Published in final edited form as:

Stat Med. 2014 December 20; 33(29): 5192–5208. doi:10.1002/sim.6314.

EM for Regularized Zero Inflated Regression Models with Applications to Postoperative Morbidity after Cardiac Surgery in Children

Zhu Wang^{a,*}, Shuangge Ma^b, Ching-Yun Wang^c, Michael Zappitelli^d, Prasad Devarajan^e, and Chirag Parikh^f

^aDepartment of Research, Connecticut Children's Medical Center, Department of Pediatrics, University of Connecticut School of Medicine, Hartford, CT, USA

^bDepartment of Biostatistics, School of Public Health, Yale University, New Haven, CT, USA

^cPublic Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

^dDivision of Nephrology, Department of Pediatrics, Montreal Childrens Hospital, McGill University Health Centre, Montreal, Quebec, Canada

^eDepartment of Nephrology, and Hypertension, Center for Acute Care Nephrology, Cincinnati Childrens Hospital Medical Center, Cincinnati, OH, USA

^fSection of Nephrology and Program of Applied Translational Research, Department of Medicine, Yale University School of Medicine, New Haven, CT, Clinical Epidemiology Research Center, Veterans Affairs Medical Center, West Haven, CT, USA

Abstract

This paper proposes a new statistical approach for predicting postoperative morbidity such as intensive care unit length of stay and number of complications after cardiac surgery in children. In a recent multi-center study sponsored by the National Institutes of Health, 311 children undergoing cardiac surgery were enrolled. Morbidity data are count data in which the observations take only non-negative integer values. Often the number of zeros in the sample cannot be accommodated properly by a simple model, thus requiring a more complex model such as the zero-inflated Poisson (ZIP) regression model. We're interested in identifying important risk factors for postoperative morbidity among many candidate predictors. There is only limited methodological work on variable selection for the zero-inflated regression models. In this paper, we consider regularized ZIP models through penalized likelihood function, and develop a new expectation-maximization (EM) algorithm for numerical optimization. Simulation studies show that the proposed method has better performance than some competing methods. Using the proposed methods, we analyzed the postoperative morbidity, which improved the model fitting and identified important clinical and biomarker risk factors.

*Correspondence to: Department of Research, Connecticut Children's Medical Center, Department of Pediatrics, University of Connecticut School of Medicine, Hartford, CT, USA. zwang@connecticutchildrens.org.

SUPPLEMENTARY MATERIAL

Supplementary material includes numerical algorithms for fitting the regularized generalized linear models, statistical properties of the proposed methods, and extensive simulation results for scenarios 2–5.

Keywords

ZIP; variable selection; LASSO; MCP; SCAD

1. Introduction

About two million cardiac surgeries are performed worldwide each year. As mortality rates from cardiac surgery (CS) have declined, research on morbidity as an outcome has become increasingly more relevant. Prolonged length of stay (LOS) and post-surgical complications are examples of important measures of morbidity after CS. These morbidity outcomes are associated with high mortality, high costs and burdens to patients and families.

Complication rate and LOS are used as measures of hospital efficiency and a reasonable proxy of resource consumption. Accurate prediction of these outcomes can help determine indications for surgery, provide patient counseling, allocate resources, and monitor quality of surgeons and institutions. There is extensive research on predicting post-CS LOS and complications in adults [1, 2]. However, these risk models are not applicable to children with congenital heart disease because the underlying diagnoses and co-morbidities drastically differ [3, 4]. There is very limited research on morbidity outcomes in children [5, 6, 7]. Moreover, the existing morbidity outcome research in adults and children suffers from methodological drawbacks since the stepwise procedure was typically used to develop models but the stepwise approach is well known for its unstable performance.

We propose new statistical methodology for evaluating morbidity outcomes, using clinical and biomarker data from patients enrolled in the NIH sponsored Transnational Research Investigating Biomarker Endpoints (TRIBE) study, which includes a cohort of children undergoing CS. The TRIBE study was designed to identify novel biomarkers for early acute kidney injury (AKI) diagnosis. In a prospective cohort study, 311 children between 1 month and 18 years old undergoing CS at three academic medical centers were enrolled from July 2007–December 2009. The study collected pre-, intra-, and postoperative variables. Postoperative complications include operative, renal, neurologic, pulmonary and others. The TRIBE study also collected preoperative and postoperative urine and plasma biomarkers of AKI. The presence of AKI and these AKI biomarkers were strongly associated with several outcomes, including LOS. In this paper, we will identify risk factors influencing ICU LOS and complications of CS using the TRIBE study data. What makes the TRIBE study data unique is the inclusion of novel biomarkers of AKI. These biomarkers increase the specificity of capturing AKI diagnosis, and the use of AKI biomarkers is at the forefront of research evaluating the effect of AKI on outcome.

Evaluating morbidity outcomes evokes a new complexity to studying CS outcomes due to the statistical characteristics of these events. LOS and the number of complications are characterized by (i) nonnegative integers, (ii) a skewed distribution and (iii) a large proportion of zeros. Figure 1 illustrates these issues within the TRIBE cohort of patients undergoing CS. Often the number of zeros in the sample cannot be handled properly by a simple statistical model. Zero-inflated data refers to data with a heterogeneous distribution, contrary to the simple homogeneous assumptions of standard statistical analysis. For

mathematical convenience, a value of one is subtracted from observed LOS, allowing for the minimum LOS value to be zero and for a unified analytic design for both complications and LOS. Another methodological challenge is the variable selection problem. When building statistical models to predict outcomes, a large number of predictors are usually introduced at the initial model-building stage to reduce possible biases.

Variable selection is fundamental for developing a parsimonious model, but is challenging when the sample size is small, as is characteristic of pediatric studies. Variable selection is the most commonly conducted using stepwise regression [5, 6, 7]. Stepwise regression builds a model sequentially which introduces inherent drawbacks including inconsistencies among model selection algorithms and the problem of multiple hypothesis testing [8, 9]. Alternatively, Akaike information criterion (AIC) [10] and Bayesian information criterion (BIC) [11] can be used. These are computationally expensive and may be infeasible even for moderate numbers of variables [9]. Statistical models for zero-inflated count data have applications on LOS [12, 13], as well as on complications [14, 15]. Proposed strategies of variable selection in zero-inflated models include AIC, BIC and stepwise procedures [16, 5]. However, the state-of-the-art approach is to select variables and estimate coefficients simultaneously [17], which optimizes the likelihood function with an additional penalty function to avoid overfitting and to reduce coefficient variance. These penalty functions include the least absolute shrinkage and selection operator (LASSO) [18], smoothly clipped absolute deviation (SCAD) [9] and minimax concave penalty (MCP) [19], and are particularly useful for problems with small sample size and large number of predictors. Despite wide applications in many statistical problems, these penalized methods have not been well studied with zero-inflated count data. One exception is the one-step LASSO and SCAD [20].

In this paper, we propose a new EM algorithm for the zero-inflated Poisson models. Compared to [20], there are at least three different important directions. First, we consider the MCP which has attracted great attention recently due to its solid theoretic foundation and excellent performance in data analysis. Second, Buu *et al.* [20] considered a local linear approximation (LLA) and employed the LARS algorithm, while we employ an EM algorithm and coordinate descent algorithm. Third, when selecting two penalty tuning parameters, our computationally efficient strategy does not pose any constraint on the tuning parameters. The last point may help explain why the new proposed method outperforms the LLA algorithm in the simulation studies. In addition, for the tuning parameter selection, the one-step estimators require the standard errors of the unpenalized maximum likelihood estimates. This extra task has proved to be difficult for some data with the Newton-Raphson method, particularly when the sample size is small or the number of predictors is large. In fact, we fail to fit unpenalized ZIP models for some simulated and real data. In contrast, the new proposed EM algorithm avoids fitting unpenalized ZIP models, consequently is a robust tool even when the sample size is small. The coordinate descent algorithms have been shown to be competitive with the LARS algorithm [21, 22]. Breheny and Huang [23] proposed the numerically stable coordinate descent algorithm for logistic regression with MCP and SCAD. We adopt the algorithm and further extend to the Poisson regression. The EM algorithm is thus built on these coordinate descent algorithms. We have evidence that

our new algorithm is generally reliable, even for small sample sizes. The numerical convergence problem, however, can become severe with the LLA algorithm. Buu *et al.* [20] reported that up to half of the simulation replications failed to converge with small sample sizes. Determining the tuning parameters is crucial in real data analysis and can be even more difficult with two penalty parameters. In this paper, we do not constrain these penalty parameters, which is different from [20]. This strategy helps obtain optimal tuning parameters and consequently more accurate estimates for the model parameters.

