conjugate prior to the normal likelihood $p(y|\boldsymbol{\gamma}, \sigma^2)$ is used. More precisely, we use a weakly informative multivariate normal prior distribution for $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma} \sim \text{N}(\mathbf{0}, 0.1\boldsymbol{I}) \tag{6}$$

and an inverse Gamma prior for the error variance σ^2 (with mean 5 and variance 25) which is assumed to be independent from $\boldsymbol{\gamma}$.

For the version with parameter shrinkage, we apply a lasso regression model (Park and Casella, 2008) with reversible

Table 1
Prediction approaches and techniques applied to the underlying classification problem.

Approach	Technique	Shortcut
One-stage model with variable selection/weighting	Multinomial model with AIC variable selection	A1
	Random Forest	A2
One-stage model with dimension reduction	Multinomial model based on factor analysis	B1
	Polyclass model based on factor analysis	B2
Two-stage model	Bayesian latent variable model without shrinkage	C1
	Bayesian latent variable model with shrinkage	C2

jump mechanism (Green, 1995; Troughton and Godsill, 1997) implemented in the blasso function in the R package monomvn (Gramacy et al., 2007). The reversible jump algorithm modifies the acceptance probability of the applied Metropolis–Hastings algorithm by introducing moves between parameter spaces of different dimensionality (Troughton and Godsill, 1997).

The lasso penalization is implemented by including the shrinkage parameter λ which controls the degree of parameter shrinkage in the hierarchical model representation. We then assume an exponential power distribution as prior for $\boldsymbol{\gamma}$ which is no longer independent from the error variance σ^2:

$$p(\boldsymbol{\gamma}|\sigma^2) \propto \prod_{l=1}^{r} e^{-\lambda(|\gamma_l|/\sqrt{\sigma^2})}. \quad (7)$$

For λ (more precisely, λ^2), we use a non-informative Jeffrey's hyperprior (Gramacy et al., 2007).

After having updated the second stage models for the first time, we use the updated parameter vector $\boldsymbol{\gamma}^{(1)}$ to calculate first estimations of the latent variable scores $v_{\text{PUP}}^{(1)}, \ldots, v_{\text{NBI}}^{(1)}$ for all observations in the subsample. Based on these estimations, an update of the first stage model can be performed.

For updating the first stage model, we use the MCMCmnl function in the R package MCMCpack (Martin et al., 2011). We assume a multivariate normal distribution with expectation equal to zero and infinite variance for $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$, i.e. an improper non-informative prior distribution. We receive first updates $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\beta}^{(1)}$ by independent Metropolis–Hastings sampling (Chib et al., 1998). The Metropolis proposal distribution is centered at the current value of $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ and has the covariance matrix TCT. T is a diagonal positive definite matrix depending on the tuning parameter t that controls the acceptance rate and C is the large sample covariance matrix of the maximum likelihood estimate of $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$.

In each further iteration s ($s = 1, \ldots, S = 250$) of the fitting process, we use $\boldsymbol{\gamma}^{(s-1)}$ on the second stage and $\boldsymbol{\alpha}^{(s-1)}$ and $\boldsymbol{\beta}^{(s-1)}$ on the first stage as initial parameter vectors of the sampling functions. We stabilize the prediction results by calculating the median of all $\boldsymbol{\alpha}^{(s)}$, $\boldsymbol{\beta}^{(s)}$ and $\boldsymbol{\gamma}^{(s)}$ after burn-in (i.e. $s = 51, \ldots, 250$) over all subsamples denoted by $\boldsymbol{\alpha}_{\text{med}}$, $\boldsymbol{\beta}_{\text{med}}$ and $\boldsymbol{\gamma}_{\text{med}}$. By plugging in these final estimates as well as the observed reporting results and invoice parameters related to a new invoice k into the model Eqs. (2) and (3), we get a prediction $\boldsymbol{\pi}_k^* = (\pi_{k\text{UP}}^*, \pi_{k\text{UJ}}^*, \pi_{k\text{BI}}^*, \pi_{k\text{NI}}^*)$ of the fraud and abuse probabilities for this invoice.

