Published in final edited form as:

Can J Stat. 2011 September 1; 39(3): 488–497. doi:10.1002/cjs.10117.

The effect of misspecification of random effects distributions in clustered data settings with outcome-dependent sampling

John M. Neuhaus* and Charles E. McCulloch

Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143-0560, USA

Abstract

Genetic epidemiologists often gather outcome-dependent samples of family data to measure within-family associations of genetic factors with disease outcomes. Generalized linear mixed models provide effective methods to estimate within-family associations but typically require parametric specification of the random effects distribution. Although misspecification of the random effects distribution often leads to little bias in estimated regression coefficients in standard, prospective clustered data settings, some recent studies suggest that such misspecification will impact parameter estimates from outcome-dependent cluster sampling designs. Using analytic results, simulation studies and fits to example data, this study examines the effect of misspecification of random effects distributions on parameter estimates in clustered data settings with outcome-dependent sampling. We show that the effects are consistent with results from prospective cluster sampling settings. In particular, ascertainment corrected mixed model methods that assume normally distributed random intercepts and conditional likelihood approaches provide accurate estimates of within-family covariate effects even under a misspecified random effects distribution.

Keywords

Ascertainment correction; conditional likelihood; generalized linear mixed models; misspecified mixing distributions

1. INTRODUCTION

In studying associations of genetic factors with disease outcomes such as the occurrence of cancer, genetic epidemiologists often gather family-based data, selected on the basis of values of the outcomes. When the disease outcomes are rare, sampling observations conditional on outcome status can ensure that the sample includes sufficient outcomes of interest for meaningful analysis. This is known as *outcome-dependent sampling*. Family-based samples are used to avoid bias that can occur from what is called population stratification (e.g., Knowler et al., 1988).

As an example of such a genetic study, Rybicki et al. (2004) assessed the associations of predictors with the presence of prostate cancer using a sample of brothers. The investigators first recruited a sample of men with prostate cancer into the study and from them observed predictors such as demographic and genetic information, as well as levels of prostate-

specific antigen. The investigators also recruited the brothers of the initial sample of subjects and observed the above predictors and prostate cancer outcomes among the brothers in the secondary sample. The sampling strategy was therefore outcome dependent because it contained no sets of brothers with no cancer cases, which is a common feature of family-based genetic epidemiology studies. The objective of the study of Rybicki et al. (2004) was to relate genetic differences between brothers to differences in prostate cancer status.

Statistical analysis of outcome-dependent clustered data must address both differential sampling of clusters as well as dependence of responses within clusters to obtain valid inference. Epstein, Lin, & Boehnke (2002) presented simulation findings that showed that correct specification of both the model for the clustered responses and the sampling process yielded accurate estimation of model parameters while gross misspecification, such as the failure to acknowledge the correlation of responses within clusters, produced biased estimation. This study considers a subtler form of model misspecification, namely incorrect assumptions about the form of the random effects distribution in the generalized linear mixed model used to analyze such data.

We show below that the outcome-dependent sampling process alters both the within-cluster correlation structure of the responses and the distribution of the covariates and also affects other aspects of the random effects distribution, which Neuhaus & McCulloch (2006) showed are central to consistent estimation of parameters of interest. After sampling using an outcome-dependent sampling plan, the shape of the distribution of the random effects used in standard analyses of such data is not the same as in the original population. In addition, previous work by Glidden & Liang (2002) and Noh, Lee, & Pawitan (2005) has shown sensitivity in estimation of some model parameters. All of these considerations suggest that the effects of model misspecification in outcome-dependent sampling settings may have more impact than in prospective cluster sampling settings.

1.1. Outcome-Dependent Sampling

We assume that the data of interest consist of clustered or longitudinal binary responses Y_{ij} along with p -dimensional covariates X_{ij} , where i indexes clusters ($i = 1, \dots, m$) and j indexes units within clusters ($j = 1, \dots, n_i$), and that we want to assess the association of changes in X with a known function of $E(Y)$. In addition, we assume that the study samples clusters based on the outcomes $Y_i = (Y_{i1}, \dots, Y_{in_i})$, as in Rybicki et al. (2004).

