Genetic Epidemiology 23:209–213 (2002)

Comment on "Ascertainment Adjustment in Complex Diseases"

Michael P. Epstein*

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan

INTRODUCTION

Glidden and Liang [2002] have raised important issues regarding ascertainment adjustment in the framework of variance-components modeling for complex genetic traits. While the structure of the authors' logistic variance-component model is simple, ascertainment issues arising with this model are likely analogous to ascertainment issues in more complex variance-component models commonly used for the analysis of either genetic disease data [Duggirala et al., 1997; Burton et al., 1999] or quantitative trait data [Amos, 1994; Almasy and Blangero, 1998]. Therefore, the results of Glidden and Liang [2002] are of importance both to investigators who design gene mapping studies, and to analysts who use variance-component methods to study genetic trait data.

The authors first demonstrate that, if the ascertainment scheme is correctly modeled, ascertainment-adjusted parameter estimates from their logistic variance-component model for analyzing disease data reflect the true values of the population-based parameter values rather than the sample-based parameter values. These results are analogous to those of Epstein et al. [2002], who used a similar logistic variance-component model initially proposed by Burton et al. [2000]. de Andrade and Amos [2000] showed similar results for the traditional linear variance-component method that assumed major gene, polygene, and environmental effects. In their example, de Andrade and Amos [2000] selected families in which one sibling had a trait value more extreme than 90% of the population, and properly accounted for ascertainment by dividing the unconditional likelihood by the likelihood that the selected

Grant sponsor: University of Michigan Rackham Predoctoral Fellowship.
*Correspondence to: Michael P. Epstein, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: mepstein@umich.edu
Received for publication 30 April 2002; Revision accepted 23 May 2002
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/gepi.10197

sibling had a trait value greater than or equal to the 90th percentile of the population.

While these results support the idea that ascertainment-adjusted parameter estimates reflect the true population-based values, one's ability to obtain these estimates can be complicated by many factors. For disease data, the population prevalence of the disease can have an impact on the ability to obtain unbiased ascertainment-adjusted parameter estimates. In particular, if the disease is rare, one can show that ascertainment-adjusted parameter estimates from the logistic variance-components model of Glidden and Liang [2002] are biased with respect to the population-based values unless one uses enormous sample sizes.

To demonstrate this, I first show how one calculates the population prevalence of a disease simulated from the logistic variance-component model in equation 1 of Glidden and Liang [2002]. Using Gauss-Hermite integration, the population prevalence of the simulated disease is

$$\begin{aligned} P(D_{ij}=1) &= \int \frac{\exp(\alpha + C_i)}{1+\exp(\alpha + C_i)} \frac{\exp\left(-\frac{C_i^2}{2\sigma_c^2}\right)}{\sqrt{2\pi\sigma_c^2}} dC_i \\ &= \int \frac{\exp\left(\alpha + \sqrt{2\sigma_c^2}x\right)}{1+\exp\left(\alpha + \sqrt{2\sigma_c^2}x\right)} \frac{\exp(-x^2)}{\sqrt{\pi}} dx \\ &\approx \sum_{k=1}^{N} w_k \frac{\exp\left(\alpha + \sqrt{2\sigma_c^2}a_k\right)}{1+\exp\left(\alpha + \sqrt{2\sigma_c^2}a_k\right)} \frac{1}{\sqrt{\pi}}, \end{aligned} \tag{1}$$

where w_k and a_k are predefined weights and abscissas, and N denotes the number of quadrature points.

Glidden and Liang [2002] simulated disease data, assuming $\alpha = -5$ and $\sigma_c^2 = 4.5$ which, using (1), corresponds to a disease with a population prevalence of approximately 0.04. To simulate a rarer disease, I investigated a disease model in which $\alpha = -10$ and $\sigma_c^2 = 4.5$. Using (1), this model corresponds to a disease with a population prevalence of approximately 0.0004. I performed analyses to determine whether ascertainment-adjusted parameter estimates from the logistic variance-component method of Glidden and Liang [2002] were unbiased for this rarer disease. I simulated datasets of 1,000 sibships of size 5 under complete ascertainment (collecting all sibships with at least one affected sibling) and obtained ascertainment-adjusted parameter estimates of the two parameters, using adaptive Gaussian quadrature [Pinheiro and Bates, 1995]. Over 200 simulated datasets, the mean estimates of α and σ_c^2 are -6.81 (SD = 1.53) and 2.04 (SD = 1.30), respectively, which are severely biased with respect to the true simulated values.