Now, we can, for instance, use the maximum of $\boldsymbol{\pi}_k^*$ to determine the predicted class y_k^* and compare it with the real class y_k of invoice k in the test sample. Based on this comparison, a misclassification matrix and further measures of predictive performance (see Section 3) can be calculated. Of course, Bayesian modeling also allows deriving credibility intervals for the parameter estimates and predicted probabilities.

Both versions of the algorithm, with shrinkage and without shrinkage, lead to a reasonable number of reports which have measurable influence on the final predictions, where of course, the shrunk version produces sparser models. In general, the methodology described yields consistent parameter estimates and prediction results, not sensitive to changes in the data input, which is indicated by a relatively small variation of parameter estimates across the subsamples. This stability is a beneficial property that increases the transferability of the model.

2.3. Benchmarking

In order to benchmark the predictive performance of the developed Bayesian Latent Variable Model in the existing data situation, we compare it to several other prediction techniques. As explained, the Bayesian approach involves two stages to predict class probabilities for all response categories. In both steps we relate the covariates to the real response category using different distributional assumptions for the response of both model parts (normal and multinomial). An alternative strategy is to first reduce the dimension of the covariate space without consideration of the target variable and use the resulting factors or principle components as input for the class prediction model. A third possible strategy is to combine variable selection/weighting and class prediction in one step.

Table 1 summarizes the applied approaches together with concrete prediction techniques that we use for benchmarking. All benchmark models have been calculated based on the subsampling approach described in Section 1.3.

Models A1 and A2 assume a direct relationship between the reporting results and class affiliation of corresponding invoices without any latent variables or factors in between. Model A1 is a multinomial logit model fitted via Neural Networks as implemented in the function multinom of the R package nnet (Venables and Ripley, 2002). To ensure model convergence and avoid overfitting in view of the high number of covariates, we combine this technique with a stepwise (forward) variable selection algorithm based on the AIC criterion. As an alternative direct prediction method, we use a Random Forest which is a standard method for class prediction. The forest is calculated (and tuned) with the R function randomForest in the R package of the same name (Breiman, 2001).

Like our Bayesian approach, models B1 and B2 reduce the dimension of the covariate space before performing the final classification. To reduce the dimension of the covariate space, we use a factor analysis to conserve as much as possible of the covariance between the reports resulting in 10 factors summarizing provider, member and network behavior. Based on these factors, we apply either a multinomial logit model (like for model A1) or a polyclass model as described by Kooperberg et al. (1997) and Stone et al. (1997). Polyclass models are related to the MARS (multivariate adaptive regression splines) approach first described by Friedman (1991). They include a stepwise variable selection approach that also takes into account non-linear covariate effects and potential interactions between model covariates. The polyclass models are implemented in the function polyclass of the R package polspline.

Models C1 and C2 correspond to the Bayesian approach described in Sections 2.1 and 2.2. Model C1 represents the fitting algorithm without parameter shrinkage and model C2 represents the version with parameter shrinkage.

Table 2
Category-specific and average *AUC*s of all applied classification techniques based on the test sample (overall best results marked in boldface).

Technique	AUC_{NI}	AUC_{UP}	AUC_{UJ}	AUC_{BI}	$\overline{AUC}$
A1	**0.90**	0.74	0.70	0.75	0.77
A2	0.67	0.60	0.75	0.76	0.70
B1	0.78	0.81	0.77	0.79	0.79
B2	0.74	0.81	0.84	**0.90**	0.82
C1	0.80	**0.83**	**0.92**	0.86	**0.85**
C2	0.80	**0.83**	0.91	0.86	**0.85**

Table 3
Content-specific performance measures of all applied classification techniques based on the test sample (overall best results marked in boldface).

