Editorial

Control Selection Options for Genome-Wide Association Studies in Cohorts

Sholom Wacholder and Melissa Rotunno

Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, Maryland

Abstract

Investigators planning studies within cohorts have many options for choosing an efficient sampling design for genome-wide association and other molecular epidemiology studies. Consideration of person-year and proportional hazards analyses of full cohorts may add further insight into ramifications of different designs. Empirical evidence from genome-wide association studies can supplement intuition and simulations in comparing properties of various case-control designs within cohorts. Additional theoretical and empirical work, justification of sampling choice in publications, and consideration of context and scientific aims can improve designs and, thereby, increase the scientific value and cost effectiveness of future studies. (Cancer Epidemiol Biomarkers Prev 2009;18(3):695–7)

The basic theory of sampling controls from cohorts to reduce cost of biospecimen collection and biomarker evaluation is well-established. Nonetheless, choosing the sampling design for a particular molecular epidemiology study is always challenging, and investigators choose among sampling schemes for reasons that are not always obvious or explained.

In *incidence-density sampling*, the sampling fraction, or chance of selection as a control, is the same for every member of the *risk set* of eligible cohort subjects at risk at the time of disease incidence or diagnosis of the case, regardless of whether he was previously selected as a control or will be diagnosed as a case in the future. As intended, the simulations in the paper by Wang et al. (1) should convince the skeptical genome-wide association (GWA) community of the advantages of incidence-density sampling. Wang and colleagues (1) show in a simple scenario that estimates of the relative risk from incidence-density sampling seem to be unbiased, even when estimates from two seemingly more natural sampling schemes are biased: (*a*) *Incidence-density sampling without replacement*, wherein a cohort member previously selected as a control is not eligible for further selection and (*b*) *Pure controls*, a design where controls are selected only from those who did not develop disease during follow-up. Wang et al. (1) support their case with an intuitive argument and citations to earlier work (2, 3).

Wang et al. (1) provide intuition by referring to the person-year method of analysis of a *full cohort*, i.e., one without any sampling. Considering proportional hazards analyses may add further insight (4). In full cohort studies and in efficient, unbiased designs, the estimate of the hazard ratio for the exposure of interest is based on the weighted differences between the exposure of each case and the weighted average of the exposures in the corresponding risk sets. Incidence-density sampling produces sets of controls that are independent random samples and, therefore, perfectly mimic the exposure distributions in the full cohort at the incidence times of the cases (4). On the other hand, the exposure distribution of controls in the risk sets chosen in designs that lead to bias does not reflect the exposure distribution of the risk set in the full cohort. Incidence-density sampling without replacement oversamples recent entrants to the cohort, who might be more likely or less likely to be exposed than long-term cohort members. Similarly, pure control sampling includes those alive and unaffected at the end of follow-up, notably long-term survivors, which will bias the sample in the direction of those less likely to become cases or to be censored when the exposure affects risk of disease or loss to follow-up.

Empirical studies from today's GWA studies, each with hundreds of thousands or more tests, can also complement theory, intuition, and simulation in evaluating performance of different design and analysis schemes. Empirical evaluation of bias and statistical and financial efficiency from many GWA studies and good cost estimates for various designs combined, eventually, with large numbers of solid positive findings for assessing power, may lead to better practice. Note, however, that comparing designs directly can be prohibitively expensive because different sets of controls would need to be selected and genotyped; still, empirical comparison of analytic methods is quite feasible. Just like other methods of evaluation, however, results of empirical comparisons may be inconsistent across different settings. Particularly important is the level and shape of the time-specific hazard of the study disease.

As a first effort at using empirical evidence from GWA studies, here, we present some simple, preliminary results showing the effect of ignoring time in analyses from two GWA studies in the Cancer Genetic Markers of

Received 11/20/08; accepted 11/24/08; published OnlineFirst 3/3/09.

Grant support: Intramural Research Program of the NIH, National Cancer Institute, Division of Cancer Epidemiology and Genetics.

Requests for reprints: Sholom Wacholder, NIH, National Cancer Institute, EPS 8046, 6120 Executive Boulevard, Rockville, MD 20892-7211. Phone: 301-496-3358; Fax: 301-402-0081. E-mail: sholom_wacholder@nih.gov

doi:10.1158/1055-9965.EPI-08-1114

Susceptibility project.[1] The Cancer Genetic Markers of Susceptibility prostate cancer GWA study (5) selected controls via incidence-density sampling; the breast cancer GWA study (6) used pure controls. For both analyses, we evaluated the associations between case-control status and the 530,000 single nucleotide polymorphisms (SNP) in unadjusted and age-adjusted unconditional logistic regression.