As in Neuhaus & Jewell (1990), we define a binary random variable, S_i , that indicates whether the i th cluster is included in the sample ($S_i = 1$) or not ($S_i = 0$). We assume that cluster sampling depends on the cluster outcomes y_i but not the covariates x_i ,

$$\text{pr}(S_i=1|y_i, X_i) = \text{pr}(S_i=1|y_i) \equiv \lambda(y_i), \quad (1)$$

although we can relax this assumption. Using Bayes' theorem it is easy to show that

$$\text{pr}(y_i|X_i, S_i=1) = \frac{\text{pr}(S_i=1|y_i, X_i)\text{pr}(y_i|X_i)}{\text{pr}(S_i=1|X_i)} = \frac{\lambda(y_i)\text{pr}(y_i|X_i)}{\sum_{y_k} \lambda(y_k)\text{pr}(y_k|X_i)}, \quad (2)$$

where the denominator of (2) sums overall possible cluster outcomes y_k . The prostate cancer study by Rybicki et al. (2004) is a special case where investigators only sample families with at least one $y_{ij} = 1$ so that

$$\text{pr}(y_i|X_i, S_i=1) = \frac{\text{pr}(y_i|X_i)}{1 - \text{pr}(y_{i1} = \dots = y_{in_i} = 0|X_i)}. \quad (3)$$

We are interested in fitting generalized linear mixed models to data from studies such as Rybicki et al. (2004) since these models estimate the covariate effects of interest. Generalized linear mixed models specify that, given a vector b_i of parameters specific to the i th cluster, for the j th unit, the conditional density of Y_{ij} is of the form

$$f_Y(y_{ij}|b_i, x_{ij}) = \exp[\{y_{ij}\theta_{ij} - c(\theta_{ij})\}\phi + d(y_{ij}, \phi)] \quad (4)$$

where c and d are functions of known form, ϕ is a scale parameter and θ_{ij} depends on covariates X_{ij} . In addition, one assumes that

$$\mu_{ij} = E(Y_{ij}|b_i, z_{ij}, x_{ij}) = g^{-1}(z_{ij}^T b_i + x_{ij}^T \beta) \equiv g^{-1}(\eta_{ij}), \quad (5)$$

where x_{ij}^T and z_{ij}^T are the specified covariate row vectors relating the fixed and random effects, respectively, to the observations, g is a link function and μ_{ij} is a function of θ_{ij} . Without loss of generality, we assume that $E(X) = 0$. Given b_i , we assume that the responses $Y_{i1}, \dots, Y_{in_i}$ are independent. This study focuses on random intercept models so that b_i is a scalar and $z_{ij} \equiv 1$. For convenience, we will write the covariate function in (5) as

$$\eta_{ij} = \beta_0 + \sigma_b b_i + x_{ij}^T \beta_1, \quad (6)$$

where $E(b_i) = 0$ and $\text{var}(b_i) = 1$, so we are assuming that the random effects distribution has variance equal to 1 and is scaled by σ_b in the model equation to obtain the desired magnitude of variability.

With clustered data it is useful to distinguish between covariates that vary within-clusters, such as genotypes in the prostate cancer study, and those that do not vary within clusters, that is, between cluster covariates that are measured at the cluster level. Analysis of clustered and longitudinal data often focuses on the effects of within-cluster covariates and assesses the change in response associated with changes in covariates within clusters.

Outcome-dependent sampling not only affects the distribution of the response as in (2) but can also alter the relationship between random effects and covariates. As an example, consider the case of a cluster-level covariate X_B , so designated because it varies between clusters, which is independent of the random intercept b in the population. Following the Bayes' theorem argument leading to (2) we can compute the joint density of (X_B, b) as

$$f(x_B, b|S_i=1) = \frac{\text{pr}(S_i=1|x_B, b)f(x_B, b)}{\text{pr}(S_i=1)}. \quad (7)$$

As in (2), we have

$$\text{pr}(S_i=1|x_B, b) = \sum_{y_k} \lambda(y_k) \text{pr}(y_k|X_B, b)$$

so that

$$f(x_B, b|S_i=1) = \frac{\left\{ \sum_{y_k} \lambda(y_k) \text{pr}(y_k|X_B, b) \right\} f(x_B) f(b)}{\int_{x_B} \int_b \left\{ \sum_{y_k} \lambda(y_k) \text{pr}(y_k|X_B, b) \right\} f(x_B) f(b) d_{x_B} db}. \quad (8)$$

Since the terms inside the curly brackets in the numerator of (8) involve both x_B and b , and because these terms are not, in general, separable into a component depending solely on x_B and another depending solely on b , (8) will not factor, as would be required for independence of x_B and b . This implies that the covariate and random intercept are not independent after the outcome-dependent sampling process.

