Genetic Epidemiology 33: 717–728 (2009)

# Genome-Wide Association Scans for Secondary Traits Using Case-Control Samples

**Genevieve M. Monsees,[1]* Rulla M. Tamimi[1,2] and Peter Kraft[1,3]**

[1]*Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts*
[2]*Channing Laboratory, Brigham and Women's Hospital, Boston, Massachusetts*
[3]*Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts*

Genome-wide association studies (GWAS) require considerable investment, so researchers often study multiple traits collected on the same set of subjects to maximize return. However, many GWAS have adopted a case-control design; improperly accounting for case-control ascertainment can lead to biased estimates of association between markers and secondary traits. We show that under the null hypothesis of no marker-secondary trait association, naïve analyses that ignore ascertainment or stratify on case-control status have proper Type I error rates except when *both* the marker and secondary trait are independently associated with disease risk. Under the alternative hypothesis, these methods are unbiased when the secondary trait is not associated with disease risk. We also show that inverse-probability-of-sampling-weighted (IPW) regression provides unbiased estimates of marker-secondary trait association. We use simulation to quantify the Type I error, power and bias of naïve and IPW methods. IPW regression has appropriate Type I error in all situations we consider, but has lower power than naïve analyses. The bias for naïve analyses is small provided the marker is independent of disease risk. Considering the majority of tested markers in a GWAS are not associated with disease risk, naïve analyses provide valid tests of and nearly unbiased estimates of marker-secondary trait association. Care must be taken when there is evidence that both the secondary trait and tested marker are associated with the primary disease, a situation we illustrate using an analysis of the relationship between a marker in *FGFR2* and mammographic density in a breast cancer case-control sample. *Genet. Epidemiol.* 33:717–728, 2009. 

**Key words: genome-wide association studies; inverse probability weighting; ascertainment bias; biomarker; directed acyclic graph; case-control; efficiency**

Contract grant sponsor: US National Cancer Institute; Contract grant number: P01 CA087969; Contract grant sponsor: SPORE; Contract grant number: CA089393.
*Correspondence to: Genevieve M. Monsees, Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115. E-mail: gmonsees@hsph.harvard.edu
Received 18 June 2008; Revised 3 October 2008; Accepted 4 March 2009
Published online 13 April 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20424

## INTRODUCTION

Genome-wide association studies (GWAS) require massive investments of time and money, so researchers are understandably looking to maximize return by studying multiple traits collected on the same set of subjects. For example, recent GWAS for height [Lettre et al., 2008; Sanna et al., 2008; Weedon et al., 2007] and body mass index [Loos et al., 2008] were conducted using samples from multiple GWAS which were originally conducted to find markers associated with risk of diabetes, breast and prostate cancers, and other traits. An explicit goal of the NHGRI's dbGAP and PhenX projects is to facilitate the sharing and harmonization of data on multiple traits relevant to human health collected on samples that have already been genotyped as part of a GWAS. Considering the large sample sizes necessary to minimize the chances of both the false-positive and false-negative results when conducting GWAS [Chanock et al., 2007; Hunter and Kraft, 2007], combining existing GWAS data even when the trait under study was not the primary outcome of all (or perhaps even any) of the component studies is an attractive, quick and inexpensive option.

However, many GWAS to date have adopted a case-control design that implicitly conditions on a primary disease outcome. This non-random ascertainment from the study base can in principle lead to inflated Type I error rate for tests of association between markers and a secondary trait that either ignore ascertainment (i.e. treat the sample as representative of the entire study base) or "adjust for" ascertainment by conditioning on disease status (e.g. restricting analysis to controls or including case-control status as a covariate in a regression model). These commonly used analysis approaches can also provide biased estimates of the effect of marker genotypes on the secondary trait under the alternative hypothesis where there is a direct effect of the marker on the secondary trait.

Richardson et al. [2007] recently showed that if an exposure (in the GWAS context: marker genotype) and a secondary trait are each independently associated with a disease trait, then estimates of exposure-secondary trait association are typically biased. In response, we argued [Kraft, 2007] that for binary genotypes and binary

secondary traits this is the only situation where the odds ratio relating the secondary trait to genotype in the ascertained sample differed from the same odds ratio in the general population. In other words, if the marker is not associated with the primary disease or the secondary trait is not associated with the primary disease, the secondary trait-marker genotype odds ratio from the case-control study is not biased. Here we extend this result to general (ordinal, categorical) genotypes and general (ordinal, categorical, continuous) secondary traits. We demonstrate that bias can occur if the secondary trait is a cause of the primary disease. We show that common analytical approaches (including the common practice of restricting to controls) are not always effective at eliminating bias, and may in some circumstances increase the amount of bias observed. We illustrate that the degree of this bias is dependant on the rarity of the primary disease and the strength of the association between the secondary and primary traits as well as the association between the tested marker and the primary trait.

In the setting of most GWAS, where the magnitudes of the marker-phenotype association are often low, common analytical approaches will have appropriate Type I error rates and good power. Moreover, replication in further studies that have not been ascertained on the basis of the primary trait will help reduce the risk of false positives due to ascertainment bias. When the strength of the association between the secondary and primary traits is high, however, and the marker is thought to be associated with the primary trait—see for example the recent series of articles regarding the interrelatedness of 15q24/15q25.1, smoking behavior and lung cancer [Amos et al., 2008; Chanock and Hunter, 2008; Hung et al., 2008; Thorgeirsson et al., 2008]—special care will be needed to avoid bias in estimating the effect of the marker on the secondary trait. We address how to handle this type of situation.

Richardson et al. [Richardson et al., 2007] also proposed using inverse-probability-of-sampling-weighted (IPW) regression to estimate genotype-secondary trait association when case and control sampling fractions are available. These sampling fractions will be available for case-control studies nested within a prospective cohort, for example [Langholz et al., 1999]. We illustrate why this approach provides unbiased estimates of genotype-secondary trait association even when both the genotype and secondary trait are independently associated with primary disease outcome and verify that it does so via simulation. We also investigate the relative power for IPW regression and commonly used analyses in situations where the latter provide valid (or nearly valid) tests.

The outline of the remainder of this article is as follows. In the first section, we investigate the circumstances under which commonly used analyses that either ignore case-control ascertainment or stratify on disease status provide valid tests of the null hypothesis of no direct marker-secondary trait association and the circumstances where they provide unbiased estimates of marker-secondary trait association. We also describe IPW regression and provide some motivation for why it provides valid tests and unbiased estimates. In the second section we quantify Type I error rates, power and bias for commonly used analyses and IPW regression via simulation. In the third section we illustrate the various methods by applying them to a study of the association between an SNP in fibroblast growth factor receptor 2 (*FGFR2* [MIM 176943]) and mammographic density in a sample of breast cancer cases and controls. Finally, we discuss the implications of our qualitative and quantitative results for the design and analysis of GWAS and candidate gene studies of secondary traits using samples ascertained on the basis of another trait.

# TYPE I ERROR RATE AND BIAS: QUALITATIVE RESULTS

Figure 1 presents directed acyclic graphs (DAGs) for the joint, cross-sectional distribution of marker genotype(s) $G$, secondary trait $X$, primary (disease) phenotype $D$, and ascertainment (sampling) indicator $S$ in the study base (e.g. underlying cohort for a nested case-control study). Here $S = 1$ if a subject was selected to be in the case-control study, and $S = 0$ otherwise. We use these DAGs as convenient representations of the joint $G$, $X$, $D$ and $S$ distribution, based on conditional probabilities and assumptions of conditional independence [Greenland et al., 1999]. For example, the single arrow from $D$ to $S$ implies that $S$ is independent of $G$ and $X$, conditional on $D$. This is a reasonable assumption in many cases, but may not be true. (If participation rates vary by ethnicity, for example, then ascertainment may be associated with both $D$ and $G$; or the ascertainment scheme may directly depend on both $D$ and $X$, for example if controls in a prostate cancer study are required to have low PSA levels [Eeles et al., 2008].) We also assume there are no measured or unmeasured confounders of the $G$-$X$ relationship. It is possible these exist—again perhaps due to population stratification—but our goal here is to determine when the case-control sampling design per se induces bias in measures of genotype-secondary trait association. Finally, we assume there are no arrows leading from $X$, $D$ or $S$ into $G$, based on the directionality of any causal relationships among $G$, $X$ and $D$ (genetic variation causes traits, not vice versa).

