American Journal of Epidemiology
© The Author(s) 2017. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Vol. 186, No. 9
DOI: 10.1093/aje/kwx161
Advance Access publication:
June 8, 2017

Practice of Epidemiology

A Multinomial Regression Approach to Model Outcome Heterogeneity

BaoLuo Sun, Tyler VanderWeele, and Eric J. Tchetgen Tchetgen*

* Correspondence to Dr. Eric J. Tchetgen Tchetgen, Departments of Biostatistics and Epidemiology, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Kresge Room 822, Boston, Massachusetts 02115 (e-mail: etchetge@hsph.harvard.edu).

Initially submitted January 28, 2016; accepted for publication December 5, 2016.

When a risk factor affects certain categories of a multinomial outcome but not others, outcome heterogeneity is said to be present. A standard epidemiologic approach for modeling risk factors of a categorical outcome typically entails fitting a polytomous logistic regression via maximum likelihood estimation. In this paper, we show that standard polytomous regression is ill equipped to detect outcome heterogeneity and will generally understate the degree to which such heterogeneity may be present. Specifically, nonsaturated polytomous regression will often a priori rule out the possibility of outcome heterogeneity from its parameter space. As a remedy, we propose to model each category of the outcome as a separate binary regression. For full efficiency, we propose to estimate the collection of regression parameters jointly using a constrained Bayesian approach that ensures that one remains within the multinomial model. The approach is straightforward to implement in standard software for Bayesian estimation.

constrained Bayes; etiologic heterogeneity; multinomial outcome; outcome heterogeneity

Abbreviations: CB, constrained Bayesian; CHD, coronary heart disease; SL, separate logistic.

Categorical outcomes are of common occurrence in epidemiologic practice. A standard modeling approach to evaluate risk factors in such settings involves fitting by maximum likelihood, a polytomous logistic regression for the multinomial outcome (1). In empirical studies, an important form of outcome heterogeneity arises when a given risk factor affects certain categories of the outcome but not necessarily others. This form of outcome heterogeneity, also sometimes called etiologic heterogeneity (2), has in recent years drawn considerable interest in medicine and other health sciences (3–5). In this paper, we establish that standard polytomous logistic regression is often ill suited to model this type of outcome heterogeneity, in the sense that the approach may understate the degree to which such heterogeneity may be present. Specifically, standard polytomous regression will often a priori rule out the possibility of outcome heterogeneity from its parameter space, because under the model a risk factor for a given category of the outcome must necessarily be a risk factor for all other categories of the outcome. In the following sections, we demonstrate how this phenomenon is manifested with a certain paradox that arises in the context of using polytomous logistic regression in the presence of outcome heterogeneity and propose an alternative general multinomial regression approach with constrained Bayesian (CB) estimation of the regression parameters. We investigate the finite-sample properties of the proposed estimators in a simulation study and illustrate the new methodology in an application that evaluates risk factors for death from coronary heart disease (CHD), stroke, and cancer in the original cohort of the Framingham Heart Study.

METHODS

A paradox from using polytomous logistic regression

To describe the paradox, suppose that the outcome Y takes 1 of 3 possible values $k = 0,1,2$, where $Y = 0$ denotes disease-free persons, $Y = 1$ denotes individuals with the given disease of the first subtype, and $Y = 2$ denotes diseased persons with the second subtype. Also suppose that 2 continuous risk factors (X_1 and X_2) are known to be associated with diseased individuals, that is,

$$\Pr\{Y \neq 0 \,|\, x_1, x_2\} = \pi_0(x_1, x_2) \qquad (1)$$

varies both in (x_1, x_2), where x_j denotes a possible value of X_j. A standard approach to model such data entails positing a polytomous logistic regression, such as say

$$\Pr\{Y = k \mid X\} = \frac{\exp\{\beta_{k0} + \beta_{k1}X_1 + \beta_{k2}X_2\}}{1 + \exp\{\beta_{10} + \beta_{11}X_1 + \beta_{12}X_2\} + \exp\{\beta_{20} + \beta_{21}X_1 + \beta_{22}X_2\}}, \quad k = 1, 2 \qquad (2)$$

where $X = (1, X_1, X_2)$, and

$$\Pr\{Y = 0 \mid X\} = 1 - \Pr\{Y = 1 \mid X\} - \Pr\{Y = 2 \mid X\}. \qquad (3)$$