The magnitude of the association between X_B and b depends on the true distribution of b . Neuhaus & McCulloch (2006) showed that misspecification of $\text{corr}(X_B, b)$ can yield biased estimates of parameters of interest in prospective cluster sample settings. Since the outcome-dependent sampling plan as well as the random effects distribution affect the magnitude of $\text{corr}(X_B, b)$, the impact of misspecification of the distribution of the random effects may be greater than in prospective cluster sample settings.

1.2. Estimation

Generalized linear mixed models extend the class of generalized linear models by adding random effects to the linear predictor and provide useful methods to analyze clustered and longitudinal data. Fitting generalized linear mixed models in practice typically requires parametric specification of the distribution of the random effects and there exists little information to guide this choice. There is an extensive literature on the effects of misspecifying the random effects distribution when fitting random intercepts models to prospectively sampled clustered data and most of this work shows that misspecification has little effect on analyses (e.g., Butler & Louis, 1992; Neuhaus, Hauck, & Kalbfleisch, 1992; Neuhaus, Kalbfleisch, & Hauck, 1994; Verbeke & Lesaffre, 1997; Heagerty & Kurland, 2001; Neuhaus & McCulloch, 2006).

However, Glidden & Liang (2002) and Noh, Lee, & Pawitan (2005) presented the results of simulation studies that they claimed showed that random effects misspecification can produce substantial bias in estimated model parameters. We present analytical results in Analytic Results to Assess Bias Section and Conditional Likelihood Methods Section, as well as simulation results in Simulation Studies Section and an example in Example Section to clarify the effects of misspecification of the random effects distribution in outcome-dependent cluster sampling settings. In particular, we show that the effects of random effects misspecification in outcome-dependent samples are consistent with those in prospective cluster samples, namely that misspecification of shape produces virtually no bias in within cluster coefficients, but can produce biased intercept and variance component estimators.

2. ANALYTIC RESULTS TO ASSESS BIAS

We develop analytic results by viewing outcome-dependent sampling as a modification of the random effects distribution. Prospectively we have

$$\text{pr}(y_i|X_i) = \int \prod_j \text{pr}(y_{ij}|X_{ij}, b_i) dG_T(b_i), \quad (9)$$

where $G_T(\cdot)$ is the c.d.f. of the distribution of the random effects, b_i . We use the subscript T to denote the true distribution and to distinguish it (below) from the fitted distribution, for which we will use a subscript F . Under outcome-dependent sampling we have

$$\text{pr}(y_i|X_i, S_i=1) = \frac{\lambda(y_i) \int \prod_j \text{pr}(y_{ij}|X_{ij}, b_i) dG_T(b_i)}{\sum_{y_i} \lambda(y_i) \text{pr}(y_i|X_i)}, \quad (10)$$

which we can express in the form of Equation (9) as

$$\text{pr}(y_i|X_i, S_i=1) = \int \prod_j \text{pr}(y_{ij}|X_{ij}, b_i) dH_T(b_i),$$

where

$$dH_T(b) = \frac{\lambda(y_i) dG_T(b)}{\sum_{y_k} \lambda(y_k) \text{pr}(y_k|X_i)}. \quad (11)$$

We can estimate model parameters $\xi = (\beta_0, \beta_1, \sigma_b)$ of (6) by maximizing

$$L = \prod_i \text{pr}(y_i|X_i, S_i=1) = \prod_i \int \prod_j \text{pr}(y_{ij}|X_{ij}, b_i) dH_T(b_i), \quad (12)$$

where dH_T depends on the true dG_T and the outcome-dependent sampling design. However, there exists little information to guide the choice of the random effects distribution so that the fitted model will use a distribution $dH_F(b)$, misspecified to some degree, and different from $dH_T(b)$. That is, we will incorrectly assume that the underlying random intercepts follow $dG_F(b)$ where $dG_F(b) \neq dG_T(b)$. For example, we may incorrectly assume $dG_F(b)$ is $N(0,1)$.

Thus, under the true model

$$dH_T(b) = \frac{\lambda(y_i) dG_T(b)}{\sum_{y_k} \lambda(y_k) \text{pr}(y_k|X_i)}, \quad (13)$$

while the misspecified model uses

$$dH_F(b) = \frac{\lambda(y_i) dG_F(b)}{\sum_{y_k} \lambda(y_k) \text{pr}(y_k|X_i)}. \quad (14)$$

This is a useful representation since it allows us to take advantage of existing results to assess the effects of misspecification.