We can make inferences about the independence or non-independence of $G$ and $X$ conditional on $D$ or $S$ using the topology of these DAGs and the following simple rules [Lauritzen et al., 1990; Robins et al., 2001]. We create a "moralized ancestral graph" for $X$ and $G$ and a set of conditioning variables C by first removing any variables (and any corresponding edges) not in AN[{X,G}∪C]. Here AN[{X,G}∪C] denotes the set of ancestors of variables in {X,G}∪C, where a variable is an ancestor of a second variable if there is a directed path from the first to the second. We then connect any two variables that share a common child (i.e. there is an arrow from each into a third

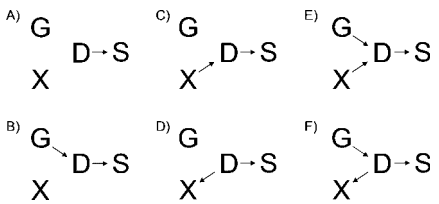


**Fig. 1. Directed acyclic graphs (DAGs) describing the joint probabilities and conditional independence structure for genotype (*G*), disease status (*D*), secondary trait (*X*) and sampling indicator (*S*) for the six scenarios described in the text.**

variable) with a non-directed edge. (The graph is ''moralized'' because all parents are ''married.'') Therefore, in scenario E the moralized ancestral graph for $X$ and $G$ conditional on $S$ would have a line connecting $G$ and $X$ to indicate that they share a common child ($D$); scenario C, however, would have no such line as $G$ is not a cause of $D$ in scenario C.

It can be shown that $G$ and $X$ are independent conditional on a set of variables C if and only if any path from $G$ to $X$ in the moralized ancestral graph passes through some variable in C [Lauritzen et al., 1990; Robins et al., 2001]. In scenarios A–D in Figure 1, there is no path from $G$ to $X$ in the moralized ancestral graph, so commonly used analyses that condition on $S$ (by restricting to a case-control sample but otherwise ignoring ascertainment) or condition on $D$ and $S$ (by including case-control status as a covariate in a regression model, or by restricting analysis to cases or controls) provide valid tests of the null hypothesis of $G$-$X$ independence. In scenario $E$, $G$ and $X$ are connected by an edge in the moralized ancestral graphs conditional on $S$ or {S,D}, so none of the commonly used analyses of case-control data provide valid tests of the null hypothesis. In scenario F, the analysis that conditions on $S$ but not $D$ (e.g. regressing $X$ on $G$ in a case-control sample but ignoring ascertainment) does not provide a valid test of the null hypothesis of no direct effect of $G$ on $X$, while analyses that condition on both $S$ and $D$ (by stratifying on case-control status or restricting to cases or controls) do provide valid tests of this hypothesis.

We note that for prospectively collected $X$ (e.g. for nested case-control studies where information on dietary patterns or anthropometry or plasma biomarkers etcetera are collected at baseline on a cohort of disease-free subjects) scenarios D and F are unlikely (but not impossible: some of the cohort may have had undiagnosed disease at baseline). These two scenarios are more likely when $X$ is collected retrospectively, after a subject's diagnosis (as in a classic case-control study); ''reverse causality'' is a concern in this context. Subject matter considerations may also rule out some of these scenarios; for example, onset of menarche predates diagnosis with breast cancer by several decades, so scenario F where disease status causes age at menarche in the general population is unrealistic.

These qualitative results on presence or absence of bias under the null can also be derived using the rules of conditional probability—in fact the results on DAGs and conditional independence used in the previous paragraphs are based on these rules [Lauritzen et al., 1990; Robins et al., 2001]. In Appendix A we use the basic rules of probability to infer the presence or absence of bias under the alternative, when there is a direct effect of genotype on the secondary trait (an arrow from $G$ to $X$). Conclusions about the presence or absence of bias in Scenarios A–F under the null or alternative summarized in Table I. Neither these calculations nor the DAGs are able to clearly illustrate the potential magnitude of the bias; therefore, we investigate the magnitude of bias via simulation in the next section.

Finally, the DAGs in Figure 1 also provide some insight into how IPW regression removes any bias induced by sampling conditional on $D$. If $G$ and $X$ were available on the entire study base, then regressing $X$ on $G$ (and *not* conditioning on either $D$ or $S$) provides a valid test of the null hypothesis of $G$-$X$ independence in scenarios A through E, because the corresponding moralized ancestral graphs do not contain a path from $G$ to $X$. (Scenario F is a special case where $X$ and $G$ are marginally associated, because $G$ influences $X$ through the intermediate $D$.) Loosely, IPW approximates the study base by upweighting each sampled subject so that he or she stands in for multiple individuals in the study base. For example, say there are 10,000 individuals in the study base, of whom 500 have the disease. Genotype and secondary trait information are collected on a case-control sample consisting of all 500 cases and 1,000 randomly selected controls. Then in an IPW analysis, each case stands in for his- or herself, while each control represents his- or herself and 8.5 other controls. Thus IPW regression ''recreates'' the study base and allows for unbiased estimation of the $G$-$X$ association. A variance correction is needed to account for the fact that subjects in the case-control sample are ''stand-ins'' for multiple subjects in the study-base; this correction is easily implemented in standard software (Appendix B). Although our discussion of IPW regression has been rather heuristic, we note that the procedure (and corresponding variance corrections) can be placed on solid theoretical grounds by considering the case-control study as a two-stage design [Reilly and Pepe, 1995; Siegmund et al., 1999] or missing data problem [Robins, 1995; Robins et al., 1995; Wacholder, 1996].

**TABLE I. Presence or absence of bias in measures of marker-secondary trait association under the null and alternative hypotheses of commonly used analyses that condition on case-control ascertainment or disease status**

| | Conditioning event C | |
|---|---|---|
| | $S$[a] | {S,D}[b] |
| *Scenario A* | | |
| Null | No bias | No bias |
| Alternative[c] | No bias | No bias |
| *Scenario B* | | |
| Null | No bias | No bias |
| Alternative | No bias | No bias |
| *Scenario C* | | |
| Null | No bias | No bias |
| Alternative | Bias[d] | Bias |
| *Scenario D* | | |
| Null | No bias | No bias |
| Alternative | Bias | Bias |
| *Scenario E* | | |
| Null | Bias | Bias |
| Alternative | Bias | Bias |
| *Scenario F* | | |
| Null | Bias | No bias |
| Alternative | Bias | Bias |

[a]Conditions only on ascertainment, $S = 1$; e.g. analyses that ignore case-control sampling.
[b]Conditions on ascertainment (by restricting to case-control sample) and case-control status $D$, e.g. analyses restricted to controls (or cases) or stratified by case-control status.
[c]Relationships among $G$, $X$ and $D$ as in the corresponding null scenario, with the addition of a directed edge from $G$ to $X$.
[d]Measures of the $G-X$ relationship conditional on $C$ may not reflect the $G-X$ relationship in the general population.

# TYPE I ERROR RATE AND BIAS: QUANTITATIVE RESULTS (SIMULATION STUDY)

To quantify the power and bias of commonly used analyses and IPW regression for secondary traits in a case-control sample, we simulated case-control studies drawn from an underlying cohort of size $n$. Diallelic genotypes $G_i$ for subjects $i = 1, \ldots, n$ were sampled assuming the cohort was in Hardy Weinberg Equilibrium with a minor allele frequency of 0.13 (carrier prevalence of 0.25). We focused our simulation study on the prospective scenarios in Figure 1 (A–C and E), so we sampled the continuous trait $X_i$ conditional on $G_i$. $X_i$ was drawn from a normal distribution with mean $= \beta_{XG} Z(G_i)$ and standard deviation $= 1$, where $Z(G_i)$ was an additive coding for the genotype ($Z(G_i)$ = number of minor alleles carried by subject (i)). Disease status $Di$ was sampled conditional on $G_i$ and $X_i$ as a Bernoulli random variable with $\text{logit}\{P(D_i = 1 \mid X_i, G_i)\} = \beta_0 + \beta_{DG} G_i + \beta_{DX} X_i$. Case-control samples were generated by sampling all $n_1$ cases from the simulated cohort and randomly sampling $n_1$ controls from the $n_0 = n - n_1$ disease-free subjects.

We simulated 108 scenarios, varying four parameters: disease prevalence $\kappa \in \{0.001, 0.01, 0.10, 0.20\}$; the percent of variance in $X$ explained by $G = r^2_{XG} \in \{0, 0.005, 0.01\}$; the increase in log odds of disease per copy of the minor allele $= \beta_{DG} \in \{0, \log(1.7)/2, \log 1.7\}$; and the increase in log odds of disease per unit change in $X = \beta_{DX} \in \{0, \log(2)/2, \log(2)\}$. The mean change in $X$ per copy of the minor allele ($\beta_{XG}$) and the baseline odds parameter $\beta_0$ were chosen to be consistent with $r^2_{XG}$ and $\kappa$, respectively. We varied the underlying cohort size $n$ so that on average approximately $n_1 = 1{,}000$ cases and $n_0 = n_1 = 1{,}000$ controls were sampled in all scenarios. A total of 5,000 replicate data sets were simulated for each scenario.

