THE UNIVERSITY OF BRITISH COLUMBIA

DEPARTMENT OF STATISTICS

TECHNICAL REPORT #252

Bayesian inference of gene-environment interaction from incomplete data: What happens when information on environment is disjoint from data on gene and disease?

Paul Gustafson
Igor Burstyn

September 2009

Bayesian inference of gene-environment interaction from incomplete data: what happens when information on environment is disjoint from data on gene and disease?

Paul Gustafson[1,*] and Igor Burstyn[2]

1: Department of Statistics, University of British Columbia, 333-6356 Agricultural Road, Vancouver, B.C., V6T 1Z2, Canada

2: Community and Occupational Medicine Program, Department of Medicine, Faculty of Medicine and Dentistry, The University of Alberta, 13-103E Clinical Sciences Building, Edmonton, Alberta, T6G 2G3, Canada

* gustaf@stat.ubc.ca

July 23, 2009

SUMMARY. Inference in gene-environment studies can sometimes exploit the assumption of Mendelian randomization that genotype and environmental exposure are independent in the population under study. Moreover, in some such problems it is reasonable to assume that the disease risk for subjects without environmental exposure will not vary with genotype. When both assumptions can be invoked, we consider the prospects for inferring the dependence of disease risk on genotype and environmental exposure (and particularly the extent of any gene-environment interaction), without detailed data on environmental exposure. The data structure envisioned involves data on disease and genotype jointly, but only external information about the distribution of the environmental exposure in the population. This is relevant as for many environmental exposures individual-level measurements are costly and/or highly error-prone. Working in the setting where all relevant variables are binary, we examine the extent to which such data is informative about the interaction, via determination of the large-sample limit of the posterior distribution. Comparisons are drawn with inferences based on joint measurements of disease, genotype, and error-prone exposure. The ideas are illustrated using data from a case-control study for bladder cancer involving smoking behaviour and the NAT2 genotype.

KEY WORDS: Bayesian inference; Case-control study; Exposure misclassification; Gene-environment interaction; Mendelian randomization; Non-identifiable model; Semi-ecological design.

1 Introduction

In epidemiology and toxicology, 'exposure' refers to a characteristic of the environment that a subject encounters through behaviour (e.g. exposure to cigarette smoke), while 'biologically effective dose' is the amount of toxic substance that reaches a target organ or tissue in the subject's body following exposure and metabolism (e.g. the mass of aromatic amines that reaches the bladder after bio-activation by *N*-acetyl transferase (NAT) enzymes). Exposure is usually assumed to be independent of person's genetic make-up, with some notable exceptions where genetics affects behavior, such as alcohol consumption in persons who are inebriated easily due to genetic susceptibility (Davey Smith & Ebrahim 2004), or cruciferous vegetable intake for persons with a particular haplotype for bitter-taste response (Sacerdote et al. 2007). In contrast, biologically effective dose, which lies exclusively on the causal pathway between exposure and disease, clearly depends on genetics for some exposures.

The Mendelian randomization assumption, which we take in the most generic form to simply assert the random allocation of alleles to individuals, has been used for varied purposes in epidemiology. Where genetics affects behaviour (i.e. exposure) but can reasonably be assumed independent of confounders (whether measured or not), there is the potential to infer the causal exposure-disease relationship by using genotype as an instrumental variable (e.g. Didelez and Sheehan 2007, Lawlor et al. 2008). Or, in rare situations where the biologically effective dose is an observable quantity, genotype can be used as an instrumental variable to estimate the relationship between biologically effective dose (also know as the intermediate phenotype) and disease. Thomas and Conti

(2004) emphasize such application when health outcome is measured on a continuous scale and point to some complications that arise when the disease state is binary.

A parallel set of methodological developments apply to situations where genotype and exposure can be assumed independent, but a gene-environment interaction may arise via a relationship between genotype and the (unobservable) biologically effective dose. Methods that utilize the assumption of gene-exposure independence can more efficiently estimate the magnitude of the interaction than those which do not invoke the assumption (Umbach and Weinberg 1997, Chatterjee and Carroll 2005, Mukherjee et al 2007, Mukherjee and Chatterjee 2008).

In yet another related line of inquiry, Burstyn et al. (2009), motivated by the study of Cherry et al. (2002), consider settings where gene and exposure can be assumed to be independent, and the lack of a main effect of gene in the model for disease given exposure and gene can also be assumed (i.e. gene alone does not confer disease risk in the absence of exposure). They give a frequentist procedure for testing the existence of gene-exposure interaction, using gene and disease data only. No detailed exposure data are used beyond an assumption that all subjects are exposed to some unknown extent by virtue of shared environment (e.g. occupation). Such a procedure appears to be especially advantageous when exposure can only be assessed with error (Burstyn et al. 2009).

Focussing on the situation where all variables are binary, the objective of this work is to develop Bayesian methods for estimating gene-environment interaction, in the situation where information on exposure is disjoint from that on gene and disease. That is, exposure is not determined for each subject, but knowledge at an ecological level is

available (e.g. exposure prevalence in the population). The assumption of gene-exposure independence is exploited, as is the assumption that gene alone, in absence of exposure, does not cause the disease. Both prospective and retrospective analyses are considered. We also explore the robustness of the proposed method to misspecified information on exposure prevalence, and provide comparison with analysis in which exposure is measured imprecisely for each subject. Lastly, we investigate the sensitivity of the procedure to violation of the assumption that gene has no effect in absence of exposure. The method is illustrated using study data from Gu et al. (2005).

2 Models and Theory

2.1 General Framework

Let Y be the binary outcome, X the binary exposure (or `environment' variable), and G the binary `gene' variable. We consider situations where two key assumptions are defensible. The first is that gene alone confers no additional disease risk in absence of exposure, so that the model

$$\text{logit } \Pr(Y=1|X,G) = \beta_0 + \beta_x X + \beta_{xg} XG \qquad (1)$$

is thought to hold in the population of interest. The second is the Mendelian randomization assumption that X and G are independent of one another in the study population. Logistical challenges in exposure assessment may imply that (Y,G,X) data are difficult or impossible to obtain (e.g. historical exposure over many years, important for a chronic disease, cannot be assessed with a reasonable degree of accuracy), whereas (Y,G) data are easier to obtain (e.g. genotyping subjects at time of diagnosis) and reliable information exists about the population prevalence of X, denoted as *r* hereafter.

Alternatively, perhaps (Y,X^*,G) data can be obtained on individuals, where X^* is an imperfect surrogate for X.

From the point of view of testing the null hypothesis $\beta_{xg} = 0$, Burstyn et al. (2009) illustrate through simulations that under the above two key assumptions, one can perform a valid test using (Y,G) data alone, since association between Y and G arises if and only if $\beta_{xg} \neq 0$. We follow this up by investigating, from an estimation point of view, the extent to which (Y,G) data are informative about β_{xg}.

2.2 Large-Sample Limit of the Posterior Distribution

Our main theoretical, proof-of-concept investigation regarding information about β_{xg} is as follows. In addition to the two key assumptions, say that *r* is known. For a given specification of prior distribution on $\beta=(\beta_0, \beta_x, \beta_{xg})$ and a given set of true values for these parameters, we determine the posterior distribution arising from an infinite sample of (Y,G) data (but no observation of X values). The extent to which this *limiting posterior distribution* (LPD) on β_{xg} is narrower than the prior distribution on β_{xg} then quantifies the extent to which (Y,G) data are informative about β_{xg}. Of course, the LPD represents the `best possible' answer achieved in the limit of the sample size tending to infinity. Therefore, a wide LPD indicates futility in using (Y,G) data to infer the target β_{xg}. On the other hand, a narrow LPD indicates plausibility for inference, but does not directly address the secondary question of how quickly the posterior based on *n* data points narrows to the LPD as *n* increases.

