Biostatistics (2013), *pp.* 1–14
doi:10.1093/biostatistics/kxt052

Analysis of counts with two latent classes, with application to risk assessment based on physician-visit records of cancer survivors

HUIJING WANG, X. JOAN HU*

Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6
joanh@stat.sfu.ca

MARY L. MCBRIDE, JOHN J. SPINELLI

Cancer Control Research, BC Cancer Agency, Vancouver, BC, Canada V5Z 1L3

Summary

Motivated by a cancer survivorship program, this paper explores event counts from two categories of individuals with unobservable membership. We formulate the counts using a latent class model and consider two likelihood-based inference procedures, the maximum likelihood estimation (MLE) and a pseudo-MLE procedure. The pseudo-MLE utilizes additional information on one of the latent classes. It yields reduced computational intensity and potentially increased estimation efficiency. We establish the consistency and asymptotic normality of the proposed pseudo-MLE, and we present an extended Huber sandwich estimator as a robust variance estimator for the pseudo-MLE. The finite-sample properties of the two-parameter estimators along with their variance estimators are examined by simulation. The proposed methodology is illustrated by physician-claim data from the cancer program.

Keywords: Efficiency vs. robustness; Mixture Poisson model; Pseudo-maximum likelihood estimation; Robust variance estimation; Supplementary information.

1. Introduction

The population of cancer survivors has been increasing rapidly because of improvements in cancer treatments. These survivors are often at risk of subsequent and ongoing problems that are mainly treatment-related. The evaluation or development of strategies for long-term management requires risk assessment, particularly for those diagnosed with cancer at a young age. The Childhood, Adolescent, Young Adult Cancer Survivorship (CAYACS) research program at the British Columbia (BC) Cancer Agency (http://www.cayacs.ca), using existing population-based data sets and record-linkage methodology, has been conducting a series of epidemiologic, clinical, and health-service studies relating to the survivorship issues of cancer survivors diagnosed at age 0 to 19; see McBride *and others* (2010). A recent CAYACS project, summarized in McBride *and others* (2011), reports an analysis of the physician claims associated

*To whom correspondence should be addressed.

with a cohort of young cancer survivors. It shows that the demand for physician care among these survivors is considerably greater than that of a similar age and sex group in the general population.

The analysis of McBride *and others* (2011) provides insights into the physician-visit patterns of the survivors and also raises further issues. For example, the comparison of the cancer survivors as a group to the general population may implicitly reveal whether the number of survivors in the cohort at risk of later and ongoing problems is significant. It does not explicitly relate this risk to the consequences of the original cancer diagnoses. Moreover, the analysis of McBride *and others* (2011) indicates that the physician-visit frequency of the females in the cohort is significantly higher than that of the males. It is not clear whether this identifies sex as an important risk factor or simply reflects an overall pattern of physician visits. In fact, such a pattern is also seen in the general population.

Preliminary analyses indicated that, while many cancer survivors visit physicians rather frequently, some survivors in the cohort have physician-visit patterns similar to those of the general population. A perception among researchers in the field is that some cancer survivors can live as normally as people without a cancer diagnosis. This motivated us to model the survivor cohort as a mixture of two latent classes: the groups "at-risk" and "not-at-risk" of later effects of the original cancer diagnoses. The individuals in the *at-risk* group have a potentially higher frequency of physician visits, while the individuals in the *not-at-risk* group have the same physician-visit patterns as the general population.

Goodman (1974) formalizes the latent class model introduced by Lazarsfeld and Henry (1968) and derives the maximum likelihood estimation (MLE) procedure. The latent class model has had a wide range of applications; see, for example, Magidson and Vermunt (2002), Pepe and Janes (2007), and Vermunt (2008). The formulation provides us with a convenient framework to study the features of physician visits due to the later and ongoing treatment-related problems of cancer survivors. It allows us to evaluate separately the visit frequencies of the two latent groups in the cohort and may better assess the risk of later and ongoing problems. The model also leads to a natural comparison of the survivors in the *at-risk* group to the general population, if the *not-at-risk* group in the cohort is defined as the class that has the same physician-visit frequency as the general population.

In an analysis with a latent class model, one usually needs to specify the underlying probability model in a parametric form for each of the latent classes to avoid non-identifiability problems in general. Moreover, in addition to issues such as computational robustness when implementing likelihood-based procedures with latent class models (e.g. Hall and Shen, 2010), the efficiency of the MLE will drop considerably because of the increased number of parameters. A model with two latent classes has almost three times as many parameters as a comparable marginal model. On the other hand, in many practical situations, information is readily available on one of the two latent classes. In the CAYACS case, the provincial medical insurance system collects rich information on the general population. These considerations yielded a pseudo-MLE procedure, an alternative way to estimate the model parameters using additional/supplementary information. The procedure is potentially more efficient and robust as well as relatively easy to implement.

In this paper, we motivate and illustrate the proposed model and associated inference procedures using the CAYACS program. The methodology is not limited to the program and can be applied more broadly. The rest of this paper is organized as follows. Section 2 introduces the notation and a mixture Poisson model for the physician-visit records of the CAYACS cohort. In Section 3, we first present the MLE for the model parameters with the primary data and an application of the expectation–maximization (EM) algorithm to compute the MLE. We then propose a pseudo-MLE procedure using the additional information on the *not-at-risk* group, namely the physician-visit records from a collection of individuals selected from the general population. We establish the consistency and asymptotic normality of the pseudo-MLE and derive its asymptotic variance. Two variance estimators for the pseudo-MLE are presented. Section 4 reports the simulation studies of efficiency and robustness that we conducted to examine the finite-sample properties of the inference procedures together with the two variance estimators. An analysis of the CAYACS

physician-visit data via the proposed methodology is presented in Section 5. Some final remarks are given in Section 6.

2. Notation and model

Let N represent a subject's count of physician visits over the time period $(0, T]$ and Z be his/her covariate vector. The observation period in the CAYACS application is the time interval starting when a BC resident with a cancer diagnosis at a young age becomes a survivor, and ending at his/her death or the end of the data collection. Here, a survivor is a person who has survived at least five years since his/her original cancer diagnosis. We allow the observation period to vary from subject to subject.