Now, suppose also that, reflecting the presence of outcome heterogeneity, the first risk factor X_1 only affects the first disease subtype and X_2 only affects the second disease subtype, that is,

$$\Pr\{Y = 1 \mid x_1, x_2\} = \pi_1(x_1) \qquad (4)$$

for all x_2 and each x_1 and

$$\Pr\{Y = 2 \mid x_1, x_2\} = \pi_2(x_2) \qquad (5)$$

for all x_2 and each x_1. Then, for equation 4 to hold under the polytomous regression model, it must be that

$$\beta_{12} = \beta_{22} = 0, \qquad (6)$$

so that the right-hand side of equation 2 does not depend on X_2. Likewise, for equation 5 to hold under the polytomous regression model, it must be that

$$\beta_{11} = \beta_{21} = 0, \qquad (7)$$

so that the right-hand side of equation 2 does not depend on X_2. However, both equations 6 and 7 would imply that

$$\Pr\{Y \neq 0 \mid X\} = \frac{\exp\{\beta_{10}\} + \exp\{\beta_{20}\}}{1 + \exp\{\beta_{10}\} + \exp\{\beta_{20}\}}$$

depends on neither X_1 nor X_2, which contradicts the fact that X is a risk factor for Y as given by equation 1, giving rise to the paradox. The above paradox stems from the fact that a standard polytomous logistic regression of the form given in expression 2 cannot simultaneously encode assumption 2 and assumptions 4 and 5. This is because such models do not have a specific parameter or set of parameters that can be set to a value that solely implies either assumption 4 or 5 without also implying that Y is altogether independent of either X_2 or X_1, respectively. Note that incorporating interactions and nonlinearities in X_1 and X_2 would in principle make the regression model somewhat more flexible; however, this would not necessarily resolve the above paradox unless a genuine nonparametric model were used in place of a parametric model. Even under a nonparametric polytomous regression framework, it is unclear whether one could easily encode assumption 4. Note also that this form of paradox will become even more ubiquitous when multiple risk factors are being considered, in which case nonparametric regression may no longer be practical. We may conclude that polytomous logistic regression is generally ill suited to either detect or model outcome heterogeneity of the type described above. In the next section, we describe a simple alternative approach that circumvents this difficulty.

A general multinomial regression approach to model outcome heterogeneity

The proposed approach involves modeling each category of the outcome (other than a reference level) with a separate binary regression model. To fix ideas, let us reconsider the example from the previous section. Suppose that instead of equation 2, one posits the following pair of logistic regressions:

$$\operatorname{logit} \Pr\{Y = 1 \mid x_1, x_2\} = \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 \qquad (8)$$

$$\operatorname{logit} \Pr\{Y = 2 \mid x_1, x_2\} = \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2. \qquad (9)$$

As before, $\Pr\{Y = 0 \mid x_1, x_2\}$ is given by equation 3. The intercept β_{k0} may be interpreted as the log-odds that $Y = k$, $k = 1{,}2$ when $x_1 = x_2 = 0$. The regression coefficient β_{kj} corresponds to a difference in the log-odds of the event $I\{Y = k\}$ versus its complement $I\{Y \neq k\}$ per unit increment in X_j conditional on the value of the other covariate; that is, β_{kj} captures the association between X_j and the risk of disease subtype k. The degree of outcome heterogeneity as it relates to X_1 is therefore measured by the difference in the regression coefficients β_{11} and β_{21}, which are the associations of X_1 with disease subtype 1 and 2, respectively. Likewise, the degree of outcome heterogeneity as it relates to X_2 can be captured by comparing β_{12} and β_{22}. Notably, the hypothesis corresponding to equations 4 and 5 is readily encoded without imposing further restriction by setting $\beta_{12} = \beta_{21} = 0$.

