The International Journal of Biostatistics

Volume 8, Issue 2

2012

Article 2

CAUSAL INFERENCE IN HEALTH RESEARCH

Bias Analysis to Guide New Data Collection

Timothy L. Lash, *Aarhus University Hospital*

Thomas P. Ahern, *Brigham and Women's Hospital and
Harvard Medical School*

Recommended Citation:

Lash, Timothy L. and Ahern, Thomas P. (2012) "Bias Analysis to Guide New Data Collection,"
The International Journal of Biostatistics: Vol. 8: Iss. 2, Article 2.

DOI: 10.2202/1557-4679.1345

Bias Analysis to Guide New Data Collection

Timothy L. Lash and Thomas P. Ahern

Abstract

Bias analysis serves multiple objectives in epidemiologic data analysis. The objectives most often emphasized are quantification of uncertainty due to systematic errors and reduction in overconfidence by specifying hypotheses that compete with the causal hypothesis. A third objective is the utility of bias analysis to identify strategies for new data collection that will be productive in evaluating the validity of an association. The authors illustrate the value of this objective using two examples. The first example examines the value of comprehensive CYP2D6 genotyping in a study of tamoxifen resistance. Tamoxifen is metabolized primarily by CYP2D6 to more active forms. More than thirty polymorphisms in the CYP2D6 gene reduce its function. We genotyped the most prevalent CYP2D6 polymorphism and found a null association between genotype and breast cancer recurrence in a Danish population. One possibility is that incomplete genotyping of the multiple functional polymorphisms introduced non-differential misclassification and biased the association toward the null. We used bias analysis to evaluate the plausibility of this explanation and to guide a decision about devoting study resources toward more comprehensive genotyping of other polymorphisms in the CYP2D6 gene. The second example examines the association between vitamin K antagonist (VKA) therapy and the incidence of 24 site-specific cancers, using heart valve replacement as an instrumental variable. Earlier studies suggested a protective association between VKA anticoagulants and the incidence of cancer. We observed a null-centered distribution of associations, which may be due to non-differential misclassification of VKA therapy by the instrument. We used bias analysis to evaluate whether this misclassification was likely to explain the null-centered distribution of associations and to guide decisions about conducting a more expensive validation study. In the first example, the bias analysis showed that new data collection would be required to resolve the uncertainty, whereas the second example showed that new data collection was unlikely to be a productive use of scarce study resources.

KEYWORDS: epidemiologic methods, bias analysis, causal inference

Author Notes: Timothy L. Lash, Department of Clinical Epidemiology, Aarhus University Hospital. Thomas P. Ahern, Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School. This work was supported by grants from the Karen Elise Jensen Foundation, the Regional Clinical Epidemiological Monitoring Initiative for Central and North Denmark Regions, the US National Cancer Institute at the National Institutes of Health (grant numbers R01 CA118708 and T32 CA09001-35) and the Congressionally Directed Medical Research Programs (grant number BC073012). The authors thank the Danish Breast Cancer Cooperative group for access to its registry data and for preparing the initial dataset for the first example. The granting agencies had no role in the design of the study; the collection,

analysis, and interpretation of the data; the writing of the manuscript; or the decision to submit the manuscript for publication.

Introduction

Bias analysis is the process of quantifying the direction, magnitude, and uncertainty of the bias affecting an estimate of association (Greenland and Lash, 2008; Lash *et al.*, 2009a). This process is also sometimes called “sensitivity analysis,” but we prefer to reserve that term to describe the process of varying the assumptions made as part of the bias analysis (*i.e.*, the process of evaluating the sensitivity of the bias analysis results to the underlying assumptions). Several objectives of quantitative bias analysis have been proposed. The most common objective is to obtain an estimate of the direction and magnitude of the bias induced by a source of systematic error (Lash *et al.*, 2009a). A second objective is to reduce the human tendency towards overconfidence in a particular study’s result (Lash *et al.*, 2009a). Quantitative bias analysis requires specification of alternative non-causal explanations for a given association. The act of specifying these non-causal (bias) explanations reduces the tendency towards overconfidence in the causal hypothesis, regardless of the results of the bias analysis (Lash, 2007).

A third objective of bias analysis is to guide new data collection. This objective has been less often emphasized, although it has been implicitly stated throughout the history of bias analysis literature. In 1954, Bross recognized that time and cost factors might require a cruder measure of exposure classification than a more accurate measure (Bross, 1954), and in 1967 he wrote about the utility of methods that distinguish between confounders that could, or could not, plausibly account for a given association (Bross, 1967). Left unsaid was that, in the former case, new data collection would be required to evaluate the role of the potential confounder. This deficiency was shared, for example, by Greenland’s 1996 review of bias analysis methods (Greenland, 1996). By 2008, Greenland and Lash hinted at the value of bias analysis in guiding new data collection, writing that “...comparisons (of alternative assumptions) allow observers to isolate more easily sources of disagreement...helping to move debates beyond qualitative assertions and counterassertions” (Greenland and Lash, 2008). Once again, the destination of the movement beyond “qualitative assertions and counterassertions” was left unsaid, but presumably included new data collection. It was not until 2009 that this objective was explicitly stated (Lash *et al.*, 2009a): “A second utility of bias analysis is as an aid to identifying points of departure in interpretation among stakeholders...(and thereby to) provide guidance for further research.”