To formulate the two strata with unobservable membership, corresponding to the *at-risk* and *not-at-risk* groups in the survivor cohort, we introduce a latent binary variable η to indicate whether a subject belongs to the *at-risk* group. Denote $E(\eta|Z) = P(\eta = 1|Z)$ by $p(Z)$, and the conditional expectations of N for the *at-risk* and *not-at-risk* groups by $E(N|\eta, T, Z) = \Lambda_\eta(T, Z)$ for $\eta = 1$ and 0, respectively. Thus, the expectation of N conditional on T and Z is $E(N|T, Z) = \Lambda_1(T, Z)p(Z) + \Lambda_0(T, Z)[1 - p(Z)]$. This latent class model is further specified into a finite mixture Poisson model as follows.

We assume that the counts N of the two groups follow the Poisson distribution with the conditional expectations $\Lambda_\eta(T, Z)$ for $\eta = 1, 0$. This formulation includes the popular zero-inflated Poisson (ZIP) model (e.g. Lambert, 1992) as a special case with $\Lambda_0(T, Z) \equiv 0$. This paper adopts the commonly used parametric specifications for $p(Z)$ and $\Lambda_\eta(T, Z)$, the logistic and loglinear regression models:

$$\text{logit}\{p(Z; \alpha)\} = \alpha_0 + \alpha_1' Z \tag{2.1}$$

and

$$\log\{\Lambda_1(T, Z; \beta)\} = \beta_0 + \beta_1' Z + \beta_2 \log T, \quad \log\{\Lambda_0(T, Z; \theta)\} = \theta_0 + \theta_1' Z + \theta_2 \log T. \tag{2.2}$$

Our estimation procedures and discussions are applicable to other parametric specifications with little modification.

Our primary interest lies in estimating the parameters $\alpha = (\alpha_0, \alpha_1')'$, $\beta = (\beta_0, \beta_1', \beta_2)'$, and $\theta = (\theta_0, \theta_1', \theta_2)'$ in $p(Z; \alpha)$, $\Lambda_1(T, Z; \beta)$, and $\Lambda_0(T, Z; \theta)$ as given by (2.1) and (2.2) with the primary data $\{(N_i, T_i, Z_i) : i = 1, \ldots, n\}$, a set of n independent and identically distributed realizations of (N, T, Z). For the CAYACS application, a consistent estimator of α gives a consistent estimator of the risk probability $p(Z; \alpha)$ and then yields a measure on how likely the survivors with covariates of Z having the later effects of the original cancer diagnoses. The estimator of α can also be used to identify risk factors associated directly with the later effects. Consistent estimators of β and θ, on the other hand, can be used to identify factors associated with the high visit frequency of the *at-risk* group and the low frequency of the *not-at-risk* group. Moreover, comparisons of β and θ based on their estimates can detect differences in the physician-visit frequency between the two groups. These applications are illustrated in Section 5 with the CAYACS physician-claim data.

3. Likelihood-based inference procedures

With the model in Section 2, the event count N conditional on T and Z follows the mixture Poisson distribution

$$P(N \mid T, Z; \alpha, \beta, \theta) = P(N \mid \eta = 1, T, Z; \beta)p(Z; \alpha) + P(N \mid \eta = 0, T, Z; \theta)[1 - p(Z; \alpha)], \tag{3.1}$$

where $P(N \mid \eta, T, Z)$ is the probability mass function of the Poisson distribution with a mean of $\Lambda_\eta(T, Z)$. We consider the MLE procedure based on the primary data. The EM algorithm (Dempster *and others*, 1977) is adapted to compute the MLE of the parameters. We then assume that there is a consistent estimator for θ in $\Lambda_0(T, Z; \theta)$, the event frequency model for the *not-at-risk* group. A pseudo-MLE procedure is then proposed for estimating the parameters α in the risk model (2.1) and β in the event frequency model for the *at-risk* group in (2.2).

3.1 *Maximum likelihood estimation*

Under the mixture Poisson model (3.1), the likelihood function of (α, β, θ) based on the primary data $\{(N_i, T_i, Z_i) : i = 1, \ldots, n\}$ is

$$L(\alpha, \beta, \theta; \mathbf{N} \mid \mathbf{T}, \mathbf{Z}) = \prod_{i=1}^{n} P(N_i \mid T_i, Z_i; \alpha, \beta, \theta). \tag{3.2}$$

The MLE of (α, β, θ) may be attained by directly maximizing (3.2) or its log-transformation. With the usual regularity conditions, the MLE $(\hat{\alpha}, \hat{\beta}, \hat{\theta})$ has asymptotic normality. That is, as $n \to \infty$, $\sqrt{n}(\hat{\alpha} - \alpha, \hat{\beta} - \beta, \hat{\theta} - \theta)'$ converges in distribution to the multivariate normal distribution with mean zero and variance $\mathrm{FI}(\alpha, \beta, \theta)^{-1}$. Here, $\mathrm{FI}(\alpha, \beta, \theta)$ is the Fisher information matrix; it can be consistently estimated by $-n^{-1}\partial^2 \log L(\alpha, \beta, \theta; \mathbf{N} \mid \mathbf{T}, \mathbf{Z})/\partial(\alpha, \beta, \theta)^2$ with the MLE plugged in.

Applying the EM algorithm gives us an alternative procedure for finding the MLE of (α, β, θ); this algorithm is potentially more intuitive and easier to implement. The estimation procedure is presented in Section A of supplementary material available at *Biostatistics* online. In particular, we consider the "full data" as $\{(N_i, \eta_i, T_i, Z_i) : i = 1, \ldots, n\}$ in the application. We can verify the conditions that ensure the resulting sequence of estimates converges to the MLE $(\hat{\alpha}, \hat{\beta}, \hat{\theta})$ from $L(\alpha, \beta, \theta; N \mid T, Z)$ in (3.2). This procedure with the ZIP model coincides with the estimation procedure presented in Hall and Shen (2010). We may follow their discussion to provide a variation of the EM algorithm.