The rest of the article is organized as follows. In section 2, we describe a penalized ZIP model with common penalty functions. In section 3, we propose an EM algorithm for the penalized ZIP model. Section 4 outlines the implementations of the EM algorithm and strategies for tuning parameter selection. In section 5, a simulation study is conducted to evaluate and compare the proposed algorithm and other relevant methods. In section 6, the proposed methods are applied to the TRIBE morbidity data. Section 7 concludes with discussions. The technical details and statistical properties of the proposed methods may be found in the online supplementary material.

2. Penalized ZIP model

Let Y_i be the response observed for the i th subject, $i = 1, 2, \dots, n$. In this research, Y_i is the ICU LOS or number of postoperative complications reported by the i th patient. Assume Y_i is generated from two latent components, the zero state and Poisson state. Let z_i be a latent variable such that $z_i = 1$ if y_i is from the zero state, and $z_i = 0$ if y_i is from the Poisson state. Let $z_i = 1$ with probability π_i , and $z_i = 0$ with probability $1 - \pi_i$. The ZIP model is given by

$$P(Y_i=0)=\pi_i+(1-\pi_i)\exp(-\mu_i), P(Y_i=y_i)=(1-\pi_i)\exp(-\mu_i)\mu_i^{y_i}/y_i!, y_i=1, 2, \dots \quad (1)$$

To correlate LOS with predictor, x_i , $i = 1, 2, \dots, n$, denote $d_1 + 1$ and $d_2 + 1$ -dimensional subvectors B_i and G_i of x_i . The parameters $\mu = (\mu_1, \mu_2, \dots, \mu_n)^\top$ and π_i are modeled with $\log(\mu) = B\beta$ and $\text{logit}(\pi_i) = \log(\pi_i/(1 - \pi_i)) = G_i\zeta$ for covariate matrix B and G , which can be different. To include an intercept, let $B_{i0} = G_{i0} = 1$. The predictor variables are standardized

so that $\sum_{i=1}^n B_{ij}=0$, $\frac{1}{n}\sum_{i=1}^n B_{ij}^2=1$, for $j = 1, \dots, d_1$, $\sum_{i=1}^n G_{ik}=0$, $\frac{1}{n}\sum_{i=1}^n G_{ik}^2=1$, for $k = 1, \dots, d_2$. Define $\Phi = (\beta^\top, \zeta^\top)^\top$ with length $d = d_1 + d_2 + 2$. The log-likelihood function $\ell_{ZIP}(\Phi; y)$ is given by

$$\ell_{ZIP}(\Phi; y) = \sum_{y_i=0} \log(\exp(G_i\zeta) + \exp(-e^{B_i\beta})) + \sum_{y_i>0} (y_i B_i\beta - \exp(B_i\beta)) - \sum_{i=1}^n \log(1 + \exp(G_i\zeta)) - \sum_{y_i>0} \log(y_i!). \quad (2)$$

We define a penalized log-likelihood function for ZIP model:

$$p\ell_{ZIP}(\Phi; y) = \ell_{ZIP}(\Phi; y) - p(\zeta, \beta), \quad (3)$$

with the penalty function

$$p(\zeta, \beta) = n \sum_{j=1}^{d_1} p(\lambda^{(1)}; |\zeta_j|) + n \sum_{k=1}^{d_2} p(\lambda^{(2)}; |\beta_k|), \quad (4)$$

where $p(\cdot)$ is a nonnegative and nondecreasing function, and $\lambda^{(1)}$ and $\lambda^{(2)}$ are the penalty parameters. Maximizing the penalized log-likelihood function (3) results in estimates of parameters which can be shrunk to zero if they are small. The corresponding predictors of the estimates that are zero are less important for the outcome (morbidity), whereas the estimates that are not zero correspond to those more important. Therefore, the procedure can select significant predictors and estimate parameters simultaneously. We employ three penalty functions:

- i. the LASSO penalty [18], for $\lambda \geq 0$

$$p(\lambda; |\theta|) = \lambda |\theta|. \quad (5)$$

- ii. the SCAD penalty [9]

$$p'(\lambda, \theta) = \lambda \left\{ I(\theta < \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta \geq \lambda) \right\} \quad (6)$$

for $\lambda \geq 0$ and $a > 2$, where $p'(\cdot)$ denotes a derivative, t_+ denotes the positive part of t , and often $a = 3.7$ is taken.

- iii. the MCP penalty [19], for $\lambda \geq 0$ and $a > 1$

$$p'(\lambda; \theta) = (\lambda - \theta/a) I(\theta \leq a\lambda). \quad (7)$$

Originally proposed for the least squares regression, the LASSO penalty is a convex function which enjoys a computational advantage. However, the LASSO estimator tends to over-select positive variables [17]. The SCAD is a nonconvex function and holds theoretical properties: unbiasedness, sparsity and continuity [9]. The MCP is also a nonconvex function which introduces a continuous and nearly unbiased method of penalized variable selection in linear regression [19].

3. The EM algorithm

The EM algorithm has been commonly applied for mixture models. It is convenient to program when the computer algorithm for each component of the mixture model is available. In the setting of the regularized ZIP model, we have additional advantages on tuning parameter selection with the EM algorithm (cf. section 4). Maximization of the penalized log-likelihood function of the ZIP models can be achieved by the EM algorithm [24, 25, 26]. In mixture models, the EM algorithm is set up by imposing missing data into the problem. Suppose we could observe which zeros came from the zero state and which came from Poisson state; i.e., suppose we knew $z_i = 1$ when y_i is from zero state, and $z_i = 0$ when y_i is from the Poisson state. Denote $z = (z_1, z_2, \dots, z_n)^\top$. The complete data (y, z) log-likelihood function can be written as

$$\ell_{ZIP}^c(\Phi; y, z) = \sum_{i=1}^n \{z_i G_i \zeta - \log(1 + \exp(G_i \zeta))\} + \sum_{i=1}^n (1 - z_i) \{y_i B_i \beta - \exp(B_i \beta) - \log(y_i!)\}.$$

The penalized complete data log-likelihood function is then given by

$$\begin{aligned} p\ell_{ZIP}^c(\Phi; y, z) &= \sum_{i=1}^n \{z_i G_i \zeta \\ &\quad - \log(1 + \exp(G_i \zeta))\} - n \sum_{j=1}^{d_1} p(\lambda^{(1)}; |\zeta_j|) + \sum_{i=1}^n (1 - z_i) \{y_i B_i \beta - \exp(B_i \beta) \\ &\quad - \log(y_i!)\} \\ &\quad - n \sum_{k=1}^{d_2} p(\lambda^{(2)}; |\beta_k|) \equiv p\ell^c(\zeta; y, z) \\ &\quad + p\ell^c(\beta; y, z), \end{aligned} \quad (8)$$

Thus the log-likelihood is relatively easy to maximize, since $p\ell^c(\zeta; y, z)$ and $p\ell^c(\beta; y, z)$ can be maximized separately with the penalized logistic regression and penalized weighted Poisson regression, respectively. There are fast algorithms for these optimization problems, in particular, the coordinate descent algorithm has been demonstrated to be successful [27, 21, 26]. To make the coordinate descent algorithm stable for MCP and SCAD, we employ the adaptive rescaling proposed by [23]. With the EM algorithm the log-likelihood of model (3) is maximized iteratively by alternating between estimating z_i by its conditional expectation under the current estimates Φ (E step) and maximizing $p\ell_{ZIP}^c(\Phi; y, z)$ (M-step), with z_i fixed from the E step. The iteration stops until the estimated Φ converges. In more detail, starting with an initial value $\Phi^{(0)}$, the EM algorithm iteratively proceeds in the following three steps until convergence.