Technique	TPR_{FA}	CCR_{FA}	PPV_{NI}	AAR_{NI}	PPV_{FA}	$n_{95\%}$
A1	0.74	0.50	**0.99**	0.37	0.14	20
A2	0.74	0.54	0.96	0.38	0.13	22
B1	0.81	0.57	0.89	0.37	0.15	18
B2	0.83	0.60	0.85	0.45	**0.17**	**16**
C1	0.83	0.62	0.97	0.51	0.16	17
C2	**0.84**	**0.64**	0.97	**0.53**	**0.17**	**16**

3. Results

In the following, we present the results of different prediction techniques (see Section 2.3) based on the original dataset from the Middle East using the test sample described in Section 1.3. As already mentioned in Section 1.3, it needs to be considered that both training and test dataset are based on a biased pre-selection of invoices. For this reason, the relative frequency of fraud and abuse cases is very high and the relative frequency of abuse cases compared to fraud cases is comparably low. Also, several other characteristics of the dataset (e.g. age distribution) and the market background are not comparable to most other, especially US and European, markets. Consequently, no generalization of results in terms of absolute predictive quality is possible. Target of the benchmarking is to provide an indication of relative predictive quality for the full set of invoices of the same company and datasets of other insurance companies based on the subsampling approach described in Section 1.3.

First, we compare the overall predictive quality of all applied techniques measured by *AUC* (Pearce and Ferrier, 2000; Japkowicz and Shah, 2011) based on the test sample (see Section 1.3). Table 2 shows both the category-specific *AUC*s (using binary class indicators) and the (unweighted) average *AUC* over all categories. In order to also visualize the results of the analysis, Fig. 3 summarizes all ROC curves of the applied classification techniques.

In general, the one-step approaches with variable selection/weighting (A1 and A2) perform worst in terms of overall predictive quality. The best results in the existing data environment are obtained from the Bayesian latent variable model (C1 and C2), shortly followed by the regression approaches with preliminary dimension reduction (B1 and B2), which confirms the assumption of a latent behavioral effect. In addition, the two-stage approach with both stages relating the covariates to the target variable does not seem to cause overfitting. It also seems to be a reasonable choice here, because *AUCs* are relatively high across all response categories. This indicates that the efficiency of the future investigation process can also be increased with regard to abuse and other billing issues.

The one-stage AIC selection model (A1) performs best in terms of NI cases, but clearly worse than most other techniques in the identification of fraud and abuse cases. The Random Forest (A2) performs worst on the test sample, but could potentially be improved with further parameter tuning. Regarding the techniques with preliminary factor analysis (B1 and B2), it is noticeable that the non-linear polyclass model (B2) performs better in the categories UJ and BI compared to the standard multinomial model (B1). Here, we observe some non-linear effects as well as interactions which improve the predictive quality. The Bayesian latent variable models without and with shrinkage (C1 and C2) deliver very similar results. We assume that the effect of variable shrinkage may be more significant for smaller sample sizes.

Beside the overall predictive quality, we also evaluate some specific performance indicators relevant from the perspective of insurance companies and other healthcare payers. For this, we introduce the additional category "fraudulent or abusive" (FA) meaning that the invoice belongs to one of the categories UP, UJ or BI. Table 3 summarizes all content-related measures based on the test sample.

The true positive rate (Japkowicz and Shah, 2011) of the category FA TPR_{FA} gives an indication of the percentage of real fraud and abuse cases that can be identified as such. Also here, it must be considered that the figures are assumed to be substantially lower for the full set of invoices of the insurance company due to the mentioned pre-selection bias. Regarding the results based on the test sample, the Bayesian latent variable model with parameter shrinkage performs best with a true positive rate of the category FA of 84%.

Also in terms of the percentage of correct class allocation within all actual fraud and abuse cases CCR_{FA}, the Bayes models delivers the best results with a correct class allocation rate of above 60%. Regarding the positive predictive value (Japkowicz and Shah, 2011) of NI cases PPV_{NI}, i.e. the precision with which "clean" cases are identified, the AIC selection model (A1) performs best with a precision of 99% (see also the high *AUC* value for the category NI). However, the auto-adjudication rate AAR_{NI}, i.e. the percentage of predicted NI invoices among all invoices, is only about 37%. Here, the Bayes models are clearly preferable with a similar precision (97%), but a clearly higher auto-adjudication potential of more than 50%. Even though these figures only hold for the test dataset biased through pre-selection, the comparison of techniques indicates that the Bayes models yield the highest savings potential for insurers in terms of operational costs.

