true confounder r_2 is known. This indicates to us that there is much additional work that can be done to investigate the question of model selection for the stratification score.

Michael P. Epstein,[1,*] Andrew S. Allen,[2] and
Glen A. Satten[3]
[1]Department of Human Genetics, Emory University, Atlanta, GA 30322, USA; [2]Department of Biostatistics and Bioinformatics and Duke Clinical Research Institute, Duke University, Durham, NC 27708, USA; [3]Centers for Disease Control and Prevention, Atlanta, GA 30341, USA
*Correspondence: mepstein@genetics.emory.edu

Acknowledgments

Any opinions expressed in this paper are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

References

1. Epstein, M.P., Allen, A.S., and Satten, G.A. (2007). A simple and improved correction for population stratification in case-control studies. Am. J. Hum. Genet. *80*, 921–930.
2. Akey, J.M., Zhang, G., Zhang, K., Jin, L., and Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome Res. *12*, 1805–1814.
3. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
4. Campbell, C.D., Ogburn, E.L., Lunetta, K.L., Lyon, H.N., Freedman, M.L., Groop, L.C., Altshuler, D., Ardlie, K.G., and Hirschhorn, J.N. (2005). Demonstrating stratification in a European American population. Nat. Genet. *37*, 868–872.

DOI 10.1016/j.ajhg.2007.11.010. ©2008 by The American Society of Human Genetics. All rights reserved.

XMCPDT Does Have Correct Type I Error Rates

To the Editor: In the January 2007 issue of the *Journal*, Chung et al.[1] compared X-APL proposed by them to XMCPDT proposed by Ding et al.[2] Based on their simulation results, they stated that with use of allele frequencies estimated from observed parental genotypes, XMCPDT would give inflated type I error rates. Here we wish to point out that use of estimated allele frequencies is not the cause of inflated type I error rates. Rather, the actual cause was the severe violation of the XMCPDT assumption in their simulation settings, which was discussed at length in Ding et al.[2] As explicitly stated there, one assumption for XMCPDT to be a valid test for association under linkage is that "the pedigrees in a study are assumed to be drawn from a population of (extended) families, each of which has at least one affected offspring." They went on to say, "Otherwise, bias may exist, especially when all families have the same structure and affection pattern, which, fortunately, is not the case in a genetic study that collects pedigrees of all shapes and sizes and affection patterns." To study the robustness of the test statistic to departure from the assumption, Ding et al.[2] investigated trios as well as families with six children and concluded that "in a genetic study with pedigree data, bias should be negligible, and the proposed test statistic may be safely used." However, the simulation settings in Chung et al.,[1] which fixed the affection statuses of the offspring, severely violated the assumption, leading to appreciable bias.

A fuller dissection of the assumption of Ding et al.[2] is needed in order to facilitate understanding of why the settings in Chung et al.[1] constitute severe violations. The sampling assumption treats affection status of a given family structure as a random event, and as such, all sorts of affection patterns are permitted. For example, for nuclear families with three children (a setting in Table 4 of Chung et al.[1]), under the assumption, one would expect some families having one, some having two, and some having all three children being affected. However, Chung et al.[1] only allow exactly two of the three children in each of the nuclear families to be affected, thus severely violating the assumption. Such a restriction on the affection status appears to be rather unrealistic in a genetic epidemiological study, as it is unlikely that a family with three children would only be included in the study if exactly two of the three children were affected. With inclusion of one-affected and three-affected families, the power is expected to increase substantially. More importantly, as demonstrated below through simulations, it is in fact X-APL that gave inflated type I error rates when the XMCPDT assumption was roughly satisfied, especially when data from extended families were included.

Our first simulation setting made use of the same family structure, discussed above, as that of Chung et al.,[1] but ours allowed for one-affected and three-affected families to be included in addition to the two-affected ones. One hundred nuclear families, each with two parents and three offspring, were simulated in each replicate. Among those 100 families, 25 had three male offspring, 25 had two male and one female offspring, 25 had one male and two female offspring, and the remaining 25 had three female offspring. Furthermore, parents in 50 of the families had observed genotypes, and those in the other 50 families did not. The disease models were the same as those in Table 1 of Chung et al.[1] For each of the four family types, we simulated the data until we had 25 families, each with at least one affected offspring. The disease locus was used to calculate powers. In addition to the disease locus, a marker with the same allele frequencies and in complete linkage and linkage equilibrium was also simulated and used to calculate type I error rates. The second simulation setting had