We conducted the following seven analyses for each simulated data set:

1. Full cohort analysis: simple linear regression of $X$ on $G$ in the full cohort.
2. "Unadjusted" case-control analysis: simple linear regression of $X$ on $G$ in the case-control sample (ignoring $D$).
3. "Adjusted" case-control analysis: regress of $X$ on $G$ and an indicator for case-control status D in the case-control sample.

**TABLE II. Type I error rates for six analyses of marker-secondary trait association in a case-control sample (and an analysis in the full underlying cohort); rare disease (1% prevalence)**

| | $OR_{DG} = 1$ | | | $OR_{DG} = 1.3$ | | | $OR_{DG} = 1.7$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\beta_{DX} = 0$ $OR_{DX} = 1$ | $\beta_{DX} = \ln(2)/2$ $OR_{DX} = 1.4$ | $\beta_{DX} = \ln 2$ $OR_{DX} = 2$ | $\beta_{DX} = 0$ $OR_{DX} = 1$ | $\beta_{DX} = \ln(2)/2$ $OR_{DX} = 1.4$ | $\beta_{DX} = \ln 2$ $OR_{DX} = 2$ | $\beta_{DX} = 0$ $OR_{DX} = 1$ | $\beta_{DX} = \ln(2)/2$ $OR_{DX} = 1.4$ | $\beta_{DX} = \ln 2$ $OR_{DX} = 2$ |
| 1. *Full cohort* | | | | | | | | | |
| $\alpha = 0.05$ | 0.049 | 0.051 | 0.049 | 0.051 | 0.053 | 0.051 | 0.049 | 0.057 | 0.053 |
| $\alpha = 0.01$ | 0.009 | 0.010 | 0.008 | 0.010 | 0.011 | 0.013 | 0.011 | 0.011 | 0.013 |
| $\alpha = 0.001$ | 0.001 | 0.001 | 0.001 | 0.002 | 0.001 | 0.001 | 0.002 | 0.001 | 0.002 |
| 2. *Unadjusted* | | | | | | | | | |
| $\alpha = 0.05$ | 0.052 | 0.053 | 0.049 | 0.050 | 0.078 | *0.152* | 0.051 | *0.172* | *0.469* |
| $\alpha = 0.01$ | 0.010 | 0.011 | 0.009 | 0.011 | *0.022* | *0.050* | 0.011 | *0.059* | *0.251* |
| $\alpha = 0.001$ | 0.001 | 0.001 | 0.001 | 0.001 | *0.003* | *0.010* | 0.002 | *0.013* | *0.081* |
| 3. *Adjusted* | | | | | | | | | |
| $\alpha = 0.05$ | 0.052 | 0.052 | 0.051 | 0.050 | *0.049* | *0.050* | 0.049 | *0.053* | *0.049* |
| $\alpha = 0.01$ | 0.010 | 0.009 | 0.008 | 0.011 | *0.011* | *0.009* | 0.011 | *0.013* | *0.010* |
| $\alpha = 0.001$ | 0.001 | 0.001 | 0.002 | 0.002 | *0.001* | *0.001* | 0.002 | *0.001* | *0.001* |
| 4. *Controls only* | | | | | | | | | |
| $\alpha = 0.05$ | 0.050 | 0.052 | 0.047 | 0.048 | *0.046* | *0.049* | 0.053 | *0.052* | *0.052* |
| $\alpha = 0.01$ | 0.012 | 0.010 | 0.010 | 0.010 | *0.009* | *0.007* | 0.011 | *0.010* | *0.011* |
| $\alpha = 0.001$ | 0.001 | 0.001 | 0.002 | 0.002 | *0.001* | *0.001* | 0.001 | *0.001* | *0.001* |
| 5. *Cases only* | | | | | | | | | |
| $\alpha = 0.05$ | 0.056 | 0.043 | 0.048 | 0.046 | *0.052* | *0.052* | 0.051 | *0.049* | *0.052* |
| $\alpha = 0.01$ | 0.011 | 0.010 | 0.010 | 0.009 | *0.008* | *0.013* | 0.010 | *0.010* | *0.009* |
| $\alpha = 0.001$ | 0.001 | 0.001 | 0.001 | 0.001 | *0.001* | *0.001* | 0.001 | *0.001* | *0.001* |
| 6. *Joint analysis* | | | | | | | | | |
| $\alpha = 0.05$ | 0.056 | 0.049 | 0.046 | 0.046 | *0.048* | *0.050* | 0.053 | *0.054* | *0.052* |
| $\alpha = 0.01$ | 0.010 | 0.010 | 0.010 | 0.012 | *0.009* | *0.008* | 0.011 | *0.010* | *0.010* |
| $\alpha = 0.001$ | 0.001 | 0.001 | 0.002 | 0.002 | *0.001* | *0.001* | 0.001 | *0.001* | *0.000* |
| 7. *IPW regression* | | | | | | | | | |
| $\alpha = 0.05$ | 0.058 | 0.046 | 0.048 | 0.049 | 0.054 | 0.055 | 0.054 | 0.049 | 0.051 |
| $\alpha = 0.01$ | 0.010 | 0.010 | 0.011 | 0.010 | 0.009 | 0.014 | 0.012 | 0.011 | 0.010 |
| $\alpha = 0.001$ | 0.001 | 0.002 | 0.002 | 0.002 | 0.001 | 0.002 | 0.001 | 0.001 | 0.001 |

Italicized cells represent scenarios that qualitative arguments suggest may be biased.

4. Controls only: simple linear regression of $X$ on $G$ among controls.
5. Cases only: simple linear regression of $X$ on $G$ among cases.
6. Joint analysis: regress $X$ on $G$, an indicator for $D$, and the product interaction term $D \times G$; test the joint null hypothesis that there is no main effect of $G$ and no $D \times G$ interaction (leads to a 2 d.f. test for association between $X$ and $G$ [Kraft et al., 2007]).
7. IPW regression: regress $X$ on $G$ using weights $w_1 = 1$ for cases and $w_0 = n_0/n_1$ for controls and apply appropriate variance correction (Appendix B).

The bias of each of the seven methods for each condition was obtained by subtracting the expected value $\beta_{XG}$ from the mean estimated $\beta_{XG}$. The probability of rejecting the null hypothesis under each of the methods and scenarios was estimated when applying nominal significance thresholds of $\alpha = 0.05$, 0.01 and 0.001.

## TYPE I ERROR AND POWER

Tables II and III summarize the Type I error rates across 18 of the null scenarios ($\beta_{XG} = 0$) we considered. (Results for $\kappa = 0.001$ and 0.20 are not shown but similar to those for $\kappa = 0.01$ and 0.10, respectively). As predicted by the theoretical considerations in the previous section, all of the analysis methods for the case-control data had appropriate Type I error rate as long as either $\beta_{DX} = 0$ or $\beta_{DG} = 0$, and the IPW regression analysis had appropriate Type I error rates for all of the scenarios considered. For rare disease ($\kappa = 0.01$; Table II), we did not detect any inflation in Type I error for analyses that condition on case-control status (Analyses 3–6) even when both X and $G$ directly influence disease; for more common disease ($\kappa = 0.10$; Table III) we did detect an inflation in Type I error rates for these analyses when $X$ and $G$ each directly influence disease. ''Unadjusted'' tests (Analysis 2) had noticeably inflated Type I error rates whenever $\beta_{DX} \neq 0$ and $\beta_{DG} \neq 0$, although the inflation decreases as the disease becomes more common (and the case-control sample becomes more representative of the cohort as a whole).