Finally, the reliability of a decision to filter out an invoice is analyzed. Here, the models B1 and B2 with preliminary dimension reduction are on the same level as the Bayesian latent variable models with a positive predictive value (Japkowicz and Shah, 2011) of fraud and abuse PPV_{FA} above 15%. This means – at least for the existing data situation – that less than 20 invoices need to be analyzed to uncover one fraud or abuse case with a probability of 95% (see column $n_{95\%}$ in Table 3).

Looking at the overall picture, the results are quite similar to the *AUC* evaluation: the Bayes models perform best, shortly followed by the dimension reduction models. The content-related measures give a rough indication of the saving potential for insurers and healthcare payers in the existing data situation. Due to the stability of results, we recommend applying the Bayesian latent variable model and the dimension reduction techniques in similar situations, even though we assume that the absolute predictive quality cannot be generalized. Especially, the results in the abuse category require further validation based on other datasets, due to the small number of cases in the existing situation.

4. Summary and outlook

This article transfers the idea of Bayesian latent variable modeling known from other contexts, such as social science (Fahrmeir and Raach, 2007; Fahrmeir and Steinert, 2006), to the

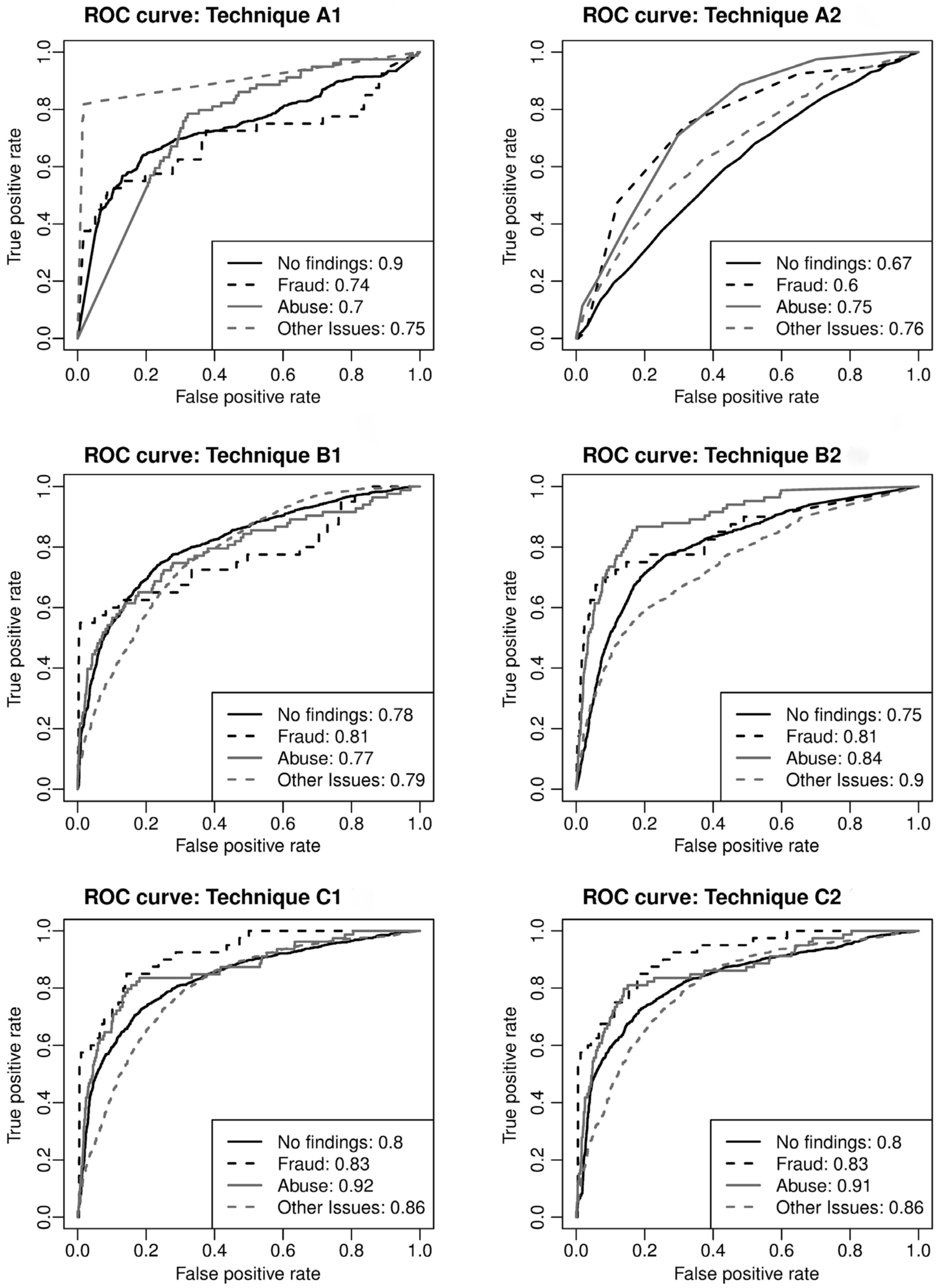


Fig. 3. ROC curves of all applied classification techniques based on the test sample.

problem of fraud and abuse detection. To handle the challenges of this specific problem, we developed a Bayesian latent variable model for multinomial response variables. The idea behind the approach is that the report results do not directly affect the class affiliation of an invoice, but indirectly via latent variables which summarize behavioral aspects of each reporting perspective. In this way, we aspire to exploit the full information potential of the underlying expert system. The developed fitting algorithm involves a lasso parameter shrinkage option intended to control overfitting to the training sample. To further increase the transferability of the model, a subsampling approach is used to balance the skewed class distribution in training data.

The prediction results indicate that the introduced Bayes approach improves the relative predictive quality compared to several benchmarking techniques. Therefore, we assume that the methodology can successfully be applied to the full set of invoices of the same insurance company and to datasets of other insurance companies. However, the observed absolute predictive quality is not realistic due to a pre-selection bias affecting training and test data. Especially those approaches which directly use the report results as input do not exploit the full potential of the data in terms of predictive quality. Regression techniques with preliminary dimension reduction (factor analysis) yield a predictive quality similar to the Bayesian approach. As these techniques are computationally faster and based on standard routines of analytical software packages, they may be preferred in practical situations, especially in the case of very large datasets.