- **E step.** Estimate z_i by its conditional mean $z_i^{(m)}$, given the data and assuming that the current estimates $\Phi^{(m)}$ provide the true parameters of the model

$$z_i^{(m)} = \begin{cases} [1 + \exp(-G_i \zeta^{(m)} - \exp(B_i \beta^{(m)}))]^{-1}, & \text{if } y_i = 0, \\ 0, & \text{if } y_i = 1, 2, \dots \end{cases} \quad (9)$$

- **M step for ζ .** The $(m + 1)$ th iteration finds $\zeta^{(m+1)}$ by maximizing $p\ell^c(\zeta; y, z)$, which is a penalized logistic log-likelihood function.
- **M step for β .** The updated $\beta^{(m+1)}$ can be obtained by maximizing $p\ell^c(\beta; y, z)$, which is a penalized weighted Poisson regression.

4. Implementation

The EM algorithm requires fitting regularized weighted Poisson regression and logistic regression for the corresponding component in the mixture model. In a broad sense, the

Poisson regression model and logistic regression model belong to the generalized linear model (GLM). Fitting a GLM can be achieved by a standard iteratively reweighted least squares (IRLS) algorithm. However, we need to fit a regularized GLM, which may be achieved by a modified IRLS algorithm (see the supplementary material). Within the modified IRLS algorithm, we utilize the coordinate descent algorithm to optimize the penalized weighted least squares. By repeating the penalized IRLS iterations, the algorithm eventually optimizes the penalized log-likelihood function. Details may be found in the supplementary material. This is different from the traditional EM algorithm for fitting the ZIP model [28] which deals with the weighted least squares (without penalty function on parameters) and the solution is well known. With the above building blocks, we can now proceed to present the pathwise coordinate descent and discuss tuning parameter selection strategies.

We compute the solutions for a pair of decreasing sequence of values for $(\lambda^{(1)}, \lambda^{(2)})$, starting with the smallest value $(\lambda_{\max}^{(1)}, \lambda_{\max}^{(2)})$ for which $\zeta_j = 0, \beta_k = 0, j = 1, \dots, d_1, k = 1, \dots, d_2$. For the candidate values of $\lambda^{(1)}$, we consider a decreasing sequence of $\lambda_m^{(1)}, m = 1, \dots, M$ values from $\lambda_{\max}^{(1)}$ to $\lambda_{\min}^{(1)}$ on the log scale with $\lambda_{\min}^{(1)} = \varepsilon_1 \lambda_{\max}^{(1)}$. The starting value $\lambda_{\max}^{(1)}$ is the smallest value such that $\hat{\zeta}_j = 0, j = 1, \dots, d_1$. Similarly, $\lambda_{\max}^{(2)}$ is chosen such that $\hat{\beta}_k = 0, k = 1, \dots, d_2$. Following [21, 29], $\lambda_{\max}^{(1)}$ and $\lambda_{\max}^{(2)}$ can be determined for the LASSO penalty. In addition, since the adaptive rescaling was utilized in the coordinate descent algorithm, these results can be used for the SCAD and MCP penalty as well. However, for the mixture model such as the ZIP model, some modifications are required. To determine $\lambda_{\max}^{(1)}$ in the EM algorithm, let $\zeta_j^{(0)} = 0, j = 1, \dots, d_1$ and $\beta_k^{(0)} = 0, k = 1, \dots, d_2$. Then compute $z_i^{(0)}$ with formula (9) and compute $\lambda_{\max}^{(1)}$ for a penalized logistic regression. Similarly, $\lambda_{\max}^{(2)}$ can be determined for a penalized weighted Poisson regression, and a decreasing sequence of $\lambda_m^{(2)}, m = 1, \dots, M$ values can be constructed from $\lambda_{\max}^{(2)}$ to $\lambda_{\min}^{(2)}$ on the log scale with $\lambda_{\min}^{(2)} = \varepsilon_2 \lambda_{\max}^{(2)}$. We then form M pair of penalization parameters $(\lambda_m^{(1)}, \lambda_m^{(2)}), m = 1, \dots, M$. Our algorithm starts with the initial solutions which are well defined and have closed form expressions. We implement a warm-start for each subsequent solution from the last solution, which typically generates stable solutions.

In practice, we often have to determine a pair of penalty parameters $(\lambda^{(1)}, \lambda^{(2)})$ leading to an optimal model for the data. For this purpose, it may be useful to select ε_2 different from ε_1 since $\lambda_{\max}^{(2)}$ and $\lambda_{\max}^{(1)}$ may have different magnitudes. Typically we first pick a (relatively small) value for ε_1 , say 0.0001. Then, we select a grid of values for ε_2 , say 0.1, 0.01, 0.001, 0.0001. Therefore, from the $4 \times M$ pairs of candidate values, the chosen $(\lambda_m^{(1)}, \lambda_m^{(2)})$ is the one giving the smallest BIC:

$$\text{BIC} = -2\ell(\hat{\Phi}; \lambda^{(1)}, \lambda^{(2)}) + \log(n)d_e, \quad (10)$$

where $\hat{\Phi}$ denotes the estimator using the parameters $\lambda^{(1)}, \lambda^{(2)}$, and $-\ell$ is the negative log-likelihood. Furthermore, the effective number of parameters $d_e = \sum_{j=0, \dots, d_1} 1\{\zeta_j \neq 0\} + \sum_{k=0, \dots, d_2} 1\{\beta_k \neq 0\}$ for the zero-inflated Poisson model.

5. Simulation Study

5.1. Models and data generation

The simulation study was motivated to evaluate performance of the proposed methods, and compare with the existing methods. Specifically, we adopted the data generation schemes in [20]. The predictor variables were randomly drawn from multivariate normal distributions $N_{20}(0, \Sigma)$, where Σ has elements $\rho^{i-j}(i, j = 1, \dots, 20)$. The correlation among predictor variables was controlled by ρ . The data were generated with sample size $n = 100$ and 300 . We generated 100 replications for each of the set of parameters described below. These parameters were chosen to have a different number of effective predictors and different levels of zero inflations.

Scenario 1:

The parameters were chosen to have about 57% zero inflation:

$$\zeta = (1.10, 0, 0, 0, -0.36, 0, 0, 0, 0, 0, 0, 0, -0.32, 0, 0, 0, 0, 0, 0)$$

$$\beta = (0.30, -0.48, 0, 0, 0, 0.4, 0, 0, 0, 0, 0.44, 0, 0.44, 0, 0, 0, 0, 0, 0, 0)$$

Scenario 2:

We kept all elements of ζ in Scenario 1 and only changed the first element of β from 0.30 to -1.20 , leading to about 26% zero inflation.

Scenario 3:

We kept all elements of ζ in Scenario 1 and only changed the first element of β from 0.30 to -0.35 , leading to about 42% zero inflation.

Scenario 4:

The number of ineffective predictors was doubled from the previous scenarios, and the parameters were set to obtain about 30% zero inflation:

$$\zeta = (1.50, 0, 0, 0, -0.22, 0, 0, 0, -0.25, 0, 0.20, 0, 0.30, -0.32, 0, 0, 0, 0.20, 0, 0)$$

$$\beta = (-1.05, 0, 0.45, 0, -0.3, 0, 0, 0, 0, 0, -0.33, 0, -0.39, 0, 0, 0.3, 0, 0, 0, 0.36, 0)$$

Scenario 5:

Compared to Scenario 4, the first element of ζ was changed from 1.50 to 1.10 and the first element of β was changed from -1.05 to -0.41 while the other elements were the same. The parameters were set to obtain about 40% zero inflation.