Our approach uses the theory of inference with misspecified models (Akaike, 1973; White, 1982) and evaluates the asymptotic expectation of “maximum likelihood” estimator, $\hat{\xi}^*$, of ξ^* from incorrect model that assumes $b \sim H_F$. This theory shows that the estimator $\hat{\xi}^*$ converges to the value ξ^* that minimizes the Kullback–Leibler divergence (Kullback, 1959) between the true and false models. That is, ξ^* minimizes

$$E_X E_{Y|X} \log \left\{ f_{H_T}(y|\xi, X) / f_{H_F}(y|\xi^*, X) \right\}. \quad (15)$$

Differentiating (15) with respect to ξ^* yields a system of equations to solve for ξ^* . The equations for β_0^* and β_1^* are

$$E_X E_{Y_{ix}} \sum_{i=1}^m \frac{\int_b \left\{ \sum_{j=1}^{n_i} r_{ij}^* \right\} \left\{ \prod_{j=1}^{n_i} f_Y^*(y_{ij}|x_{ij}, b) \right\} dH_F(b)}{\int_b \prod_{j=1}^{n_i} f_Y^*(y_{ij}|x_{ij}, b) dH_F(b)} = 0 \quad (16)$$

$$E_X E_{Y_{ix}} \sum_{i=1}^m \frac{\int_b \left\{ \sum_{j=1}^{n_i} X_{ij} r_{ij}^* \right\} \left\{ \prod_{j=1}^{n_i} f_Y^*(y_{ij}|x_{ij}, b) \right\} dH_F(b)}{\int_b \prod_{j=1}^{n_i} f_Y^*(y_{ij}|x_{ij}, b) dH_F(b)} = 0, \quad (17)$$

where $r_{ij}^* = (y_{ij} - \mu_{ij}^*) / [V(\mu_{ij}^*)g'(\mu_{ij}^*)]$, $V(\cdot)$ is the variance function of the generalized linear model, and the asterisks denote quantities evaluated with parameters equal to ξ^* . The equation resulting from differentiation with respect to σ_b^* is of similar form.

General analytic solutions to the system (16–17) are not available. We consider instead the special case when $\beta_1 = 0$ and argue that $\beta_1^* = 0$ is a solution, which would imply that $\widehat{\beta}_1^*$ consistently estimates β_1 at $\beta_1 = 0$. Consistency results at $\beta_1 = 0$ are useful since they are a prerequisite for tests of the hypothesis $H_0 : \beta_1 = 0$ using estimators $\widehat{\beta}_1^*$ and based on likelihoods that misspecify the random effects distribution to have the correct type I error rate.

When $\beta_1 = \beta_1^* = 0$, r_{ij}^* , f_Y^* , and dH_F of (16–17) are free of X_{ij} (β_1 appears in y_{ij}) so that the left hand side of (17) is a linear function of X_{ij} . Sweeping in the expectation with respect to X and recalling that X_{ij} has expectation zero shows that the left hand side is zero so that $\widehat{\beta}_1^*$ consistently estimates β_1 at $\beta_1 = 0$.

For $\beta_1 \neq 0$, analytic evaluation of the expectations of (17) is typically intractable. However, Taylor approximations in Neuhaus, Hauck, & Kalbfleisch (1992) indicate that $\beta_1^* \approx \beta_1$ when b is uncorrelated with X in prospective cluster sampling settings. This suggests that the same will hold true approximately for values other than $\beta_1 = 0$. As more direct verification, Section 4 presents the results of that examine the performance of estimators from binary mixed effects models that misspecify the distribution of the random effects b .

3. CONDITIONAL LIKELIHOOD METHODS

With canonical link, random intercept GLMMs, the conditional likelihood approach treats intercepts b_i as fixed constants and eliminates them from the likelihood by conditioning on their sufficient statistics $\sum_{j=1}^{n_i} y_{ij}$. The conditional likelihood depends only on covariates that vary within clusters and has form

$$CL(\beta_w) = \prod_{i=1}^m \int_Y \left(y_{i1}, \dots, y_{in_i} | x_{i1}, \dots, x_{in_i}, b_i, \sum_{j=1}^{n_i} y_{ij} \right). \quad (18)$$

Conditioning on $\sum_{j=1}^{n_i} y_{ij}$ removes all between-cluster effects, including the cluster-specific intercepts and any associated distributional assumptions. Neuhaus & Jewell (1990) showed that the conditional likelihood approach yields consistent estimates of within-cluster

coefficients, β_W , in outcome-dependent sampling designs with $\text{pr}(S_i=1|y_i)=\lambda \left(\sum_j y_{ij} \right)$. We note that the prostate cancer study by Rybicki et al. (2004) used such a design.