We also note that the large-sample limit of inference based on (Y,G) data plus knowledge of the population exposure prevalence can be equivalently regarded as the limit of inference arising from data on (Y,G) and (X) marginally rather than (Y,G,X)

jointly (see Umbach and Weinberg 1997 for a discussion of a somewhat related "scrambled data" problem). Of course, inference based on (Y,X,G) jointly will enjoy regular properties, such as the posterior on any parameter converging to a single and correct point as the sample size grows. Consequently, our results can be regarded as describing how much information about β_{xg} is lost when only (Y,G) and (X) marginal distributions are available.

The evaluation of the LPD for β_{xg} is somewhat involved in this setting, because regular asymptotic arguments do not apply. The model for (Y|G) implied by model (1) plus knowledge of *r* is not identified, and consequently the posterior distribution for β_{xg} does not concentrate to a single value as the sample size grows. Therefore, we apply recently developed techniques for determining LPDs from non-identified models (Gustafson 2005). The approach is based on reparameterizing in such as way as to separate parameters appearing in the likelihood function from those that do not. The main features of the LPD are summarized as follows, with full details appearing in Appendix A, and R (R Development Core Team, 2007) code for LPD determination posted at www.stat.ubc.ca/~gustaf.

We presume a prior under which the three components of β are independent, with $\text{logit}(\beta_0) \sim \text{Unif}(0,1)$, while β_x and β_{xg} have mean-zero normal priors with standard deviations σ_x and σ_{xg}, respectively. The choice of prior is uninformative with respect to disease prevalence, but admits prior suppositions that the magnitudes of X and XG effects on Y are unlikely to exceed investigator-specified thresholds. Thus, we have an algorithm which takes as inputs *r*, the components (`true values') of β, and hyperparameters σ_x and σ_{xg}; the output is the LPD for β_{xg}. Consequently, we see what

would happen if an infinite sample of (Y,G) values was generated under the input values of r and β, and the posterior distribution of β_{xg} was formed from these data, assuming correct knowledge of r and using the prior distribution corresponding to the given hyperparameters. Note that because we are investigating how informative knowledge of the (Y|G) distribution is for a parameter in the (Y|X,G) distribution, the LPD for β_{xg} does not depend on the underlying prevalence of G in the population.

The LPD determination is facilitated by reparameterizing from $\beta=(\beta_0, \beta_x, \beta_{xg})$ to $\theta=(p_0, p_1, q)$, where $p_i=\Pr(Y=1|G=i)$ and $q=\text{expit}(\beta_0)$. The salient features are as follows.

1. The determination can be carried out in two steps. First, the prior conditional distribution of $(q| p_0, p_1, r)$ is determined numerically. From this, the prior conditional distribution of $(\beta_{xg} | p_0, p_1, r)$ follows via change of variables under a non-monotonic transformation of q. The LPD is identically this conditional distribution evaluated at the true values of (p_0, p_1, r).

2. Assume, without loss of generality, that the true value of β_{xg} is positive. Then the LPD will have support $[b, \infty)$, with $b>0$. Thus, (Y,G) data plus knowledge of r can rule out small values of the gene-exposure interaction effect (including zero, in line with Burstyn et al., 2009), but not large values.

3. In the setting of an outcome that is not overly common even amongst the exposed, expressed as $\text{expit}(\beta_0 + \beta_x) + \text{expit}(\beta_0 + \beta_x + \beta_{xg}) < 1$ (i.e., $\Pr(Y=1|X=1,G=0) + \Pr(Y=1|X=1,G=1) < 1$), the left-endpoint b can be expressed in terms of the true parameter values as

$$b = 2 \text{ logit}[\{1 + \text{expit}(\beta_0 + \beta_x + \beta_{xg}) - \text{expit}(\beta_0 + \beta_x) \} /2], \quad (2)$$

in the lower exposure prevalence case of $r/(1-r) < 2\text{expit}(\beta_0)/\{ 1 - \text{expit}(\beta_0 + \beta_x) - \text{expit}(\beta_0 + \beta_x + \beta_{xg})\}$. When the exposure prevalence exceeds this threshold, the commensurate expression is

$$b = \text{logit}[\ \{(1-r)/r\}\text{expit}(\beta_0) + \text{expit}(\beta_0 + \beta_x + \beta_{xg})\ ]\ -$$

$$\text{logit}[\ \{(1-r)/r\}\text{expit}(\beta_0) + \text{expit}(\beta_0 + \beta_x)\ ]. \qquad (3)$$

It is interesting to note that (2) does not depend on r, and that (3) tends to β_{xg} as r tends to one. The interpretation is that whereas β_{xg} itself is not completely determined by (p_0, p_1, r), expressions (2) and (3) provide a lower bound for β_{xg} which is completely determined by (p_0, p_1, r). Thus, the tightness of this bound will be one key part of how well (Y,G) data plus knowledge of r inform the target parameter.

4. Still in the `not too common outcome' setting, we have the following general finding about the shape of the LPD. In the lower prevalence case with left endpoint b defined by (1), the LPD density will be infinite at b. Thus, there is a tendency for the posterior to put a lot of weight near b, which is a further mechanism by which the LPD for β_{xg} can be informative.

Some specific cases of the LPD for β_{xg} are given in Figure 1. Here σ_x=1, while several values for each of r, σ_{xg}, β_0, β_x, β_{xg} are considered. Each LPD is summarized by its 0-th, 50-th and 95-th percentiles, i.e. the 0-th percentile is the left-endpoint b described above. We first observe that there is little difference between the 50-th percentile of the LPD and true value of β_{xg} in all cases. Next, as expected, the 0-th percentile of the LPD is positive in all cases. This implies that with sufficient data one could infer that the gene-environment interaction exists. Absence or presence of a main effect of exposure

(β_x =0 or β_x = 0.2) has a negligible effect on the width of the LPD for β_{xg}, and the impact of changing the hyperparameter (σ_{xg} =1 or σ_{xg} = 2) is also quite slight. Overall, the LPD for β_{xg} tends to be wider when exposure and outcome are rarer, though the theoretically predicted effect of lower exposure prevalence on the shape of the LPD is also clearly manifested.

2.3 Knowledge of Exposure Prevalence

The LPDs presented in Figure 1 arise when the investigator correctly knows *r,* the population prevalence of exposure. To examine robustness to the specification of *r*, we also present some LPDs which arise when the investigator specifies an incorrect value. That is, whereas the true value of *r* plays a role in determining the observable (Y,G) relationship, the investigator-assumed value of *r* is used to calculate the LPD for β_{xg} arising from the (Y,G) relationship. Results for selected underlying parameter values appear in Figure 2. Note that several underlying values for β_0 are considered. Results for alternate values of β_x and β_{xg} are qualitatively similar to those presented, and hence these are not shown.