Figures 2 and 3 summarize the power of tests based on these seven analyses over all scenarios with $\kappa = 0.01$ and 0.10. Not surprisingly, analyses restricted to cases or controls generally have less power than any of the analyses that use the entire case-control sample. Due to the relatively large variance in parameters estimates from IPW regression, tests from Analysis 7 generally have the lowest power of the analyses that use the entire case-control sample. For more common disease ($\kappa = 0.10$ or 0.20 [not shown]), there are scenarios (namely when $\beta_{DX} \neq 0$ and $\beta_{DG} \neq 0$) where IPW has greater power than analyses that adjust for case-control status (Analyses 3 and 6), because the adjusted analyses are biased toward the null in these scenarios (Fig. 4). However, the IPW regression analysis still has lower power than the unadjusted analysis

**TABLE III. Type I error rates for marker-secondary trait association; common disease (10% prevalence)**

| | $OR_{DG} = 1$ | | | $OR_{DG} = 1.3$ | | | $OR_{DG} = 1.7$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\beta_{DX} = 0$ $OR_{DX} = 1$ | $\beta_{DX} = \ln(2)/2$ $OR_{DX} = 1.4$ | $\beta_{DX} = \ln 2$ $OR_{DX} = 2$ | $\beta_{DX} = 0$ $OR_{DX} = 1$ | $\beta_{DX} = \ln(2)/2$ $OR_{DX} = 1.4$ | $\beta_{DX} = \ln 2$ $OR_{DX} = 2$ | $\beta_{DX} = 0$ $OR_{DX} = 1$ | $\beta_{DX} = \ln(2)/2$ $OR_{DX} = 1.4$ | $\beta_{DX} = \ln 2$ $OR_{DX} = 2$ |
| 1. *Full cohort* | | | | | | | | | |
| $\alpha = 0.05$ | 0.052 | 0.051 | 0.052 | 0.052 | 0.045 | 0.049 | 0.051 | 0.051 | 0.047 |
| $\alpha = 0.01$ | 0.010 | 0.012 | 0.010 | 0.010 | 0.010 | 0.011 | 0.009 | 0.012 | 0.008 |
| $\alpha = 0.001$ | 0.002 | 0.002 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.002 |
| 2. *Unadjusted* | | | | | | | | | |
| $\alpha = 0.05$ | 0.051 | 0.052 | 0.047 | 0.055 | *0.059* | *0.077* | 0.052 | *0.084* | *0.148* |
| $\alpha = 0.01$ | 0.009 | 0.008 | 0.009 | 0.010 | *0.012* | *0.025* | 0.013 | *0.021* | *0.046* |
| $\alpha = 0.001$ | 0.001 | 0.001 | 0.000 | 0.001 | *0.001* | *0.003* | 0.001 | *0.003* | *0.009* |
| 3. *Adjusted* | | | | | | | | | |
| $\alpha = 0.05$ | 0.050 | 0.050 | 0.050 | 0.056 | *0.056* | *0.065* | 0.050 | *0.081* | *0.170* |
| $\alpha = 0.01$ | 0.010 | 0.009 | 0.010 | 0.010 | *0.011* | *0.015* | 0.012 | *0.020* | *0.058* |
| $\alpha = 0.001$ | 0.001 | 0.000 | 0.001 | 0.001 | *0.001* | *0.003* | 0.001 | *0.003* | *0.009* |
| 4. *Controls only* | | | | | | | | | |
| $\alpha = 0.05$ | 0.048 | 0.055 | 0.053 | 0.048 | *0.051* | *0.064* | 0.050 | *0.071* | *0.136* |
| $\alpha = 0.01$ | 0.011 | 0.013 | 0.008 | 0.011 | *0.010* | *0.013* | 0.009 | *0.015* | *0.043* |
| $\alpha = 0.001$ | 0.001 | 0.001 | 0.001 | 0.001 | *0.000* | *0.002* | 0.001 | *0.002* | *0.007* |
| 5. *Cases only* | | | | | | | | | |
| $\alpha = 0.05$ | 0.045 | 0.050 | 0.049 | 0.049 | *0.051* | *0.051* | 0.054 | *0.062* | *0.084* |
| $\alpha = 0.01$ | 0.008 | 0.011 | 0.010 | 0.011 | *0.010* | *0.011* | 0.011 | *0.016* | *0.020* |
| $\alpha = 0.001$ | 0.002 | 0.001 | 0.001 | 0.001 | *0.001* | *0.001* | 0.002 | *0.001* | *0.003* |
| 6. *Joint analysis* | | | | | | | | | |
| $\alpha = 0.05$ | 0.046 | 0.053 | 0.041 | 0.050 | *0.052* | *0.051* | 0.053 | *0.064* | *0.116* |
| $\alpha = 0.01$ | 0.010 | 0.010 | 0.006 | 0.010 | *0.008* | *0.010* | 0.011 | *0.014* | *0.033* |
| $\alpha = 0.001$ | 0.001 | 0.001 | 0.000 | 0.001 | *0.000* | *0.002* | 0.001 | *0.002* | *0.004* |
| 7. *IPW regression* | | | | | | | | | |
| $\alpha = 0.05$ | 0.047 | 0.051 | 0.052 | 0.053 | 0.052 | 0.048 | 0.051 | 0.051 | 0.050 |
| $\alpha = 0.01$ | 0.009 | 0.011 | 0.010 | 0.012 | 0.012 | 0.011 | 0.012 | 0.010 | 0.009 |
| $\alpha = 0.001$ | 0.002 | 0.001 | 0.001 | 0.001 | 0.001 | 0.002 | 0.001 | 0.001 | 0.002 |

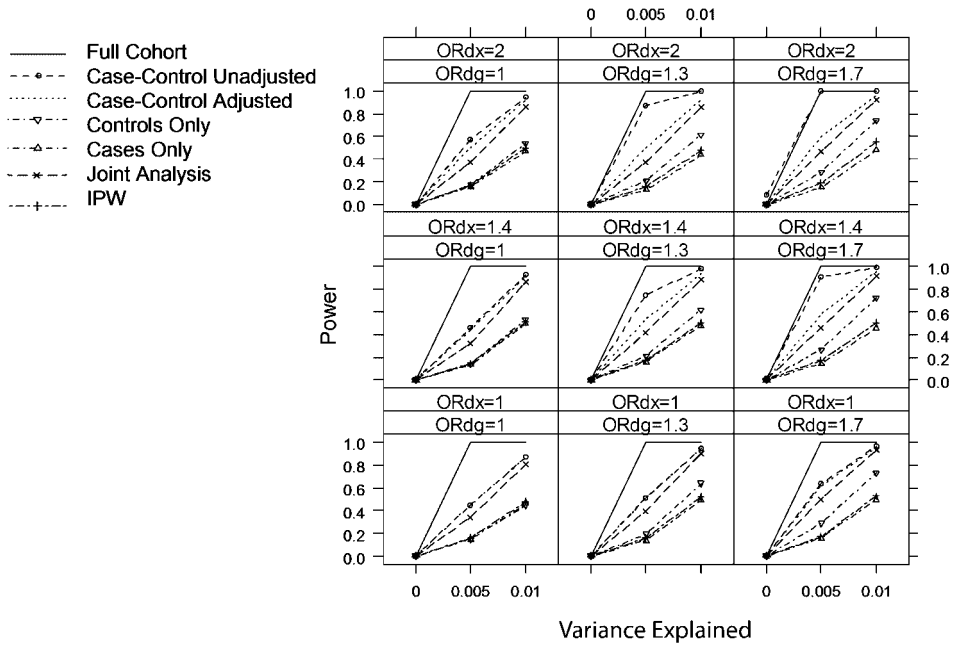


**Fig. 2. Power of six analyses of marker-secondary trait association in a case-control sample (and an analysis in the full underlying cohort); rare disease (prevalence $\kappa = 0.01$), nominal Type I error rate $\alpha = 0.001$.**