3.2 *Pseudo-MLE*

Suppose that a set of independent observations from the general population, denoted by $\{(N_j, T_j, Z_j) : j = 1, \ldots, m\}$, is available in addition to the primary data from the survivor cohort. One may estimate (α, β, θ) with the likelihood function based on the primary data in combination with the supplementary information, which is the product of $L(\alpha, \beta, \theta; \mathbf{N} \mid \mathbf{T}, \mathbf{Z})$ in (3.2) and $L_{\mathrm{supp}}(\theta) = \prod_{j=1}^{m} P(N_j \mid \eta_j \equiv 0, T_j, Z_j; \theta)$. The efficiency of the MLE for the combined data is presumably higher than that of the MLE discussed in Section 3.1 based on the primary data. However, the computational issues remain.

In many practical situations, the sample size m can be quite large relative to the size n of the primary data, and thus the supplementary data alone can lead to a consistent estimator of θ with sufficient efficiency. To make a comparison between the survivor cohort and the general population, for example, the CAYACS program collected data from the general population with a sample size (m) 10 times the size of the primary data (McBride *and others*, 2011); in fact, m could be larger in the application if necessary. We propose the following pseudo-MLE for estimating (α, β) using such an estimator of θ from the supplementary data, to achieve an easily implementable estimator with reasonably high efficiency.

Assume that the available supplementary data yield $\tilde{\theta}$, an estimator for the parameters in the frequency model associated with the *not-at-risk* group, and $\sqrt{m}(\tilde{\theta} - \theta)$ converges in distribution to the normal distribution with zero mean and variance $\mathrm{AV}_{\tilde{\theta}}(\theta)$ as $m \to \infty$. The MLE of θ from the aforementioned $L_{\mathrm{supp}}(\theta)$ based on the supplementary data, for example, satisfies the assumptions about $\tilde{\theta}$. It yields a pseudo-MLE of (α, β), denoted by $(\tilde{\alpha}, \tilde{\beta})$, maximizing the likelihood function (3.2) with respect to (α, β) and with θ

fixed at $\tilde{\theta}$. This estimation procedure is considerably simpler than the procedure for computing the MLE $\hat{\alpha}$ and $\hat{\beta}$ jointly with $\hat{\theta}$ in Section 3.1. The computational intensity is reduced by roughly one-third in general. The pseudo-MLE can be found by applying the adapted EM algorithm in Section A of supplementary material available at *Biostatistics* online, with $\theta = \tilde{\theta}$ throughout the algorithm.

Following the arguments in Gong and Samaniego (1981), we establish the consistency and the asymptotic normality of $(\tilde{\alpha}, \tilde{\beta})$. Specifically, as $n \to \infty$ and $m \to \infty$, and assuming that $n/m \to k > 0$ and $\tilde{\theta}$ is independent of the primary data, $\sqrt{n}(\tilde{\alpha} - \alpha, \tilde{\beta} - \beta)'$ converges to the normal distribution with mean zero and variance

$$\mathrm{AV}_{(\tilde{\alpha},\tilde{\beta})}(\alpha, \beta, \theta) = I_{11}^{-1} + kI_{11}^{-1} I_{12} \mathrm{AV}_{\tilde{\theta}}(\theta) I_{21} I_{11}^{-1}. \tag{3.3}$$

The derivation is outlined in Section B of supplementary material available at *Biostatistics* online. The matrices I_{11}, I_{12}, and I_{21} in (3.3) are the blocks in the partitioned Fisher information matrix associated with the likelihood function (3.2) as given by (2.6) in the Appendix of supplementary material available at *Biostatistics* online. The expression for the asymptotic variance in (3.3) shows that the efficiency of the pseudo-MLE $(\tilde{\alpha}, \tilde{\beta})$ can be close to that of the MLE of (α, β) with a known θ when either k or $\mathrm{AV}_{\tilde{\theta}}$ is small. This indicates that the efficiency of the pseudo-MLE $(\tilde{\alpha}, \tilde{\beta})$ may exceed the efficiency of the MLE of (α, β) jointly obtained with the MLE of θ using the primary data only.

Note that the corresponding blocks of $-n^{-1}\partial^2 \log L(\alpha, \beta, \theta; \mathbf{N} \mid \mathbf{T}, \mathbf{Z})/\partial(\alpha, \beta, \theta)^2$ are consistent estimators for the matrices I_{11}, I_{12}, and I_{21} with the pseudo-MLE plugged in. They, together with a consistent estimator of $\mathrm{AV}_{\tilde{\theta}}(\theta)$, naturally form a consistent estimator of $\mathrm{AV}_{(\tilde{\alpha},\tilde{\beta})}(\alpha, \beta, \theta)$. The derivation of (3.3) and the aforementioned consistent variance estimator require the underlying model specification. In practice, a more robust variance estimator is often preferable, as the Huber sandwich variance estimator for the variance of the MLE is preferred to anticipate possible model misspecification (Huber, 1967). This consideration leads us to estimate I_{11}^{-1}, the first term in (3.3), with the corresponding Huber sandwich estimator, which results in an extended Huber sandwich estimator. The details of this alternative variance estimator are presented in Section C of supplementary material available at *Biostatistics* online.

4. Simulation study

We conducted simulation studies to examine the finite-sample properties of the MLE and pseudo-MLE in terms of efficiency and robustness to model misspecification. The numerical studies in this section and the next were carried out using the *R* package for statistical computing (http://www.r-project.org).

We simulated n independent individuals from the two latent classes: the *at-risk* and *not-at-risk* groups. We followed the analysis outcomes reported by McBride *and others* (2011) to choose the parameter values for the data generation in the simulations. Specifically, we simulated two potential risk factors: a binary variable sex as the indicator of a male subject, and a continuous variable (age) as the standardized age of an individual at the beginning of the study. These two risk factors together with the latent indicator η of the *at-risk* group and the individual observation time T were generated as follows. For the ith individual in the study, (i) $\mathrm{sex}_i \sim \mathrm{Bin}(1, \frac{1}{2})$ (the Bernoulli distribution with a success probability of $\frac{1}{2}$), (ii) $\mathrm{age}_i \sim \mathrm{Beta}(0.7, 0.8)$ (the Beta distribution with the parameter values chosen to follow the distribution of the standardized age variable in the CAYACS program), (iii) $\eta_i \sim \mathrm{Bin}(1, p_i)$, where $\mathrm{logit}(p_i) = 1 - \mathrm{sex}_i - 0.8\mathrm{age}_i$, and (iv) $T_i \sim \mathrm{Beta}(2, 1)$. The event counts N_i were then generated in the following two settings, designed to assess the efficiency and robustness of the estimators.

