Generalized Monotone Incremental Forward Stagewise Method for Modeling Count Data: Application Predicting Micronuclei Frequency

Mateusz Makowski and Kellie J. Archer

Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA.

Supplementary Issue: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

ABSTRACT: The cytokinesis-block micronucleus (CBMN) assay can be used to quantify micronucleus (MN) formation, the outcome measured being MN frequency. MN frequency has been shown to be both an accurate measure of chromosomal instability/DNA damage and a risk factor for cancer. Similarly, the Agilent 4 × 44k human oligonucleotide microarray can be used to quantify gene expression changes. Despite the existence of accepted methodologies to quantify both MN frequency and gene expression, very little is known about the association between the two. In modeling our count outcome (MN frequency) using gene expression levels from the high-throughput assay as our predictor variables, there are many more variables than observations. Hence, we extended the generalized monotone incremental forward stagewise method for predicting a count outcome for high-dimensional feature settings.

KEYWORDS: micronuclei, Poisson regression, high-throughput, gene expression, penalized model

SUPPLEMENT: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

CITATION: Makowski and Archer. Generalized Monotone Incremental Forward Stagewise Method for Modeling Count Data: Application Predicting Micronuclei Frequency. *Cancer Informatics* 2015:14(S2) 97–105 doi: 10.4137/CIN.S17278.

RECEIVED: November 13, 2014. **RESUBMITTED:** January 14, 2015. **ACCEPTED FOR PUBLICATION:** January 22, 2015.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

TYPE: Methodology

FUNDING: Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM011169. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

DISCLAIMER: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

CORRESPONDENCE: makowskims@vcu.edu

Introduction

Micronuclei (MNs) are small nuclear bodies that are formed in dividing cells but are not part of the nucleus. Therefore, MNs can only be found in cells that have undergone nuclear division at least once and appear as small extranuclear bodies. When two daughter nuclei are formed during cell division, these bodies are placed into a smaller nucleus that is not part of the main nuclei, hence the term "micronuclei."[1] Once the MNs are formed, the cell has several different response options. MNs can remain within the cell, if they have functional DNA, as separate entities or be reabsorbed into the main nucleus. If the DNA is nonfunctional, the MNs may be expelled from the cell or the whole cell may be destroyed through apoptosis. Because MNs can be expelled from the cell, they can be used as a mechanism to remove extra chromosomes from the cell.[1]

MNs can form spontaneously or they can be induced by mutagens. Some spontaneous MNs are actually beneficial to the organism. An example is in the mouse cerebral cortex, wherein MN formation adds diversity to the nervous system.[1] However, a large majority of MNs are caused by mutagens and may play a role in carcinogenesis. Depending on the fate of the MN, the result could be a variety of different DNA and chromosome cell contents. This variety could result in an accumulation of DNA changes and instability that could result in cancer.[1] Several studies have shown that higher MN counts result in a higher risk of cancer in the future.[1] Thus, using the cytokinesis-block micronucleus (CBMN) assay as a risk assessment tool for cancer has potential clinical benefits. Further, combining CBMN with other high-throughput technologies such as gene expression and methylation analyses may help identify factors related to micronucleation.

Quantifying MNs in patient samples has been shown to be a good measure of genetic damage. MN scoring, ie, counting the number of MNs present in a sample, is a popular tool for testing genotoxicity mostly because of its simplicity, accuracy, applicability to different cell types, and ease of automation. Cancer cells show a loss of genetic control, which can be caused by DNA damage; so, they are good candidates for MN testing. The CBMN assay has successfully been used and validated to score MNs. The CBMN assay uses cytochalasin-B, which stops cells from performing cytokinesis but does not stop nuclear division, giving rise to cells that are binucleated.[2,3]

Furthermore, the Organisation for Economic Co-operation and Development has developed a set of guidelines for running the CBMN assay to obtain the most consistent and reliable results.[1]

Guidelines for the process of scoring MNs have been presented by the HUman MicroNucleus (HUMN) project. This is an international collaborative project aimed at improving the application of the CBMN assay. One of the main goals of the HUMN project is to identify methodological variables in the scoring of the assay to minimize confounding effects.[4] The HUMN project compiled a list of 6583 subjects from 25 laboratories in 16 countries and looked at background MN frequency using the CBMN assay. The goal of the study was to identify variables that affect the background MN frequency. Scoring criteria were found to account for 47% of the observed variability; thus, standardized scoring criteria were developed and described by Fenech et al.[4] The guideline includes scoring 2000 cells to accurately estimate MN frequency.

Because these guidelines were developed for assay performance, they do not address how to statistically analyze the data generated by the assay. This has led to the application of various statistical methods that may render different interpretations and conclusions. In a review article examining analytical methods, Ceppi et al.[5] reviewed 63 studies that statistically analyzed MN data and developed recommendations for selecting an appropriate analytical method. The review included studies that applied both parametric and nonparametric tests. The nonparametric tests included Kruskal–Wallis, Friedman, Wilcoxon, and Mann–Whitney *U*-tests. Although these tests do not require an underlying distributional assumption, they are unable to adjust for confounding factors. There were a variety of parametric tests performed that assume normality, such as analysis of variance, analysis of covariance, and multivariable linear models, which can adjust for confounding factors. Other methods such as correlations and Student's *t*-test were also used. However, applying these methods to MN data, which are rarely normally distributed, could result in inappropriate inferences. Although the data could be transformed to better adhere to a Gaussian distribution before applying such parametric tests, few studies applied any type of transformation. Further, Student's *t*-tests and Pearson's correlation cannot adjust for confounding variables. The common non-Gaussian models used were log-linear, Poisson, negative binomial, and logistic regressions. The logistic and log-linear models account for categories, whereas Poisson and negative binomial models directly model count data. For this reason, Ceppi et al.[5] recommend using negative binomial or Poisson models for MN data analysis. Another advantage of these count models is that they can adjust for confounding variables such as age, gender, and smoking status. Finally, Ceppi et al.[5] recommended that 2000 or more cells be scored for best model performance. If <2000 cells are scored, a zero-inflated Poisson model is recommended.[5]

When trying to identify molecular features related to MN frequency, high-throughput genomic assays can be used. However, the previously described methods cannot be applied in settings wherein there are more predictor variables than samples. Therefore, in this study, we extended the generalized monotone incremental forward stagewise (GMIFS) method to the Poisson regression setting and applied it to a cord blood study, the MN frequency of which we were interested in predicting using features from the Agilent 4 × 44k human oligonucleotide microarray.