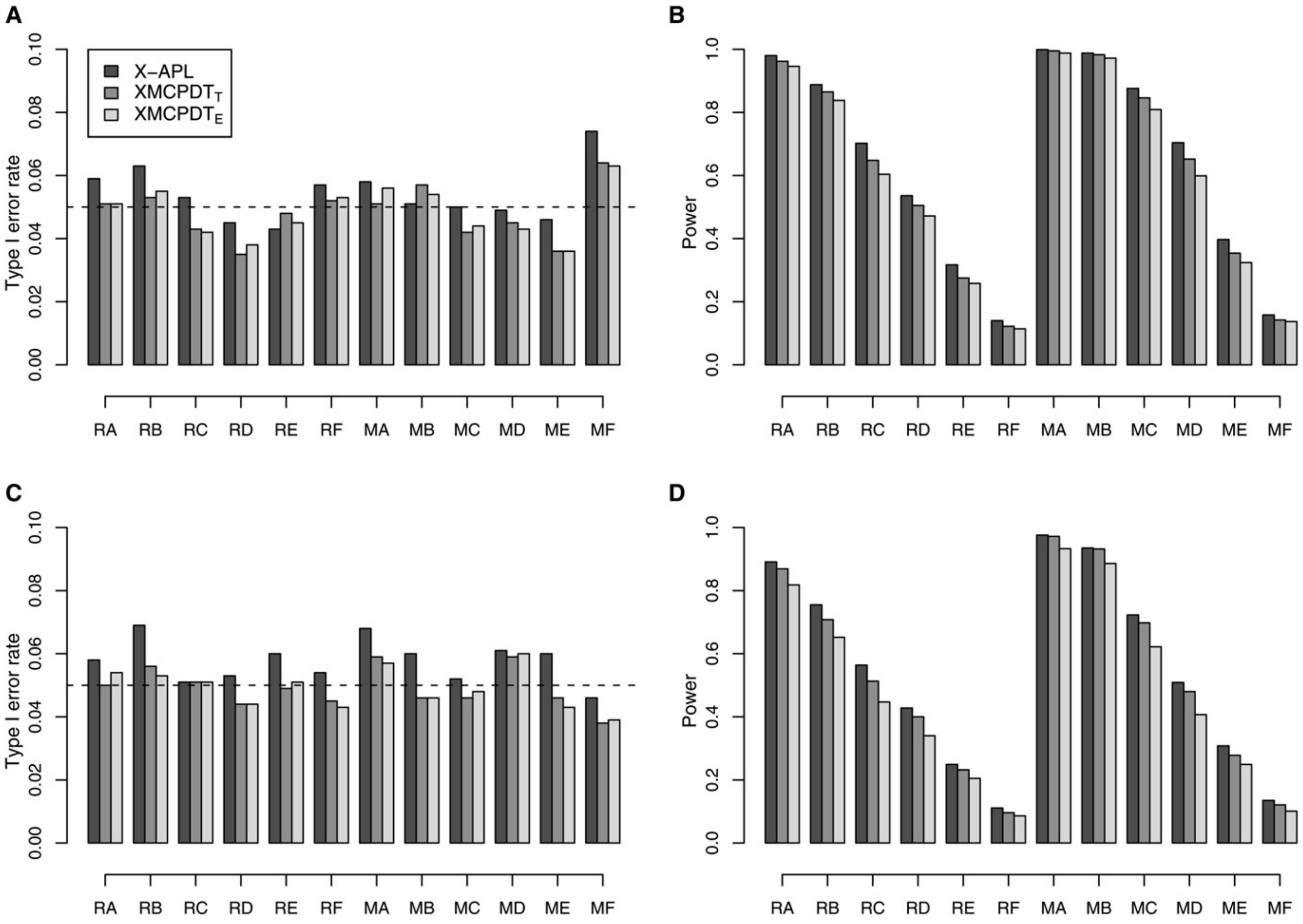

Figure 1. Comparisons of Type I Error Rates and Powers between X-APL and XMCPDT

(A and B) Type I error and power, respectively, for the setting with three-children families with at least one affected.

(C and D) Type I error and power, respectively, for the setting with OSUMS family structures.

The horizontal dashed line in (A) and (C) marks the nominal level of type I error rate. A total of 12 models were considered: recessive models RA–RF and multiplicative models MA–MF.[1] XMCPDT_T and XMCPDT_E are XMCPDT with true or estimated allele frequencies.

the same disease models but used the Ohio State University multiple sclerosis (OSUMS) pedigree structures.[2] There were a total of 81 pedigrees, with both nuclear and extended families. The total number of individuals was 386, and among them 102 were assumed to have missing genotypes. The same setups as described above were used for studying the type I error rates and powers. For each fixed family structure, we simulated genotypes and phenotypes for each member of the family. If there was no affected offspring in the family, we performed the same simulation again until the requirement was met. Then the genotypes of individuals that were missing in OSUMS were removed from our simulated data before performing the analyses. A total of 1000 replicates were simulated under each setting. Type I error rates and powers were calculated with X-APL and XMCPDT with the nominal level α set to be 0.05.

The results are shown in Figure 1. Under the first setting with only nuclear families, although X-APL had slightly higher powers than XMCPDT with either the true allele frequencies or the allele frequencies estimated from observed founder genotypes (Figure 1B), it also had larger actual type I error rates under most disease models (Figure 1A). In comparison of the two XMCPDT approaches, use of true allele frequencies gave slightly higher powers than use of estimated allele frequencies, but both gave very similar type I error rates, all around the nominal level. Under the second setting with the OSUMS pedigree structures, which included extended pedigrees, the power comparisons among the three methods were similar to those under the first setting (Figure 1D). On the other hand, although the type I error rates from XMCPDT were still very close to and around the nominal level, those from X-APL were higher than the nominal level for almost all of the disease models (Figure 1C).