A general advantage of the Bayesian approach is its adaptability to other data and market environments. The expected influence of new and already existing reports can be adjusted by corresponding (informative) prior specification. In situations where no or only few reviewed cases are available as learning data, the transfer of the model including (modified) parameter estimates is assumed to be still more efficient than the application of an unsupervised system only focused on outlier detection. Prerequisite for such transfers are comparable data availability and quality in the original and the new market. On the other hand, specification of wrong informative prior distributions may decrease the predictive quality in case of small training datasets which is a drawback compared to frequentist approaches.

The developed model can be applied in real time based on a score card approach. This means that the model does not need to be re-fitted for every new invoice, but only on a monthly or quarterly basis. This frequency is assumed to be sufficient to keep pace with the dynamic adaptation of behavior of providers and insured persons. The high stability of the described modeling approach and the prediction results (i.e. low dependency on data input) ensures the validity of the scoring for this period. Another quality of the Bayesian approach is its high interpretability. In particular, it allows backtracking of a high probability to the report results which caused it. This gives the investigator a starting point for further evaluation and allows a more targeted investigation process.

The efficiency of the claims adjudication process can further be increased by establishing specific investigation units which are trained to take appropriate action based on the scoring results. It is also important to note that usually only a small percentage of the saving potential with regard to fraudulent and abusive behavior is related to direct recovery. The larger part is assumed to arise from a measurable deterrent effect as well as a better position of the insurance company in provider network negotiations.

Statements and conclusions presented in this article mainly refer to the existing data situation. Therefore, the subject of further research will be to test the (absolute) predictive performance and stability of the model based on other insurance or healthcare datasets. Based on validated results, it further needs to be assessed from an economic perspective if the increase in predictive quality can be translated into cost savings which justify the implementation effort. Especially interesting is whether the shrinkage option will lead to a stronger differentiation of predictive quality in smaller datasets. Also, further benchmarking techniques, like a recent boosting approach for multi-class problems with high dimensional covariate space suggested by Zahid and Tutz (2013), may be considered. A potential extension of the model is the inclusion of a spatial term in the linear predictors of first stage models (Adebayo et al., 2011), as fraudulent and abusive behavior is known to be strongly dependent on geographic location. Besides, it would be interesting to evaluate whether more latent variables can improve the predictive quality, e.g. the allocation of provider reports to several behavioral aspects represented by own latent variables/first-stage models.