Methods

Data. The cord blood data were collected as part of the Norwegian Mother and Child Cohort Study (MoBa).[6] The target population of MoBa comprised all women who gave birth in Norway. The overall goal of this study was to collect data on pregnant women and their children to estimate the association between exposures and diseases. Specifically, the data are taken from a subcohort called BraMat, which translates to "good food" in English. This subcohort concentrates on what effect a pregnant woman's diet has on her child. Umbilical cord blood samples were collected immediately after birth from 200 babies. After quality control and other exclusions, 111 samples were hybridized to Agilent 4 × 44k human oligonucleotide microarrays to measure gene expression. Of the 111 subjects, 29 also had MN data collected. The MNs were scored using the procedure described by Decordier et al.[7] Further, demographics such as gender, were collected for all subjects. Data were downloaded from Gene Expression Omnibus (GSE31836). Sample processing, image analysis, normalization, background correction, and filtering are described in the study by Hochstenbach et al.[8] For this analysis, the data were further filtered to only include genes that had no missing values, leaving 8497 genes for statistical analysis.

Statistical methods. There are many available methods that can model count data. However, these methods require independence of explanatory variables (p) and that the number of samples (n) does not exceed the number of explanatory variables. The incremental forward stagewise regression method for linear regression and the GMIFS for a logistic regression model have been previously described.[9] The GMIFS method for modeling ordinal response data has also been described.[10] To assist in our extension to the Poisson regression setting, we first review Poisson regression. We subsequently describe our GMIFS method for fitting Poisson regression models when $n < p$.

Poisson regression. Poisson regression is commonly used to model count data. Let $i = 1,\ldots, n$ be the number of observations and y_i represent a Poisson-distributed random variable. Let the expected value of y_i be written as

$$\mathrm{E}\left(y_i\right) = \lambda_i.$$

Then, the conditional probability is given by

$$P\left(y_i \mid \lambda_i\right) = \frac{\mathrm{e}^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

for each observation i. The likelihood is represented by

$$L(\lambda \mid \mathrm{y}) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}.$$

Mathematically, it is easier to maximize the log-likelihood, which is given by

$$\ell(\lambda \mid \mathrm{y}) = \sum_{i=1}^{n} \left(y_i \log \lambda_i - \lambda_i - \log(y_i!) \right).$$

Thus, we are looking for the value of λ that maximizes the log-likelihood above. Further, an offset is used if the response variable can be considered a rate. For example, MN frequency is scored from a larger number of total cells. Therefore, if the total number of cells examined varies by subject, an offset is appropriate. In this case, the expected value is

$$\mathrm{E}(y_i) = t_i \lambda_i$$

where t_i is the offset value. The conditional probability is then given by

$$P(y_i \mid \lambda_i) = \frac{\mathrm{e}^{-t_i \lambda_i} (t_i \lambda_i)^{y_i}}{y_i!}$$

for each observation i. The likelihood is represented by

$$L(\lambda \mid \mathrm{y}) = \prod_{i=1}^{n} \frac{\mathrm{e}^{-t_i \lambda_i} (t_i \lambda_i)^{y_i}}{y_i!}.$$

Again, mathematically, it is easier to maximize the log-likelihood, which is given by

$$\ell(\lambda \mid \mathrm{y}) = \sum_{i=1}^{n} \left(y_i \log(t_i \lambda_i) - t_i \lambda_i - \log(y_i!) \right).$$

Once again, we are looking for the λ value that maximizes the log-likelihood. These log-likelihoods are used to model predictor variables. In Poisson regression, the model assumes that the expected value can be modeled by a linear combination of predictors. In this case, the natural log of t_i is entered as an offset in the model estimation. The natural log of the expected value is

$$\log\left(\mathrm{E}(y_i \mid \mathbf{x}_i)\right) = \log(t_i) + \boldsymbol{\theta}' \mathbf{x}_i$$

where $\mathbf{x}_i$ is a vector of predictor variables and $\boldsymbol{\theta}$ is a vector of coefficients. The estimated coefficients can be exponentiated to determine how the response changes with the predictor. By using the estimated linear combination of coefficient estimates and taking the exponent, we can calculate the estimated response of that particular subject.

GMIFS Poisson model. The GMIFS method was previously described for the logistic regression scenario by Hastie et al.[9] but can be adapted to a Poisson regression model. For the proposed method, we consider three types of parameters that $\boldsymbol{\theta}$ from the section "Poisson regression" can be separated into along with an offset (t_i). The parameters are the intercept (α), those corresponding to an unpenalized subset of predictors ($\boldsymbol{\gamma}$), and those corresponding to a set of penalized predictors ($\boldsymbol{\beta}$). The design matrix, $\mathbf{x}$, consists of two parts, $\mathbf{x}_j$ and $\mathbf{x}_k$, where $j = 1,\ldots,J$ is the number of unpenalized predictors, $k = 1,\ldots,K$ is the set of penalized predictors, and $J + K = P$ is the total number of predictors. The unpenalized predictors are those that we wish to force into the model, such as gender, age, and smoking status, which researchers consider important predictors of MN frequency[5] and their values are in the $\mathbf{x}_{ij}$ design matrix for subject i. The penalized variables (thousands of features from a high-throughput genomic experiment) are those that the model will choose for us and are considered to be the investigative predictors and their values are in the $\mathbf{x}_{ik}$ design matrix for subject i.