For inference, one could in principle estimate $\beta_k = (\beta_{k0}, \beta_{k1}, \beta_{k2})'$ by separately maximizing the likelihood function for the corresponding logistic regression in equations 8 and 9, with binary outcome $I\{Y = k\}$. However, such a strategy has 2 potential limitations that make it unattractive. First, the approach is potentially inefficient because it does not respect the multinomial nature of Y and therefore does not make use of all available information in estimating β_k separately. A second concern is that although the logit link function in equations 8 and 9 guarantees that the resulting estimate of the predicted probability $\Pr\{Y = k \mid X_1, X_2\}$ for each person in the sample falls within the unit interval (0, 1), it does not ensure that the resulting estimate of $\Pr\{Y = 0 \mid X_1, X_2\}$ given by equation 3 also falls within the unit interval.

In order to resolve these limitations, we propose that the collection of regression parameters be jointly estimated using the following CB approach, which ensures model coherence and maximizes efficiency. The approach basically entails specifying a prior distribution $\pi(\beta)$ for the vector of unknown parameters $\beta = (\beta'_1, \beta'_2)'$, which, combined with the observed data likelihood, gives rise to a posterior distribution proportional to

$$\pi(\beta) \prod_i f(Y_i \mid X_i; \beta) I\{\Pr\{Y = 1 \mid X_i; \beta_1\} + \Pr\{Y = 2 \mid X_i; \beta_2\} < 1\}, \qquad (10)$$

where $f(k \mid X_i; \beta) = \Pr\{Y = k \mid X_i; \beta\}$ and the indicator function ensures that posterior samples are restricted to values of β for which the multinomial model is well defined, that is, $0 < \Pr\{Y = 0 \mid X_i; \beta\} = 1 - \Pr\{Y = 1 \mid X_i; \beta_1\} - \Pr\{Y = 2 \mid X_i; \beta_2\} < 1$. Noninformative prior can be used for β in the case of lack of knowledge on the parameter values, for example, uniform prior or zero-mean normal prior with large variance. Conditional on the observed data, sampling from the posterior distribution yields the posterior mode (or mean) that provides an efficient estimate of β, and 95% credible intervals can likewise be obtained from the resulting posterior sample. Adaptive Gibbs sampling (6) may be implemented through BRugs, the R interface (R Foundation for Statistical Computing, Vienna, Austria) to the OpenBUGS MCMC software (7). Sample OpenBUGS code for posterior estimation in the simulation study is included in Web Appendix 1 (available at https://academic.oup.com/aje). One may then assess convergence by visually inspecting the trace plots, as well as through the Gelman-Rubin convergence statistic (8).

The approach is easily extended to handle a multinomial outcome $K > 3$ levels. As before, we simply define $K - 1$ logistic regression models as

$$\text{logit}\,\Pr\{Y = k \mid x\} = \beta'_k \boldsymbol{X},\ \ k = 1,\ \ldots, K - 1,$$

where $\boldsymbol{X}$ is a vector of J risk factors, with first component set to 1 for the intercept. The density $\pi(\beta)$ is again a diffuse prior for $\beta = \{\beta_{jk}: j = 1, \ldots, J;\ k = 1, \ldots, K - 1\}$. The posterior distribution for the general case is proportional to

$$\pi(\beta)\prod_i f(Y_i \mid \boldsymbol{X}_i;\ \beta) I\left\{\sum_{k>0} \Pr\{Y = k \mid \boldsymbol{X}_i;\ \beta_k\} < 1\right\},$$

where the indicator function constrains the posterior sampling space so that $0 < \Pr\{Y = 0 \mid \boldsymbol{X}_i;\ \beta\} < 1$. Implementation of the approach is as described above. We note that one could in principle attempt to maximize the log-likelihood

$$\sum_i \log f(Y_i \mid \boldsymbol{X}_i;\ \beta),\ \text{ subject to the constraints}$$
$$\sum_{k>0} \Pr\{Y = k \mid \boldsymbol{X}_i;\ \beta_k\} < 1\ \forall\ i.$$

However, this is potentially computationally prohibitive because there may be as many nonlinear constraints as sample size.