5.2. Performance evaluation

Estimation performance from the proposed EM approach under the ZIP regression and penalized Poisson regression (PR) models were compared for different penalty functions. In the subsequent tables, we used EM- and PR- to label the two different models. We also

computed the oracle estimator ZIP-ORACLE assuming that the zero coefficients are known. Results from the one-step estimators in [20] were listed whenever published data are available. The backward elimination (BE) procedure was also studied, with significance level $\alpha = 0.1573, 0.05$ and 0.01 following [30]. BE starts from the full ZIP model including all predictor variables. In each step, a variable is eliminated with the largest p-value (if exceeding a prespecified significance level) from the current fitted model, and a model is refit with the reduced variables. This process is repeated until all variables that remain in the model are significant. The performances of comparison were based on the following quantities. Denote the mean squared error for the count component

$$\text{MSE} = \frac{1}{d_1 + 1} \sum_{j=0}^{d_1} (\hat{\beta}_{ij} - \beta_j)^2$$
, where β_{ij} is the j -th element (including intercept) of an estimator vector in the i -th simulation. We then computed the ratio of MSE obtained from a penalized method against the corresponding non-penalized ZIP estimator. Similarly, we defined a ratio of MSE for the zero component estimator of ζ . A smaller MSE indicates a better estimator. To evaluate variable selection performances, we also compared the true positive rate (sensitivity) and the true negative rate (specificity), where

$$\text{sensitivity} = \frac{\text{number of correctly selected predictors}}{\text{number of effective predictors}}, \quad (11)$$

and

$$\text{specificity} = \frac{\text{number of correctly unselected predictors}}{\text{number of ineffective predictors}}. \quad (12)$$

A good estimator should have both sensitivity and specificity close to 1. Finally, we generated 1,000 test observations from each of the models, and used the test data to compute the predictive log-likelihood value of each model estimated.

5.3. Simulation results

We compare the new proposed EM-MCP with other regularized ZIP models including the EM-LASSO and EM-SCAD. We also compare with the regularized Poisson models. Finally, we compare the proposed EM algorithm with the one-step estimators in [20]. The simulation results for scenario 1 with $\rho = 0$ and $\rho = 0.4$ are summarized in Table 1 and 2, respectively. Also see Figures 2 and 3 for the predictive log-likelihood values, divided by the absolute log-likelihood values of the full ZIP model. For brevity, we only generated boxplots for the ZIP models since the Poisson models have poor performance. Other results can be found in the supplementary material. Some observations follow.

1. For the Poisson component, the Poisson regression methods (PR-LASSO, PR-MCP and PR-SCAD) had poor performance with large values of MSE and low values of specificity. When the zero inflation increased, the Poisson regression methods had even poorer performances with decreased specificity. Apparently the ZIP methods outperformed the Poisson regression. This was expected since the latter methods failed to model the inflated zeros and tended to select more noisy variables.

2. Among the ZIP regression methods, the MCP and SCAD performed much better than the LASSO on MSE and specificity, and had equal performance on sensitivity. There is no clear winner between the EM-MCP and EMSCAD methods. These conclusions held for both the Poisson component and the zero component. Results from the BE procedure varies depending on the significance level. With a smaller significance level, for instance $\alpha = 0.01$, the BE procedure generated more accurate estimates with smaller MSEs, but may select fewer variables resulting in less accurate sensitivity. BE(0.01) was competitive to EM-MCP or EM-SCAD in terms of MSE. However, BE(0.01) had the smallest sensitivity for the zero component across all simulation scenarios, particularly noticeable with small sample sizes. The MCP and SCAD estimators had slightly better or comparable predictive log-likelihood values against BE(0.05) or BE(0.01), and only slightly inferior to the oracle models. In this respect, for small sample size $n = 100$, BE(0.1573) had the poorest values. One important drawback of the BE procedure is that a standard computer routine may fail to invert the Hessian matrix when fitting an unpenalized ZIP model, which is required when computing the p-values. The failure rate is particularly high with a small sample size or a large number of predictors. In our simulation study, for instance, the BE procedure for significance level $\alpha = 0.1573$, 0.05 and 0.01 encountered 28 to 30 failures out of 100 replications in scenario 1 with $n = 100$ and $\rho = 0$.
3. As sample size increased, all the methods had better performances with reduced MSE values (data not shown). Note the MSE in the tables was defined to be a ratio, not the absolute value. The sensitivity increased for the zero component as the sample size increased. Our limited results in Table 1 and 2 didn't indicate noticeable changes in sensitivity when the correlation increased.
4. Compared to the one-step estimators in [20], our method generally had better performances although some differences may be due to the sampling error. For the Poisson component, the EM-SCAD had comparable sensitivity and specificity to one-step-SCAD. Similar conclusions hold for the LASSO estimator. The dramatic difference occurred for the zero component: (a) The EM-SCAD had smaller MSEs than the one-step-SCAD, whose MSEs can be as large as double (Table 1 and 2, $n = 300$). The EM-SCAD had much improved sensitivity than the one-step-SCAD. (b) The EM-LASSO outperformed the one-step-LASSO as well. The one-step-LASSO generated very low sensitivity for the zero component [20]. For instance, in scenario 1 and $n = 300$, it was reported that the sensitivity was 0.000 for the one-step-LASSO for both $\rho = 0$ and 0.4 while the EM-LASSO had mean sensitivity 0.742 and 0.8, respectively. Indeed, except for one case with average sensitivity 0.001, the one-step-LASSO had all sensitivity 0 in scenarios 1–5, including sample size $n = 600$ [20].

6. Analysis of the TRIBE data

The TRIBE study is a prospective study to evaluate the biomarkers of AKI and clinical outcomes. In the children cohort, half of the patients were below two years old and 55% were male. Most surgeries were elective (91%) and utilized cardiopulmonary bypass (99%).

Demographic and clinical characteristics were representative of the global CS populations of each site. The average ICU and hospital LOS were 4.1 (SD, 6.3) and 8.5 (SD, 10.8) days, respectively. Sixteen percent developed one complication, 8% had two complications and 4.7% had more than two complications. The first postoperative biomarkers were collected at a median of 0.5 (IQR, 0.25 to 2) hours after arrival in the ICU. We have characterized their ICU time-dependent concentrations extensively [31, 32]. Biomarkers including NGAL (neutrophil gelatinase-associated lipocalin), IL-18 and Cystatin C were all linearly associated with longer hospital stay, ICU stay, as well as duration of mechanical ventilation after adjustment for other factors [31, 32]. The ICU and hospital stays were longer in those with AKI, suggesting that CS-associated AKI is associated with increased hospital resource use and increased overall health cost [33].

We applied the standard ZIP method and developed methods (i.e., EM-LASSO, EM-MCP and EM-SCAD) to outcomes including ICU LOS and number of postoperative complications to a sample size $n = 281$ which had complete data on the preselected risk factors, as in Table 3 to 5. Of note, eGFR (estimated glomerular filtration rate or renal function) and serum creatinine were measured pre-operatively and first postoperative biomarkers serum creatinine, serum Cystatin C, urine IL18, NGAL, creatinine, KIM-1 (kidney injury molecule 1) and LFABP (liver-type fatty acid binding protein) were included as well. There was a preliminary procedure for the risk factors as we applied the logarithmic transformation to the right skewed variables including all the biomarkers. Estimating methods including the penalized Poisson regression (PR-LASSO, PR-MCP, PR-SCAD) and non-penalized Poisson regression were also utilized to fit the data. Since the penalized Poisson regression methods didn't fit the data adequately with smaller likelihood values, we only compared with the standard non-penalized Poisson regression in the sequel. The BE procedure was also conducted for significance level $\alpha = 0.1573, 0.05$ and 0.01 although the results were only shown for a level with the smallest BIC value. For the penalized EM estimators, the standard errors were estimated using the nonparametric 0.632 bootstrap [34]. We randomly sampled 177 subjects from the total sample size $n = 281$ without replacement and computed the bootstrap estimators with the same tuning parameters. The bootstrap procedure was repeated 500 times and the standard deviation of bootstrap estimators was calculated.