The algorithm proceeds in an iterative fashion and updates one of the penalized covariates by a small incremental amount at each step. To determine which penalized covariate is to be updated next, the largest negative gradient is used. Thus, we need to calculate the first derivative of the log-likelihood corresponding to each penalized predictor. The log-likelihood written in terms of α, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ is

$$\ell(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathrm{y}, \mathbf{x}_j, \mathbf{x}_k) = \sum_{i=1}^{n} \Big(y_i \left(\alpha + \log(t_i) + \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{x}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_{ik} \right) - \exp\left(\alpha + \log(t_i) + \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{x}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_{ik} \right) - \log(y_i!) \Big)$$

and the first derivative written in terms of α, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ in matrix notation is

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{x}' \left(\mathrm{y} - \exp\left(\alpha + \log(t_i) + \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{x}_{ij} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_{ik} \right) \right).$$

Once we know which covariate to update, we need to determine in what direction to update the covariate. To know the direction of the update, the second derivative would need to be calculated, which is a cumbersome process. Hastie et al.[9] showed that to avoid having to calculate the second derivative, an expanded covariate space can be used. For example, let $\beta_1,\ldots,\beta_p$ be the positive coefficient estimates and $\beta_{p+1},\ldots,\beta_{2p}$ be the negative coefficient estimates. Then, the original estimates are calculated by subtracting the pairs, $\beta_1 - \beta_{p+1},\ldots,\beta_{2p} - \beta_2$. Thus, using the notation mentioned previously, where $\mathbf{x}_j$ are the unpenalized variables and $\mathbf{x}_k$ are the variables in the penalized subset, the expanded covariate space is $\tilde{\mathbf{x}} = \left[\mathbf{x}_j : \mathbf{x}_k : -\mathbf{x}_k \right]$. The proposed GMIFS algorithm using the expanded covariate set is

1. Initialize the components of $\hat{\boldsymbol{\beta}}^{(s)} = 0$ at step $s = 0$.
2. Initialize the intercept α and the unpenalized coefficients γ_j where $j = 1,\ldots, J$ using a maximization algorithm of the log-likelihood.
3. Considering α and $\boldsymbol{\gamma}$ fixed, find the predictor $\mathbf{x}_m$ where $m = \underset{K}{\operatorname{argmin}}\left(-\frac{\partial l}{\partial \beta_k}\right)$ at the current estimate $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(s)}$.
4. Update the corresponding coefficient $\hat{\beta}_m^{(s+1)} = \hat{\beta}_m^{(s)} + \varepsilon$ to yield a new vector of parameter estimates.
5. Update α and the unpenalized coefficients, γ_j, by maximum likelihood considering the $\hat{\boldsymbol{\beta}}^{s+1}$ from step 4 as fixed.
6. Repeat steps 3–5 until the difference between successive log-likelihoods is less than a prespecified tolerance, τ.

The defaults for the GMIFS algorithm are $\varepsilon = 0.001$ and $\tau = 0.00001$.

Comparative method: penalized linear regression. A penalized linear regression model can be fit by adding a penalty term to the sums of squares. Specifically, the glmpath algorithm uses a linear combination of the L_1 and L_2 norm penalizations. The generalized linear model path (glmpath) algorithm is based on a previous algorithm called least absolute shrinkage and selection operator (LASSO). LASSO minimizes the typical sum of squares with an added constraint. Specifically, for linear regression, LASSO minimizes[11]

$$\sum_{i=1}^{N}\left(y_i - \sum_j x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

where x_{ij} are the standardized predictors and y_i is the set of centered responses for $i = 1,\ldots, N$ and $j = 1,\ldots,p$. Because of the form of the constraint, LASSO does both variable selection and shrinkage. The glmpath algorithm modifies this slightly by first considering the typical generalized linear model formula

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} L(\mathrm{y};\boldsymbol{\beta})$$

where L denotes the appropriate likelihood function. The glmpath algorithm then adds an analagous LASSO penalty term to help with variable selection when $p > n$:

$$\hat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{-\log L(\mathrm{y};\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1\right\}$$

where $\lambda > 0$ is the regularization parameter. The glmpath algorithm computes coefficient estimates as λ varies. The algorithm starts with the largest λ that makes $\hat{\boldsymbol{\beta}}(\lambda)$ nonzero, with each step using a smaller λ. Each optimization consists of three parts: determining the step size in λ, predicting the corresponding change in the coefficients, and correcting the error in the previous prediction.[12] The algorithm continues finding the next largest λ that will change the coefficient estimates until no further predictors can be found. However, when the predictors are strongly correlated, the coefficient estimates become highly unstable using the L_1 norm penalization.[9] Thus, the glmpath algorithm adds a quadratic penalty term and computes the solution to

$$\hat{\boldsymbol{\beta}}(\lambda_1) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{-\log L(\mathrm{y};\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2}\|\boldsymbol{\beta}\|_2^2\right\}$$

where $\lambda_1 \in (0, \infty)$ and λ_2 is a fixed, small, positive constant. By adding this quadratic penalty, the effects of the strong correlations do not affect the stability of the fit. Further, when the correlations are not strong, the effects of the quadratic penalty are neglible.[9] Thus, the glmpath algorithm uses both the L_1 and L_2 penalties as its default method.

The glmpath algorithm uses a default binomial distribution with a logit link and $\lambda_2 = 0.00001$. The algorithm also allows for a Poisson distribution with a log link and Gaussian distribution with an identity link. The algorithm then computes the regularization path for generalized linear models with L1 penalty.