The scatterplot of the adjusted on the unadjusted logarithms of the *P* values shows only small perturbations from a straight line with slope 1 through the origin (data not shown). Figure 1 shows the cumulative distribution of the ratio of the age-adjusted and unadjusted estimates. Age-adjustment had little effect in the analysis; for example, only 1.7% and 5.2% of SNPs reported 10% or higher differences in *P* values after age adjustment for prostate and breast cancer studies, respectively. Use of incidence density sampling design for prostate cancer might explain the smaller apparent effect of age-adjustment on *P* value in prostate cancer than in breast cancer; alternatively, the difference between the ratios for the two cancer sites may reflect differences in the patterns of age-specific incidence rates for the two diseases.

We note that the magnitude of the changes in *P* value for the adjusted estimate compared with the unadjusted estimate seem greater when the direction of the change is toward a higher unadjusted *P* value. We interpret this observation as artifact of the shape of the cumulative normal distribution: an increase in a positive normal random variable has greater effect on the *P* value than does a decrease of the same magnitude. *In toto*, the empirical data seem to suggest that ignoring time in the analysis does not have much effect.

The small number of established SNP-disease associations precludes proper evaluation of the key question of power here. This empirical work has other very important limitations: lack of generalizability, inability to compare alternative designs directly and fully, and the reliance on only two studies.

Incidence-density sampling is valid under weak assumptions. The distribution of common, low-penetrance SNPs in GWA studies will not change substantially as the risk sets change with time unless the SNPs are strongly associated with the study disease, overall mortality, or enrollment or exit from the cohort. Thus, even theoretically flawed designs are unlikely to have much bias in the GWA context. Note that the relative risks of 1.5 and 2 in Wang et al.'s simulations (1) are greater than most of the estimates we see for markers with extremely low *P* values in GWA studies (7). Furthermore, routinely used q-q plots (5;6) may detect a seriously flawed design leading to an excess (or lack) of false positives.

A word about matching and conditional and unconditional regression. The analyses in Wang et al. (1) and in the empirical analyses here use unconditional logistic regression, although unbiased estimation of the hazard ratio requires conditional logistic regression. The incomplete adjustment via modeling age as a continuous or categorical variable may have small effects on bias and performance. Conditional logistic regression would provide unbiased estimates of the effects of a SNP when age is measured without error, regardless of the effect of age on risk, when age does not modify the effect of the SNP.