The findings from Figure 2 are mixed. In some cases the LPD is somewhat insensitive to a misspecified value of *r* (e.g. with a higher true value of *r* and a lower value of β_0), and in some cases it is highly sensitive (e.g. with a lower true value of *r* and a higher value of β_0). In practice, of course, information about *r* would likely be provided in the form of a prior distribution, rather than a single value that is assumed to be correct, which may guard against erroneous conclusions due to one poor guess of *r* by appropriately inflating the posterior of β_{xg}. However, our findings suggest that such a

prior would need to be both reasonably narrow and consistent with the true value, in order to obtain useful and reliable inference about β_{xg} from (Y,G) data alone.

2.4 Impact of Exposure Misclassification

One motivation for considering inference based on (Y,G) data only is that it may not be possible to measure exposure X well. Indeed, exposure misclassification in the gene-environment context has been discussed by a number of authors, including Wong et al. (2003, 2004) and Zhang et al. (2008). To explore the issue in the present context, we consider what happens when model (1) is fit to (Y,X^*,G) data rather than (Y,X,G) data, where X^* is a misclassified surrogate for X. In particular, say X^* is a non-differential surrogate (i.e. conditionally independent of Y and G given X), characterized by its sensitivity, $SN=\Pr(X^*=1|X=1)$, and specificity, $SP=\Pr(X^*=0|X=0)$. To evaluate the impact of unacknowledged misclassification then, we examine the discrepancy between coefficients from logistic regression of Y on $(1, X^*, X^*G)$ and coefficients arising from logistic regression of Y on (1, X, XG).

In Appendix B we outline a simple, but approximate, approach to determining the discrepancy between coefficients, as well as a computational approach to determining the discrepancy exactly (i.e. computing the large-sample limit of the estimated coefficients when an intercept plus X^* and X^*G terms are fit). The approximate calculation indicates that both the X^* and X^*G coefficients will be attenuated toward zero compared to the X and XG coefficients in the true relationship. Moreover, the multiplicative attenuation factor in both cases will be $(PPV + NPV - 1)$, where $PPV = \Pr(X=1|X^*=1) = r\text{SN}/\{r\text{SN}+(1-r)(1-\text{SP})\}$ is the *positive predictive value* of the exposure classification, and $NPV = \Pr(X=0|X^*=0) = (1-r)\text{SP}/\{(1-r)\text{SP}+r(1-\text{SN})\}$ is the *negative predictive value*.

While attenuation of coefficients commonly results from exposure misclassification (see, for instance, Gustafson 2004, Ch. 3), we emphasize that the approximate attenuation factor of ($PPV + NPV - 1$) is obtained via the two key assumptions stated in 2.1, and hence is particular to the present setting. Note also that for fixed sensitivity and specificity, low exposure prevalence r will induce low PPV. Thus, we expect to see severe attenuation in the rare exposure case. It should also be noted that the calculations in Appendix B reveal that misclassification induces a main effect of gene, i.e. the lack of a main effect for G in the (Y|X,G) relationship does *not* guarantee the lack of a main effect for G in the (Y,X^*,G) relationship.

Figure 3 displays the exactly computed attenuation in a variety of settings, and these values agree quite closely with the approximate attenuation factor. The attenuation is confirmed to be particularly severe when the population prevalence of exposure is low. To be clear, note that ignoring misclassification corresponds to fitting an identified but misspecified model, i.e. the posterior distribution on regression coefficients will shrink to a single point as the sample size grows, but it will be the wrong point. Therefore, inference about the true gene-exposure interaction effect which ignores the misclassification will be both biased in terms of where the posterior distribution is centered, and falsely precise in terms of the width of the posterior distribution.

At least indirectly, the impact of misclassification of X seen in Figure 3 can be compared to the impact of misspecified r in Figure 2. In particular, we can consider an assumed value of r that arose from the population prevalence of X^* (whereas the true value of r is by definition the population prevalence of X). That is the assumed value r^* and the true value r are linked via $r^* = r\,SN + (1-r)(1-SP)$. For instance, when r=0.1,

SN=*SP*=0.9 yield r^*=0.18, and *SN*=*SP*=0.8 yield r^*=0.26. Similarly, when r=0.25, *SN*=*SP*=0.9 yield r^*=0.30, and *SN*=*SP*=0.8 yield r^*=0.35. Via this link, comparison between Figures 2 and 3 shows that mistaken knowledge about exposure prevalence in the (Y,G) data analysis is less damaging than unchecked misclassification of exposure in the full analysis based on (Y,X^*,G) data.

Of course, the most desirable strategy in the face of (Y, X^*, G) data would be to explicitly recognize the misclassification, and `adjust' for it in the analysis. To do so, however, requires knowledge about the nature and extent of misclassification, say via a prior distribution on (*SN*,*SP*), and assumptions about the extent to which the misclassification might be differential. In some situations such knowledge may be very difficult to acquire, whereas reasonable estimates of population exposure prevalence may be readily available.

2.5 Impact of Incorrectly Assuming that Gene Does Not Confer Risk of Disease

Suppose that the assumption in model (1) is wrong in a sense that a main effect, perhaps small, of G (i.e. a β_gG term with $\beta_g \neq 0$) is missing. For instance, such a main effect might arise if an unobserved exposure acts via G to cause Y. In practice, there could be lingering doubt about biological mechanisms and the possible presence of unobserved exposures that cause the disease of interest and whose risk is also modified by the given genotype. To consider the impact of a small gene effect which is missed in the analysis, we consider the LPD arising under the assumption that β_g=0 when the true value is non-zero. In particular, Figure 4 reproduces the settings in Figure 1, but with true relationship involving a main G effect of β_g=0.025.

An unacknowledged main effect of G is seen to be very damaging, even though the magnitude of this effect is small. For instance, in some cases in Figure 4 the left endpoint of the LPD (*b*) lies above the true value of β_{xg}. Thus, it seems inadvisable to make the assumption that gene alone does not confer risk unless it is strongly warranted and supported. Further evidence along these lines is as follows. Say model (1) is fit to cohort data, when in fact the true relationship involves $\beta_g \neq 0$ (hence the fitted model is violated) as well as $\beta_{xg} = 0$ (so that in fact there is no gene-environment interaction). In Appendix C we give a Taylor series argument to show that, at least approximately, all of the main effect of G in the true relationship is transferred to a spurious effect of XG in the fitted relationship. This again speaks to the method being non-robust to violations of the assumption that G alone confers no risk.

3 Extension to Retrospective Analysis

The theoretical results in Section 2 are given in the framework of a cohort study, i.e. they describe what happens when an infinitely large sample of (Y,G) values are drawn from the study population. Given the well established approach of analyzing retrospective data as if they were prospective, we might anticipate that the findings are relevant to case-control studies as well. However, there are reasons to consider `fully retrospective' analysis in the present context. In particular, when (Y,X,G) data are analyzed prospectively with model (1), the supposition that X and G are independent does not come into play (as emphasized by Chaterjee and Carroll 2005, for instance). That is, the same logistic regression fit of (1) would result whether or not one assumed this independence. However, the assumption matters when retrospective analysis is applied. That is, starting with a model for (Y|X,G), the model arising for (X,G|Y) will differ

depending on whether or not X and G are assumed independent in the population. In particular, the literature cited in Section 1 points to efficiency gains in inference when the assumption is appropriately made.

Given this, it also seems important to consider a properly retrospective analysis in the case of having (Y,G) data only. It no longer seems feasible to study the large-sample limiting behaviour as was done in the prospective case. However, Bayesian analysis for finite samples is readily implemented, for both full (Y,X,G) data and reduced (Y,G) data. Modelling and implementation details for these analyses appear in Appendix D.