Simulations. Simulations are a useful technique to test how well a new methodology performs. In this case, we wished to quantify how accurately the GMIFS method estimated true nonzero coefficients and predicted count data. Furthermore, we wished to determine how the GMIFS method compared relative to the glmpath method in predicting the count outcome, and simulations provide a good platform to accomplish this comparison. Several general steps must be considered in the simulation process: how to simulate the response, how to simulate the predictors associated with the response, and how to simulate the predictors not associated with the response. Furthermore, we wished to examine how the methods perform under ideal situations and nonideal situations, such as when distributional assumptions are met and are not met, respectively. Note that all simulations were performed using the R programming environment (version 3.1.1).[13]

First, we considered the situation where the response is Poisson distributed and the user fits a Poisson regression model. Then, we generated the response to follow a Poisson distribution where an offset was either used or not used. The uniform distribution was used to generate the predictors. The steps involved in simulating the data under these conditions were as follows:

1. Randomly generate P variables, x_{i1}, x_{i2},…, x_{iP} where $i = 1\ldots n$, using the uniform distribution on the [0,1] interval.
2. Choose P_1 of the P variables to be associated with the response.

3. Assign the P_1 β values associated with the response and the intercept value, α. If the offset is to vary, then a uniform distribution was used with maximum 2200 and minimum 1800 and subsequently rounded to the nearest integer. This range was selected because it is recommended to score MNs using 2000 cells.
4. Generate the λ values for the Poisson distribution using the following formula:

$$\lambda_i = \exp\left(\alpha + \log\left(t_i\right) + \sum_{k=1}^{P_1} \beta_k x_{ik}\right).$$

5. Randomly generate $Y_i \sim \text{Poisson}(\lambda_i)$.
6. Fit a Poisson GMIFS model and fit a glmpath model.
7. Repeat steps 1–6 r times.

This simulation method was adjusted in several places. In this case, we chose $n = 30$ and $n = 80$. We studied the models letting $P = 100$ predictor variables and $P_1 = 5$ predictor variables associated with the response; $r = 100$ simulations were used. The intercept (α) and the five predictor variables associated with the response (β_1, β_2, β_3, β_4, and β_5) were set to –5, 0.3, 0.2, –0.7, 0.5, and 0.1, respectively, for data simulated using no offset. For data simulated using an offset, α was set to –7. This was done to keep λ values low so that the Gaussian approximation for the Poisson distribution is not appropriate. To compare the two different statistical models, the following three outcomes were examined:

1. The number of true predictors that have a nonzero coefficient;
2. the number of false predictors that have a nonzero coefficient;
3. accuracy of count predictions from the model (sum of squared residuals) when applied to an independent test set.

The methods were compared with and without the use of an offset during the simulation process. Furthermore, the glmpath method allows for the use of Gaussian and Poisson distributions. Thus, those options were also used to see what effects user error had on the results. Thus, a total of three models were compared when the true distribution was Poisson:

1. Poisson GMIFS model;
2. glmpath using "poisson" family option and $\lambda_2 = 0$ which fits a LASSO model; and
3. glmpath using "gaussian" family option and $\lambda_2 = 0$ which fits a LASSO model.

Results

Simulations. Simulations were performed as described in "Simulations" of the Methods section, and Figures 1–3 show the results of the simulations. Figure 1 shows the distribution of the number of predictors correctly identified as nonzero over 100 simulations and the types of models used. The data were generated using both $n = 30$ and $n = 80$ observations. The median number of correctly identified nonzero coefficients with no offset using GMIFS is 1 (range = 0, 3) for $n = 30$ and 2 (range = 0, 4) for $n = 80$. Similarly, the median number of correctly identified nonzero coefficients with no offset using glmpath with Poisson family is 1 (range = 0, 5) for $n = 30$ and 2 (range = 0, 4) for $n = 80$. This number increases slightly when using the glmpath with Gaussian family to a median of 2 (range = 0, 5) for $n = 30$ and 4 (range = 2, 5) for $n = 80$. All the numbers are similar when an offset is used to generate the data. The median number of correctly identified non-zero coefficients using GMIFS is 0 (range = 0, 3) for $n = 30$ and 1 (range = 0, 4) for $n = 80$. The median number of correctly identified non-zero coefficients using glmpath with Poisson family is 0 (range = 0, 3) for $n = 30$ and 1 (range = 0, 4) for $n = 80$. Once again the medians increase when using the glmpath with Gaussian family to 2 (range = 0, 5) for $n = 30$ and 4 (range = 2, 5) for $n = 80$.

Figure 2 shows the distribution of the number of predictors incorrectly identified as nonzero over 100 simulations and the types of models used. The data were generated using both $n = 30$ and $n = 80$ observations. The median number of incorrectly identified nonzero coefficients with no offset using GMIFS is 3 (range = 0, 15) for $n = 30$ and 7 (range = 0, 28) for $n = 80$. Similarly, the median number of incorrectly identified nonzero coefficients with no offset using glmpath with Poisson family is 3 (range = 0, 17) for $n = 30$ and 7 (range = 0, 41) for $n = 80$. This number increases when using the glmpath with Gaussian family to a median of 26 (range = 23, 28) for $n = 30$ and 74 (range = 73, 76) for $n = 80$. All results are similar when an offset is used to generate the data. The median number of incorrectly identified non-zero coefficients using GMIFS is 2 (range = 0, 14) for $n = 30$ and 5 (range = 0, 26) for $n = 80$. The median number of incorrectly identified non-zero coefficients using glmpath with Poisson family is 2 (range = 0, 24) for $n = 30$ and 4.5 (range = 0, 31) for $n = 80$. Once again the medians increase when using the glmpath with Gaussian family to 26 (range = 23, 28) for $n = 30$ and 74 (range = 72, 76) for $n = 80$.