Efficiency Study. Conditional on $(\eta_i, T_i, \mathrm{sex}_i, \mathrm{age}_i)$, the event count N_i was generated from the Poisson distribution as follows: (i) for $\eta_i = 1$, the mean is $\Lambda_1(T_i, \mathrm{sex}_i, \mathrm{age}_i) = T_i \exp(1.8 - 0.6\mathrm{sex}_i - 0.5\mathrm{age}_i)$; (ii) for $\eta_i = 0$, the mean is $\Lambda_0(T_i, \mathrm{sex}_i, \mathrm{age}_i) = T_i \exp(0.5 - 0.3\mathrm{sex}_i - 0.25\mathrm{age}_i)$.

Robustness Study. For individual i, ξ_i was generated from the gamma distribution with mean 1 and variance γ_i: $\xi_i \sim \text{Gamma}(1, \gamma_i)$. Conditional on $(\eta_i, T_i, \text{sex}_i, \text{age}_i, \xi_i)$, N_i was generated from the Poisson distribution with mean $\xi_i \Lambda_{\eta_i}(T_i, \text{sex}_i, \text{age}_i)$, where $\Lambda_{\eta_i}(T_i, \text{sex}_i, \text{age}_i)$ was the same as in the *Efficiency Study* for $\eta_i = 1$ or 0. Note that if $\gamma_i > 0$, the variance of the simulated event count N_i conditional on $(\eta_i, T_i, \text{sex}_i, \text{age}_i)$ is $(1 + \gamma_i)\Lambda_{\eta_i}(T_i, \text{sex}_i, \text{age}_i)$. Three model misspecification scenarios were simulated: Case (i) $\gamma_i = \gamma > 0$ regardless of η_i; Case (ii) $\gamma_i = \gamma > 0$ if $\eta_i = 1$ and $\gamma_i = 0$ if $\eta_i = 0$; Case (iii) $\gamma_i = 0$ if $\eta_i = 1$ and $\gamma_i = \gamma > 0$ if $\eta_i = 0$. We chose the parameter γ to be $\frac{1}{2}$, 1, or 2 to simulate mild, medium, or severe overdispersed counts, respectively.

We formed the observed (primary) data as $\{(N_i, T_i, \text{sex}_i, \text{age}_i) : i = 1, \ldots, n\}$ in the simulations. The supplementary information was generated independently as realizations of $(N, T, \text{sex}, \text{age})$ from a group of m independent individuals with the same distribution as the *not-at-risk* group in each of the simulation settings.

Each of the experimental settings described above was repeatedly simulated 250 times. For each simulated data set, we evaluated both the MLE and the pseudo-MLE for the parameters in the latent class model, the mixture Poisson model in Section 2. We also evaluated the standard error estimators of the MLE and the pseudo-MLE based on the conventional variance estimator for the MLE and the Huber sandwich variance estimator, and the two variance estimators for the pseudo-MLE given in Section 3 and Section C in supplementary material available at *Biostatistics* online. The evaluations of $\tilde{\theta}$ used in the pseudo-MLE procedure, the estimates of the parameters θ in the frequency model for the *not-at-risk* group based on the supplementary information, were computed using the *R* function *glm*. Both the MLE and pseudo-MLE procedures were implemented by (a) maximizing the observed data likelihood and the pseudo-likelihood functions via an *R* optimization function and (b) applying the EM algorithm described in Section 3. The resulting estimates from different optimizers were close to the estimates from the EM algorithm. The estimates via the EM algorithm are discussed below.

Table 1 presents a summary of the parameter estimates and the asymptotic standard error estimates in the *Efficiency Study* with $n = 500$ and $m = 5000$ based on 250 replicates. The sample means (sm) of all the parameter estimators are close to the corresponding true values of the parameters: the relative differences range from 0% to 3.7%. This verifies the consistency of both the MLE and the pseudo-MLE. The sample standard errors (sse) of the pseudo-MLE estimators overall appear smaller than those of the MLE estimators. That is, the supplementary information along with the smaller number of parameters to be estimated may compensate for the pseudo-MLE's potential loss of efficiency, leading to better efficiency than that for the evaluable MLE with the primary data. Table 1 also presents the sm of the two standard error estimators, the conventional and sandwich estimators, for both the MLE and the pseudo-MLE. The two sets of sm of the estimated standard errors, sm_{se} (or sm_{pse}) and $\text{sm}_{\text{se.sw}}$ (or $\text{sm}_{\text{pse.sw}}$), are essentially the same. They are close to the corresponding sse of the estimators, with the absolute differences ranging from 0.2% to 3.6%. This shows that the accuracy of both the standard error estimators is satisfactory in practice.

We considered additional simulation settings in the *Efficiency Study*. For comparison, we evaluated the MLE of (α, β, θ) based on the primary data combined with the realizations of the latent indicator η. The sm and sse were close to those associated with the pseudo-MLE. To further explore the contribution of the supplementary information, we evaluated two other sets of estimators for α and β: the MLEs with θ fixed at the true value and the pseudo-MLE with θ estimated based on the supplementary information with size $m = 500$. As anticipated, the sse of the MLEs with the true θ were smaller than the sse of the MLEs with θ jointly estimated, and the sse of the pseudo-MLEs for $m = 500$ were slightly larger than those for $m = 5000$, which were close to those for the MLE with the true θ. We also evaluated the estimators with the size of the observed data set to $n = 100$. The findings were the same.