4 Illustrative Example: Bladder Cancer, Cigarette Smoking and NAT2 Genotype

To give an empirical illustration, we analyze the data from a bladder cancer study published by Gu et al. (2005). The version of the data we use is for the NAT2 genotype (G=1 corresponds to slow acetylator), with exposure X being heavy-smoking (compared to never or light smoking), from their Table 3. The pertinent data summaries are reproduced here in Table 1. We re-analyze these data using both prospective and retrospective Bayesian analysis. We also consider both full and reduced data. The full data are simply the (Y,X,G) data. The reduced data are the (Y,G) data along with information on the population prevalence of X. For illustrative purposes, this information is taken directly from the data at hand, i.e. the observed exposure status of controls is used to inform the prior specification $r \sim \text{Beta}(c+1,d+1)$, where c=110 and d=402 are the counts of exposed (X=1) and unexposed (X=0) controls. Thus, we *mimic* a situation where we have (Y,G) data on a different set of subjects than those for whom we have X data. The priors we use are $\text{logit}(\beta_0) \sim \text{Unif}(0,1)$, $\beta_x \sim N(0,1)$, $\beta_{xg} \sim N(0,1)$. Again, the

first specification is intended to be uninformative about the disease prevalence. (Incidentally, we see almost no prior-to-posterior updating of β_0 under the retrospective analysis, in accord with intuition.) The latter two specifications are made in light of $\exp(\pm 2)$ being extreme odds ratios in the present context.

The posterior distributions on β_{xg} arising from each analysis appear in Figure 5. Immediately we see that these data yield indistinguishable results under prospective and retrospective analysis, notwithstanding the discussion in Section 3. The fact that the full-data analyses are virtually the same both retrospectively and prospectively, despite the former making using of the Mendelian randomization assumption, is perhaps slightly surprising in view of the literature. However, Umbach and Weinberg (1997) consider a saturated eight-parameter log-linear model for cell counts arising without the assumption to an unsaturated seven-parameter model with the assumption. In the present instance, however, we are also assuming that there is no main effect of G in the disease model. Thus, in the log-linear model framework the comparison is now between two unsaturated models with seven and six parameters, respectively. Since the assumption of no main effect of G alters the nature of the models being compared, we do not necessarily expect the findings of Umbach and Weinberg (1997) to apply in the present setting. We also note that our full-data estimate of β_{xg} is somewhat attenuated compared to that reported by Gu et al. (2005). This is likely the result of the prior distribution used here, compared to the non-Bayesian analysis in the original publication.

The reduced-data posterior is considerably more concentrated than the prior, but considerably less concentrated than the full-data posterior. Both provide strong evidence for a gene-environment interaction, with posterior probability of $\beta_{xg} < 0$ on the order of

0.03 with reduced data, and 0.001 with full data. In line with our theoretical findings, we see that the reduced data are effective in providing evidence against smaller values of β_{xg} (above and beyond providing evidence against the null value). On the other hand, the right-tail of the posterior seems to fall in step with the prior, i.e. in this particular example it appears that the prior distribution is relied upon to provide evidence against rather large values.

5 Discussion

In the context of binary variables, we have considered the setting where Mendelian randomization of genetic traits is known to hold (i.e. X and G are independent), and that gene is known not to convey increased disease risk in the absence of exposure (i.e. the distribution of Y|X=0,G=g does not depend on g). We have seen that in this setting the distribution of Y|G and the distribution of X are partially informative for the distribution of Y|X,G. In particular, there is some scope for learning about a gene-environment interaction from a gene-disease study augmented with external (i.e. ecological) knowledge about the exposure. This partial information has been characterized theoretically via the *limiting posterior distribution* on the gene-exposure interaction coefficient. We also gave an example of accessing this information from real case-control data, both with the pretence of modeling the data prospectively and by using retrospective analysis. We showed that information lost when data are scrambled, i.e. (Y,G) and (X) data are disjointed, can be partially recovered through an analytical strategy that takes advantage of biologically justifiable assumptions. Whereas there is an existing literature on using the assumption of gene-exposure independence to improve the efficiency of the estimators for gene-environment interaction from case-control studies,

this is apparently the first work to consider using the assumption to help mitigate a lack of individual-level information about either exposure or the even harder to measure biologically effective dose.

The procedure we propose is especially advantageous when exposure is common, as would be the case in risk-enriched cohorts, such as those assembled in workplaces, where is possible to select subjects in such a way that nearly all of them are "exposed" (or in the case of continuous exposure, most are exposed to some non-negligible extent). However, even in such settings the efficiency deteriorates for rare outcomes. One of the pivotal strengths of the proposed method is its partial robustness to misspecification of exposure prevalence, and the fact that it is certainly preferable to ignoring misclassification of exposure in analysis. This is important because measurement error in exposure is recognized as one of the main threats to efficiency of gene-environment interaction studies in epidemiology (Davey Smith and Ebrahim 2003, Vineis 2004, Burstyn et al. 2009). Furthermore, the method we propose may indirectly mitigate the concern about spurious associations due to multiple comparisons that plague many gene-environment interaction studies (see for example Wakefield et al., 2009). Specifically, our method promotes hypothesis-driven analysis and study design that focuses on just a few associations for which the key assumptions can be supported.

The method we developed is vulnerable to violations of the two fundamental assumptions. The conditions under which Mendelian randomization of genetic traits is violated were extensively discussed by Davey Smith and Ebrahim (2003). Although not formally addressed here, it is clear that inter-dependence of genetic susceptibility and exposure would invalidate the proposed method. As well, it may be challenging to rule

out a direct effect of gene on disease. Unfortunately, it would appear that even a small direct effect of gene on disease biases inference about gene-environment interaction. Therefore, for our proposed method to apply, it is paramount to ensure during the design of a study that an effect of gene on the disease in absence of exposure can be ruled out with reasonable certainty. It should be mentioned that the related problem of using gene as an instrumental variable to deal with confounding, mentioned in Section 1, also invokes an assumption that gene does not directly influence disease. A commensurate lack of robustness to violation of this assumption is seen in that setting as well.

The applicability of our approach for disease, exposure and gene-environment interaction models other than those considered here (even when the two key assumptions are satisfied) all constitute fruitful future directions of research. However, given that there is always uncertainty about model specification, our methodology offers clear practical approach to studying gene-environment interactions when the two key assumptions can be invoked.

Acknowledgements

Igor Burstyn was supported by the Population Health Investigator salary award from the Alberta Heritage Foundation for Medical Research. Paul Gustafson was supported by grants from the Natural Sciences and Engineering Research Council of Canada, and the Canadian Institutes of Health Research (Funding Reference Number 62863). The authors wish to thank Xifeng Wu and Maosheng Huang for clarifying some aspects of their published data used to illustrate the methodology.

References

Burstyn I, Kim HM, Yasui Y, Cherry NM (2009). The virtues of a deliberately misspecified disease model in demonstrating a gene-environment interaction. *Occupational and Environmental Medicine* **66**(6), 374-80.

Chatterjee N, Carroll RJ (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrics* **92**, 399-418.

Cherry N, Mackness M, Durrington P, Povey A, Dippnall, M, Smith T, Mackness B (2002). Paraoxonase (PON1) polymorphisms in farmers attributing ill health to sheep dip. *Lancet* **359** (9308), 763-4.

Davey Smith G, Ebrahim S (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32,** 1–22

Davey Smith G, Ebrahim S (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology* **33**, 30-42.