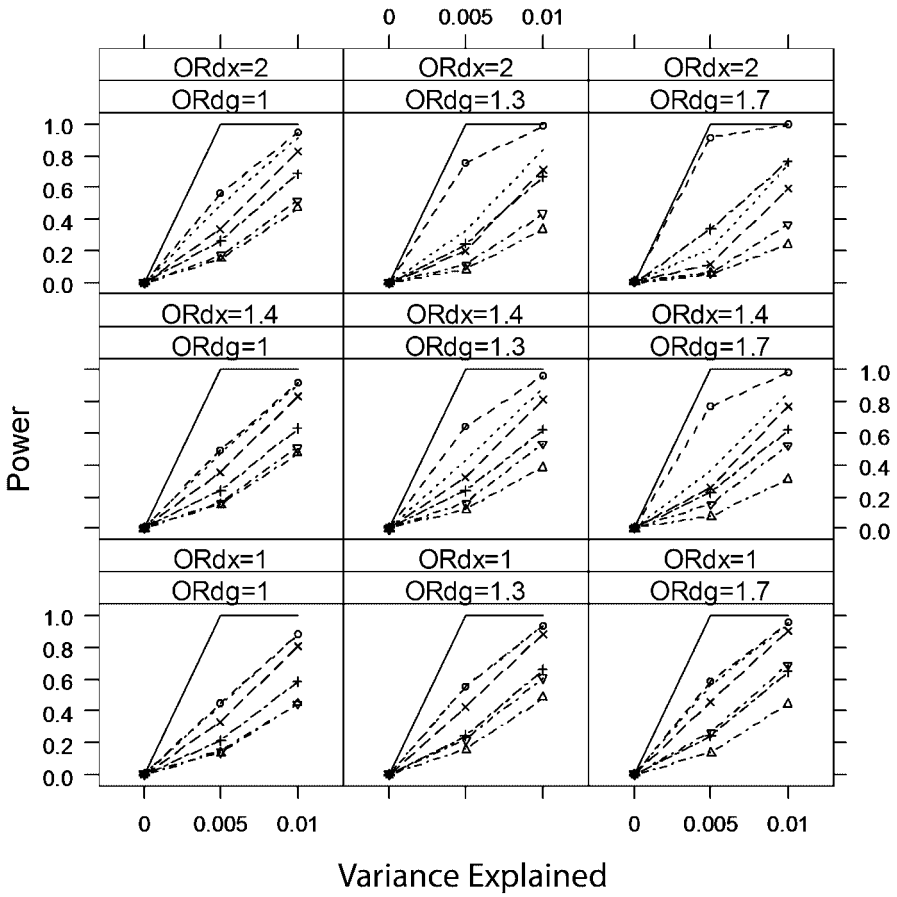


**Fig. 3. Power of six analyses of marker-secondary trait association in a case-control sample (and an analysis in the full underlying cohort); common disease (prevalence $\kappa = 0.10$), nominal Type I error rate $\alpha = 0.001$.**

(Analysis 2) in these scenarios. This is an unfair comparison, as the unadjusted analysis has inflated Type I error rate for tests of $\beta_{XG} = 0$ in these scenarios, but we note that the inflation is small (Table II) and the unadjusted analysis provides a valid test of the joint null hypothesis $\beta_{XG} = 0$ and $\beta_{DG} = 0$.

Among the commonly used analyses, the unadjusted analysis (Analysis 2) always has greater power than any adjusted analyses (Analyses 3-6), although we caution that for rare disease ($\kappa = 0.01$) the unadjusted analysis can have a strikingly inflated Type I error rate whenever both $\beta_{DX} \neq 0$ and $\beta_{DG} \neq 0$.

### BIAS

Figures 4 and 5 show the average bias over the simulated scenarios for rare ($\kappa = 0.01$) and common ($\kappa = 0.10$) disease, respectively. As predicted, none of the analyses are biased when $\beta_{DX} = 0$. When $\beta_{DX} \neq 0$, the commonly used analyses can be biased, although for rare disease, methods that condition on case-control status (Analyses 3–6) have no perceptible bias ($P > 0.05$) when $\beta_{DG} = 0$. The bias for the unadjusted method is small (magnitude $< 3.5\%$ of the true $\beta_{XG}$) when $\beta_{DX}$ is modest (a 1.3-fold increase in disease odds per standard deviation change in X). When both the secondary trait and the marker are independently associated with disease risk ($\beta_{DX} \neq 0$ and $\beta_{DG} \neq 0$), the unadjusted analysis can be quite biased.

For common disease, all methods except IPW regression have detectable bias when both $\beta_{DX} \neq 0$ and $\beta_{DG} \neq 0$. The unadjusted analysis overestimates $\beta_{XG}$, while the adjusted analyses underestimate $\beta_{XG}$. In particular, the analysis restricted to controls has the greatest magnitude of bias among the adjusted analyses. This may seem counter-intuitive; but we note that while for rare disease controls are fairly representative of the underlying population (even when $\beta_{DX} \neq 0$ and $\beta_{DG} \neq 0$), for common disease controls will not be representative of the underlying population if $\beta_{DX} \neq 0$ and $\beta_{DG} \neq 0$.

IPW regression has no detectable bias in any of the situations we simulated.

## EXAMPLE: FGFR2 AND MAMMOGRAPHIC DENSITY

To illustrate the application of these methods, we present several tests for association between the *FGFR2* SNP rs2981582 and pre-diagnostic mammographic density in a set of 790 breast cancer cases and 1,140 controls from the Nurses' Health Study. This study was approved by the Committee on the Use of Human Subjects in Research at Brigham and Women's Hospital.

Both *FGFR2* rs2981582 and mammographic density are known to be risk factors for breast cancer, and in fact both are significantly associated with risk in this data set