In the special case in which all outcomes are rare compared with the baseline level $Y = 0$ for all values of (x_1, x_2), the paradox may not be as relevant because expression 2 can then be approximated by

$$\Pr\{Y = k \mid \boldsymbol{X}\} \approx \exp\{\beta_{k0} + \beta_{k1}X_1 + \beta_{k2}X_2\},\ \ k = 1, 2, \tag{11}$$

and the coefficients β_{kj} in 10 are approximately equal to the log risk ratios per unit change in the corresponding covariate x_j, $j = 1, 2$. When the disease outcomes are not necessarily rare across levels of (x_1, x_2), the coefficients in 2 and 10 have different interpretations, with the former relating to the log relative risk ratios, for example,

$$\beta_{k1} = \log\left\{\frac{\Pr(Y = k \mid X_2, X_1 = x_1 + 1)/\Pr(Y = k \mid X_2, X_1 = x_1)}{\Pr(Y = 0 \mid X_2, X_1 = x_1 + 1)/\Pr(Y = 0 \mid X_2, X_1 = x_1)}\right\},$$
$$k = 1, 2,$$

and the latter relating to the familiar log odds ratio interpretation

$$\beta_{k1} = \log\left\{\frac{\begin{array}{c}\Pr(Y = k \mid X_2, X_1 = x_1 + 1)/\\ [1 - \Pr(Y = k \mid X_2, X_1 = x_1 + 1)]\end{array}}{\begin{array}{c}\Pr(Y = k \mid X_2, X_1 = x_1)/\\ [1 - \Pr(Y = k \mid X_2, X_1 = x_1)]\end{array}}\right\},$$
$$k = 1, 2.$$

Relative risk ratios may potentially be difficult to interpret in practice because they are dependent on the choice of referent category. In addition, significant associations based on the polytomous logit model does not necessarily mean that the corresponding risk factor is associated with the risk of a given outcome type, which can be misleading in the presence of outcome heterogeneity.

SIMULATION

In this section, we report a simulation study to evaluate the finite-sample properties of the proposed CB estimator compared with the polytomous and separate logistic estimators. Full data consists of n independent and identically distributed (Y_i, X_{1i}, X_{2i}), $i = 1, \ldots, n$, where Y denotes the categorical outcome and (X_1, X_2) the 2 risk factors. The vector $(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ is generated from a bivariate standard normal distribution with correlation coefficient $\rho = -0.3$ and $X_1 = \Phi(\boldsymbol{Z}_1)$, $X_2 = \Phi(\boldsymbol{Z}_2)$, where $\Phi(\cdot)$ is the CDF of the standard normal distribution. The categorical outcome is generated as

$$\Pr(Y = 1) = \{1 + \exp[-(\beta_{10} + \beta_{11}X_1 + \beta_{12}X_2)]\}^{-1}$$

$$\Pr(Y = 2) = \{1 + \exp[-(\beta_{20} + \beta_{21}X_1 + \beta_{22}X_2)]\}^{-1}$$

$$\Pr(Y = 3) = \{1 + \exp[-(\beta_{30} + \beta_{31}X_1 + \beta_{32}X_2)]\}^{-1}$$

$$\Pr(Y = 0) = 1 - \Pr(Y = 1) - \Pr(Y = 2) - \Pr(Y = 3),$$

with true parameter values $(\beta_{10}, \beta_{11}, \beta_{12}) = (-1.1, 0.3, 0.0)$, $(\beta_{20}, \beta_{21}, \beta_{22}) = (-0.9, 0.0, -0.4)$ and $(\beta_{30}, \beta_{31}, \beta_{32}) = (-1.1, 0.3, -0.3)$ for $n = 200, 500$ with 1,000 simulation replicates at each sample size. Table 1 shows the results of polytomous logistic regression based on the model