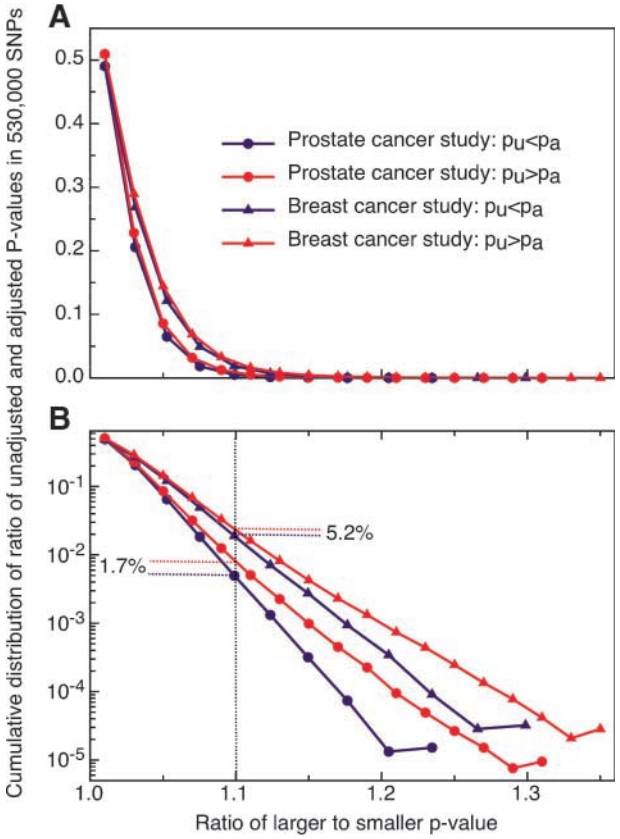

Figure 1. Empirical comparison of *P* values from unadjusted (p_u) and age-adjusted (p_a) *P* values for 530,000 SNPs from Cancer Genetic Markers of Susceptibility GWA studies of prostate (5) and breast (6) cancers. The figure displays the cumulative distribution of the ratio of the larger to the smaller *P* value. *Triangles,* graphs for SNPs from the breast cancer analysis; *dots,* graphs for SNPs from the prostate cancer analysis. Blue display is used when $p_u < p_a$; red display is used when $p_a < p_u$. **A.** Plot on a natural scale. **B.** The same data on a logarithm scale in base 10. The *P* value ratios tend to be higher for the breast cancer analysis than for the prostate cancer analysis. Data from builds 1 and 8[1] for prostate cancer and breast cancer studies, respectively.

The goal in GWA studies is identification of SNPs associated with disease. Type I error and coverage probabilities of the confidence intervals (Table 3 of Wang et al.; ref. 1) seem accurate for all designs. The bias away from the null gives increased power (Table 2 of Wang et al.; ref. 1) for incidence-density sampling without replacement and pure controls, which is an advantage for test-focused GWA studies; estimates of magnitude for significant effects are less important in GWA studies, and, anyway, estimates of effect highlighted because they met a *P* value threshold tend to be exaggerated and to

[1] http://cgems.cancer.gov/data/

regress to the mean (8) in independent replications; the phenomenon is sometimes called winner's curse (9, 10) in this context. Any difference in power, however, is likely inconsequential in GWA studies, where the relatively common genetic markers convincingly associated with disease seem to have weak effects, at least for cancer and diabetes (7).

The choice of the most appropriate design in any given setting is influenced by factors other than limited bias or efficiency of estimators under idealized assumptions. The choice of best design might be different when the same controls are used for studies of germ-line DNA and of biomarkers that change diurnally, monthly, seasonally, or with age, or have large measurement error (11). Furthermore, many sampling options may be more suitable at times. For example, case-cohort studies (12) have practical advantages: they are simplest to design, provide a random sample, and are most easily reusable for other disease end points, thereby lowering long-term costs (13).

The cost of large-scale genotyping continues to decrease. Cohorts are accumulating increasing numbers of relatively rare cancers and of other disease end points. The work of Wang et al. (1) helps understand design options for sampling from within cohorts. Additional theoretical and empirical work, justification of sampling choice in publications, and consideration of context and scientific aims can improve designs and, thereby, increase the scientific value and cost effectiveness of future studies.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

References

1. Wang MH, Shugart YY, Cole SR, Platz EA. A Simulation Study of Control Sampling Methods for Nested Case-Control Studies of Genetic and Molecular Biomarkers and Prostate Cancer Progression. Cancer Epidemiol Biomarkers Prev 2009;18:706–11.
2. Lubin JH, Gail MH. Biased selection of controls for case-control analyses of cohort studies. Biometrics 1984;40:63–75.
3. Robins JM, Gail MH, Lubin JH. More on ''Biased selection of controls for case-control analyses of cohort studies''. Biometrics 1986;42:293–9.
4. Wacholder S, Gail M, Pee D. Selecting an efficient design for assessing exposure-disease relationships in an assembled cohort. Biometrics 1991;47:63–76.
5. Yeager M, Orr N, Hayes RB, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007; 39:645–9.
6. Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007;39:870–4.
7. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest 2008;118:1590–605.
8. Yu K, Chatterjee N, Wheeler W, et al. Flexible design for following up positive findings. Am J Hum Genet 2007;81:540–51.
9. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology 2008;19:640–8.
10. Kraft P. Curses-winner's and otherwise-in genetic epidemiology. Epidemiology 2008;19:649–51.
11. Rundle AG, Vineis P, Ahsan H. Design options for molecular epidemiology research within cohort studies. Cancer Epidemiol Biomarkers Prev 2005;14:1899–907.
12. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 1986;73:1–11.
13. Wacholder S. Practical considerations in choosing between the case-cohort and nested case-control designs. Epidemiology 1991;2:155–8.

Control Selection Options for Genome-Wide Association Studies in Cohorts

Sholom Wacholder and Melissa Rotunno

Cancer Epidemiol Biomarkers Prev 2009;18:695-697.