6.1. ICU length of stay

First, we compared a standard non-regularized Poisson model with the ZIP regression models. The standard Poisson regression with all aforementioned predictors had the log-likelihood value -735.1 and BIC 1583.0 , with degree of freedom (df) 20. The standard non-regularized ZIP model with an additional intercept in the zero-inflation model had the log-likelihood value -729.4 and BIC 1577.1 with $df=21$. The Vuong test [35] comparing the ZIP model with the standard Poisson regression model had p -value 0.21 indicating the simple ZIP model is only slight better than the standard Poisson model. The full ZIP model with all predictors in both the Poisson model and the zero-inflation model ($df=40$) had log-likelihood value -683.6 and BIC 1592.8 . The Vuong test had p -value $=0.0005$ showing that the full ZIP model significantly improved over the standard Poisson model, despite some evidence of overfitting with a larger BIC value. Therefore, we utilized the penalized ZIP models to

conduct variable selection and avoid overfitting. The penalized ZIP methods in Table 3 generated larger log-likelihood values and smaller BIC values, indicating better model fitting than the standard Poisson regression model. Among the penalized ZIP methods, the EM-LASSO selected many more predictors and appeared to overfit data. This is consistent with findings from the simulation study as the EM-LASSO had lower specificity than the EM-MCP and EM-SCAD. The likelihood ratio test was utilized to compare the penalized ZIP model with the full ZIP model. The EM-LASSO is significantly different from the full ZIP model ($\chi^2=46.76$, $df=14$, $p\text{-value} < 0.0001$). The EM-MCP is not statistically different from the full ZIP ($\chi^2=21.42$, $df=17$, $p\text{-value} = 0.21$), so is the EM-SCAD ($\chi^2 = 19.65$, $df=15$, $p\text{-value}=0.19$). BE(0.05) generated a model similar to the EM-MCP on BIC value with one more variable selected. Apparently BE(0.05) had advantage on AIC than other penalized methods. As in the simulation study, the results from the BE(α) procedure depend on the significance level α . Table 3 only displayed results for $\alpha = 0.05$ and other results are briefly mentioned here. With a small significance level $\alpha = 0.01$, the BE procedure generated a fair model with BIC 1519.2, although the data were not adequately fitted. The BE(0.01) generated a model with the smallest log-likelihood value -714.5 and large AIC value 1461, most likely because the model is too sparse. The BE(0.1573) resulted in a complex model with more risk factors. Although the data were adequately fitted with log-likelihood value -686.1 and AIC value 1428.3, there is some evidence of overfitting with large BIC value 1530.1.

To further compare BE(0.05) and EM-MCP, we conducted BE(0.05) for 500 replications of 0.632 bootstrap. In this process, BE(0.05) failed in 360 replications when inverting a Hessian matrix. Thus, we restrict our comparison within the remaining 140 replications. The median values of log-likelihood value, BIC and AIC for BE(0.05) are -419.2 , 940.5 and 876.1, respectively, while the EM-MCP had the corresponding median values -406.3 , 933.3 and 859.7, respectively. Note that because of random sampling, the results are representative. Indeed, with all 500 replications, the EM-MCP had the median values -406.6 , 933.4, 860.0, respectively. For the corresponding 3 paired values between BE(0.05) and EM-MCP, all Wilcoxon signed rank tests had $p\text{-values} < 0.0001$ indicating significantly better model fitting of the EM-MCP method. We also compared the frequencies of variable selection as reported in Table 4. Among those variables selected by BE(0.05) in Table 3, only 12 out of 22 (54.5%) variables had selection frequencies above 50% for the 0.632 bootstrap. In contrast, for the EM-MCP, there were 18 out of 21 (85.7%) variables that were selected over 50% times. Among those selected in Table 3, the medians were 0.55 (range 0.21–0.99) for BE(0.05) and 0.73 (range 0.36 – 0.99) for EM-MCP. Our analysis suggested that the new proposed methods are quite different from the classical BE procedure, and can be a worthwhile addition to the ZIP models.

For the TRIBE data, we briefly interpret the EM-MCP analysis. The selected predictors were not surprising. For instance, the estimated negative coefficient of weight (-0.94) in the Poisson component indicates that a lighter child could have prolonged ICU LOS. The LOS was associated with age, RACHS (Risk Adjustment for Congenital Heart Surgery), CPB (cardiopulmonary bypass) time, use of aspirin, and study sites. Increased early biomarkers urine NGAL and urine KIM-1 were associated with prolonged LOS while decreased early

urine creatinine was associated with prolonged LOS. The analysis results were consistent with findings from previous analysis [31].

6.2. Number of complications

Like the analysis of LOS, we compared a standard Poisson with ZIP regression models. The standard Poisson regression with all predictors had the log-likelihood value -200.1 and BIC 513.0 with $df=20$. The standard ZIP model with an additional intercept in the zero-inflation model had the log-likelihood value -199.7 and BIC 517.8 with $df=21$. The Vuong test comparing the ZIP model with the standard Poisson regression model had p -value 0.41 indicating the ZIP model is only slight better than the standard Poisson model. However, the standard ZIP model with predictor study site in the zero-inflation model had log-likelihood value -189.4 and BIC 508.4 with $df=23$, and the Vuong test had p -value $=0.03$ showing a significant improvement over the standard Poisson model. We consider this model as the benchmark ZIP model. The computer software failed to fit a full ZIP model with all predictors in both the Poisson component and zero component. This was not unusual when fitting complex statistical models. In other words, it also justifies the need for variable selection when developing ZIP models. The penalized ZIP methods in Table 5 had larger log-likelihood values and smaller BIC values than the standard Poisson model, indicating better model fitting with the ZIP models. Again, the EM-LASSO selected many more predictors than other two penalized methods. The EM-MCP and EM-SCAD had the very similar estimation results. The likelihood ratio test was utilized to compare the penalized ZIP model with the standard ZIP model. The EM-LASSO is not significantly different from the bench ZIP model ($\chi^2=13.56$, $df=13$, p -value $=0.41$), as well as the EM-MCP and EM-SCAD ($\chi^2=6.94$, $df=16$, p -value $=0.97$). Unlike the models for the ICU LOS, the fitted models were even more parsimonious. It showed that the post-operative number of complications were mostly associated with study site, RACHS and early urine IL18.

7. Discussion

There are numerous studies on the biomarkers of AKI. Published studies were typically focused on individual biomarkers. Combining biomarkers in one statistical model requires sound statistical methodologies. This can be very challenging in studies with small sample sizes. Postoperative morbidity has increased impact to the patients, society and beyond. This paper advocates to utilizing modern variable selection methodologies to predict postoperative morbidity. Our methodology may have applications for other outcomes such as duration of AKI.

In the literature, influential variable selection methods including the LASSO and SCAD have been extended to the zero inflated models. The major contributions of this paper are three folds. First, we extended the MCP method in this area. Second, we proposed alternative computing algorithms. Third, we developed new tuning parameter selection strategies. In real data analysis, the three elements are equally important to fit a sound statistical model. The numerical studies suggested that the proposed methods are competitive to the existing methods. There was substantial improvement of sensitivity when modeling the extra zeros. In all the methods studied, sample size and the number of

predictors are important factors for accurate variable selection. In the simulation study, when sample size increases, more effective variables are selected in the Poisson and zero components. The classical BE procedure has been a popular technique in applications. Our simulations indicate that the prediction accuracy varies among different significance levels for $BE(\alpha)$. In particular, $BE(0.1573)$ is less accurate on prediction when sample size is small. With appropriate choice of α level, the predicted log-likelihood values from BE are similar to those from the penalized methods.

In the TRIBE data analysis, we used nonparametric bootstrap resampling technique to evaluate the standard errors for the estimated parameters in the penalized ZIP models. A numerical study would shed some light on how reliable the bootstrap is for this problem, as a reviewer suggested. Given that the EM algorithm itself is time-consuming, bootstrap SEs in a simulation study is a highly computational demanding task. This issue remains open and deserves future research attention. As pointed out by a reviewer, a sandwich formula can be constructed as an estimator for the covariance of the selected variables, as outlined in [9]. In this paper, the EM-SCAD or EM-MCP estimator utilizes the adaptive rescaling of the penalty parameter a in (6) or (7) to match the continually changing scale of the covariates [23, section 3.2]. How to implement a sandwich formula for the proposed algorithm is beyond the scope of the current paper. The sample size in the bootstrap subsamples is 177, which can be challenging to the traditional ZIP model with 19 variables on the Poisson and the zero component. Due to relatively small sample size and large number of variables, computing the inverse of the Hessian matrix for the ZIP model has proved to be difficult in many bootstrap resampling data. However, the EM algorithm for the penalized ZIP model is robust for this type of data.