4. SIMULATION STUDIES

We carried out a set of simulation studies to evaluate the performance of mixed-effects model methods in outcome-dependent sampling settings with a misspecified random effects distribution. We generated populations of clustered binary responses from logistic models with a random intercept along with a within-cluster covariate x_{Wij} :

$$\text{logit}[\text{pr}(y_{ij}=1|b_i, x_{ij})]=\beta_0+\exp(\log\sigma_b)b_i+\beta_W x_{Wij}, \quad (19)$$

where $b_i \sim G_T$, $i = 1, \dots, 1,000$, and $j = 1, 2, 3$. We parametrized the random intercept variance in terms of $\log\sigma_b$ to improve computation. We constructed the within-cluster covariate, x_{Wij} , to take on values $-1, 0$, and 1 in every cluster so that $E(X_W) = 0$. We used four true random intercept distributions, G_T :

1. $b \sim \text{normal}$
2. $b \sim t$ with 3 degrees of freedom
3. $b \sim \text{lognormal}$
4. $b \sim \text{exponential}(1)$,

then standardized to have $E(b) = 0$ and $\text{var}(b) = 1$. The models used the same true values of $\beta_0 = -5.0$ and $\sigma_b^2 = 4.5$ as Glidden & Liang (2002) and Noh, Lee, & Pawitan (2005), as well as $\beta_W = 1.0$. We note that a value of $\sigma_b^2 = 4.5$ is a large random effects variance since it corresponds to a covariate effect with odds ratio of $\exp(\sqrt{4.5}) \approx 8.3$. Each simulation contained 1,000 replications. To mimic the outcome-dependent sampling strategy of Rybicki et al. (2004) we started “prospectively” with 4,000 clusters and then we dropped all clusters with $y_i = (0, 0, 0)$ to yield our outcome-dependent sample, which was typically around 400 clusters.

To each outcome-dependent sample we fit a mixed effects logistic of form (19) with a variety of distributional assumptions. For each of the above true distributions we fit a model with assumed normal random effects, that is, the correct distribution for the first scenario and a misspecified distribution G_F for the last three. For the lognormal and exponential models we also fit models assuming the correct distribution. In addition, for each scenario we calculated the conditional maximum likelihood estimates, which do not involve any distributional assumption for the random effects.

Our mixed-effects logistic fits were corrected for ascertainment based on (3) and (9) under the assumed random effects distribution (normal, exponential, or lognormal). We fit the mixed models using Proc NLMIXED in SAS, parameterizing the model in terms of $\log(\sigma_b)$. We calculated $\text{pr}(y_i = 0)$ for the ascertainment correction using Gauss–Hermite quadrature with 30 quadrature points within NLMIXED to construct dH_F . Our simulation code is available upon request. We fit the conditional likelihood approach using Proc PHREG in SAS. The distributions of the estimators from the simulations, particularly of $\log \hat{\sigma}_b$, were highly skewed. Thus, we present simulation results as medians and interquartile ranges instead of the more standard means and standard deviations.

Table 1 gives the simulation results and shows that misspecification of the shape of the random effects distribution can yield biased estimates of $\hat{\beta}_0$ and $\log \hat{\sigma}_b$. However, and more

importantly, the estimates of the within-cluster coefficient, $\hat{\beta}_W$, are little affected. This is consistent with the theory of Section 2. Consistent with the theory of Section 3, the medians of the conditional likelihood estimators in Table 1 closely corresponded to the true value of 1.0. The interquartile ranges for β_W based on the conditional likelihood approach were nearly identical to the corresponding ranges for the correctly specified models, indicating that the conditional likelihood approach lost very little efficiency with respect to the full model. Also, the interquartile ranges for β_W based on the correct and misspecified models were nearly identical indicating little efficiency lost due to misspecification of the random effects distribution. This is consistent with the efficiency results of Neuhaus, McCulloch, & Boylan (2011) in prospective cluster sample settings. Glidden & Liang (2002) and Noh, Lee, & Pawitan (2005) studied only models with random intercepts and therefore came to the conclusion that parameter estimates could be drastically influenced by model misspecification. However, our results clarify that the parameters most often of interest in family-based studies, namely within cluster effects, are highly robust to misspecification, even in this scenario with a large random effects variance.