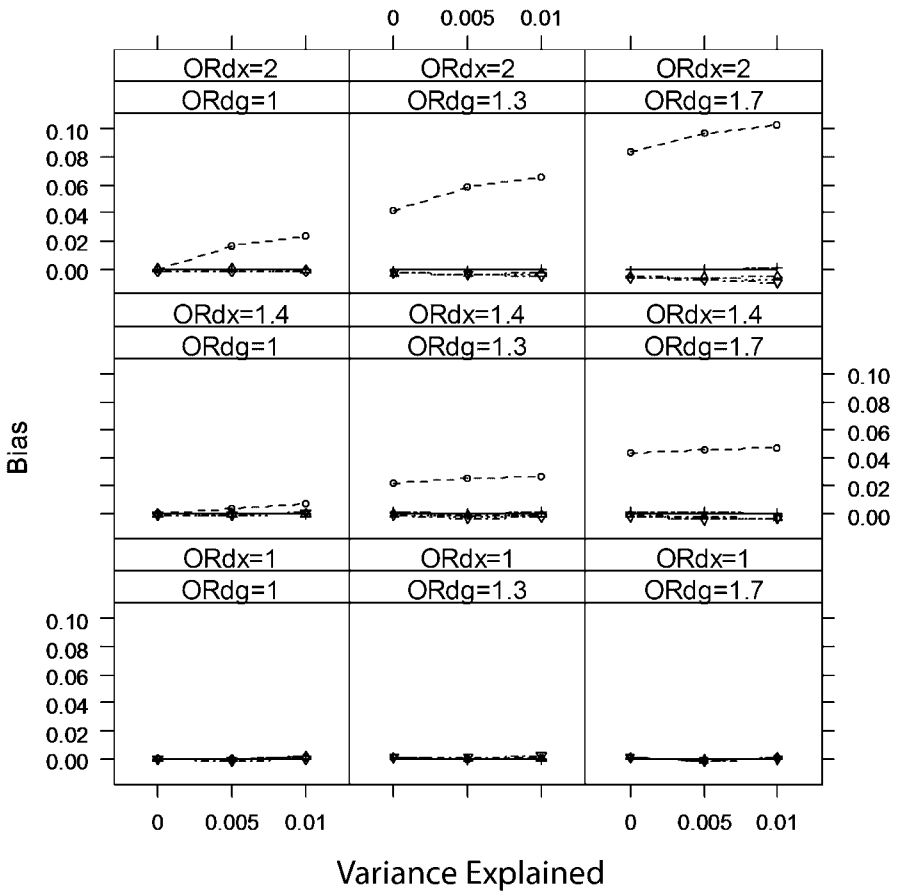


**Fig. 4. Average bias of analyses of marker-secondary trait association; rare disease (prevalence $\kappa = 0.01$).**

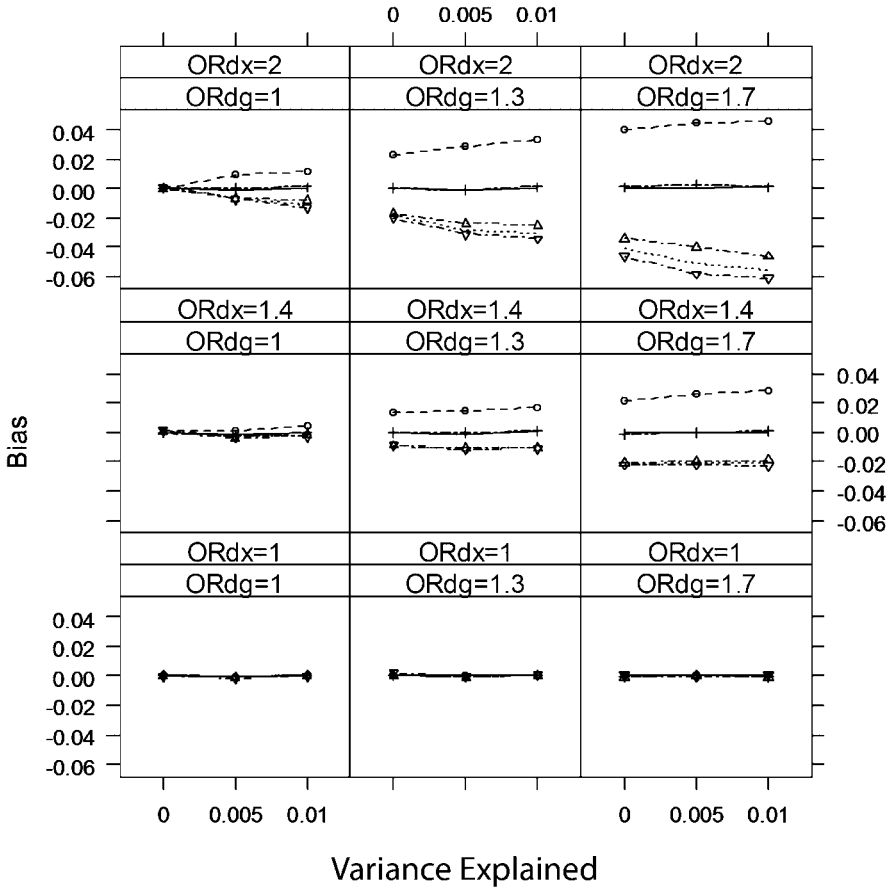


**Fig. 5. Average bias of analyses of marker-secondary trait association; common disease (prevalence $\kappa = 0.10$).**

[Easton et al., 2007; Hunter et al., 2007]. The crude, marginal per-minor-allele odds ratio of breast cancer risk for rs2981582 is 1.32 (CI: 1.16–1.51), and the crude, marginal odds ratio comparing the 80th percentile of mammographic density in controls to the 20th percentile was 2.33 (CI: 1.91–2.86). The association between rs2981582 and breast cancer risk remains significant ($P < 0.0001$) after adjusting for mammographic density, and the per-allele odds ratio changes little (OR: 1.34; CI: 1.17–1.52). Thus, based on our qualitative and quantitative results, this is a situation where the commonly used methods may provide biased estimates of the association between *FGR2* rs2981582 and mammographic density. However, as this sample was nested within a cohort study, we can estimate the case and control sampling fractions and conduct an IPW regression analysis.

The underlying cohort for this nested case-control study consisted of 29,625 women in the Nurses' Health Study who were cancer-free when they gave a blood sample from 1989 to 1990. Of the 934 eligible women in this cohort who developed breast cancer by 1998, 790 had available genotype and mammographic density data. Controls selected for genotyping were matched to individual cases on age and menopausal status at blood draw and were required to be breast-cancer-free up to the case's date of diagnosis; 1,440 genotyped controls had available mammographic density data. The sampling fractions by age, menopausal status at blood draw and case-control status are given in Table IV. Because controls were matched to cases on age and menopausal status, we use these stratified sampling fractions in the IPW regression analysis. (Controls were also matched on post-menopausal hormone use at blood draw and other variables related to plasma biomarker ascertainment—e.g. time of day of blood draw; for simplicity we ignore these factors in this illustration.)

We regressed square-root transformed mammographic density on rs2981582 under an additive model using six different methods (Analyses 2–7, listed above); the results are summarized in Table V. All of these methods yielded similar effect estimates, suggesting no direct association between rs2981582 and mammographic density. This is consistent with the results from our simulation study, which suggest that for a rare outcome and little or no association between the tested marker and the secondary trait, Analyses 3–7 have no detectable bias, while the bias for unadjusted case-control analysis (Analysis 2) is small, leading to a slight inflation in the Type I error rates. As expected, Analyses 2 and 3 (unadjusted and naively adjusted analyses) had the smallest standard errors, while methods that reduce sample size (case-only and control-only analyses) or adjust for weighting (IPW) yielded much larger standard errors. It is important to note that not applying the proper variance correction (Appendix B) in the IPW analysis would have falsely deflated our standard error from 0.096 to 0.073.

**TABLE IV. Sampling fractions for the *FGFR2*-mammographic density association analysis**

| | Cases in blood cohort | | Cases in blood cohort with genotypes and mammographic density | | Case sampling fractions | | Controls in blood cohort | | Controls in blood cohort with genotypes and mammographic density | | Control sampling fractions | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | Pre | Post | Pre | Post | Pre | Post | Pre | Post | Pre | Post | Pre | Post |
| 40–44 | 10 | 4 | 8 | 4 | 0.80 | 1.00 | 577 | 121 | 14 | 1 | 0.02 | 0.01 |
| 45–49 | 96 | 26 | 83 | 21 | 0.86 | 0.81 | 3710 | 1,309 | 117 | 31 | 0.03 | 0.02 |
| 50–54 | 95 | 80 | 83 | 73 | 0.87 | 0.91 | 2840 | 2,954 | 94 | 89 | 0.03 | 0.03 |
| 55–59 | 26 | 183 | 23 | 157 | 0.88 | 0.86 | 1037 | 5,045 | 28 | 227 | 0.03 | 0.04 |
| 60–64 | 1 | 218 | 1 | 183 | 1.00 | 0.84 | 396 | 5,445 | 0 | 285 | – | 0.05 |
| 65– | 0 | 195 | 0 | 154 | – | 0.79 | 179 | 5,054 | 0 | 254 | – | 0.05 |

**TABLE V. Association between *FGFR2* rs2981582 and mammographic density in a Nurses' Health Study breast-cancer case-control sample**

| Analytic approach | Mean difference | Standard error | *P*-value |
|---|---|---|---|
| 2. Unadjusted analysis | −0.0017 | 0.0681 | 0.97 |
| 3. Adjusted analysis | −0.0577 | 0.0672 | 0.39 |
| 4. Controls only | −0.0339 | 0.0922 | 0.71 |
| 5. Cases only | −0.0864 | 0.0979 | 0.37 |
| 6. Joint analysis | | | |
| (controls) | −0.0339 | 0.0922 | 0.14 |
| (cases) | −0.0864 | 0.0979 | |
| 7. IPW regression | 0.0055 | 0.0960 | 0.95 |

Mean difference is the per-minor-allele change in mean square-root transformed mammographic density, estimated using each of the six approaches outlined in the text.

# DISCUSSION

We have shown analytically (and verified through simulation) that under the null hypothesis of no direct association between marker genotypes and a secondary trait, several simple, commonly used analyses of genotype-secondary trait association in case-control samples have appropriate Type I error rates as long as either the marker or the secondary trait is not associated with disease risk in the study base. In particular, the naive analysis that ignores case-control ascertainment and treats the sample as if it were cross-sectional provides a valid test in these situations, as does the analysis that simply includes case-control status as a covariate in a regression of a secondary trait on marker genotype, and the analysis restricted to controls. Moreover, when the secondary trait is independent of disease risk, these simple analyses provide unbiased estimates of measures of marker-secondary trait association.

These results have important implications for genome-wide association scans of secondary traits, both those that have already been conducted—which have typically used one of these simple analyses—and those yet to come. For those secondary traits that are not associated with disease risk, these simple analyses are easily conducted using standard software for genome-wide association analyses [e.g. PLINK; Purcell et al., 2007] and provide valid tests and unbiased estimates of the marker-secondary trait association in the study base for all markers, even those associated with disease risk. There is no need to restrict analysis to controls to preserve test validity or remove bias—which is encouraging, as restricting to controls greatly reduces sample size and power.

For those secondary traits that are associated with disease risk—often the case, as researchers are interested in discovering the variants that influence risk factors for disease (e.g. mammographic density or age at menarche for breast cancer)—these standard approaches provide valid tests as long as the tested marker is not independently associated with disease risk (Fig. 1, Scenarios A, C and D). Considering that the vast majority of tested markers are in fact not independently associated with disease risk, these simple analyses will provide near-nominal control of the expected number of false-positive tests across all tested markers. For example, if 500 out of 500,000 markers are truly associated with disease risk, and we assume a 10-fold increase in the nominal Type I error rate of 0.001 for these 500 markers (larger than any inflation we observed in our simulation studies), then we expect $499{,}500 \times 0.001 + 500 \times 0.010 = 504.5$ significant tests instead of 500. Furthermore, any association seen in the initial scan should be replicated in an appropriate independent sample [Chanock et al., 2007]—preferably a cross-sectional sample, or perhaps a case-control study for a disease not associated with the secondary trait. Similarly, the fact that these simple analyses provide biased estimates of the marker-secondary trait association in the general population for those few makers that are associated with both disease risk and the secondary trait is not of vital importance in the genome-wide scan context, where the primary goal is to discover loci associated with the secondary trait. Larger, appropriately designed follow-up studies will be required to characterize the marker-secondary trait association in different populations.