Figure 3 shows the distribution of the sum of residuals squared as a measure of the model prediction accuracy. The data were generated using both $n = 30$ and $n = 80$ observations. For both sample sizes, a learning data set was used to estimate coefficients and then the model was applied to an independent test data set. The median accuracy with no offset using GMIFS is 133 (range = 68, 240) for $n = 30$ and 325 (range = 188, 699) for $n = 80$. Similarly, the median accuracy with no offset using glmpath with Poisson family is 142 (range = 55, 254) for $n = 30$ and 333 (range = 185, 1666) for $n = 80$. The median accuracy with no offset using glmpath with Gaussian family is 206 (range = 90, 383) for $n = 30$ and

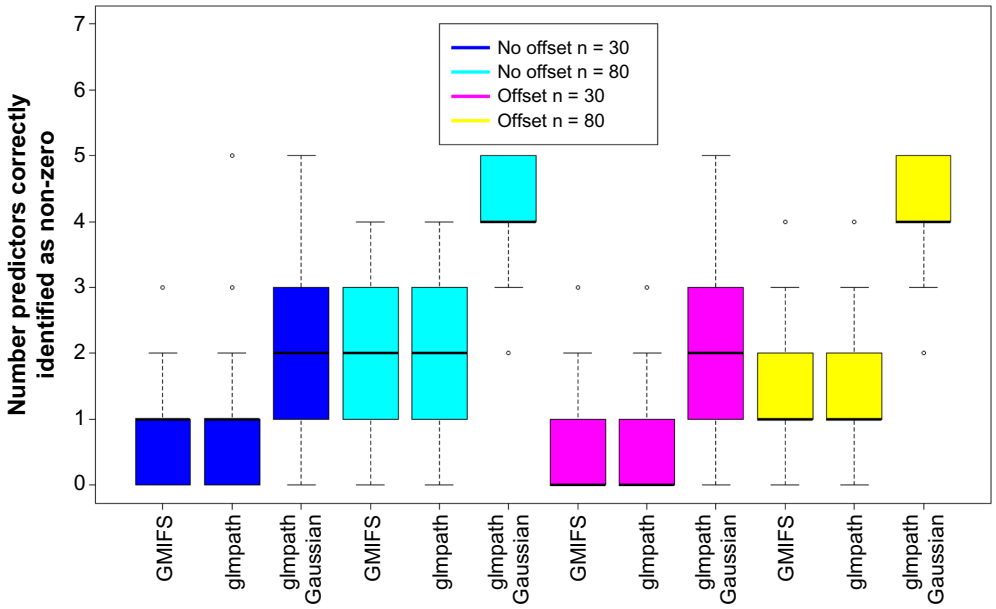

Figure 1. Number of predictors correctly identified as nonzero. This figure shows the distribution of the number of predictors correctly identified as nonzero over 100 simulations. There were five predictors that were set as nonzero. Boxplots are separated by the type of distribution used to generate the data and the number of observations.

1503 (range = 535, 3772) for $n = 80$. The numbers are different when an offset is used to generate the data. The median accuracy using GMIFS is 80 (range = 30, 185) for $n = 30$ and 205 (range = 137, 367) for $n = 80$. The median accuracy using glmpath with Poisson family is 80 (range = 33, 805) for $n = 30$ and 206 (range = 126, 339) for $n = 80$. The median accuracy with an offset using glmpath with Gaussian family for both sample sizes is above 50000.

Gene expression analysis. Both GMIFS and glmpath models were applied to the cord blood gene expression data set described under "Data" of Methods section. For glmpath, the Poisson family option was used and the lambda2 option was set to zero. For GMIFS, the default options were chosen. The response in the model was MN counts, and the predictors were the gene expression intensities. Gender was included in the model as part of the unpenalized subset. Based on Figure 4, a

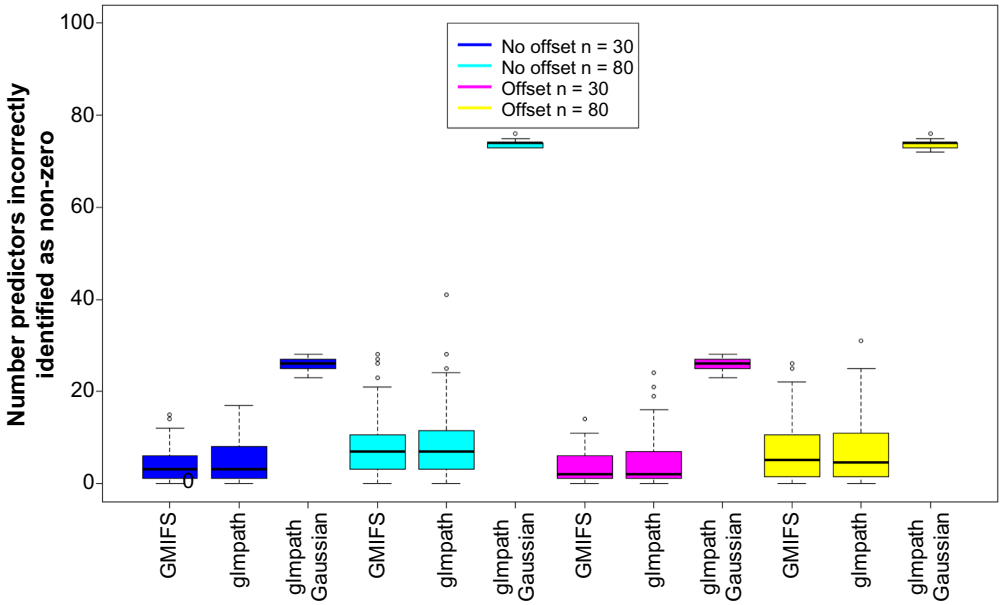

Figure 2. Number of predictors incorrectly identified as nonzero. This figure shows the distribution of the number of predictors incorrectly identified as nonzero over 100 simulations. There were 95 predictors for which their coefficients were set to zero. Boxplots are separated by the type of distribution used to generate the data and the number of observations.