5. EXAMPLE

We further illustrate our results by fitting several approaches to data from Rybicki et al. (2004). The data set consists of 526 prostate cancer cases along with 611 brothers (affected and unaffected). The analysis assesses the within-family association of the RNase L gene, RNASEL, with prostate cancer adjusted for the age of the brother. We code the three RNASEL genotypes AA, AG, and GG as two dummy indicators, X_{AA} and X_{AG} to denote the comparisons of genotypes AA and AG with the baseline GG genotype. Since the covariate age varies both between and within families, we followed Neuhaus & McCulloch (2006) and decomposed age into between- and within-family components by calculating mean age for each family and computing the deviation of each brother's age value from the family mean to yield covariates $\text{age}_{\text{between}}$ and $\text{age}_{\text{within}}$. We coded the covariates to provide useful interpretations of covariate effects and did not further manipulate them so that they had zero means in the sample, although we note that the covariate $\text{age}_{\text{within}}$ has this property. We fit three approaches to these data: (1) an ascertainment corrected random intercept logistic model (2) that assumed normally distributed random intercepts; (2) a standard conditional likelihood approach; and (3) a standard random intercept logistic model with $b \sim N(0, 1)$ that ignored ascertainment. Table 2 presents the results. Consistent with the theory of Section 2 and Section 3 and simulations results of Section 4, the ascertainment corrected estimates corresponded well with the conditional likelihood estimates. Table 2 also illustrates that ignoring ascertainment can yield extremely misleading estimates. For example, comparing the AA to the GG genotype the odds of prostate cancer were estimated to increase approximately twofold ($\exp(0.73) \approx 2.08$) and were statistically significant using the conditional maximum likelihood or ascertainment corrected mixed model methods. However, the odds was only estimated to be elevated by about 40% and was not statistically significant when ignoring ascertainment.

6. DISCUSSION

Although misspecification of the shape of the random effects distribution typically produces little bias in the estimated regression coefficients of generalized linear mixed models in prospective cluster sample settings, outcome-dependent sampling plans alter both the relationship of covariates with the response, as in (10), and the relationship of covariates with random effects so that results from the prospective sample setting do not immediately carry over. Furthermore, previous work looking at intercept only models has shown sensitivity to the parametric random effects assumptions involved with using generalized linear mixed models.

Via theory, which cast the effects of the outcome-dependent sampling plan as a misspecification of the random effects distribution, and via simulation, we have shown there is little bias in within cluster coefficients due to misspecification of the shape of the random effects distribution. Therefore, the robustness of within cluster covariate effects that has been demonstrated in prospective sampling situations continues to hold in outcome-dependent settings. This is of considerable practical importance because within cluster covariate effects are often of paramount importance in clustered data settings and covariates can be decomposed to isolate their within cluster components. Inference for these covariate effects can be expected to be highly robust to the parametric assumption for the random effects distribution.

Acknowledgments

A grant from the U.S. National Institutes of Health supported this research. We thank Dr. John Witte of the University of California, San Francisco for providing the prostate cancer data and Dr. Ross Boylan of the University of California, San Francisco for assistance with computing.