Regardless of the value of the overdispersion parameter γ, the simulation outcomes in the three cases of the *Robustness Study* show that the MLE is sensitive to model misspecification overall, but the robustness

Table 1. *Simulation outcomes: Efficiency Study*

(Primary data $n = 500$; repetition number $= 250$)											
Parameter	α_0	α_1	α_2	β_0	β_1	β_2	β_3	θ_0	θ_1	θ_2	θ_3
True value	1	−1	−0.8	1.8	−0.6	−0.5	1	0.5	−0.3	−0.25	1
	MLE of (α, β, θ)										
sm†	1.000	−1.026	−0.770	1.790	−0.599	−0.500	1.006	0.487	−0.295	−0.261	1.009
sse‡	0.231	0.291	0.397	0.081	0.053	0.070	0.056	0.192	0.090	0.142	0.129
$\text{sm}^{\dagger}_{\text{se}}$	0.243	0.302	0.420	0.084	0.058	0.069	0.060	0.197	0.094	0.141	0.133
$\text{sm}^{\dagger}_{\text{sw.se}}$	0.248	0.312	0.433	0.083	0.057	0.068	0.058	0.198	0.096	0.142	0.135
								Supplementary data $m = 5000$			
	Pseudo-MLE of (α, β)							MLE of θ			
sm	1.003	−1.011	−0.775	1.791	−0.602	−0.499	1.005	0.503	−0.301	−0.251	1.000
sse	0.220	0.266	0.350	0.080	0.048	0.066	0.055	0.030	0.013	0.020	0.021
sm_{pse}	0.231	0.252	0.382	0.083	0.051	0.066	0.059	0.029	0.014	0.022	0.021
$\text{sm}_{\text{sw.pse}}$	0.232	0.252	0.383	0.081	0.050	0.064	0.058	0.029	0.014	0.022	0.021

†The sample means of the parameter estimates (sm), the conventional standard error estimates (sm_{se}), and the sandwich standard error estimates ($\text{sm}_{\text{sw.pse}}$).
‡The sample standard errors (sse) of the parameter estimates.

of the pseudo-MLE varies. The sm for the MLE reveal some serious biases in the simulated situations, especially for the regression coefficients in the risk model. The differences of the sm for the pseudo-MLE from the true parameter values are considerably smaller. Particularly in Case (iii), which simulated situations where only the underlying frequency model for the *not-at-risk* group (i.e. the group where $\eta = 0$) was misspecified, the pseudo-MLE estimates are basically unbiased. In all three cases, the sm of the standard error estimates based on the conventional variance estimator for MLE have discrepancies compared with the sse associated with both the MLE and the pseudo-MLE estimators. The sm of the corresponding sandwich standard error estimator, on the other hand, is close to the sse. This verifies the robustness of the sandwich estimator. We summarize the simulation results of Case (iii) with $m = 5000$, $\gamma = 1$ in Table 2. The other results are presented by Table 1 in Section D of supplementary material available at *Biostatistics* online.

Another simulation study was conducted to explore the difference in robustness between the MLE and pseudo-MLE in situations similar to Case (iii). We substituted the mixed Poisson model with a mixture of two Poisson models for the group $\eta = 0$: the mean of one component was the same as the mean of the group $\eta = 0$ in the *Efficiency Study*, and the mean of the second component was close to the mean of the group $\eta = 1$. We varied the proportion of the second component in the mixture from 10% to 80%, and observed that the corresponding bias with the MLE of (α, β) changed from minor to major, while the pseudo-MLE of the parameters remained close to the true values. This further suggests the benefit of using supplementary information. See Table 2 for a summary of the simulation outcomes in Section D of supplementary material available at *Biostatistics* online.

5. Analysis of CAYACS physician visits

This section presents an analysis of the CAYACS physician-visit records using the methodology described in the previous sections.

Table 2. *Simulation outcomes: Case* (iii) *of Robustness Study*

(Primary data $n = 500$; repetition number $= 250$)											
Parameter	α_0	α_1	α_2	β_0	β_1	β_2	β_3	θ_0	θ_1	θ_2	θ_3
True value	1	−1	−0.8	1.8	−0.6	−0.5	1	0.5	−0.3	−0.25	1
	MLE of (α, β, θ)										
sm†	1.305	−0.623	−0.462	1.762	−0.606	−0.499	1.003	−0.384	−0.513	−0.418	1.140
sse‡	0.253	0.247	0.381	0.085	0.051	0.073	0.059	0.575	0.259	0.420	0.422
$\text{sm}^{\dagger}_{\text{se}}$	0.231	0.221	0.351	0.066	0.031	0.047	0.047	0.326	0.122	0.188	0.226
$\text{sm}^{\dagger}_{\text{sw.se}}$	0.244	0.250	0.394	0.083	0.051	0.072	0.060	0.526	0.257	0.407	0.370
								Supplementary data $m = 5000$			
	Pseudo-MLE of (α, β)							GEE estimate of θ			
sm	1.190	−1.119	−0.845	1.800	−0.507	−0.436	0.980	0.494	−0.297	−0.252	1.003
sse	0.237	0.225	0.358	0.080	0.044	0.066	0.056	0.061	0.035	0.048	0.041
sm_{pse}	0.243	0.242	0.382	0.079	0.042	0.059	0.056	0.067	0.032	0.050	0.047
$\text{sm}_{\text{sw.pse}}$	0.230	0.229	0.361	0.078	0.045	0.062	0.056	0.067	0.032	0.050	0.047

†The sample means of the parameter estimates (sm), the conventional standard error estimates (sm_{se}), and the sandwich standard error estimates ($\text{sm}_{\text{sw.pse}}$).
‡The sample standard errors (sse) of the parameter estimates.

5.1 *Study description and preliminary analysis*

The CAYACS program that motivated this research is primarily concerned with people in BC under the age of 20, diagnosed with cancer from 1980 to 1999, who survived five years or longer after the diagnosis. The survivor cohort has $n = 1962$ subjects; see McBride *and others* (2010) for more information. We consider one of CAYACS's objectives, to evaluate the cohort's physician-visit frequency and patterns from 1986 to 2006, to compare the cohort with the general population, and to identify risk factors for later effects. To avoid potential collinearity in the regression analysis, we chose the following six variables as covariates from the list of potential risk factors identified by the study team: sex (male vs. female), *age at study entry* (five years after the cancer diagnosis), *socioeconomic status* (SES, high vs. low based on the neighborhood income of residence at start of follow-up), *relapse or second cancer* (yes vs. no relapse or second cancer status at start of follow-up), *cancer diagnosis period* (1990s vs. 1980s), and *cancer treatment* (chemo only, radiation only, both chemo and rad, or others). A standardized age value $(\text{age} - 5)/20$ was used in the analysis. To focus on the primary interest of this paper, we excluded individuals either missing information for the six variables or with an observation period of zero length, and a few outliers. This reduced the size of the primary data to $n = 1628$. The summary statistics of the six covariates are presented in Table 3.