Didelez V, Sheehan N (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* **16**, 309-330.

Gu J, Liang D, Wang Y, Lu C, Wu X (2005). Effects of N-acetyl transferase 1 and 2 polymorphisms on bladder cancer risk in Caucasians. *Mutation Research* **581**(1-2), 97-104.

Gustafson P (2004). *Measurement Error and Misclassification in Statistics and Epidemology: Impacts and Bayesian Adjustments*. Chapman and Hall/CRC Press.

Gustafson P (2005). On model expansion, model contraction, identifiability, and prior information: two illustrative scenarios involving mismeasured variables" (with discussion). *Statistical Science* **20**, 111-140.

Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27**, 1133-1163.

Mukherjee B, Zhang L, Ghosh M, Sinha S (2007). Semiparametric Bayesian analysis of case-control data under conditional gene-environment independence. *Biometrics* **63**, 834-844.

Mukherjee B, Chatterjee N (2008). Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* **64**, 685-694.

R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sacerdote C, Guarrera S, Davey Smith G, Grioni S, Krogh V, Masala G, Mattiello A, Palli D, Panico S, Tumino R, Veglia F, Matullo G, Vineis P (2007). Lactase persistence and bitter taste response: Instrumental variables and Mendelian randomization in epidemiologic studies of dietary factors and cancer risk. *American Journal of Epidemiology* **166**, 576-581.

Thomas DC, Conti C (2004). Commentary: The concept of `Mendelian randomization.' *International Journal of Epidemiology* **33**, 21-25.

Umbach DM, Weinberg CR (1997). Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine* **16,** 1731-1743.

Vineis P (2004). A self-fulfilling prophecy: are we underestimating the role of the environment in gene-environment interaction research*? International Journal of Epidemiology* **33**(5), 945-6.

Wakefield J, De Vocht F, Hung RJ (2009). Bayesian mixture modeling of gene-environment and gene-gene interactions. *Genetic Epidemiology*. Jun 2. [Epubahead of print] PubMed PMID: 19492346.]

Wong MY, Day NE, Luan JA, Chan KP, Wareham NJ (2003). The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *International Journal of Epidemiology* **32**, 51-57.

Wong MY, Day NE, Luan JA, Wareham NJ (2004). Estimation of magnitude in gene-environment interactions in the presence of measurement error. *Statistics in Medicine* **23**, 987-998.

Zhang L, Mukherjee B, Ghosh M, Gruber S, Moreno V (2008). Accounting for error due to misclassification of exposures in case-control studies of gene-environment interaction. *Statistics in Medicine* **27**, 2756-83.

Appendix A: Limiting Posterior Distribution in the Prospective Analysis

The LPD for β_{xg} when r is known can be determined via a two-step procedure. First, we re-parameterize from $\beta=(\beta_0, \beta_x, \beta_{xg})$ to $\theta=(p_0, p_1, q)$, where $p_i=\Pr(Y=1|G=i)$ and $q=\text{expit}(\beta_0)$. More specifically, note that

$$p_0 = (1-r)\,\text{expit}(\beta_0) + r\,\text{expit}(\beta_0 + \beta_x), \qquad (4)$$

$$p_1 = (1\text{-}r)\ \text{expit}(\beta_0) + r\ \text{expit}(\beta_0 + \beta_x + \beta_{xg}). \quad (5)$$

Note also that this mapping is explicitly invertible, via

$$\beta_0 = \text{logit}(q) \quad (6)$$

$$\beta_x = \text{logit}[r^{-1}\{p_0 - (1\text{-}r)q\}] - \text{logit}(q), \quad (7)$$

$$\beta_{xg} = \text{logit}[r^{-1}\{p_1 - (1\text{-}r)q\}] - \text{logit}[r^{-1}\{p_0 - (1\text{-}r)q\}]. \quad (8)$$

Moreover, the Jacobian of the mapping (in either direction) is readily determined. Thus, starting with a specified prior density for β, it is straightforward to compute the prior density for θ. In turn, this yields the prior conditional density for $(q \mid p_0, p_1)$. This conditional distribution, evaluated at the true values of (p_0, p_1), characterizes the uncertainty which remains about θ after observation of an infinite (Y,G) sample. It is important to note that the support of the prior conditional for q may be smaller than the unit interval. In particular, say without loss of generality that $p_0 < p_1$. Then, from the form of the mapping above we have the support as $\max\{0,q_0\} < q < \min\{q_1,1\}$, where $q_0 = (1\text{-}r)^{-1}(p_1\text{-}r)$ and $q_1 = (1\text{-}r)^{-1}p_0$. This speaks to knowledge of parameters inside the reduced-data likelihood function, p_0 and p_1, being able to impart some information about the parameter q not involved in the likelihood.

The second part of the calculation is to determine the conditional prior for $(\beta_{xg} \mid p_0, p_1)$ from the conditional prior for $(q \mid p_0, p_1)$. This is straightforward since given (p_0, p_1) we have $\beta_{xg} = h(q)$, where $h(q)$ is defined by (8) and can be re-expressed as

$$h(q) = \text{logit}\{s - k(q\text{-}q^*)\} + \text{logit}\{s + k(q\text{-}q^*)\},$$

where $s = \frac{1}{2} + (p_1\text{-}p_0)/(2r)$, $k=(1\text{-}r)/r$, and $q^*=(q_0 + q_1)/2$. Upon noting that $h()$ is symmetric about, and minimized at, q^*, we have

$$\Pr\{ q^* - \epsilon < q < q^* + \epsilon \mid p_0, p_1\} = \Pr\{\beta_{xg} < \text{logit}(s - k\epsilon) + \text{logit}(s + k\epsilon) \mid p_0, p_1\},$$

which allow computation of the prior conditional distribution for β_{xg} from the prior conditional distribution for q.

To gain further qualitative insight, we specialize to the situation where q^* lies below the true value of q. Translated back to the original parameters, this condition is expressed as $\text{expit}(\beta_0 + \beta_x) + \text{expit}(\beta_0 + \beta_x + \beta_{xg}) < 1$, which corresponds to an outcome which is not too common. Then we have that q^* is positive when

$$r/(1-r) < 2\text{expit}(\beta_0)/\{1 - \text{expit}(\beta_0 + \beta_x) + \text{expit}(\beta_0 + \beta_x + \beta_{xg})\}, \qquad (9)$$

and negative when the inequality is reversed. In the former case then, the LPD for β_{xg} must have $h(q^*)$ as the left-endpoint of its support. Expressing $h(q^*)$ in terms of the original parameters then directly gives the left-endpoint b expressed in equation (2). Also in this case, the fact that $h'(q^*)=0$ implies immediately that the density of the LPD will be infinite at b. Conversely, in the case that r is sufficiently large to violate (9), the left-endpoint must be $h(0)$, which translates to equation (3).