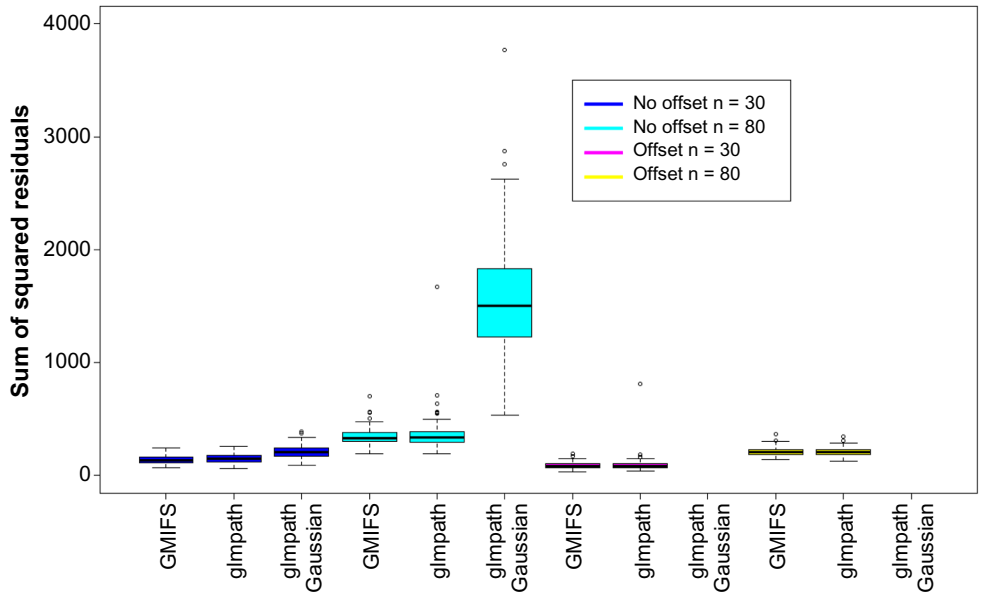

Figure 3. Accuracy of count predictions. This figure shows the distribution of the sum of residuals squared over 100 simulations using a learning data set and an independent test data set. Boxplots are separated by the type of model fit to the data and the number of observations. The results for glmpath with Gaussian family using an offset are not displayed because both values are above 50000.

Poisson distribution was assumed for both models because the data appear skewed. The final model parameters were chosen using the minimum Akaike information criterion. The GMIFS model identified 17 nonzero gene expression coefficients as associated with MN count and the glmpath with Poisson family identified 23. Out of the genes that were identified, 10 were common to both models. Figures 5 (sum of squared residuals = 101.7) and 6 (sum of squared residuals = 1.8) show that both models seem to predict MNs relatively well. Table 1 shows the genes that both models identified as being associated with MN count and the types of cancer with which they are linked. Nine out of the 10 genes in common between both models are linked to some type of cancer.

Discussion

We have described the GMIFS method for modeling a count response when we want to (1) coerce some variables into the model and (2) perform automatic variable selection and model estimation by penalizing predictors. High-throughput data contain more predictors than there are samples, so traditional methods are not appropriate in this setting. The GMIFS method was compared to glmpath, a popular penalization algorithm. Simulations showed that both methods performed similarly when identifying predictors known to be nonzero. GMIFS appeared to slightly outperform glmpath in the sense that GMIFS included fewer predictors that are truly unimportant in the model. Similarly, when applied to an independent data set, GMIFS appeared to have higher predictive accuracy. Thus, it appears that GMIFS is more generalizable than glmpath to independent data sets.

Finally, both methods were applied to a cord blood gene expression data set. Gene expression profiles were used

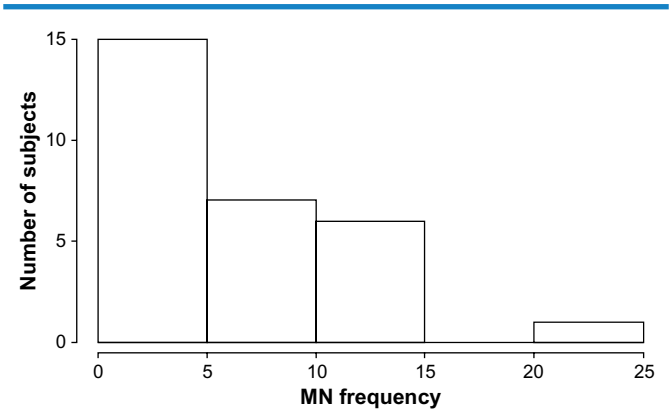

Figure 4. Histogram of MN counts.

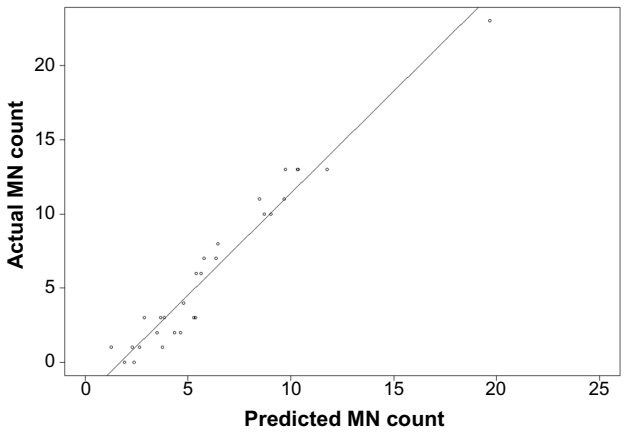

Figure 5. Plot of actual MN counts versus predicted MN counts using GMIFS.

Table 1. Genes identified as associated with MN count by both GMIFS and glmpath.