The CAYACS program selected from the BC population 19 620 people, 10 times the number in the survivor cohort, matching the cohort in sex and birth year. It obtained their physician claims after the age of five years from 1986 to 2006 (McBride *and others*, 2011). We removed those people who were older in 1986 than $(20 + 5) = 25$ years, the oldest possible age at that time of the survivors, and who had missing covariates. We also excluded a few outliers. This gave a set of physician-claim data from the general population of size $m = 16\,494$. The summary statistics of sex, SES, and *age at entry* for the sample of the general population are presented in Table 3. The population sample has distributions that match the survivor cohort for sex and SES but not *age at entry*.

We analyzed the physician-claim records from the cohort supplemented with the records from the sample of the general population. Visits to both GPs and specialists were considered in the analysis. The

Table 3. *Characteristics summary: survivor cohort vs. general population sample in CAYACS*

Risk factor	Sex		SES		Relapse or second cancer		Diagnosis period		Treatment			
	Male	Female	High	Low	Yes	No	1990s	1980s	Chemo Only	Rad Only	Both	Others
Survivor cohort ($n = 1628$)	912	716	667	961	172	1456	981	647	670	140	403	415
General population Sample ($m = 16\,494$)	9040	7434	6050	10 444	–	–	–	–	–	–	–	–
Explanatory variable	Age at entry (in years)						Observation length (in years)					
	The five numbers†				Mean (SD)		The five numbers†					Mean (SD)
Survivor cohort ($n = 1628$)	5.0, 8.5, 13.8, 20.3, 25.0				14.4 (6.3)		0.005, 5.05, 9.16, 13.95, 20.98					9.68 (5.42)
General population Sample ($m = 16\,494$)	5.0, 5.0, 6.7, 14.4, 25.0				10.0 (6.1)		0.003, 7.46, 13.08, 19.25, 21.00					12.73 (6.58)

†The sample minimum, first quartile, median, third quartile, and maximum values.

summary statistics of the observation length associated with the cohort and the population sample are given in Table 3. They indicate that the cancer survivors had shorter observation periods in general.

Table 3 in Section E of supplementary material available at *Biostatistics* online summarizes the quasi-Poisson regression analyses conducted with the physician records from the cohort and the population sample separately. Adjusted for the independent variables, the frequency of physician visits appears significantly higher in the cohort. In both data sets, male subjects had many fewer physician visits than female subjects had. This is in agreement with the results reported in McBride *and others* (2011). In addition, the analysis found that, in contrast with the significantly lower visit frequency associated with the high SES group in the general population, there is no significant difference between the two SES groups in the cohort. It also revealed a rather different pattern for the frequency increase trend of the survivors. The analysis indicates that the visit counts are highly overdispersed: the estimates for the overdispersion parameter are $\hat{\phi} = 23.40$ and 31.90 for the general population and the survivor cohort, respectively. The larger overdispersion for the survivor cohort, along with its higher overall visit frequency, signals potential strata of physician visits in the cohort.

5.2 *CAYACS data analysis under a latent class model*

We then used the latent class model in Section 2 to formulate the physician-claim data of the survivor cohort. We evaluated the MLE and pseudo-MLE presented in Section 3. Table 4 summarizes the two sets of analysis outcomes. Both the MLE and pseudo-MLE analyses identified several significant risk factors for later effects: (i) *relapse or second cancer*, (ii) *diagnosis in* 1980*s rather than* 1990*s*, and (iii) *treatment with radiation only or both radiation and chemo therapies rather than other treatments*. The pseudo-MLE also found a significantly higher risk rate associated with female survivors.

For illustration, we present in Figure 1 the estimated *at-risk* probability functions of age at entry together with pointwise approximate 95% confidence intervals from the MLE and pseudo-MLE for three typical groups: Group A—females diagnosed in the 1980s, with relapse/second cancer, who received radiation treatment; Group B—females diagnosed in the 1980s, without relapse/second cancer, who received radiation treatment; and Group C—males diagnosed in the 1990s, without relapse/second cancer, and with treatment other than chemo/radiation. The risk of later effects for the three groups is found by both the MLE and pseudo-MLE to be significantly different. People in Group A seem likely to suffer such effects, and those in Group C have a low risk.

The MLE and pseudo-MLE analyses are consistent with the findings of a significantly lower visit rate associated with male survivors across the two risk groups and a similar association with the length of the observation period in the *not-at-risk* group. The two sets of outcomes indicate different magnitudes of the visit frequency in the two strata, the *at-risk* and *not-at-risk* groups. See Figure 2 for the estimated means of the visit counts over time from the MLE and pseudo-MLE for the two risk strata. The MLE analysis showed that the visit frequency was not significantly associated with either *age at study entry* or SES across the two risk strata. This is not in agreement with the quasi-Poisson regression outcomes. However, the pseudo-MLE analysis, using the quasi-Poisson estimates for the general population for the *not-at-risk* group, yielded results for the visit frequency of the *at-risk* group similar to those of the MLE analysis. Figure 2 presents the estimated means of the cumulative visit counts of the two risk strata over time, along with pointwise approximate 95% confidence intervals, from the MLE and pseudo-MLE for female and male subjects with low SES and average age of entry.