Appendix B: Misclassification of Exposure at the Individual Level

Say that model (1) holds, and that X and G are independent. Also, say that X^* is a non-differentially misclassified surrogate for X. *Approximately*, then:

$$\begin{aligned}
\Pr(Y=1|X^*,G) &= E\{ \Pr(Y=1|X,G) \mid X^*, G \} \\
&= E \{ \Pr(Y=1|X,G) \mid X^* \} \\
&\approx \text{expit}[E\{ \text{logit} \Pr(Y=1|X,G) \mid X^* \}] \\
&= \text{expit}\{ \beta_0 + (\beta_x + \beta_{xg}G) E(X|X^*) \}
\end{aligned}$$

$$= \text{expit}[\ \beta_0 + (\beta_x + \beta_{xg}G)\{\ (1\text{-}NPV) + (PPV\text{+}NPV\text{-}1)\ X^*\}\ ]$$

$$= \text{expit}[\ \{\beta_0 + \beta_x(1\text{-}NPV)\} + \{\beta_x(PPV + NPV\ \text{-}1)\}X^*\ + \{\beta_{xg}(1\text{-}NPV)\}G +$$

$$\{\beta_{xg}(PPV + NPV - 1)\}X^*G\ ].$$

Thus, we see that both the X^* and X^*G coefficients are attenuated by a factor of (PPV + NPV - 1) relative to the X and XG coefficients in the true relationship. We also see the spurious introduction of a main effect for G.

More formally now, let $\alpha = (\alpha_0, \alpha_x, \alpha_{xg})$ be the large-sample limit of coefficients resulting from the logistic regression of Y on $(1, X^*, X^*G)$. In passing we note that this will in fact be a misspecified model, particularly because the misclassification will induce a main effect for G even though it is absent in the true relationship for (Y|X,G). Nonetheless, we can define $(\alpha_0, \alpha_x, \alpha_{xg})$ via

$$E\ [\{Y\text{-expit}(\alpha_0 + \alpha_x X^* + \alpha_{xg}\ X^*G)\}\ (1,\ X^*,\ X^*G)'\ ] = 0.$$

Upon noting that

$$E\{Y(1\ \ X^*\ X^*G)'\} = E[Y\{1,\ (1\text{-}SP) + (SN\text{+}SP\text{-}1)X,\ (1\text{-}SP)G\text{+}(SN\text{+}SP\text{-}1)XG\}]',$$

and that by definition

$$E\{Y(1, X, XG, G)'\} = E\ \{\ \text{expit}(\beta_0\text{+}\beta_x X\text{+}\beta_{xg})\ (1, X, XG, G)'\ \},$$

we have α as a function of β via

$$E\ \{\text{expit}(\alpha_0 + \alpha_x X^* + \alpha_{xg}\ X^*G)\ (1,\ X^*,\ X^*G)'\ \} =$$

$$E[\text{expit}(\beta_0\text{+}\beta_x X\text{+}\beta_{xg}XG)\ \ \{1, (1\text{-SP})\text{+}(SN\text{+}SP\text{-}1)X,\ (1\text{-}SP)G\text{+}(SN\text{+}SP\text{-}1)XG\}'].\qquad (10)$$

Given β, we can solve (10) for α numerically. In particular, the derivative of the left-hand side with respect to α is readily computed, so that the Newton-Raphson algorithm

can be applied. Here, the attenuation of inference on the exposure-gene coefficient, given by α_{xg}/β_{xg}, is of interest.

Appendix C: Violation of the Gene-Exposure Independence Assumption

Say that the true relationship is given by $\text{logit}\{\Pr(Y=1|X,G)\} = \alpha_0 + \alpha_x X + \alpha_g G$, whereas $(\beta_0, \beta_x, \beta_{xg})$ are the large-sample limiting coefficients arising from fitting model (1). Thus β is determined from α according to

$$E[\{\text{expit}(\alpha_0 + \alpha_x X + \alpha_g G) - \text{expit}(\beta_0 + \beta_x X + \beta_{xg} XG)\}\ (1, X, XG)'] = 0.$$

If we fix (α_0, α_x), differentiate this expression with respect to α_g, and evaluate the resulting expression at $\alpha_g=0$ we obtain

$$E\ \{\text{expit}'(\alpha_0 + \alpha_x X)\ (1, X, XG)'\ (1, X, XG)\}v = E\ \{\text{expit}'(\alpha_0 + \alpha_x X)\ (G, XG, XG)'\ \},$$

where v is the derivative of β with respect to α_g, evaluated at $\alpha_g = 0$. Upon taking expectation with respect to G (under the assumption that G and X are independent), the solution to this equation is seen to be $v = (E(G), -E(G), 1)'$. In particular, the derivative of β_{xg} with respect to α_g, evaluated at α_g, is one. Locally then the spurious interaction coefficient in the fitted relationship is the same as the main effect of G in the true relationship.

Appendix D: Retrospective Analysis

Let p_{xgy} be the proportion of the study population with $(X=x, G=g, Y=y)$. Under the two key assumptions we have:

$$p_{xgy} = r^x(1-r)^{1-x}\ t^g(1-t)^{1-g}\ \{\text{expit}(\beta_0 + \beta_x x + \beta_{xg} xg)\}^y \{1 - \text{expit}(\beta_0 + \beta_x x + \beta_{xg} xg)\}^{1-y},$$

where r and t are population prevalences of X and G, respectively. Thus, the parameter vector is comprised of (r, t, β), and in contrast to the prospective analysis, prior distributions are required for r and t, in addition to β. Absent of substantive information, each prevalence might be assigned Unif(0,1) prior distributions. The joint prior on the parameter vector is informed by the (X,G) data for controls which arise via multinomial sampling with cell probabilities ($p_{xg0} / p_{\bullet\bullet 0}$), and the corresponding data for cases having cell probabilities ($p_{xg1} / p_{\bullet\bullet 1}$). (Here the 'dot' notation indicates summation.) Note that without the assumption of exposure-gene independence, a further parameter would be required, with the corresponding potential for less efficient estimation.

This approach is easily adapted to the reduced-data situation involving an informative prior for r along with (Y,G) case-control data. This informative prior would be encoded as part of a joint prior on unknown parameters (r, t, β). This prior is then updated via a binomial sampling model for G amongst controls and cases, with $\Pr(G=1|Y=y) = (p_{\bullet 1y} / p_{\bullet\bullet y})$. R code implementing both the full-data and reduced data retrospective analyses via Markov Chain Monte Carlo with random walk proposals is posted at www.stat.ubc.ca/~gustaf.

	controls ($Y=0$)		cases ($Y=1$)	
	$G=0$	$G=1$	$G=0$	$G=1$
$X=0$	172	230	106	156
$X=1$	58	52	83	157

Table 1: Data from Gu et. al. on the effect of NAT2 genotype and smoking status on bladder cancer risk. Genotype is coded as 0=rapid acetylator, 1=slow acetylator. Smoking status is coded as 0=never/light, 1=heavy.