PROBE ID	GENE SYMBOL	GENE NAME	ASSOCIATED WITH CANCER	GMIFS	GLMPATH
A-23-P100196	USP10	ubiquitin specific peptidase 10	Glioblastoma multiforme[14]	X	X
A-23-P138967	SDHD	succinate dehydrogenase complex	Tumor Suppressor[15]	X	X
A-23-P42331	HMGA1	high mobility group AT-hook 1	Pancreatic Adenocarcinoma[16]	X	X
A-23-P9293	TJP2	tight junction protein 2	Breast[17]	X	X
A-24-P19410	CBX7	chromobox homolog 7	Carcinomas[18]	X	X
A-24-P214858	TREML2	triggering receptor expressed on myeloid cells-like 2	Pancreatic[19]	X	X
A-24-P2463	WHSC1	Wolf-Hirschhorn syndrome candidate 1	Carcinogenesis[20]	X	X
A-24-P397584	TBCC	tubulin folding cofactor C	None Found	X	X
A-24-P398064	KIAA0258	KIAA0258	Colorectal[21]	X	X
A-32-P18547	C21ORF57	chromosome 21 open reading frame 57	Breast[22]	X	X
A-23-P103824	FAU	Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed	None Found	X	
A-23-P209394	CFLAR	CASP8 and FADD-like apoptosis regulator	Human cancers[23]	X	
A-23-P79911	PSMF1	proteasome (prosome, macropain) inhibitor subunit 1 (PI31)	Breast[24]	X	
A-24-P202567	ITPKC	inositol 1,4,5-trisphosphate 3-kinase C	Cervical[25]	X	
A-24-P31235	EIF5A	eukaryotic translation initiation factor 5A	Chronic myeloid leukemia[26]	X	
A-24-P405054	C1ORF144	chromosome 1 open reading frame 144	Mantle cell lymphoma[27]	X	
A-32-P156549	C1ORF144			X	
A-23-P118313	GABARAPL2	GABA(A) receptor-associated protein-like 2	Lung[28]		X
A-23-P143817	MYLK	myosin, light polypeptide kinase	Gastric[29]		X
A-23-P156809	LOC642880	similar to FKSG62	None Found		X
A-23-P394304	PDZK1IP1	PDZK1 interacting protein 1	Thyroid[30]		X
A-23-P39665	SLC11A1	solute carrier family 11, member 1	Esophageal[31]		X
A-23-P67529	KCNN4	potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4	Colorectal[32]		X
A-24-P594683	LOC645592	similar to peptidylprolyl isomerase A isoform 1			X
A-24-P708161					X
A-24-P98086	GNA12	guanine nucleotide binding protein (G protein) alpha 12	Oral[33]		X
A-32-P10067					X
A-32-P137849					X
A-32-P169754	LOC145221	EST			X
A-32-P208078	MTHFR	5,10-methylenetetrahydrofolate reductase (NADPH)	Breast[34]		X

to predict MN frequency. Both models identified a similar number of genes as related to MN frequency. Further, 10 of those genes were common to both models. Nine out of the 10 genes have been shown to be associated with different types of cancers. Because MN count is a measure of DNA damage, genes associated with MN frequency would be expected to be linked to cancer.

Both models appear to identify genes linked to cancer. As in the simulations, glmpath identified more genes as nonzero compared to GMIFS. In the simulations, this was because glmpath was including more predictors incorrectly. However, there is no way to know whether this is also the case in the cord blood data set, given that these data are observational and no further confirmatory studies can be performed on the samples.

Author Contributions

Conceived and designed the experiments: KJA. Analyzed the data: MSM. Wrote the first draft of the manuscript: MSM. Contributed to the writing of the manuscript: KJA. Agree with manuscript results and conclusions: MSM, KJA. Jointly developed the structure and arguments for the paper: MSM, KJA. Made critical revisions and approved final

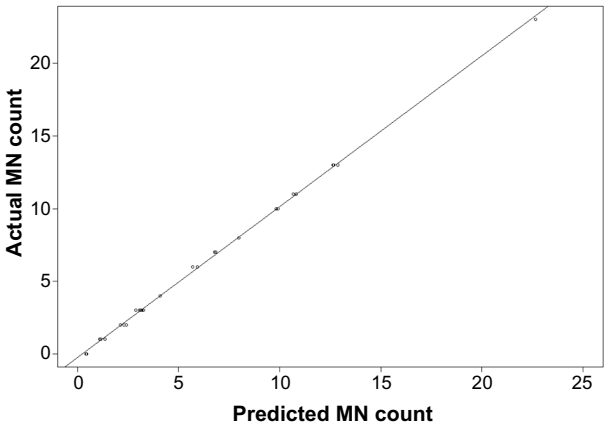

Figure 6. Plot of actual MN counts versus predicted MN counts using glmpath.

version: KJA. Both authors reviewed and approved of the final manuscript.