The problem of increased Type I error rate for these simple analyses is arguably even less of an issue for meta-analyses of multiple genome-wide studies, some of which were originally case-control studies for different diseases related to the secondary trait, some of which were case-control studies for disease unrelated to the secondary trait, and some of which were cross-sectional or originally designed to study the secondary trait (i.e. the trait may be secondary for some studies and primary in others). In this situation, the bias under the null hypothesis will be reduced by averaging the biased estimates from some studies with the unbiased estimates from others. (Under the alternative, some of the heterogeneity in measures of marker-secondary trait association may be due to differences in ascertainment across studies.)

Throughout this article we have assumed that the probability of being sampled depends only upon case-control status. As we mentioned above, this will not be the case in some situations. However, in these situations the corresponding DAG can be drawn and inferences about the presence or absence of bias under the null hypothesis can be made using the rules we have outlined. For example, if controls are sampled conditional on the secondary trait (e.g. a prostate cancer case-control study where controls are required to have low PSA levels), the DAGs in Figure 1 can be modified by adding a directed arrow from $X$ to $S$.

An earlier letter [Kraft, 2007] examining the presence of bias in measures of association between marker genotypes and a secondary trait in case-control studies was restricted to the special case where the marker genotype and secondary trait were binary, and the measure of interest was the genotype-secondary trait odds ratio. The qualitative arguments we have made in this article (based either on DAG theory or conditional probabilities) make no assumptions about the distributional forms of the marker genotypes $G$ or secondary trait $X$. In particular, $G$ could represent an additive, dominant or general coding of a marker genotype, or indeed some coding for a multi-marker genotype, while $X$ could be binary, categorical or continuous; for continuous $X$ we did not assume normality or any other particular form for the distribution.

In contrast to scans testing many markers for association with a secondary trait, the bias induced by the case-control design will be a concern when the goal of the study is characterizing the relationship between a marker and a secondary trait that are both known to influence disease risk. In particular, some loci may have pleiotropic effects and influence disease risk both through their association with the secondary trait and another independent causal mechanism. For example, it is of great current interest to determine whether variants in the 15q25.1 region are associated with lung cancer indirectly, because of their association with smoking behavior, or if they are also associated with lung cancer along a second, distinct causal pathway [Chanock and Hunter, 2008]. We have shown that when sampling fractions are known, IPW regression provides unbiased estimators of measures of association between the marker and secondary trait in case-control studies. Further work is necessary to examine the degree of bias induced using incorrect sampling fractions, or developing more efficient, unbiased methods. Of course, when characterization of the secondary marker-trait association is a primary goal, the analytic problem can be avoided at the design phase; convenience samples from an existing GWAS may not be the best choice. The savings in genotyping costs for a small number of markers may be too small to offset concerns regarding analytic validity.

## ACKNOWLEDGMENTS

The authors thank Nan Laird for helpful discussion. This research was supported by a grant from the US National Cancer Institute: P01 CA087969 "Dietary and Hormonal Determinants of Cancer in Women." Dr. Tamimi was funded by the SPORE to look at genetics and mammographic density: SPORE in Breast Cancer CA089393.

## WEB RESOURCES

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/.

The database of Genotype and Phenotype (dbGAP), http://www.ncbi.nlm.nih.gov/sites/entrez?db = gap.

Consensus Measures for Phenotypes and Exposures (PhenX), https://www.phenx.org/.

## REFERENCES

Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, Sullivan K, Matakidou A, Wang Y, Mills G, Doheny K, Tsai YY, Chen WV, Shete S, Spitz MR, Houlston RS. 2008. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet 40:616–622.

Chanock SJ, Hunter DJ. 2008. Genomics: when the smoke clears. Nature 452:537–538.

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni Jr JF, Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, Collins FS. 2007. Replicating genotype-phenotype associations. Nature 447:655–660.

Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R; SEARCH collaborators, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, Johnson N, Seal S, Stratton MR, Rahman N, Chenevix-Trench G, Bojesen SE, Nordestgaard BG, Axelsson CK, Garcia-Closas M, Brinton L, Chanock S, Lissowska J, Peplonska B, Nevanlinna H, Fagerholm R, Eerola H, Kang D, Yoo KY, Noh DY, Ahn SH, Hunter DJ, Hankinson SE, Cox DG, Hall P, Wedren S, Liu J, Low YL, Bogdanova N, Schürmann P, Dörk T, Tollenaar RA, Jacobi CE, Devilee P, Klijn JG, Sigurdson AJ, Doody MM, Alexander BH, Zhang J, Cox A, Brock IW, MacPherson G, Reed MW, Couch FJ, Goode EL, Olson JE, Meijers-Heijboer H, van den Ouweland A, Uitterlinden A, Rivadeneira F, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Hopper JL, McCredie M, Southey M, Giles GG, Schroen C, Justenhoven C, Brauch H, Hamann U, Ko YD, Spurdle AB, Beesley J, Chen X, kConFab; AOCS Management Group, Mannermaa A, Kosma VM, Kataja V, Hartikainen J, Day NE, Cox DR, Ponder BA. 2007. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447:1087–1093.

Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, Mulholland S, Leongamornlert DA, Edwards SM, Morrison J, Field HI, Southey MC, Severi G, Donovan JL, Hamdy FC, Dearnaley DP, Muir KR, Smith C, Bagnato M, Ardern-Jones AT, Hall AL, O'Brien LT, Gehr-Swain BN, Wilkinson RA, Cox A, Lewis S, Brown PM, Jhavar SG, Tymrakiewicz M, Lophatananon A, Bryant SL; UK Genetic Prostate Cancer Study Collaborators; British Association of Urological Surgeons' Section of Oncology; UK ProtecT Study Collaborators, Horwich A, Huddart RA, Khoo VS, Parker CC, Woodhouse CJ, Thompson A, Christmas T, Ogden C, Fisher C, Jamieson C, Cooper CS, English DR, Hopper JL, Neal DE, Easton DF. 2008. Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet 40:316–321.

Greenland S, Pearl J, Robins JM. 1999. Causal diagrams for epidemiologic research. Epidemiology 10:37–48.

Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P,

Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Chen C, Goodman G, Field JK, Liloglou T, Xinarianos G, Cassidy A, McLaughlin J, Liu G, Narod S, Krokan HE, Skorpen F, Elvestad MB, Hveem K, Vatten L, Linseisen J, Clavel-Chapelon F, Vineis P, Bueno-de-Mesquita HB, Lund E, Martinez C, Bingham S, Rasmuson T, Hainaut P, Riboli E, Ahrens W, Benhamou S, Lagiou P, Trichopoulos D, Holcátová I, Merletti F, Kjaerheim K, Agudo A, Macfarlane G, Talamini R, Simonato L, Lowry R, Conway DI, Znaor A, Healy C, Zelenika D, Boland A, Delepine M, Foglio M, Lechner D, Matsuda F, Blanche H, Gut I, Heath S, Lathrop M, Brennan P. 2008. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 452:633–637.

Hunter DJ, Kraft P. 2007. Drinking from the fire hose–statistical issues in genomewide association studies. N Engl J Med 357:436–439.

Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni Jr JF, Hoover RN, Thomas G, Chanock SJ. 2007. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39:870–874.

Kraft P. 2007. Analyses of genome-wide association scans for additional outcomes. Epidemiology 18:838.

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. 2007. Exploiting gene-environment interaction to detect genetic associations. Hum Hered 63:111–119.

Langholz B, Rothman N, Wacholder S, Thomas DC. 1999. Cohort studies for characterizing measured genes. J Natl Cancer Inst Monogr 26:39–42.

Lauritzen S, Dawid A, Larsen B, Leimer H. 1990. Independence properties of directed Markov fields. Networks 20:491–505.

Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, Eyheramendy S, Voight BF, Butler JL, Guiducci C, Illig T, Hackett R, Heid IM, Jacobs KB, Lyssenko V, Uda M; Diabetes Genetics Initiative; FUSION; KORA; Prostate, Lung Colorectal and Ovarian Cancer Screening Trial; Nurses' Health Study; SardiNIA, Boehnke M, Chanock SJ, Groop LC, Hu FB, Isomaa B, Kraft P, Peltonen L, Salomaa V, Schlessinger D, Hunter DJ, Hayes RB, Abecasis GR, Wichmann HE, Mohlke KL, Hirschhorn JN. 2008. Identification of ten loci associated with height highlights new biological pathways in human growth. Nat Genet 40:584–591.

Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, Inouye M, Freathy RM, Attwood AP, Beckmann JS, Berndt SI; Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, Jacobs KB, Chanock SJ, Hayes RB, Bergmann S, Bennett AJ, Bingham SA, Bochud M, Brown M, Cauchi S, Connell JM, Cooper C, Smith GD, Day I, Dina C, De S, Dermitzakis ET, Doney AS, Elliott KS, Elliott P, Evans DM, Sadaf Farooqi I, Froguel P, Ghori J, Groves CJ, Gwilliam R, Hadley D, Hall AS, Hattersley AT, Hebebrand J, Heid IM; KORA, Lamina C, Gieger C, Illig T, Meitinger T, Wichmann HE, Herrera B, Hinney A, Hunt SE, Jarvelin MR, Johnson T, Jolley JD, Karpe F, Keniry A, Khaw KT, Luben RN, Mangino M, Marchini J, McArdle WL, McGinnis R, Meyre D, Munroe PB, Morris AD, Ness AR, Neville MJ, Nica AC, Ong KK, O'Rahilly S, Owen KR, Palmer CN, Papadakis K, Potter S, Pouta A, Qi L; Nurses' Health Study, Randall JC, Rayner NW, Ring SM, Sandhu MS, Scherag A, Sims MA, Song K, Soranzo N, Speliotes EK; Diabetes Genetics Initiative, Syddall HE, Teichmann SA, Timpson NJ, Tobias JH, Uda M; SardiNIA Study, Vogel CI, Wallace C, Waterworth DM, Weedon MN; Wellcome Trust Case Control Consortium, Willer CJ; FUSION, Wraight, Yuan X, Zeggini E, Hirschhorn JN, Strachan DP, Ouwehand WH, Caulfield MJ, Samani NJ, Frayling TM, Vollenweider P, Waeber G, Mooser V, Deloukas P, McCarthy MI, Wareham NJ, Barroso I, Jacobs KB, Chanock SJ, Hayes RB, Lamina C, Gieger C, Illig T, Meitinger T, Wichmann HE, Kraft P, Hankinson SE, Hunter DJ, Hu FB, Lyon HN, Voight BF, Ridderstrale M, Groop L, Scheet P, Sanna S, Abecasis GR, Albai G, Nagaraja R, Schlessinger D, Jackson AU, Tuomilehto J, Collins FS, Boehnke M, Mohlke KL. 2008. Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet 40:768–775.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575.

Reilly M, Pepe MS. 1995. A mean score method for missing and auxiliary covariate data in regression models. Biometrika 82:299–314.

Richardson DB, Rzehak P, Klenk J, Weiland SK. 2007. Analyses of case-control data for additional outcomes. Epidemiology 18:441–445.

Robins JM. 1995. Semiparametric efficiency in multivariate regression models with missing data. J Am Stat Assoc 90:122–129.

Robins JM, Rotnitzky A, Zhao L. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc 90:106–121.

Robins JM, Smoller JW, Lunetta KL. 2001. On the validity of the TDT test in the presence of comorbidity and ascertainment bias. Genet Epidemiol 21:326–336.

Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, Bonnycastle LL, Shen H, Timpson N, Lettre G, Usala G, Chines PS, Stringham HM, Scott LJ, Dei M, Lai S, Albai G, Crisponi L, Naitza S, Doheny KF, Pugh EW, Ben-Shlomo Y, Ebrahim S, Lawlor DA, Bergman RN, Watanabe RM, Uda M, Tuomilehto J, Coresh J, Hirschhorn JN, Shuldiner AR, Schlessinger D, Collins FS, Davey Smith G, Boerwinkle E, Cao A, Boehnke M, Abecasis GR, Mohlke KL. 2008. Common variants in the GDF5-UQCC region are associated with variation in human height. Nat Genet 40:198–203.

Siegmund KD, Whittemore AS, Thomas DC. 1999. Multistage sampling for disease family registries. J Natl Cancer Inst Monogr 26:43–48.

Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, Manolescu A, Thorleifsson G, Stefansson H, Ingason A, Stacey SN, Bergthorsson JT, Thorlacius S, Gudmundsson J, Jonsson T, Jakobsdottir M, Saemundsdottir J, Olafsdottir O, Gudmundsson LJ, Bjornsdottir G, Kristjansson K, Skuladottir H, Isaksson HJ, Gudbjartsson T, Jones GT, Mueller T, Gottsäter A, Flex A, Aben KK, de Vegt F, Mulders PF, Isla D, Vidal MJ, Asin L, Saez B, Murillo L, Blondal T, Kolbeinsson H, Stefansson JG, Hansdottir I, Runarsdottir V, Pola R, Lindblad B, van Rij AM, Dieplinger B, Haltmayer M, Mayordomo JI, Kiemeney LA, Matthiasson SE, Oskarsson H, Tyrfingsson T, Gudbjartsson DF, Gulcher JR, Jonsson S, Thorsteinsdottir U, Kong A, Stefansson K. 2008. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 452:638–642.

Wacholder S. 1996. The case-control study as data missing by design: estimating risk differences. Epidemiology 7:144–150.

Weedon MN, Lettre G, Freathy RM, Lindgren CM, Voight BF, Perry JR, Elliott KS, Hackett R, Guiducci C, Shields B, Zeggini E, Lango H, Lyssenko V, Timpson NJ, Burtt NP, Rayner NW, Saxena R, Ardlie K, Tobias JH, Ness AR, Ring SM, Palmer CN, Morris AD, Peltonen L, Salomaa V; Diabetes Genetics Initiative; Wellcome Trust Case Control Consortium, Davey Smith G, Groop LC, Hattersley AT, McCarthy MI, Hirschhorn JN, Frayling TM. 2007. A common variant of HMGA2 is associated with adult and childhood height in the general population. Nat Genet 39:1245–1250.

# APPENDIX A

In general, the probability of $X$ conditional on $G$ and $S$ in scenarios where there is no edge connecting $X$ to $D$ in the Directed Acyclic Graph is

$$\Pr(X|G,S) = \frac{\sum_D \Pr(S|D)\Pr(D|X,G)\Pr(X|G)}{\sum_{D,X} \Pr(S|D)\Pr(D|X,G)\Pr(X|G)}. \quad \text{(A1)}$$

Under the assumption that $\Pr(D \mid X,G) = \Pr(D \mid G)$ (i.e. in scenarios A and B, or when there is a directed arrow from $G$ to $X$ but no edge connecting $X$ and $D$) expression (A1) reduces to $\Pr(X \mid G)$—so the distribution of $X$ in the case-control sample is identical to that in the underlying study base. Under the assumption that $\Pr(X \mid G) = \Pr(X)$ and $\Pr(D \mid X,G) = \Pr(D \mid X)$ (Scenario C), expression (A1) simplifies to

$$\frac{\sum_D \Pr(S|D)\Pr(D|X)\Pr(X)}{\sum_{D,X} \Pr(S|D)\Pr(D|X)\Pr(X)}.$$

Although this is not identical to the distribution of $X$ in the study base, $X$ remains independent of $G$, so tests of $G-X$ independence should have valid Type I error rates. However, when there is a directed edge from $G$ to $X$, expression (A1) becomes

$$\frac{\sum_D \Pr(S|D)\Pr(D|X)\Pr(X|G)}{\sum_{D,X} \Pr(S|D)\Pr(D|X)\Pr(X|G)}, \qquad \text{(A2)}$$

which does not in general simplify to $\Pr(X \mid G)$. Thus, under the alternative the distribution of $G$ and $X$ in the ascertained sample is not identical to the distribution in the study base, and measures of the $G-X$ association in the case-control sample (or restricted to either controls or cases) will in general be biased, i.e. they will not reflect the $G-X$ relationship in the study base. (The odds ratio relating a binary exposure $X$ to a binary genotype $G$ is one exception to this rule [Kraft, 2007].) The magnitude of the bias is not clear from expression (A2); we investigate the magnitude of bias via simulation.

Similar calculations lead to conclusions about the presence or absence of bias in Scenarios A–F under the null or alternative summarized in Table I.

# APPENDIX B

The following SAS and R code perform IPW regression with appropriate variance correction (via the weight statement in PROC GENMOD or weights option in the geeglm() function. Here id is a unique subject identifier; $X$ is a continuous trait; $G$ is some coding for marker genotypes (e.g. count of minor alleles); and weight is the (known) inverse probability of being sampled into the case-control data set.

**SAS**

```
proc genmod data = fgfrw;
  class id;
  model X = G;
  weight weight;
  repeated subject = id;
run;
```

**R**

```
library(geepack)
ipw.fit ← geeglm(X–G,weights = weight,id = id)
```