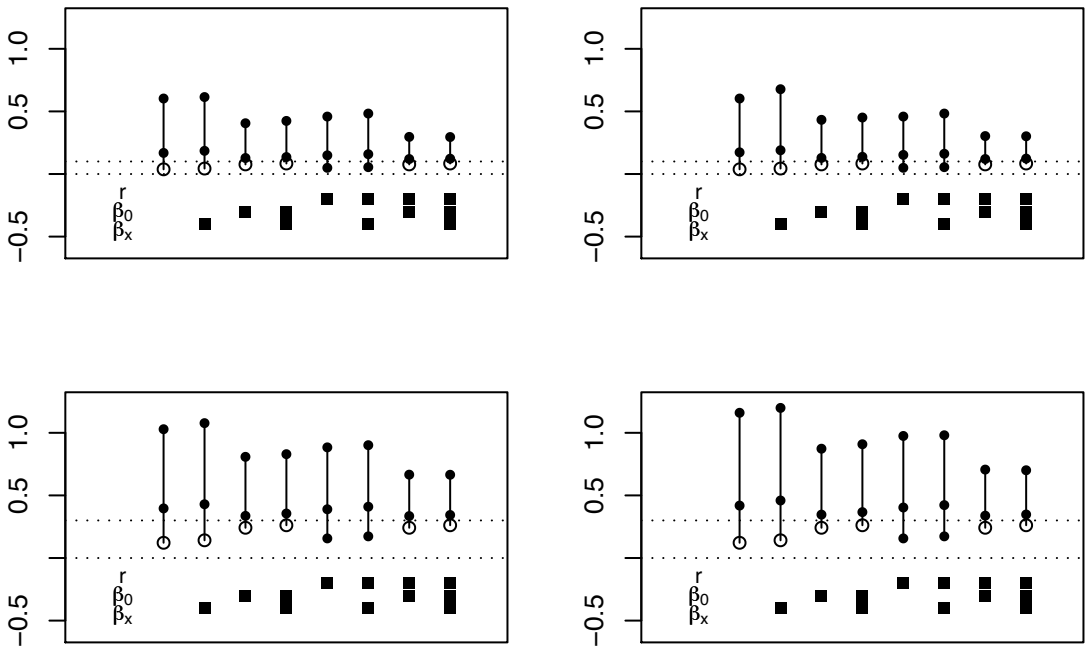

Figure 1: LPD for β_{xg}. The top (bottom) panels correspond to true values $\beta_{xg} = 0.1$ ($\beta_{xg} = 0.3$). The left (right) panels correspond to hyperparameters $\sigma_{xg} = 1$ ($\sigma_{xg} = 2$). Within each panel, all combinations of $\beta_0 = \text{logit}0.1, \text{logit}0.5$, $\beta_x = 0, 0.2$, $r = 0.1, 0.4$ are considered. In each case, ■ in the legend indicates the second of the two values. The 0th, 50th, and 95th percentiles of the LPD are displayed. The open circles indicate cases where the LPD has infinite density at the 0th percentile. The dotted horizonal lines indicate zero and the true value of β_{xg}.

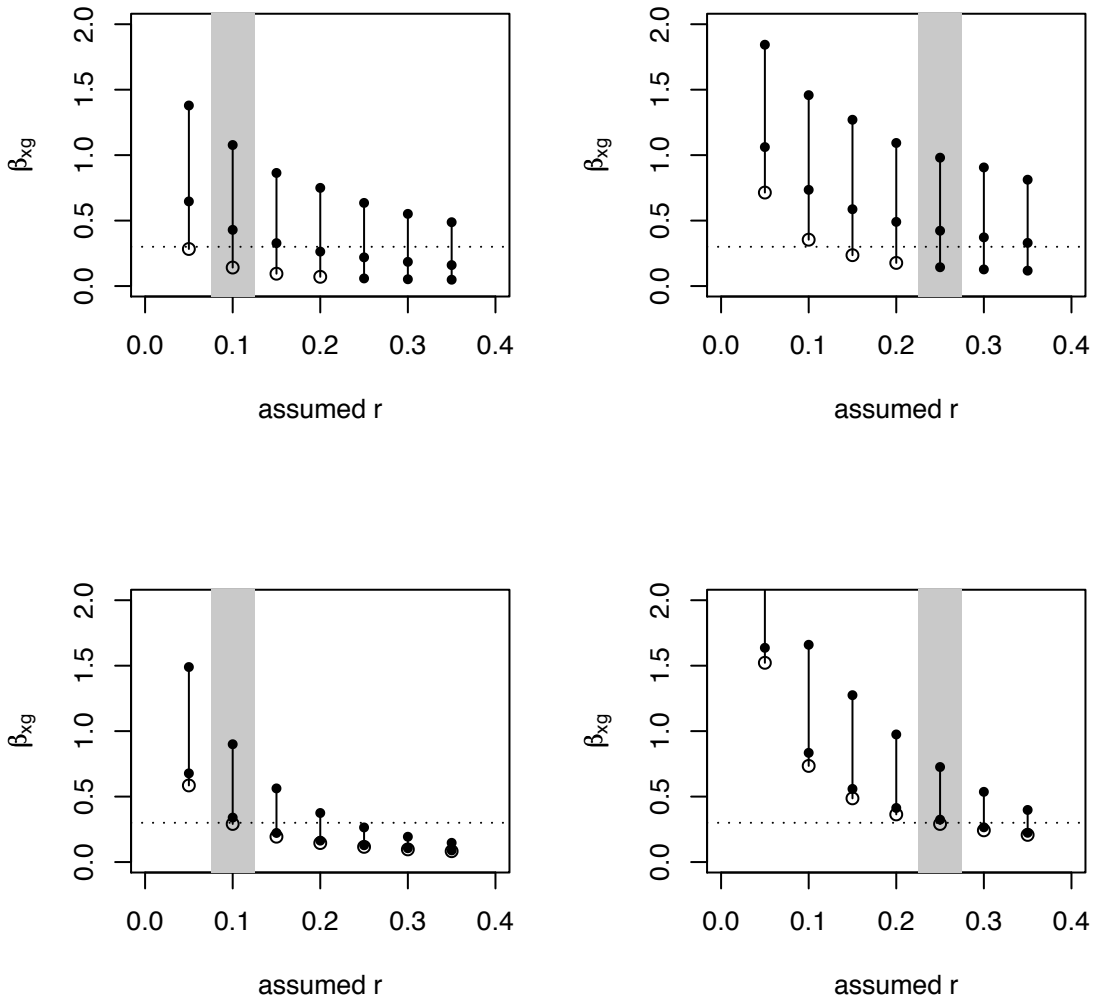

Figure 2: LPD for β_{xg} when the value of r is misspecified. The LPD is displayed as per Figure 1 (0th, 50th, 95th percentiles). True parameter values are fixed at $\beta_x = 0.2$, $\beta_{xg} = 0.3$ and either $\beta_0 = \text{logit}0.1$ (top panels) or $\beta_0 = \text{logit}0.5$ (bottom panels). The hyperparameter setting is $\sigma_{xg} = 1$. In each panel the LPDs for various assumed values of $r = Prev(X)$ are given, with the correct value being $r = 0.1$ (left panels) or $r = 0.25$ (right panels). Cases where the assumed and true prevalences match are highlighted.

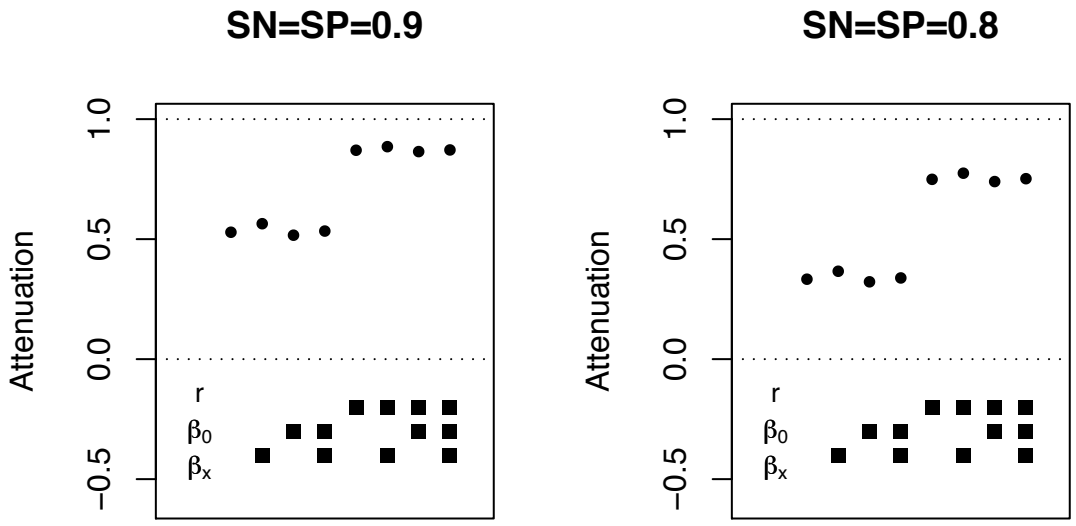

Figure 3: Attenuation in estimating β_{xg} when nondifferential misclassification of X is ignored. The true value of β_{xg} is set at 0.3, though results are similar for other values. The attenuation is reported as a ratio of estimated value (in the large sample limit) to true value. Settings for r, β_0, β_{xg} are as per Figure 1. In some cases the attenuated estimate is lower than the corresponding left endpoint of the LPD reported in Figure 1.