In this example, the ascertainment-adjusted analysis of the rare disease yielded biased results due to small sample effect. In a simulated dataset of 1,000 sibships of size 5 selected under complete ascertainment, the average proportion of sibships with only one affected sibling is approximately 0.97. As these sibships are selected based on that affected sibling (under complete ascertainment), they provide little or no information for statistical inference. Inference instead is based primarily on the other approximate 3% of the sample that has more than one affected sibling; hence, the small sample effect. When ascertainment-adjusted analyses were repeated for a much

larger sample of 10,000 sibships, estimates of α and σ_c^2 over 200 simulated datasets were, on average, unbiased with respect to the simulated values (data not shown).

Another factor that can complicate one's ability to obtain ascertainment-adjusted parameter estimates is that, for small family sizes, the parameter estimates may be nonidentifiable (Glidden, personal communication). For example, if one had analyzed sibships of size 2 instead of size 5, using the model of Glidden and Liang [2002] (sibships still ascertained under complete ascertainment), one would be unable to obtain unique maximum likelihood estimates of α and σ_c^2. The ascertained-adjusted likelihood would have the form $(p_1)^{n_1}(p_2)^{n_2}$, where p_j denotes the probability of j affected siblings in a sibship, and n_j denotes the observed number of such sibships. Here, one can easily rewrite p_j as a function of α and σ_c^2. Because one ascertains sibships based on having at least one affected sibling, one can rewrite the likelihood as $(p_1)^{n_1}(1-p_1)^{n_2}$ and subsequently obtain the maximum-likelihood estimate $\hat{p}_1 = n_1/(n_1+n_2)$. As multiple values of α and σ_c^2 can solve $\hat{p}_1$, the maximum-likelihood estimates of $\hat{\alpha}$ and $\hat{\sigma}_c^2$ are nonidentifiable. To find identifiable estimates of $\hat{\alpha}, \hat{\sigma}_c^2$, one needs to impose a constraint on the relationship between the two parameters. If one had prior knowledge of the population disease prevalence, one could possibly use the constraint shown in Equation (18) of Breslow and Clayton [1993] to obtain $\hat{\alpha}$ and $\hat{\sigma}_c^2$. As many disease studies will consist only of ascertained sibpairs and sibtrios, future work in the area of suitable constraints would be of interest.

In the second part of their article, Glidden and Liang [2002] focused on the impact of random-effect misspecification on parameter estimates under ascertainment sampling. Their results showed that slight misspecification of random effects can lead to biased estimates of α and σ_c^2 under their logistic variance-components model, assuming complete ascertainment. While biased parameter estimates are expected under random-effect misspecification for random samples [Neuhaus et al., 1992; Heagerty and Kurland, 2001], the severity of the bias under ascertainment sampling is extraordinary, given that the misspecified random-effects distributions (logistic, $\sqrt{2.7}$ times a t-distribution with 5 degrees of freedom) are quite similar in form to the assumed multivariate normal distribution. The authors' findings lead to the issue of whether tests of $H_0 : \sigma_c^2 = 0$ have the correct size. If type I error is elevated, one needs to develop statistics that are robust to misspecified random effects. For random samples, Lin [1997] developed a variance-component score statistic that is robust to misspecified effects. One could easily extend this robust statistic to account for ascertainment.

This result from Glidden and Liang [2002] leads one to question whether the same phenomena will occur in more complicated ascertainment-adjusted variance-component models for genetic linkage analysis. Many articles have focused on the effect of random-effect misspecification on variance-components linkage analyses of quantitative traits, assuming random sampling [Allison et al., 1999; Blangero et al., 2001]. Results showed that the type I error for testing for a major gene effect increases as the kurtosis of the misspecified random-effects distribution increases. To my knowledge, no one has determined the effect of random-effect misspecification on ascertainment-adjusted variance-component analyses for selected normal traits [Hopper and Mathews, 1982; Elston and Sobel, 1979; de Andrade and Amos, 2000]. Due to the more complex structure of the ascertainment-adjusted likelihood, one

would expect that an ascertainment-adjusted variance-component linkage method would be even more sensitive to random-effect misspecification than the variance-component method under random sampling.

I performed some preliminary simulations to investigate this issue. Data were simulated for sibtrios from a trait model that assumed unmeasured effects due to a major gene, polygenes, and environment. Unmeasured effects were simulated from a Laplace distribution (similar to a normal distribution, but with a positive nonzero kurtosis). After trait data were simulated, I ascertained sibtrios that had at least one sibling with a trait value in the 90th or greater percentile of the population. I conducted ascertainment-adjusted variance-components analyses assuming multivariate normality of the random effects, as performed by de Andrade and Amos [2000]. Preliminary results indicate that misspecification of random effects leads to severely biased variance-component estimates and a large increase in type I error for testing the major gene effect. Therefore, the development of robust test statistics is required for valid inference of these ascertained samples. Such methods could include the ascertainment-based extensions of the robust score tests and LOD score methods for variance-component linkage analyses that Blangero et al. [2000, 2001] developed for random samples.