$$\Pr\{Y = k \mid \boldsymbol{X}\} = \frac{\exp\{\alpha_{k0} + \alpha_{k1}X_1 + \alpha_{k2}X_2\}}{1 + \sum_{j=1}^{3}\exp\{\alpha_{j0} + \alpha_{j1}X_1 + \alpha_{j2}X_2\}}$$
$$k = 1, 2, 3, \tag{12}$$

where $\boldsymbol{X} = (1, X_1, X_2)$ and $Y = 0$ is the referent level. Separate logistic (SL) regression estimates are based on the model

$$\text{logit}\,\Pr\{Y = k \mid \boldsymbol{X}\} = \gamma_{k0} + \gamma_{k1}X_1 + \gamma_{k2}X_2,\ \ k = 1, 2, 3, \tag{13}$$

whereas the CB estimates are the Monte Carlo mean values of the posterior distribution, which is proportional to

Table 1. Simulation Results Based on Polytomous Logistic Regression

No. and Variable	Y = 1 Versus Y = 0				Y = 2 Versus Y = 0				Y = 3 Versus Y = 0			
	log(OR)	RMSE	MCSE	%COV	log(OR)	RMSE	MCSE	%COV	log(OR)	RMSE	MCSE	%COV
200												
X_1	0.80	0.76	0.78	81.7	0.57	0.78	0.82	87.5	0.81	0.78	0.80	82.8
X_2	−0.60	0.79	0.71	87.8	−0.93	0.55	0.48	80.3	−0.80	0.66	0.57	84.9
500												
X_1	0.78	0.47	0.46	64.5	0.57	0.48	0.47	79.0	0.76	0.48	0.49	62.6
X_2	−0.61	0.47	0.48	73.7	−0.92	0.48	0.51	51.4	−0.84	0.49	0.49	60.4

Abbreviations: log(OR), mean estimated log odds ratio; MCSE, Monte Carlo standard error; %COV, empirical coverage of log(OR) = 0 by 95% confidence interval; RMSE, square root of mean estimated variance; *Y*, categorical outcome variable in simulation study.

$$\pi(\eta)\prod_{i=1}^{n}\left\{\prod_{k=0}^{3}\{\Pr\{Y_i=k\mid \boldsymbol{X}_i;\eta_k\}\}^{I(Y_i=k)} \times I\left(\sum_{j=1}^{3}\Pr\{Y_i=j\mid \boldsymbol{X}_i;\eta_j\}<1\right)\right\}, \qquad (14)$$

where

$$\operatorname{logit}\Pr\{Y_i=k\mid \boldsymbol{X}_i;\eta_k\}=\eta_{k0}+\eta_{k1}X_1+\eta_{k2}X_2,\; k=1,2,3.$$

The OpenBUGS code for fitting model 14 can be found in Web Appendix 1. The trace plots and posterior densities of η in a typical simulation replicate are shown in Web Figures 1 and 2, respectively. The results for SL and CB analyses are included in Table 2.

The simulation results for polytomous logistic regression show that the mean estimated log odds ratios for covariates X_1 and X_2 differ significantly from zero across each of the 3 comparison groups, as indicated by the empirical under-coverage of 95% confidence intervals for the parameter value log(OR) = 0,

Table 2. Simulation Results Based on Separate Logistic and Constrained Bayesian Regressions