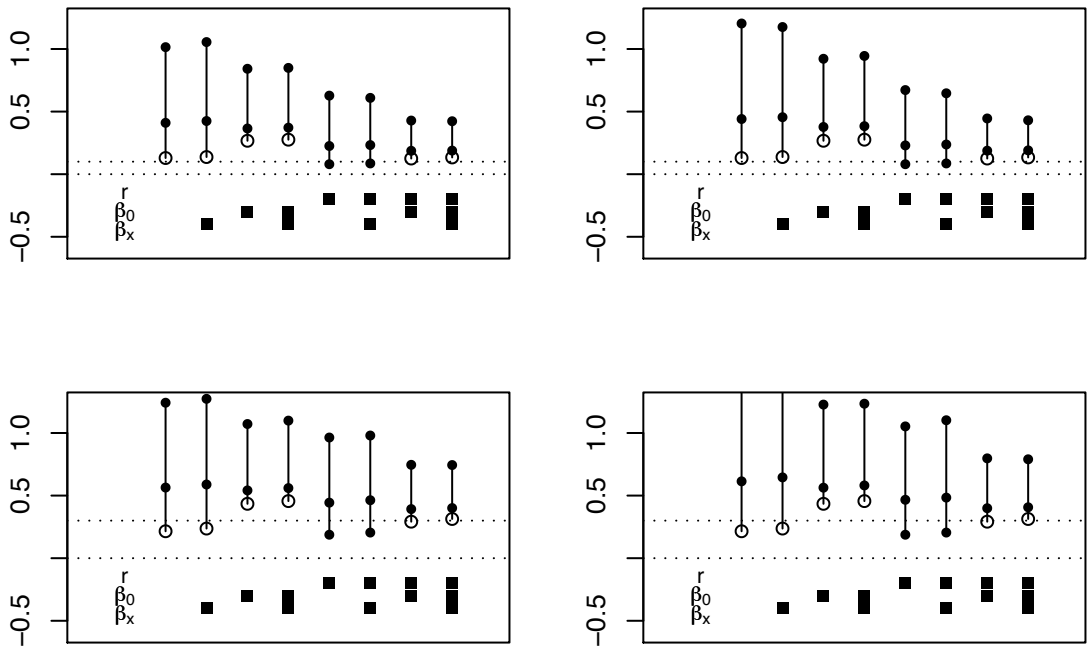

Figure 4: LPD for β_{xg} when the assumption of no main effect for gene is violated. All settings are as per Figure 1. While the model still assumes (1), the true relationship additionally involves a main effect of G, with $\beta_g = 0.025$.

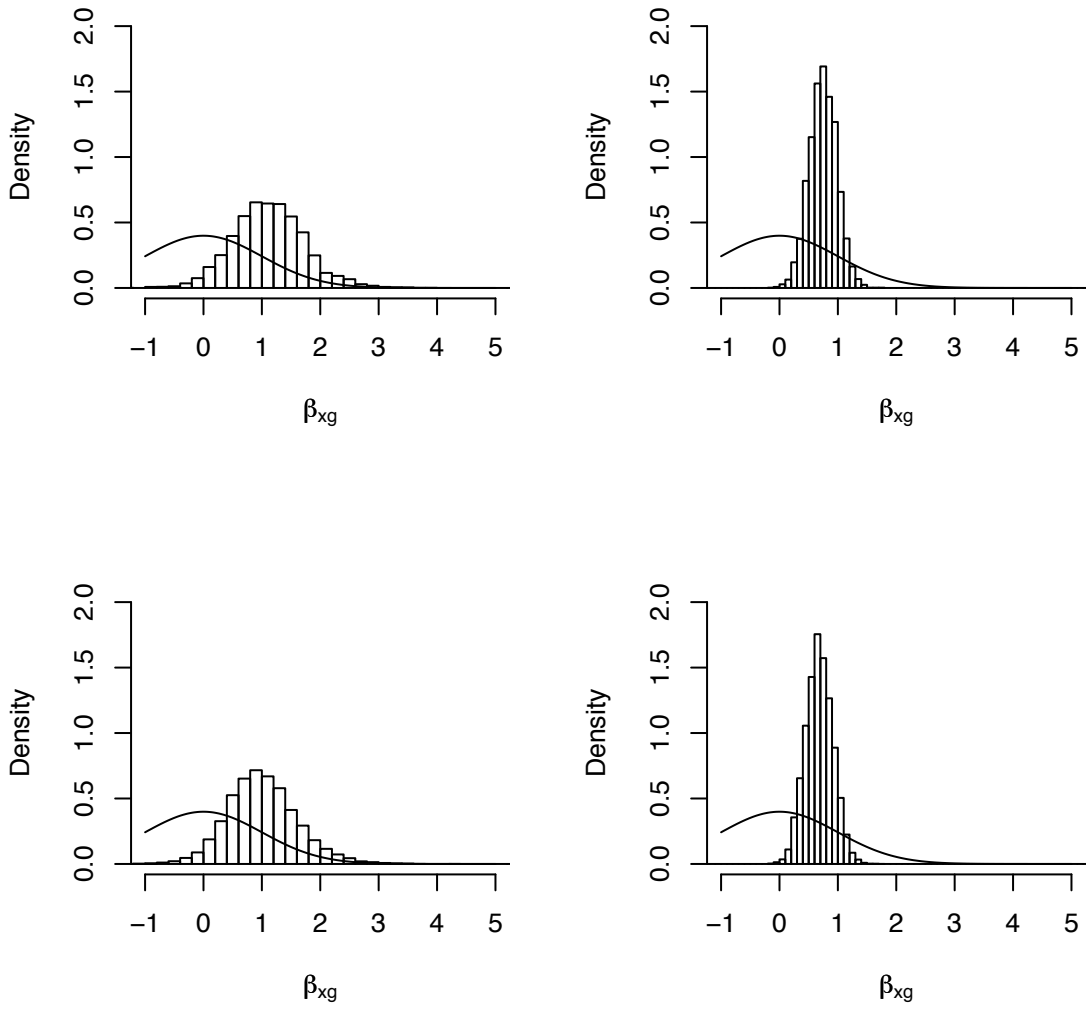

Figure 5: Posterior distribution on the interaction coefficient β_{xg} for the bladder cancer study. The top (bottom) panels correspond to prospective (retrospective) analysis. The left (right) panels correspond to reduced (full) data. The superimposed curve is the prior density.