BIBLIOGRAPHY

- Akaike, H. Petrov, BN.; Czaki, F. Second International Symposium on Information Theory. Budapest: Akademiai Kiadó; 1973. Information theory and an extension of the maximum likelihood principle; p. 267-281.
- Butler SM, Louis TA. Random effects models with non-parametric priors (Disc: p2017-2023). *Statistics in Medicine*. 1992; 11:1981–2000. [PubMed: 1480884]
- Epstein M, Lin X, Boehnke M. Ascertainment-adjusted parameter estimates revisited. *American Journal of Human Genetics*. 2002; 70:886–895. [PubMed: 11880949]
- Glidden D, Liang K-Y. Ascertainment adjustment in complex diseases. *Genetic Epidemiology*. 2002; 23:201–208. [PubMed: 12384971]
- Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*. 2001; 88:973–986.
- Knowler W, Williams R, Pettitt D, Steinberg A. Gm3-5,13,14 and type-2 diabetes mellitus—An association in American-Indians with genetic admixture. *American Journal of Human Genetics*. 1988; 43:520–526. [PubMed: 3177389]
- Kullback, S. *Information Theory and Statistics*. New York: John Wiley and Sons; 1959.
- Neuhaus JM, Jewell NP. The effect of retrospective sampling on binary regression models for clustered data. *Biometrics*. 1990; 46:977–990. [PubMed: 2085642]
- Neuhaus JM, McCulloch CE. Separating between- and within-cluster covariate effects using conditional and partitioning methods. *Journal of the Royal Statistical Society, Series B*. 2006; 68:859–872.
- Neuhaus JM, Hauck WW, Kalbfleisch JD. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika*. 1992; 79:755–762.
- Neuhaus JM, Kalbfleisch JD, Hauck WW. Conditions for consistent estimation in mixed-effects models for binary matched-pairs data. *The Canadian Journal of Statistics*. 1994; 22:139–148.
- Neuhaus JM, McCulloch CE, Boylan R. A note on type II error under random effects misspecification in generalized linear mixed models. *Biometrics*. 2011 (in press).
- Noh M, Lee Y, Pawitan Y. Robust ascertainment-adjusted parameter estimation. *Genetic Epidemiology*. 2005; 29:68–75. [PubMed: 15892092]
- Rybicki B, Conti D, Moreira A, Cicek M, Casey G, Witte J. DNA repair gene XRCC1 and XPD polymorphisms and risk of prostate cancer. *Cancer Epidemiology, Biomarkers and Prevention*. 2004; 13:23–29.
- Verbeke G, Lesaffre E. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis*. 1997; 23:541–556.
- White H. Maximum likelihood estimation of misspecified models. *Econometrica*. 1982; 50:1–25.

Table 1

Observed medians and interquartile ranges (IQR), in parentheses, of the regression coefficients of two methods for fitting generalized linear mixed models with random intercepts to simulated outcome-dependent clustered data with various true (G_T) random intercept distributions. The fitting methods were: (1) ascertainment corrected random intercept (mixed) logistic models with various assumed distributions (10) and (2) conditional maximum likelihood (18). True values: $\beta_0 = -5.0$, $\beta_W = 1.0$, $\log(\sigma_b) = \log(\sqrt{4.5}) = 0.75$.

G_T	Fitting method	β_0 (IQR)	β_W (IQR)	$\log(\sigma_b)$ (IQR)
Normal	Mixed assuming normal	-5.12 (-7.73 to -3.73)	1.00 (0.95–1.05)	0.77 (0.52–1.03)
	Conditional likelihood		1.00 (0.95–1.05)	
t_3	Mixed assuming normal	-10.43 (-10.88 to -10.07)	0.99 (0.93–1.06)	1.35 (1.30–1.40)
	Conditional likelihood		1.00 (0.94–1.07)	
Exp(1)	Mixed assuming Exp(1)	-5.28 (-5.46 to -5.06)	1.00 (0.95–1.06)	0.77 (0.73–0.81)
	Mixed assuming normal	-8.36 (-8.68 to -8.06)	1.00 (0.95–1.06)	1.46 (1.41–1.51)
	Conditional likelihood		1.00 (0.95–1.06)	
Lognormal	Mixed assuming lognormal	-4.56 (-4.77 to -4.39)	1.01 (0.95–1.08)	0.84 (0.80–0.88)
	Mixed assuming normal	-8.99 (-9.42 to -8.59)	1.00 (0.93–1.07)	1.57 (1.51–1.63)
	Conditional likelihood		1.00 (0.94–1.07)	

Table 2

Log odds ratio and standard error estimates, in parentheses, from three methods for fitting mixed-effects logistic models to the prostate cancer data (Rybicki et al., 2004). The fitted models were: (1) ascertainment corrected random intercept logistic model that assumed normally distributed random intercepts; (2) conditional likelihood approach (CML); and (3) standard random intercept logistic model that ignored ascertainment.

	CML	Mixed-effects	Ignore Ascer.
Age _{within}	−1.03 (0.17)	−1.05 (0.17)	−1.47 (0.20)
Age _{between}		0.57 (0.26)	0.13 (0.09)
AA versus GG	0.74 (0.33)	0.73 (0.30)	0.32 (0.21)
AG versus GG	0.31 (0.19)	0.32 (0.17)	0.18 (0.14)