No., Method, and Variable	Y = 1 Versus Y ≠ 1			Y = 2 Versus Y ≠ 2			Y = 3 Versus Y ≠ 3		
	log(OR)	RMSE	MCSE	log(OR)	RMSE	MCSE	log(OR)	RMSE	MCSE
200									
Separate logistic									
Intercept	−1.13	0.50	0.50	−0.89	0.51	0.52	−1.16	0.51	0.51
X_1	0.32	0.58	0.59	−0.01	0.60	0.63	0.32	0.61	0.61
X_2	0.00	0.58	0.58	−0.44	0.60	0.60	−0.26	0.61	0.60
Constrained Bayesian									
Intercept	−1.09	0.45	0.48	−0.96	0.46	0.49	−1.18	0.47	0.49
X_1	0.20	0.54	0.57	0.01	0.54	0.59	0.30	0.56	0.57
X_2	0.01	0.52	0.56	−0.36	0.54	0.58	−0.28	0.55	0.58
500									
Separate logistic									
Intercept	−1.11	0.31	0.30	−0.91	0.32	0.32	−1.10	0.32	0.32
X_1	0.32	0.36	0.35	0.01	0.38	0.37	0.28	0.38	0.38
X_2	0.00	0.36	0.36	−0.41	0.38	0.39	−0.30	0.38	0.38
Constrained Bayesian									
Intercept	−1.10	0.29	0.30	−0.93	0.29	0.32	−1.12	0.30	0.32
X_1	0.27	0.34	0.35	0.01	0.34	0.36	0.29	0.35	0.37
X_2	0.01	0.34	0.36	−0.38	0.34	0.38	−0.31	0.35	0.38

Abbreviations: log(OR), mean estimated log odds ratio; MCSE, Monte Carlo standard error; RMSE, square root of mean estimated variance; *Y*, categorical outcome variable in simulation study.

Table 3. Estimated Odds Ratio[a] of Mortality From Various Causes by Risk Factors Based on Polytomous Logistic Regression

Variable	CHD		Stroke		Cancer	
	OR	95% CI	OR	95% CI	OR	95% CI
Age	1.113	1.093, 1.133[b]	1.182	1.152, 1.212[b]	1.105	1.090, 1.119[b]
Female sex	0.191	0.141, 0.258[b]	0.574	0.406, 0.813[b]	0.521	0.428, 0.633[b]
Serum cholesterol	1.006	1.003, 1.009[b]	0.998	0.994, 1.002	0.998	0.995, 1.000
Body mass index	1.032	0.995, 1.070	0.999	0.956, 1.043	0.994	0.968, 1.020
High blood pressure	2.257	1.697, 3.003[b]	2.213	1.552, 3.158[b]	1.258	1.019, 1.553[b]

Abbreviations: CHD, coronary heart disease; CI, confidence interval; OR, odds ratio.

[a] Odds ratio of mortality from specific cause versus survival by the end of the follow-up period.

[b] Denotes significance with $P < 0.05$.

which is more severe as sample size increases. Based on model 12, this implies that (X_1, X_2) are risk factors for each of the 3 levels of outcome in Y, and therefore appears to contradict the outcome heterogeneity of risk factors (X_1, X_2) with Y under the true model. In order to be a coherent model for outcome heterogeneity in the present data-generating mechanism, polytomous logistic regression must depend on neither X_1 nor X_2 ($\log(\text{OR}) = 0$), as argued in the paradox previously described. Model 12 is therefore a misspecified model for the full data law incorporating outcome heterogeneity, and the odds ratio estimates suggest that it is unable to differentiate between risk factors that influence a particular outcome category and those risk factors that do not. Estimates of variance tend to be conservative compared with the empirical variance in finite samples. Note that estimates and corresponding estimated standard errors for polytomous logistic regression summarized in Table 1 are not directly comparable to those from the separate logistic or CB approach in Table 2 because they have different interpretations.

The log odds ratio estimates for SL and CB regressions are consistent for the true log odds ratios, with vanishing biases as sample size increases. The CB regression estimator appears to be slightly less biased than the SL estimator in finite samples. The variance estimator based on the posterior distribution in the CB approach performs well, whereas those for SL tend to be conservative in finite samples. In addition, the asymptotic relative efficiency comparing SL and CB estimates (i.e., $\mathrm{Var}(\widehat{log\,(OR)}_{CB})/\mathrm{Var}(\widehat{log\,(OR)}_{SL})$) varies between 0.81–0.88. This is in agreement with theory, because the CB estimation incorporates all available information from the data by simultaneously estimating all parameters.