References

Adebayo, S.B., Fahrmeir, L., Seiler, C., Heumann, C., 2011. Geoadditive latent variable modeling of count data on multiple sexual partnering in nigeria. Biometrics 67, 620–628.

Aral, K.D., Güvenir, H.A., Sabuncuoğlu, İ, Akar, A.R., 2012. A prescription fraud detection model. Comput. Methods Progr. Biomed. 106, 37–46.

Bermúdez, L., Pérez, J.M., Ayuso, M., Gómez, E., Vázquez, F.J., 2008. A bayesian dichotomous model with asymmetric link for fraud in insurance. Insurance Math. Econom. 42 (2), 779–786.

Berwick, D.M., Hackbarth, A.D., 2012. Eliminating Waste in US Health Care. JAMA 307, 1513–1516.

Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J.C., 2011. Data mining for credit card fraud: A comparative study. Decis. Support Syst. 50, 602–613.

Bolton, R.J., Hand, D.J., 2002. Statistical fraud detection: A review. Statist. Sci. 17, 235–255.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Busch, R.S., 2008. Healthcare Fraud: Auditing and Detection Guide. John Wiley & Sons.

Chib, S., Greenberg, E., Chen, Yu., 1998. MCMC methods for fitting and comparing multinomial response models. Econometrics, EconWPA, http://EconPapers.repec.org/RePEc:wpa:wuwpem:9802001.

Dua, P., Bais, S., 2014. Supervised learning methods for fraud detection in healthcare insurance. In: Machine Learning in Healthcare Informatics. Springer, pp. 261–285.

Ekina, T., Leva, F., Ruggeri, F., Soyer, R., 2013. Application of Bayesian methods in detection of healthcare fraud. Chem. Eng. Trans. 33.

EU-Commission. Study on corruption in the healthcare sector. http://www.ehfcn.org/images/EHFCN/Documents/EHFCN_ECORYS_20131219_study_on_corruption_in_the_healthcare_sector_en.pdf, October, 2013, HOME/2011/ISEC/PR/047-A2.

Ezekiel, J.E., Fuchs, V.R., 2008. The perfect storm of overutilization. JAMA 299, 2789–2791.

Fahrmeir, L., Raach, A., 2007. A Bayesian semiparametric latent variable model for mixed responses. Psychometrika 72, 327–346.

Fahrmeir, L., Steinert, S., 2006. A geoadditive Bayesian latent variable model for Poisson indicators. Technical report, Discussion paper Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Statist. 1–67.

Gee, J., Button, M., 2014. The financial cost of healthcare fraud 2014. Technical report. BDO LLP.

Goldberger, A.S., 1972. Structural equation methods in the social sciences. Econometrica 979–1001.

Gramacy, R.B., Lee, J.H., Silva, R., On estimating covariances between many assets with histories of highly variable length, 2007. arXiv preprint arXiv:0710.5837.

Green, P.J., 1995. Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika 82, 711–732.

He, H., Wang, J., Graco, W., Hawkins, S., 1997. Application of neural networks to detection of medical fraud. Expert Syst. Appl. 13, 329–336.

Japkowicz, N., Shah, M., 2011. Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press.

Jin, Y., Rejesus, R.M., Little, B.B., 2005. Binary choice models for rare events data: a crop insurance fraud application. Appl. Econ. 37, 841–848.

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., Arab, M., 2014. Using data mining to detect health care fraud and abuse: A review of literature. Global J. Health Sci. 7, 194–202.

Kooperberg, C., Bose, S., Stone, C.J., 1997. Polychotomous regression. J. Amer. Statist. Assoc. 92 (437), 117–127.

Li, J., Huang, K.-Y., Jin, J., Shi, J., 2008. A survey on statistical methods for health care fraud detection. Health Care Manag. Sci. 11, 275–287.

Major, J.A., Riedinger, D.R., 2002. EFD: A hybrid knowledge/statistical-based system for the detection of fraud. J. Risk Insurance 69, 309–324.