As described here and in Glidden and Liang [2002], problems can arise when one properly adjusts a variance-component analysis for ascertainment. Therefore, the development of variance-component methods that are independent of the ascertainment criterion could prove useful in genetic analyses of ascertained samples. For variance-component linkage analyses of selected samples, one such approach consists of using a retrospective likelihood, as proposed by Whittemore [1996]. For this approach, instead of maximizing the typical prospective likelihood of the trait data conditional on both the ascertainment criterion and the major gene data, one instead maximizes the likelihood of the major gene data conditional on the trait data. By using this retrospective likelihood, one avoids ascertainment bias and also may resolve the identifiability issues of the parameter estimates, given small families. Li and Zhong [2002] applied such a retrospective likelihood approach to the analysis of survival data, using genetic frailities. I am in the process of extending this retrospective likelihood approach to variance-component analyses of genetic data. In the future, I will investigate the impact of random-effect misspecification on such retrospective likelihood approaches and develop robust approaches, if necessary.

ACKNOWLEDGMENTS

I thank Drs. Michael Boehnke and Xihong Lin for their helpful comments. I also thank Dr. David Glidden for his valuable insights into the topic.

REFERENCES

Allison DB, Neale MC, Zannolli R, Schork NJ, Amos CI, Blangero J. 1999. Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Am J Hum Genet 65:531–44.

Almasy L, Blangero J. 1998. Multipoint QTL analysis in pedigrees. Am J Hum Genet 62:1198–211.

Amos CI. 1994. Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54:535–43.

Blangero J, Williams JT, Almasy L. 2000. Robust LOD scores for variance component-based linkage analysis. Genet Epidemiol [Suppl] 19:8–14.

Blangero J, Williams JT, Almasy L. 2001. Variance component methods for detecting complex trait loci. In: Rao DC, Province MA, editors. Genetic dissection of complex traits. London: Academic Press. p 151–82.

Breslow NE, Clayton DG. 1993. Approximate inference in generalized linear mixed models. J Am Stat Assoc 88:9–25.

Burton PR, Tiller K, Gurrin LC, Musk AW, Cookson WOCM, Palmer LJ. 1999. Genetic variance components analysis for binary phenotypes using generalized linear mixed models GLMMs and Gibbs sampling. Genet Epidemiol 17:118–40.

Burton PR, Palmer LJ, Jacobs K, Keen KJ, Olson JM, Elston RC. 2000. Ascertainment adjustment: where does it take us? Am J Hum Genet 67:1505–14.

de Andrade M, Amos CI. 2000. Ascertainment issues in variance components models. Genet Epidemiol 19:333–44.

Duggirala R, Williams JT, Williams-Blangero S, Blangero J. 1997. A variance component approach to dichotomous trait linkage analysis using a threshold model. Genet Epidemiol 14:987–92.

Elston RC, Sobel E. 1979. Sampling considerations in the gathering and analysis of pedigree data. Am J Hum Genet 31:62–9.

Epstein MP, Lin X, Boehnke M. 2002. Ascertainment-adjusted parameter estimates revisited. Am J Hum Genet 70:886–95.

Glidden DV, Liang K-Y. 2002. Ascertainment adjustment in complex diseases. Genet Epidemiol (in press).

Heagerty PJ, Kurland BF. 2001. Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika 88:973–85.

Hopper JL, Mathews JD. 1982. Extensions to multivariate normal models for pedigree analysis. Ann Hum Genet 46:373–83.

Li H, Zhong X. 2002. Multivariate survival models induced by genetic frailities, with application to linkage analysis. Biostatistics 3:57–75.

Lin X. 1997. Variance component testing in generalised linear models with random effects. Biometrika 84:309–26.

Neuhaus JM, Hauck WW, Kalbfleisch JD. 1992. The effects of mixture distribution misspecification when fitting mixed-effects models. Biometrika 79:755–62.

Pinheiro JC, Bates DM. 1995. Approximations to the log-likelihood function in the nonlinear mixed-effects model. J Comput Graph Stat 4:12–35.

Whittemore AS. 1996. Genome scanning for linkage: an overview. Am J Hum Genet 59:704–16.