The proposed methods can be extended further. With highly correlated predictors, an additional L_2 penalty may be introduced:

$$p(\zeta, \beta) = n \sum_{j=1}^{d_1} p(\lambda_1^{(1)}, \lambda_2^{(1)}; |\zeta_j|) + n \sum_{j=1}^{d_2} p(\lambda_1^{(2)}, \lambda_2^{(2)}; |\beta_j|).$$

For example, the ENET penalty [36]

$$p(\lambda_1, \lambda_2; |\theta|) = \lambda_1 |\theta| + \frac{1}{2} \lambda_2 \theta^2.$$

All of these proposals are publicly accessible in an open-source R package `mpath` (<http://cran.r-project.org>).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the associated editor and two referees who gave constructive comments on an earlier draft of this article. ZW is partially supported by a grant from the Charles H. Hood Foundation, Inc., Boston, MA. CYW is partially supported by National Institutes of Health grants CA53996, R01ES017030 and HL121347. Heather Thiessen-Philbrook provided assistance for the TRIBE data, which is greatly appreciated.

References

1. Kramer AA, Zimmerman JE. A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of stay. *BMC Medical Informatics and Decision Making*. 2010; 10(1)
2. Ettema RGA, Peelen LM, Schuurmans MJ, Nierich AP, Kalkman CJ, Moons KGM. Prediction models for prolonged intensive care unit stay after cardiac surgery: Systematic review and validation study. *Circulation*. 2010; 122(7):682–689. [PubMed: 20679549]
3. Stark JF, Gallivan S, Davis K, Hamilton JRL, Monro JL, Pollock JCS, Watterson KG. Assessment of mortality rates for congenital heart defects and surgeons' performance. *Annals of Thoracic Surgery*. 2001; 72(1):169–175. [PubMed: 11465173]
4. Gallivan S, Davis KB, Stark JF. Early identification of divergent performance in congenital cardiac surgery. *European Journal of Cardio-Thoracic Surgery*. 2001; 20(6):1214–1219. [PubMed: 11717031]
5. Brown KL, Ridout DA, Goldman AP, Hoskote A, Penny DJ. Risk factors for long intensive care unit stay after cardiopulmonary bypass in children. *Critical Care Medicine*. 2003; 31(1):28–33. [PubMed: 12544989]
6. Gillespie M, Kuijpers M, Van Rossem M, Ravishankar C, Gaynor JW, Spray T, Clark B III. Determinants of intensive care unit length of stay for infants undergoing cardiac surgery. *Congenital Heart Disease*. 2006; 1(4):152–160. [PubMed: 18377540]
7. Wheeler DS, Dent CL, Manning PB, Nelson DP. Factors prolonging length of stay in the cardiac intensive care unit following the arterial switch operation. *Cardiology in the Young*. 2008; 18(1): 41–50. [PubMed: 18093360]
8. Breiman L. Heuristics of instability and stabilization in model selection. *Annals of Statistics*. 1996; 24:2350–2383.
9. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96(456):1348–1360.
10. Akaike H. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*. 1973:267–281.
11. Schwarz G. Estimating the dimension of a model. *The Annals of Statistics*. 1978; 6(2):461–464.
12. Lee AH, Gracey M, Wang K, Yau KKW. A robustified modeling approach to analyze pediatric length of stay. *Annals of Epidemiology*. 2005; 15(9):673–677. [PubMed: 16157254]
13. Atienza N, García-Heras J, Muñoz Pichardo JM, Villa R. An application of mixture distributions in modelization of length of hospital stay. *Statistics in Medicine*. 2008; 27(9):1403–1420. [PubMed: 17680551]
14. Sharma SS, Mamtani MR, Sharma MS, Kulkarni H. A prospective cohort study of postoperative complications in the management of perforated peptic ulcer. *BMC Surgery*. 2006; 6(8)
15. Hur K, Hedeker D, Henderson W, Khuri S, Daley J. Modeling clustered count data with excess zeros in health care outcomes research. *Health Services and Outcomes Research Methodology*. 2002; 3(1):5–20.
16. Dalrymple ML, Hudson IL, Ford RPK. Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. *Computational Statistics & Data Analysis*. 2003; 41(3–4):491–504.
17. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*. 2010; 20(1):101–148. [PubMed: 21572976]
18. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.

19. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*. 2010; 38(2):894–942.
20. Buu A, Johnson N, Li R, Tan X. New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine*. 2011; 30:2326–2340. [PubMed: 21563207]
21. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 2010; 33:1–22. [PubMed: 20808728]
22. Wu TT, Lange K. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*. 2008; 2(1):224–244.
23. Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*. 2011; 5(1):232–253. [PubMed: 22081779]
24. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*. 1977; 39(1):1–38.
25. Khalili A, Chen J. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*. 2007; 102(479):1025–1038.
26. Städler N, Bühlmann P, Van De Geer S. ℓ_1 -penalization for mixture regression models. *Test*. 2010; 19(2):209–256.
27. Friedman J, Hastie T, Höfling H, Tibshirani R. Pathwise coordinate optimization. *The Annals of Applied Statistics*. 2007; 1(2):302–332.
28. Lambert D. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*. 1992; 34(1):1–14.
29. Fan J, Lv J. Nonconcave penalized likelihood with NP-dimensionality. *Information Theory, IEEE Transactions on*. 2011; 57(8):5467–5484.
30. Sauerbrei W, Holländer N, Buchholz A. Investigation about a screening step in model selection. *Statistics and Computing*. 2008; 18(2):195–208.
31. Parikh C, Devarajan P, Zappitelli M, Sint K, Thiessen-Philbrook H, Li S, Kim R, Koynier J, Coca S, Edelstein C, et al. Postoperative biomarkers predict acute kidney injury and poor outcomes after pediatric cardiac surgery. *Journal of the American Society of Nephrology*. 2011; 22(9):1737–1747. [PubMed: 21836147]
32. Zappitelli M, Krawczeski C, Devarajan P, Wang Z, Sint K, Thiessen-Philbrook H, Li S, Bennett M, Ma Q, Shlipak M, et al. Early postoperative serum cystatin C predicts severe acute kidney injury following pediatric cardiac surgery. *Kidney International*. 2011; 80(6):655–662. URL <http://www.ncbi.nlm.nih.gov/pubmed/21525851>. [PubMed: 21525851]
33. Li S, Krawczeski C, Zappitelli M, Devarajan P, Thiessen-Philbrook H, Coca S, Kim R, Parikh C. Incidence, risk factors, and outcomes of acute kidney injury after pediatric cardiac surgery: A prospective multicenter study. *Critical Care Medicine*. 2011; 39(6):1493–1499. URL <http://www.ncbi.nlm.nih.gov/pubmed/21336114>. [PubMed: 21336114]
34. Efron, B.; Tibshirani, R. *An Introduction to the Bootstrap*. New York: CRC press; 1993.
35. Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*. 1989; 57(2):307–333.
36. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society, Series B*. 2005; 67(2):301–320.

Figure 1.
Empirical distributions of ICU LOS (left) and number of complication (right) for the TRIBE study.