Martin, A.D., Quinn, K.M., Park, J.H., 2011. MCMCpack: Markov chain monte carlo in R. J. Stat. Softw. 42, 1–21.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman & Hall/CRC.

Musal, R.M., 2010. Two models to investigate medicare fraud within unsupervised databases. Expert Syst. Appl. 37, 8628–8633.

Ngai, E.W.T., Hu, Y., Wong, Y.H., Chen, Y., Sun, X., 2011. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decis. Support Syst. 50, 559–569.

Park, T., Casella, G., 2008. The Bayesian Lasso. J. Amer. Statist. Assoc. 103, 681–686. http://www.stat.ufl.edu/~casella/Papers/Lasso.pdf.

Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Model. 133, 225–245.

Peng, Y., Kou, G., Sabatka, A., Matza, J., Chen, Z., Khazanchi, D., Shi, Y., 2007. Application of classification methods to individual disability income insurance fraud detection. In: Computational Science, ICCS 2007. Springer, pp. 852–858.

Phua, C., Lee, V., Smith, K., Gayler, R., 2005. A comprehensive survey of data mining-based fraud detection research. Artifical Intell. Rev. 1–14.

Robert, C.P., Casella, G., 2004. Monte Carlo Statistical Methods. Springer, New York.

Sahin, Y., Duman, E., 2011a. Detecting credit card fraud by ANN and logistic regression. In: 2011 International Symposium on Innovations in Intelligent Systems and Applications, INISTA. IEEE, pp. 315–319.

Sahin, Y., Duman, E., 2011b. Detecting credit card fraud by decision trees and support vector machines In: International MultiConference of Engineers and Computer Scientists, Vol. 1.

Sammel, M.D., Ryan, L.M., Legler, J.M., 1997. Latent variable models for mixed discrete and continuous outcomes. J. R. Stat. Soc. Ser. B Stat. Methodol. 59, 667–678.

Skrondal, A., Rabe-Hesketh, S., 2004. Generalized Latent Variable Modeling. Chapman & Hall/CRC.

Stone, C.J., Hansen, M., Kooperberg, C., Truong, Y.K., 1997. The use of polynomial splines and their tensor products in extended linear modeling (with discussion). Ann. Statist. 25, 1371–1470.

Šubelj, L., Furlan, Š, Bajec, M., 2011. An expert system for detecting automobile insurance fraud using social network analysis. Expert Syst. Appl. 38, 1039–1052.

Thornton, D., Mueller, R.M., Schoutsen, P., van Hillegersberg, J., 2013. Predicting healthcare fraud in Medicaid: A multidimensional data model and analysis techniques for fraud detection. Procedia Technol. 9, 1252–1264.

Troughton, P.T., Godsill, S.J., 1997. A reversible jump sampler for autoregressive time series, employing full conditionals to achieve efficient model space moves. CUED/F-INFENG/TR. 304, University of Cambridge: Department of Engineering.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, fourth ed.. Springer.

Viaene, S., Derrig, R.A., Baesens, B., Dedene, G., 2002. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. J. Risk Insurance 69, 373–421.

Viaene, S., Derrig, R.A., Dedene, G., 2004a. Cost-sensitive learning and decision making for Massachusetts PIP claim fraud data. Int. J. Intell. Syst. 19, 1197–1215.

Viaene, S., Derrig, R.A., Dedene, G., 2004b. A case study of applying boosting Naive Bayes to claim fraud diagnosis. IEEE Trans. Knowl. Data Eng. 16 (5), 612–620.

Viaene, S., Dedene, G., Derrig, R.A., 2005. Auto claim fraud detection using bayesian learning neural networks. Expert. Syst. Appl. 29, 653–666.

Wallace, B.C., Small, K., Brodley, C.E., Trikalinos, T.A., 2011. Class imbalance, redux. In: 2011 IEEE 11th International Conference on Data Mining, ICDM. IEEE, pp. 754–763.

Yang, W.-S., Hwang, S.-Y., 2006. A process-mining framework for the detection of healthcare fraud and abuse. Expert. Syst. Appl. 31, 56–68.

Zahid, F.M., Tutz, G., 2013. Multinomial logit models with implicit variable selection. Adv. Data Anal. Classif. 7 (4), 393–416.