From the results of our simulation study, we can see that, when the aforementioned assumption was not severely violated, XMCPDT using estimated allele frequencies does have appropriate type I error rates. However, if the amount of data available for estimating the frequencies is extremely limited, then the results could be affected. Although a sample size of one under each pedigree structure in the OSUMS data was used, XMCPDT appears to be robust because of the random nature in which each family was sampled. In

Table 1. Type I Error Rates for Multiplex Families

	Family Types[a]	
Method[b]	Three Children	Four Children
X-APL	0.056	0.052
XMCPDT$_T$	0.049	0.042
XMCPDT$_E$	0.049	0.041

[a] Two multiplex family scenarios were considered. Three and four children refer to families with three and four children, respectively, with at least two of them being affected.

[b] XMCPDT$_T$ and XMCPDT$_E$ refer to XMCPDT with true and estimated allele frequencies, respectively.

addition to the simulation detailed above, we also considered a hypothetical study focusing on multiplex families. We considered two scenarios, one with three children, two or three of them being affected, and the other with four children, at least two of them being affected. Either scenario clearly violated the sampling assumption, but the violation was not severe because not all families were forced to have exactly the same number of affected children. For both the three-children and the four-children families, XMCPDT with either true or estimated allele frequencies gave a p value of less than 0.05 (the nominal), demonstrating once again its robustness to slight departure from the assumption (Table 1). These results were based on 100 simulated families with the RecA model[1] and 4000 replicated runs. For each run, half of the families were assumed to have missing parental genotypes. We chose to perform much longer runs to obtain more accurate estimates of the actual type I error rates.

In contrast, for datasets with extended pedigrees, X-APL tends to have inflated type I error rates. The reason might be that when handling extended pedigrees, X-APL dissects them into nuclear families and analyzes them as if they were independent. However, whether this is the main reason remains unclear because explicit explanation on how extended pedigrees were handled was not available in Chung et al.[1] It is clear, though, that X-APL is a valid test only for nuclear families, and as such, it should not come as a surprise that it has inflated type I error rates when used for analysis of data from extended pedigrees. Perhaps X-APL and XMCPDT should not be viewed as competing approaches; rather, they should be viewed as complementary, utilizing their individual strengths. In particular, X-APL could be used for analyzing data from nuclear families, whereas data from extended pedigrees might be better treated with XMCPDT. For a dataset comprising both types of family, a combined analysis utilizing the strengths of both methods would be desirable.

Jie Ding[1] and Shili Lin[1,*]
[1]Department of Statistics, Ohio State University, Columbus, OH 43210, USA
*Correspondence: shili@stat.osu.edu

Acknowledgments

This work was supported in part by National Institutes of Health grant HG002657 and the Ohio State BRTT grant Tech 05-062.

Web Resources

The URLs for data presented herein are as follows:

X-APL, http://www.chg.duke.edu/research/software.html
XMCPDT, http://www.stat.osu.edu/~statgen/SOFTWARE/MC-PDT/

References

1. Chung, R., Morris, R.W., Zhang, L., Li, Y., and Martin, E.R. (2007). X-apl: An improved family-based test of association in the presence of linkage for the X chromosome. Am. J. Hum. Genet. *80*, 59–68.
2. Ding, J., Lin, S., and Liu, Y. (2006). Monte carlo pedigree disequilibrium test for markers on X chromosome. Am. J. Hum. Genet. *79*, 567–573.

DOI 10.1016/j.ajhg.2007.11.011. ©2008 by The American Society of Human Genetics. All rights reserved.

Response to Ding and Lin

To the Editor: In Chung et al.,[1] we reported simulation results showing that when a large fraction of families are missing parental genotypes, XMCPDT[2] can exhibit an inflated type I error rate. Ding and Lin dismiss the fraction of missing parental genotypes as an explanation for excess type I error and instead attribute our observation to violation of a sampling assumption of XMCPDT. They point out that our simulations condition on a fixed number of affected and unaffected offspring and note that this violates the XMCPDT assumption that family structure is random with respect to the number of affected offspring. To investigate this further, we performed a simulation study that allowed a variety of nuclear-family structures and varied the proportion of missing parent genotypes. Replicates of 300 families, each with three siblings, were generated via SIMLA[3] under an X-linked recessive disease model (RecF[1]). To ensure a variety of family phenotypes, we set disease prevalence to 0.3 and randomly sampled families with at least one affected sibling. Among 3000 replicates, the average proportions of families with one affected and two unaffected siblings, two affected and one unaffected siblings, and three affected siblings were 48%, 42%, and 10%, respectively. We believe that this simulation model achieves the family-ascertainment assumption of Ding et al.[2]

Figure 1 plots the relationship between type I error rate and the fraction of missing parental genotypes for XMCPDT, XPDT, and X-APL. Type I error rate increases