Figure 2.
Predictive log-likelihood with scenario 1, $\rho = 0$

Figure 3.
Predictive log-likelihood with scenario 1, $\rho = 0.4$

Table 1

Simulation results with scenario 1, $\rho=0$, $n=100$ and 300. Median and robust standard deviations (in parentheses) of the ratio between the penalized model MSE and the full model MSE. Mean and standard deviations of sensitivity and specificity defined in the main text. The results for one-step estimators were extracted from [20]

Method	MSE	Sensitivity	Specificity
n=100			
Poisson component			
PR-LASSO	1.204 (0.756)	0.87 (0.272)	0.531 (0.229)
PR-MCP	1.692 (0.928)	0.825 (0.296)	0.653 (0.133)
PR-SCAD	1.684 (0.965)	0.86 (0.267)	0.571 (0.181)
EM-LASSO	0.118 (0.105)	0.72 (0.385)	0.956 (0.063)
EM-MCP	0.075 (0.086)	0.87 (0.232)	0.935 (0.093)
EM-SCAD	0.08 (0.087)	0.855 (0.278)	0.922 (0.127)
BE(0.1573)	0.59 (0.319)	0.903 (0.199)	0.701 (0.162)
BE(0.05)	0.164 (0.203)	0.864 (0.254)	0.895 (0.116)
BE(0.01)	0.069 (0.081)	0.807 (0.286)	0.985 (0.035)
ZIP-ORACLE	0.019 (0.017)	1 (0)	1 (0)
Zero component			
EM-LASSO	0.016 (0.02)	0.302 (0.247)	0.968 (0.048)
EM-MCP	0.02 (0.027)	0.42 (0.246)	0.942 (0.08)
EM-SCAD	0.019 (0.024)	0.38 (0.255)	0.946 (0.081)
BE(0.1573)	0.387 (0.378)	0.715 (0.238)	0.656 (0.206)
BE(0.05)	0.078 (0.101)	0.457 (0.233)	0.913 (0.129)
BE(0.01)	0.041 (0.049)	0.132 (0.179)	0.991 (0.022)
ZIP-ORACLE	0.007 (0.009)	1 (0)	1 (0)
n=300			
Poisson component			
PR-LASSO	12.487 (4.586)	0.975 (0.131)	0.461 (0.183)
PR-MCP	14.939 (5.004)	0.985 (0.086)	0.627 (0.14)
PR-SCAD	14.922 (5.11)	0.985 (0.086)	0.557 (0.166)
EM-LASSO	0.42 (0.32)	0.995 (0.05)	0.954 (0.061)
EM-MCP	0.117 (0.103)	0.995 (0.05)	0.982 (0.033)
EM-SCAD	0.094 (0.084)	1 (0)	0.966 (0.066)
BE(0.1573)	0.536 (0.171)	1 (0)	0.82 (0.094)
BE(0.05)	0.312 (0.189)	1 (0)	0.938 (0.059)
BE(0.01)	0.11 (0.102)	0.995 (0.05)	0.988 (0.024)
ZIP-ORACLE	0.077 (0.064)	1 (0)	1 (0)
one-step-LASSO	0.375	0.996	0.955
one-step-SCAD	0.094	0.999	0.971

	Method	MSE	Sensitivity	Specificity
Zero component				
	EM-LASSO	0.715 (0.444)	0.742 (0.265)	0.966 (0.066)
	EM-MCP	0.477 (0.327)	0.768 (0.262)	0.981 (0.043)
	EM-SCAD	0.599 (0.437)	0.762 (0.271)	0.962 (0.059)
	BE(0.1573)	0.602 (0.182)	0.958 (0.101)	0.826 (0.095)
	BE(0.05)	0.423 (0.321)	0.878 (0.153)	0.949 (0.056)
	BE(0.01)	0.537 (0.388)	0.685 (0.245)	0.991 (0.024)
	ZIP-ORACLE	0.15 (0.09)	1 (0)	1 (0)
	one-step-LASSO	1.470	0.000	1
	one-step-SCAD	1.192	0.185	1

Table 2

Simulation results with scenario 1, $\rho=0.4$, $n=100$ and 300. Median and robust standard deviations (in parentheses) of the ratio between the penalized model MSE and the full model MSE. Mean and standard deviations of sensitivity and specificity defined in the main text. The results for one-step estimators were extracted from [20]

Method	MSE	Sensitivity	Specificity
n=100			
Poisson component			
PR-LASSO	1.048 (0.623)	0.875 (0.26)	0.522 (0.231)
PR-MCP	1.502 (0.905)	0.775 (0.296)	0.66 (0.138)
PR-SCAD	1.576 (1.012)	0.805 (0.292)	0.604 (0.174)
EM-LASSO	0.1 (0.089)	0.785 (0.364)	0.941 (0.075)
EM-MCP	0.068 (0.088)	0.875 (0.25)	0.942 (0.08)
EM-SCAD	0.067 (0.081)	0.83 (0.32)	0.931 (0.116)
BE(0.1573)	0.699 (0.374)	0.833 (0.285)	0.666 (0.192)
BE(0.05)	0.162 (0.214)	0.781 (0.327)	0.881 (0.129)
BE(0.01)	0.097 (0.128)	0.731 (0.346)	0.965 (0.068)
ZIP-ORACLE	0.018 (0.017)	1 (0)	1 (0)
Zero component			
EM-LASSO	0.022 (0.027)	0.33 (0.241)	0.969 (0.05)
EM-MCP	0.023 (0.031)	0.392 (0.269)	0.951 (0.068)
EM-SCAD	0.023 (0.029)	0.37 (0.269)	0.952 (0.068)
BE(0.1573)	0.388 (0.336)	0.66 (0.251)	0.643 (0.202)
BE(0.05)	0.086 (0.1)	0.4 (0.266)	0.909 (0.101)
BE(0.01)	0.057 (0.061)	0.175 (0.196)	0.982 (0.032)
ZIP-ORACLE	0.008 (0.01)	1 (0)	1 (0)
n=300			
Poisson component			
PR-LASSO	10.567 (4.667)	0.99 (0.07)	0.493 (0.164)
PR-MCP	13.353 (6.126)	0.965 (0.128)	0.666 (0.104)
PR-SCAD	13.23 (5.919)	0.985 (0.086)	0.6 (0.133)
EM-LASSO	0.355 (0.279)	1 (0)	0.934 (0.069)
EM-MCP	0.132 (0.126)	1 (0)	0.971 (0.048)
EM-SCAD	0.113 (0.113)	1 (0)	0.949 (0.08)
BE(0.1573)	0.536 (0.233)	1 (0)	0.824 (0.107)
BE(0.05)	0.249 (0.225)	1 (0)	0.942 (0.076)
BE(0.01)	0.133 (0.142)	1 (0)	0.986 (0.028)
ZIP-ORACLE	0.079 (0.078)	1 (0)	1 (0)
one-step-LASSO	0.281	0.998	0.941
one-step-SCAD	0.086	1	0.961

	Method	MSE	Sensitivity	Specificity
Zero component				
	EM-LASSO	0.466 (0.27)	0.8 (0.254)	0.958 (0.056)
	EM-MCP	0.366 (0.22)	0.782 (0.243)	0.971 (0.045)
	EM-SCAD	0.419 (0.289)	0.798 (0.265)	0.962 (0.052)
	BE(0.1573)	0.576 (0.193)	0.925 (0.126)	0.825 (0.101)
	BE(0.05)	0.439 (0.331)	0.865 (0.161)	0.931 (0.08)
	BE(0.01)	0.403 (0.26)	0.705 (0.214)	0.985 (0.03)
	ZIP-ORACLE	0.128 (0.084)	1 (0)	1 (0)
	one-step-LASSO	1.103	0.000	1
	one-step-SCAD	0.902	0.224	0.999

Table 3

Estimation results for fitting the ICU LOS in the TRIBE data. The estimated coefficients with standard errors, log-likelihood value (logLik), BIC and AIC value.