Table 4. Estimated Odds Ratio[a] of Mortality From Various Causes by Risk Factors Based on Separate Logistic and Constrained Bayesian Regressions

Method and Variable	CHD		Stroke		Cancer	
	OR	95% Confidence Interval	OR	95% Confidence Interval	OR	95% Confidence Interval
Separate logistic						
Age	1.078	1.059, 1.096[b]	1.145	1.118, 1.173[b]	1.082	1.069, 1.096[b]
Female sex	0.230	0.171, 0.310[b]	0.820	0.584, 1.151	0.632	0.522, 0.765[b]
Serum cholesterol	1.007	1.004, 1.010[b]	0.998	0.994, 1.002	0.998	0.995, 1.000
Body mass index	1.035	0.998, 1.072	0.999	0.957, 1.043	0.994	0.969, 1.020
High blood pressure	2.032	1.534, 2.690[b]	1.920	1.351, 2.728[b]	1.079	0.878, 1.328
	CHD		Stroke		Cancer	
	OR	95% Credible Interval	OR	95% Credible Interval	OR	95% Credible Interval
Constrained Bayesian						
Age	1.065	1.048, 1.082[c]	1.128	1.107, 1.149[c]	1.079	1.068, 1.092[c]
Female sex	0.255	0.191, 0.338[c]	0.924	0.665, 1.285	0.661	0.547, 0.798[c]
Serum cholesterol	1.006	1.003, 1.009[c]	0.996	0.992, 0.999[c]	0.997	0.995, 0.999[c]
Body mass index	1.033	1.003, 1.061[c]	0.993	0.955, 1.031	0.998	0.978, 1.026
High blood pressure	1.884	1.431, 2.477[c]	1.782	1.265, 2.522[c]	1.051	0.857, 1.292

Abbreviations: CHD, coronary heart disease; OR, odds ratio.

[a] Odds ratio of mortality from specific cause versus mortality from other causes or survival by the end of the follow-up period.

[b] Denotes significance with $P < 0.05$

[c] Denotes exclusion of 1 from 95% credible interval.

Even though 13 is the correct model, the SL estimate $\widehat{\Pr}(Y = 0 | X_1, X_2)$ is negative for at least 1 sample in 11.2% and 0.15% of the simulation replicates when $n = 200$ and $n = 500$, respectively. Therefore, fitted probabilities for the reference outcome sometimes do not lie in the unit interval, which occurs despite the absence of model misspecification. Estimation with the proposed CB approach ensures that estimates $\widehat{\Pr}(Y = 0 | X_1, X_2)$ all lie within the unit interval.

EMPIRICAL ILLUSTRATION

The empirical application concerns a cohort study of community health in Framingham, Massachusetts (9). Categories of the multinomial outcome Y are different causes of death in the present analysis, with 261 (6.2%) subjects who died from CHD, 164 (3.9%) subjects who died from stroke, 539 (12.9%) subjects who died from cancer, and 3,218 (76.9%) subjects who survived by the last examination taken in the years between 1979 and 1982. Our goal was to investigate the associations of the separate causes of death with different risk factors, including sex (female coded as 1), age in years, body mass index, serum cholesterol (mg/100 mL), and high blood pressure (systolic blood pressure ≥140 mm Hg or diastolic blood pressure ≥90 mm Hg), measured at baseline during the first examination in the years 1948–1953. There were 4,060 (97%) subjects with complete information on the outcome and risk factors, and 122 (3%) subjects with missing values were excluded from the analysis. The results of polytomous logistic regression are shown in Table 3, whereas the results from separate logistic and CB logistic regressions of Y on the risk factors are shown in Table 4.

The results from polytomous logistic regression suggested that increasing values in age and serum cholesterol, as well as male sex and high blood pressure, were significantly associated with greater risks of death from at least 1 of the 3 causes (CHD, stroke, or cancer) relative to survival rates by the end of the follow-up period. Only body mass index was not significantly associated with the risks of death from any cause. Based on a main effects polytomous logistic model, the results suggested that age, serum cholesterol, sex, and high blood pressure were significant risk factors for all causes of death.