To verify the findings of the pseudo-MLE and further assess its efficiency, we evaluated the MLE with the data from the survivor cohort in combination with the sample data from the general population, described at the beginning of Section 3.2. The parameter estimates from the MLE with the combined data together with their estimated standard errors are presented in Table 4 in Section E of supplementary material available at *Biostatistics* online. They appear almost identical to the corresponding estimates from

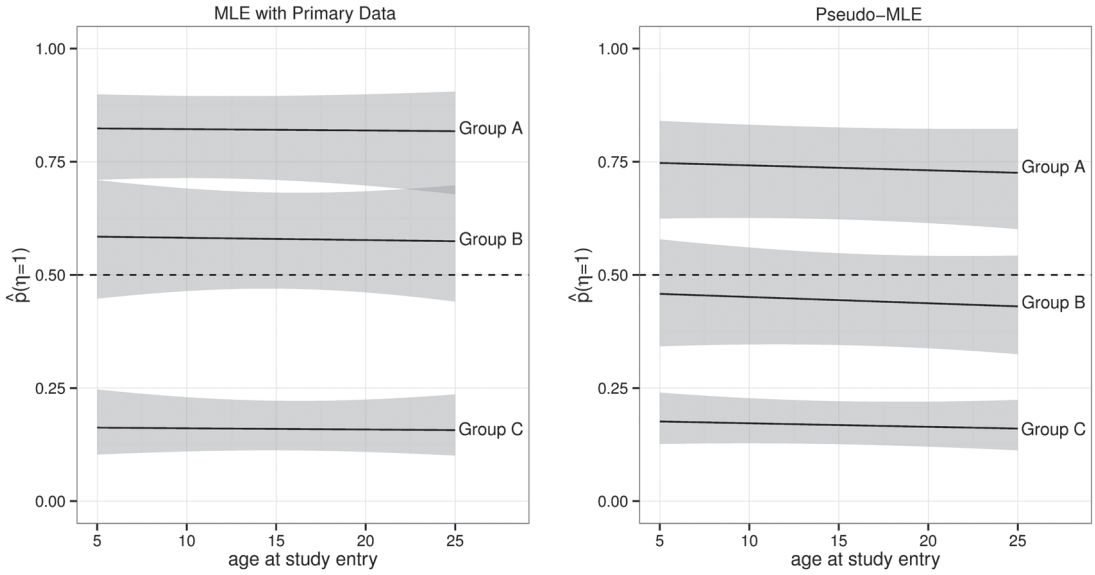

Fig. 1. Estimated risk probabilities with approximate 95% confidence intervals for three groups: Group A. Female, diagnosed in 1980s, with relapse/second cancer, and treated with radiation therapy; Group B. Female, diagnosed in 1980s, no relapse/second cancer, and treated with radiation therapy; Group C. Male, diagnosed in 1990s, no relapse/second cancer, and treated without chemo/radiation therapy.

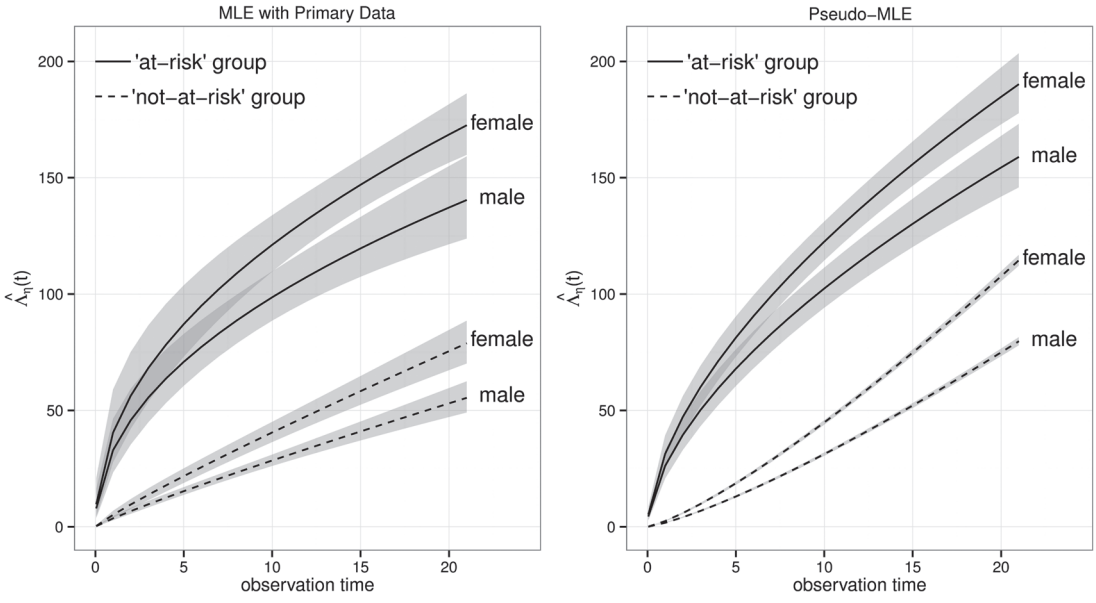

Fig. 2. Estimated mean functions of cumulative physician-visit counts with approximate 95% confidence intervals, for survivors with low SES and average age at entry.

Table 4. *Estimates of parameters and standard errors for the CAYACS data*[†]

		MLE		Pseudo-MLE	
Factor		Estimate	se.sw	Estimate	pse.sw
In the risk model					
Intercept		−0.322	(0.204)	**−0.683**	(0.188)
Male (vs. female)		−0.355	(0.207)	**−0.315**	(0.130)
Age at study entry		−0.041	(0.324)	−0.111	(0.219)
SES high (vs. low)		−0.092	(0.232)	0.028	(0.132)
Relapse/second cancer (vs. not)		**1.200**	(0.215)	**1.253**	(0.177)
Diagnosis period 1990s (vs. 1980s)		**−0.964**	(0.167)	**−0.545**	(0.129)
Treatment (vs. other)	Chemo only	0.107	(0.158)	0.145	(0.151)
	Rad only	**0.663**	(0.229)	**0.515**	(0.224)
	Both	**0.362**	(0.163)	**0.392**	(0.162)
In the frequency model for the at-risk group					
Intercept		**3.656**	(0.187)	**3.386**	(0.111)
Male (vs. female)		**−0.206**	(0.070)	**−0.180**	(0.047)
Age at study entry		0.095	(0.099)	0.125	(0.076)
SES high (vs. low)		−0.005	(0.072)	−0.023	(0.046)
ln(time length)		**0.476**	(0.066)	**0.592**	(0.041)
				GEE estimates Based	
In the frequency model for the not-at-risk group				on supp. data	
Intercept		**1.560**	(0.137)	**0.751**	(0.038)
Male (vs. female)		**−0.353**	(0.070)	**−0.362**	(0.011)
Age at study entry		0.162	(0.103)	**0.306**	(0.019)
SES high (vs. low)		0.028	(0.077)	**−0.047**	(0.011)
ln(time length)		**0.897**	(0.053)	**1.263**	(0.013)