	ZIP	BE(0.05)	EM-LASSO	EM-MCP	EM-SCAD
Poisson component					
(Intercept)	5.15(0.59)	5.48(0.50)	3.32(1.19)	4.89(1.77)	4.81(1.70)
age	0.36(0.08)	0.37(0.07)	0.22(0.18)	0.42(0.28)	0.41(0.27)
genderM	-0.17(0.08)	-0.15(0.08)	-0.17(0.19)	-0.11(0.22)	-0.15(0.21)
weight	-1.16(0.20)	-1.22(0.19)	-0.47(0.32)	-0.94(0.60)	-0.92(0.59)
site9	0.77(0.10)	0.75(0.10)	0.69(0.28)	0.77(0.31)	0.77(0.31)
site11	1.24(0.10)	1.18(0.10)	1.09(0.42)	1.17(0.45)	1.18(0.45)
RACHS	0.08(0.07)		0.13(0.20)		0.02(0.22)
CPB time	0.42(0.10)	0.41(0.10)	0.31(0.27)	0.40(0.31)	0.39(0.31)
cross-clamp time	0.15(0.10)	0.21(0.09)	0.11(0.18)	0.18(0.22)	0.17(0.23)
ACE inhibitors Yes	0.28(0.10)	0.29(0.10)	0.17(0.21)	0.29(0.23)	0.29(0.23)
aspirin Yes	-0.41(0.13)	-0.37(0.13)	-0.31(0.21)	-0.38(0.25)	-0.37(0.26)
eGFR	0.01(0.005)	0.01(0.005)			
baseline serum creatinine	1.34(0.42)	1.24(0.39)	0.22(0.44)	0.39(0.86)	0.39(0.94)
serum creatinine	-0.07(0.17)				
serum Cystatin C	0.30(0.16)		0.13(0.35)		
urine IL18	-0.02(0.03)				
urine NGAL	0.11(0.03)	0.10(0.03)	0.09(0.09)	0.10(0.10)	0.10(0.10)
urine creatinine	-0.59(0.07)	-0.59(0.07)	-0.55(0.24)	-0.55(0.26)	-0.55(0.26)
urine KIM-1	0.37(0.04)	0.37(0.04)	0.33(0.12)	0.38(0.13)	0.38(0.13)
urine LFABP	-0.06(0.03)	-0.06(0.03)	-0.02(0.08)	-0.06(0.10)	-0.06(0.10)
Zero component					
(Intercept)	1.78(4.55)	7.32(2.96)	-0.20(2.02)	-2.42(3.07)	-2.97(4.03)
age	1.98(0.83)	2.00(0.71)	0.37(0.17)	1.14(0.61)	1.13(0.72)
genderM	-0.59(0.65)				-0.003(0.56)
weight	-1.48(1.61)	-2.69(1.17)			
site9	-0.43(0.65)				

	ZIP	BE(0.05)	EM-LASSO	EM-MCP	EM-SCAD
site11	-0.11(0.89)				
RACHS	-1.02(0.50)	-1.03(0.45)	-0.28(0.22)	-1.02(0.64)	-0.97(1.03)
CPB time	6e-04(1.00)		-0.34(0.36)		
cross-clamp time	-2.50(0.99)	-1.98(0.82)	-0.75(0.43)	-2.42(3.30)	-2.48(4.71)
ACE inhibitors Yes	-2.11(1.57)	-2.55(1.19)	-1.00(0.43)	-1.52(0.86)	-2.22(1.19)
aspirin Yes	-17(3829)				
eGFR	-0.02(0.03)				
baseline serum creatinine	-3.07(3.38)				
serum creatinine	-0.61(1.64)			-3.02(1.34)	-3.11(1.72)
serum Cystatin C	-1.67(1.30)	-1.81(0.89)	-0.73(0.47)		
urine IL18	0.02(0.27)				
urine NGAL	-0.58(0.31)	-0.56(0.19)	-0.13(0.09)	-0.46(0.27)	-0.47(0.35)
urine creatinine	0.76(0.62)			0.62(0.44)	0.77(0.57)
urine KIM-1	0.17(0.39)				
urine LFABP	-0.10(0.25)		-0.03(0.07)		
Summary					
logLik	-683.6	-691.4	-707	-694.3	-693.4
BIC	1592.8	1518.1	1560.6	1518.3	1527.8
AIC	1447.2	1430.7	1466	1434.7	1436.9

Table 4

Frequencies of variable selection with 140 replication of 0.632-bootstrap for fitting the ICU LOS in the TRIBE data.

	BE(0.05)	EM-MCP
Poisson component		
age	0.79 *	0.73 *
genderM	0.49 *	0.73 *
weight	0.92 *	0.71 *
site9	0.93 *	0.96 *
site11	0.99 *	0.99 *
RACHS	0.49	0.77
CPBtime	0.79 *	0.85 *
crossclamptime	0.46 *	0.69 *
ACE inhibitorsYes	0.41 *	0.7 *
aspirinYes	0.48 *	0.77 *
eGFR	0.55 *	0.49
baseline serum creatinine	0.7 *	0.45 *
serum creatinine	0.29	0.49
serum Cystatin C	0.42	0.56
urine IL18	0.37	0.57
urine NGAL	0.67 *	0.82 *
urine creatinine	0.93 *	0.96 *
urine KIM-1	0.98 *	0.99 *
urine LFABP	0.41 *	0.58 *
Zero component		
age	0.89 *	0.73 *
genderM	0.06	0.41
weight	0.21 *	0.17
site9	0.03	0.2
site11	0.02	0.23
RACHS	0.42 *	0.74 *
CPBtime	0.19	0.44
crossclamptime	0.49 *	0.87 *
ACE inhibitorsYes	0.36 *	0.91 *
aspirinYes	0	0.41
eGFR	0.08	0.24
baseline serum creatinine	0.21	0.14
serum creatinine	0.18	0.36 *

	BE(0.05)	EM-MCP
serum Cystatin C	0.26 *	0.53
urine IL18	0.06	0.06
urine NGAL	0.55 *	0.51 *
urine creatinine	0.19	0.39 *
urine KIM-1	0.09	0.36
urine LFABP	0.17	0.31

* indicates variable selected in Table 3.

Table 5

Estimation results for fitting the number of complications in the TRIBE data. The estimated coefficients and standard errors, log-likelihood value (logLik), BIC and AIC value.

	ZIP	BE(0.1573)	EM-LASSO	EM-MCP	EM-SCAD
Poisson component					
(Intercept)	-0.37(1.80)	-3.64(0.57)	-2.21(0.65)	-3.19(1.24)	-3.20(1.20)
age	0.06(0.21)				
genderM	0.05(0.24)				
weight	-0.74(0.53)				
site9	0.95(0.42)	1.79(0.26)	0.93(0.42)	1.13(0.65)	1.14(0.66)
site11	1.73(0.45)	2.30(0.29)	1.61(0.43)	2.01(0.78)	2.01(0.75)
RACHS	0.30(0.20)	0.45(0.15)	0.35(0.23)	0.47(0.31)	0.47(0.31)
CPB time	-0.03(0.31)				
cross-clamp time	0.28(0.33)		0.09(0.17)		
ACE inhibitors Yes	0.19(0.39)				
aspirin Yes	-0.34(0.45)				
eGFR	0.03(0.01)				
baseline serum creatinine	1.88(1.14)				
serum creatinine	0.20(0.49)				
serum Cystatin C	0.22(0.46)				
urine IL18	0.18(0.13)	0.22(0.07)	0.08(0.07)	0.22(0.14)	0.22(0.13)
urine NGAL	0.02(0.10)		0.07(0.07)		
urine creatinine	-0.28(0.25)				
urine KIM-1	0.18(0.16)		0.05(0.09)		
urine LFABP	0.01(0.11)				
Zero component					
(Intercept)	0.50(0.53)	-0.73(0.39)	0.40(1.87)	0.43(3.82)	0.42(3.76)
age					
genderM					
weight					
site9	-20(4671)		-1.86(0.74)	-11.55(2.04)	-9.14(2.02)

	ZIP	BE(0.1573)	EM-LASSO	EM-MCP	EM-SCAD
site11	-0.18(0.69)				
RACHS					
CPBtime					
crossclamptime					
ACE inhibitors Yes					
aspirin Yes					
eGFR					
baseline serum creatinine					
serum creatinine					
serum Cystatin C					
urine IL18					
urine NGAL					
urine creatinine					
urine KIM-1					
urine LFABP					
Summary					
logLik	-189.4	-204.1	-196.2	-192.8	-192.8
BIC	508.4	442	448.7	425.2	425.2
AIC	424.7	420.2	412.3	399.7	399.7