Estimation using the separate logistic method suggested that increasing values in age were significantly associated with greater risks of death from CHD, stroke, and cancer. On the other hand, the risk factors sex and high blood pressure showed more heterogeneity. Being female was significantly associated with a lower risk of death from CHD and cancer but not stroke. High blood pressure was a significant risk factor for greater risk of death from CHD and stroke but not from cancer. Eighteen persons had negative estimated probabilities of surviving through the follow-up period under the separate logistic method. Results from the CB and separate logistic methods for age, sex, and high blood pressure were similar. The estimated asymptotic relative efficiency of the CB estimator compared with the separate logistic estimator varied between 0.61–0.98. More efficient estimation from the CB method identified serum cholesterol as a statistically significant risk factor for greater risk of death from CHD, but it was paradoxically significantly associated with lower risks of death from stroke and cancer. These apparent "protective" associations could essentially have been due to competing risk from death by CHD. Higher body mass index was found to be significantly associated with death from CHD but not from stroke or cancer.

Using CB estimation, it appeared that high blood pressure was associated with increased mortality from CHD and stroke but not cancer, whereas outcome heterogeneity was entirely understated by polytmous logistic regression. Likewise, using CB estimation, we found that being a female was associated with lower mortality from CHD and cancer but not stroke, another level of outcome heterogeneity that was undetected by polytomous logistic regression. The OpenBUGS code for unconstrained Bayesian estimation based on the Framingham data can be found in Web Appendix 2. We see then that the problem described in this paper with polytomous logistic regression is not simply theoretical; it can and does arise in practice. Continued use of this standard approach might perpetuate lack of detection of scientifically relevant outcome heterogeneity in epidemiologic practice.

DISCUSSION

Polytomous regression is the standard approach in the analysis of data from clinical or observational studies with polytomous outcome. However, a peculiar feature of this model is that its parameterization cannot encode or detect simple outcome heterogeneity, whereby certain risk factors contribute exclusively to the occurrence of some outcomes but not others. We propose an alternative approach to polytomous logistic regression that involves modeling each category of the outcome (other than a reference level) with a separate binary regression model. By doing so, our multinomial regression readily encodes a broad range of outcome heterogeneity of practical interest. In order to ensure coherent inferences and maximize efficiency, the collection of regression parameters are jointly estimated, which is straightforward to implement in standard software for Bayesian estimation. The CB approach should form a part of the standard statistical methods for assessing outcome heterogeneity.

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (BaoLuo Sun, Tyler VanderWeele, Eric J. Tchetgen Tchetgen); and Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (Tyler VanderWeele, Eric J. Tchetgen Tchetgen).

This work was funded by National Institutes of Health grants ES017876 and AI104459.

Conflict of interest: none declared.

REFERENCES

1. Agresti A. *Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons, Inc.; 2002.

2. Begg CB, Zabor EC. Detecting and exploiting etiologic heterogeneity in epidemiologic studies. *Am J Epidemiol*. 2012; 176(6):512–518.

3. Troester MA, Swift-Scanlan T. Challenges in studying the etiology of breast cancer subtypes. *Breast Cancer Res*. 2009;11(3):104.

4. Begg CB, Seshan VE, Zabor EC, et al. Genomic investigation of etiologic heterogeneity: methodologic challenges. *BMC Med Res Methodol*. 2014;14:138.

5. Wang M, Kuchiba A, Ogino S. A meta-regression method for studying etiological heterogeneity across disease subtypes classified by multiple biomarkers. *Am J Epidemiol*. 2015;182(3):263–270.

6. Gilks WR, Best NG, Tan KK. Adaptive rejection metropolis sampling within Gibbs sampling. *Appl Stat*. 1995;44(4): 455–472.

7. Lunn D, Spiegelhalter D, Thomas A, et al. The BUGS project: evolution, critique and future directions. *Stat Med*. 2009;28(25): 3049–3067.

8. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci*. 1992;7(4):457–511.

9. Dawber TR, Kannel WB, Lyell LP. An approach to longitudinal studies in a community: the Framingham study. *Ann NY Acad Sci*. 1963;107:539–556.