[†]Significant effect with p-value $\leqslant 0.05$ in boldface.

the pseudo-MLE. For comparison, Figures 1 and 2 in the Appendix of supplementary material available at *Biostatistics* online display the estimated risk probabilities and average physician-visit frequencies based on the three sets of parameter estimates: the MLE with only the cohort data, the MLE with the combined data, and the pseudo-MLE.

We remark that, under the mixture Poisson model assumed in the MLE and pseudo-MLE analyses, the variance of the counts conditional on T, Z is the mean $E(N \mid T, Z)$ plus $p(Z;\alpha)[1 - p(Z;\alpha)][\Lambda_1(T, Z;\beta) - \Lambda_0(T, Z;\theta)]^2$. This together with the parameter estimates under the latent class model yields estimates for the overall overdispersion parameter for the survivor cohort of 18.11 (for MLE) and 20.80 (for pseudo-MLE). Compared with the quasi-Poisson analysis for the cohort, about two-thirds of the large overdispersion of the visit counts can be attributed to the two risk strata by the mixture Poisson model. The unexplained part of the overdispersion indicates a departure of the counts from this model. In addition, Table 3 shows that the distributions of *age at entry* and *length of observation* for the sample from the general population are rather different from those for the survivor cohort. Caution is necessary in the application of these results.

6. FINAL REMARKS

Motivated by the physician-visit project of the CAYACS program, we have proposed a latent class model to formulate event counts from a cohort with two unobservable strata. In the young cancer survivor cohort,

these two classes are the *at-risk* group who suffer long-term effects of their cancer diagnosis and visit physicians more frequently and the *not-at-risk* group who have the same physician-visit frequencies as the general population. Under a mixture Poisson model, we have presented two likelihood-based inference procedures, the MLE and pseudo-MLE. The pseudo-MLE procedure employs a consistent estimator of the distribution of the *not-at-risk* group based on the general population data. Compared with the MLE with the primary data, it requires less computational effort, has consistency and asymptotic normality, and has potentially higher efficiency. As observed by a referee, one may apply the proposed methodology with little modification in situations involving more than two strata.

The simulation results show that the likelihood-based estimating procedures are quite efficient under the mixture Poisson model, but they have a lack of robustness to model misspecification. Therefore, there is a need for an inference procedure that is robust to model misspecification. This has led to an ongoing project to develop an extension of the generalized estimating equations approaches (Liang and Zeger, 1986). The new approach can be straightforwardly extended to analyze the cost data associated with physician claims in CAYACS.

Several other investigations would also be worthwhile. The model formulation assumes that the time effect for the count of interest is proportional to the length of the observation period on average. The available longitudinal data allow us to consider a semiparametric specification for the mean function of the counts over time, and thus to check this assumption. Another possibility is to extend the proposed modeling and inferential procedures to investigate the potential correlation of study individuals, similarly to, for example, the approach in Lee *and others* (2006). A third suggestion is to introduce a time-dependent risk indicator to accommodate evolving cohorts.

Supplementary material

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The BC Cancer Agency, BC Children's Hospital, and BC Ministry of Health approved access to and use of the data in the MSP Registry File and MSP Claims Data (physician-ordered services); this was facilitated by Population Data BC. We acknowledge the valuable assistance of the CAYACS team, and in particular Maria Lorenzi. We thank a referee, the Associate Editor, and the Editor for their instructive comments and suggestions. *Conflict of Interest*: None declared.

Funding

The statistical research was partially supported by the Canadian Cancer Society (CCS) Research Institute and CCS BCY Program Project Grants [grant numbers 19000, 19804], and the Natural Sciences and Engineering Research Council of Canada (NSERC) [grant number 177430].

References

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)* **39**, 1–38.

Gong, G. and Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics* **8**, 861–869.

GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61**, 215–231.

HALL, D. B. AND SHEN, J. (2010). Robust estimation for zero-inflated Poisson regression. *Scandinavian Journal of Statistics* **37**, 237–252.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume I, pp. 221–233.

LAMBERT, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14.

LAZARSFELD, P. F. AND HENRY, N. W. (1968). *Latent Structure Analysis*. Mifflin: Houghton.

LEE, A. H., WANG, K., SCOTT, J. A., YAU, K. K. W. AND MCLACHLAN, G. J. (2006). Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research* **15**, 47–61.

LIANG, K. Y. AND ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

MAGIDSON, J. AND VERMUNT, J. (2002). Latent class models for clustering: a comparison with K-means. *Canadian Journal of Marketing Research* **20**, 36–43.

MCBRIDE, M. L., LORENZI, M. F., PAGE, J., BROEMELING, A. M., SPINELLI, J. J., GODDARD, K., PRITCHARD, S., ROGERS, P. AND SHEPS, S. (2011). Patterns of physician follow-up among young cancer survivors. *Canadian Family Physician* **57**, e482–e490.

MCBRIDE, M. L., ROGERS, P. C., SHEPS, S. B., GLICKMAN, V., BROEMELING, A. M., GODDARD, K., HU, J., LORENZI, M., PEACOCK, S., PRITCHARD, S. *and others*. (2010). Childhood, adolescent, and young adult cancer survivors research program of British Columbia: objectives, study design, and cohort characteristics. *Pediatric Blood & Cancer* **55**, 324–330.

PEPE, M. S. AND JANES, H. (2007). Insights into latent class analysis of diagnostic test performance. *Biostatistics* **8**, 474–484.

VERMUNT, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research* **17**, 33–51.

[*Received July 16, 2013; revised October 18, 2013; accepted for publication October 24, 2013*]