REFERENCES

1. Kirsch-Volders M, Plas G, Elhajouji A, et al. The in vitro MN assay in 2011: origin and fate, biological significance, protocols, high throughput methodologies and toxicological relevance. *Arch Toxicol*. 2011;85(8):873–99.
2. Fenech M. The cytokinesis-block micronucleus technique: a detailed description of the method and its application to genotoxicity studies in human populations. *Mutat Res*. 1993;285(1):35–44.
3. Fenech M, Holland N, Chang WP, Zeiger E, Bonassi S. The human micronucleus projectan international collaborative study on the use of the micronucleus technique for measuring DNA damage in humans. *Mutat Res*. 1999;428(1):271–83.
4. Fenech M, Chang WP, Kirsch-Volders M, Holland N, Bonassi S, Zeiger E. Humn project: detailed description of the scoring criteria for the cytokinesis-block micronucleus assay using isolated human lymphocyte cultures. *Mutat Res*. 2003;534(1):65–75.
5. Ceppi M, Biasotti B, Fenech M, Bonassi S. Human population studies with the exfoliated buccal micronucleus assay: statistical and epidemiological issues. *Mutat Res*. 2010;705(1):11–9.
6. Magnus P, Irgens LM, Haug K, et al. Cohort profile: the Norwegian mother and child cohort study (moba). *Int J Epidemiol*. 2006;35(5):1146–50.
7. Decordier I, Papine A, Plas G, et al. Automated image analysis of cytokinesis-blocked micronuclei: an adapted protocol and a validated scoring procedure for biomonitoring. *Mutagenesis*. 2009;24(1):85–93.
8. Hochstenbach K, van Leeuwen DM, Gmuender H, et al. Global gene expression analysis in cord blood reveals gender-specific differences in response to carcinogenic exposure in utero. *Cancer Epidemiol Biomarkers Prev*. 2012;21(10):1756–67.
9. Hastie T, Taylor J, Tibshirani R, Walther G. Forward stagewise regression and the monotone lasso. *Electron J Stat*. 2007;1:1–29.
10. Archer K, Hou J, Zhou Q, Ferber K, Layne J, Gentry A. ordinalgmifs: an R package for ordinal regression in high-dimensional data settings. *Cancer Inform*. 2014;13:187–95.
11. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc Series B Stat Methodol*. 2011;73(3):273–82.
12. Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. *J R Stat Soc Series B Stat Methodol*. 2007;69(4):659–77.
13. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2014.
14. Grunda JM, Nabors LB, Palmer CA, et al. Increased expression of thymidylate synthetase (TS), ubiquitin specific protease 10 (USP10) and survivin is associated with poor survival in glioblastoma multiforme (GBM). *J Neurooncol*. 2006;80(3):261–74.
15. King A, Selak M, Gottlieb E. Succinate dehydrogenase and fumarate hydratase: linking mitochondrial dysfunction and cancer. *Oncogene*. 2006;25(34):4675–82.
16. Liau SS, Rocha F, Matros E, Redston M, Whang E. High mobility group at-hook 1 (HMGA1) is an independent prognostic factor and novel therapeutic target in pancreatic adenocarcinoma. *Cancer*. 2008;113(2):302–14.
17. Martin TA, Watkins G, Mansel RE, Jiang WG. Loss of tight junction plaque molecules in breast cancer tissues is associated with a poor prognosis in patients with breast cancer. *Eur J Cancer*. 2004;40(18):2717–25.
18. Federico A, Pallante P, Bianco M, et al. Chromobox protein homologue 7 protein, with decreased expression in human carcinomas, positively regulates e-cadherin expression by interacting with the histone deacetylase 2 protein. *Cancer Res*. 2009;69(17):7079–87.
19. Loos M, Hedderich DM, Ottenhausen M, et al. Expression of the costimulatory molecule b7-h3 is associated with prolonged survival in human pancreatic cancer. *BMC Cancer*. 2009;9(1):463.
20. Toyokawa G, Cho HS, Masuda K, et al. Histone lysine methyltransferase wolf-hirschhorn syndrome candidate 1 is involved in human carcinogenesis through regulation of the wnt pathway. *Neoplasia*. 2011;13(10):887-IN11.
21. Sasaki H, Miura K, Horii A, et al. Orthotopic implantation mouse model and cdna microarray analysis indicates several genes potentially involved in lymph node metastasis of colorectal cancer. *Cancer Sci*. 2008;99(4):711–9.
22. Smeets A, Daemen A, Bempt IV, et al. Prediction of lymph node involvement in breast cancer from primary tumor tissue using gene expression profiling and mirnas. *Breast Cancer Res Treat*. 2011;129(3):767–76.
23. Fulda S. Targeting c-flice-like inhibitory protein (CFLAR) in cancer. *Expert Opin Ther Targets*. 2013;17(2):195–201.
24. Kuznetsova E, Kekeeva T, Larin S, et al. [Novel methylation and expression markers associated with breast cancer]. *Mol Biol*. 2006;41(4):624–33.
25. Yang YC, Chang TY, Chen TC, et al. Genetic polymorphisms in the ITPKC gene and cervical squamous cell carcinoma risk. *Cancer Immunol Immunother*. 2012;61(11):2153–9.
26. Balabanov S, Gontarewicz A, Ziegler P, et al. Hypusination of eukaryotic initiation factor 5a (eIF5A): a novel therapeutic target in BCR-ABL – positive leukemias identified by a proteomics approach. *Blood*. 2007;109(4):1701–11.
27. Schraders M, Jares P, Bea S, et al. Integrated genomic and expression profiling in mantle cell lymphoma: identification of gene-dosage regulated candidate genes. *Br J Haematol*. 2008;143(2):210–21.
28. Borczuk AC, Gorenstein L, Walter KL, Assaad AA, Wang L, Powell CA. Non-small-cell lung cancer molecular signatures recapitulate lung developmental pathways. *Am J Pathol*. 2003;163(5):1949–60.
29. Chen L, Su L, Li J, et al. Hypermethylated FAM5C and MYLK in serum as diagnosis and pre-warning markers for gastric cancer. *Dis Markers*. 2012;32(3):195–202.
30. Di Maro G, Maria Orlandella F, Claudio Bencivenga T, et al. Identification of targets of Twist1 transcription factor in thyroid cancer cells. *J Clin Endocrinol Metab*. 2014;99(9):E1617–26.
31. Zaahl MG, Warnich L, Victor TC, Kotze MJ. Association of functional polymorphisms of SLC11A1 with risk of esophageal cancer in the south african colored population. *Cancer Genet Cytogenet*. 2005;159(1):48–52.
32. Lai W, Chen S, Wu H, et al. PRL-3 promotes the proliferation of lovo cells via the upregulation of KCNN4 channels. *Oncol Rep*. 2011;26(4):909.
33. Gan C, Zain RB, Abraham M, et al. P126. Expression of GNA12 and its role in oral cancer. *Oral Oncology*. 2011;47:S114–5.
34. Chen J, Gammon MD, Chan W, et al. One-carbon metabolism, MTHFR polymorphisms, and risk of breast cancer. *Cancer Res*. 2005;65(4):1606–14.