The utility of bias analysis in identifying systematic errors that could explain a given result, versus those that could not plausibly explain a given result, has therefore long been understood. Explicit illustrations of this utility have not been previously published to our knowledge. Our aim, therefore, is to provide two

examples of bias analysis used to guide new data collection. In the first example, we conclude that new data collection is required, and in the second example, we conclude that it is not.

Example 1: CYP2D6 function and tamoxifen effectiveness

Motivation

Tamoxifen halves the risk of breast cancer recurrence in non-metastatic patients with estrogen receptor-positive (ER+) tumor cells (2005). While the risk reduction is substantial, some tamoxifen-treated women develop a recurrence and many then die of the disease. Accurate markers of tamoxifen resistant tumors would allow personalization of combined therapies (Riggins *et al.*, 2007), with the goal of preventing these recurrences. Tamoxifen is metabolized mostly by the gene product of *CYP2D6* (Stearns *et al.*, 2003; Collier *et al.*, 2002) to secondary metabolites that bind the estrogen receptor 100-fold more readily than tamoxifen itself (Desta *et al.*, 2004), so they are the most important inhibitors of tumor cell growth (Lim *et al.*, 2005). In the past five years, it has been suggested that ER+ breast cancer patients with nonfunctional alleles of *CYP2D6* may be poor candidates for adjuvant tamoxifen therapy (Desta and Flockhart, 2007; Jordan, 2007).

To address limitations of earlier research (Lash *et al.*, 2009b), and to provide a precise estimate obtained from a large, well-identified study population, we conducted a study of the association between genetic markers of *CYP2D6* inhibition and breast cancer recurrence (Lash *et al.*, 2011) nested in a population-based clinical registry of Danish breast cancer patients (Blichert-Toft *et al.*, 2008). We identified 541 recurrent or contralateral breast cancers among women with estrogen receptor positive (ER+) disease treated with tamoxifen for at least one year and matched one control subject per case patient on ER status, tamoxifen treatment, menopausal status, stage, calendar time, and county. We assessed genetic inhibition by genotyping the most prevalent *CYP2D6* knockout allele among Caucasians—*CYP2D6**4 (Sachse *et al.*, 1997; Myrand *et al.*, 2008). We estimated the odds ratio associating *CYP2D6* inhibition with breast cancer recurrence and adjusted for potential confounding with logistic regression. The adjusted odds ratio associating genetically reduced function (zero or one functional *CYP2D6**4 allele, compared with two functional *CYP2D6**4 alleles) with breast cancer recurrence equaled 1.1 (95% CI 0.81, 1.4).

Although we expected *CYP2D6**4 to be the most prevalent reduced-function allele in our study population, there are more than seventy known genetic mutations in *CYP2D6* (<http://www.cypalleles.ki.se/cyp2d6.htm>), with more than half having some functional consequence. While tamoxifen-treated women with

one or two *CYP2D6**4 alleles are likely to have reduced concentrations of the most active metabolites (Stearns *et al.*, 2003; Collier *et al.*, 2002; Gjerde *et al.*, 2008; Jin *et al.*, 2005), it is also possible that women with two functional *CYP2D6**4 alleles may carry other non-functional mutations that reduce the concentration of these metabolites. From a genetic epidemiology perspective, the problem involves non-differential misclassification of metabolizer phenotype because the classification system relies on only one of the alleles with functional consequences. This misclassification, which should be non-differential and independent of other errors, is expected to bias the odds ratio associating reduced CYP2D6 function toward the null. In particular, the reference group (two functional *CYP2D6**4 alleles) is a mix of women with no genetic inhibition of CYP2D6 function and women with genetic inhibition of CYP2D6 function arising from mutations other than *4. The question is whether this misclassification is sufficient to explain our study's null result.

Other investigators have examined the same question in relation to their study's results. For example, Schroth *et al.* examined this question in their cohort of breast cancer patients (Schroth *et al.*, 2010). In that study, genotyping only the *CYP2D6**4 allele categorized 5.5% of the breast cancer patients as poor metabolizers and 33% as intermediate metabolizers, and resulted in a hazard ratio of 1.3 (95% CI 0.48, 3.7) associating reduced CYP2D6 function with recurrence. More comprehensive genotyping categorized 8.3% of the patients as poor metabolizers and 54% as intermediate metabolizers, and resulted in a hazard ratio of 2.9 (95% CI 1.4, 6.1). Thompson similarly reported that the hazard ratio in their cohort obtained from limited *CYP2D6* genotyping equaled 1.3 (95% CI 0.7, 2.2), whereas with comprehensive genotyping it equaled 2.0 (95% CI 1.1, 3.7) (Thompson *et al.*, 2011). In contrast, Abraham *et al.* reported a hazard ratio of 1.2 (95% CI 0.82, 1.9) based on only the *CYP2D6**4 allele, which changed to only 1.4 (95% CI 0.84, 2.2) based on comprehensive genotyping (Abraham *et al.*, 2010). It was crucial, therefore, that we determine whether our null result could have been explained by the fact that we genotyped only *CYP2D6**4.

Quantitative bias analysis

To evaluate the potential bias from incomplete information on CYP2D6 function, we conducted a quantitative bias analysis (Lash *et al.*, 2009a) informed by classifications based on comprehensive genotyping of the *CYP2D6* gene in the Schroth study of German breast cancer patients (Schroth *et al.*, 2010) (Table 1). We assume this cohort of breast cancer patients should be genetically similar to the Danish patients included in our case-control study. As shown in the table, we used the Schroth *et al.* study to estimate the sensitivity and specificity of CYP2D6 functional classification based on genotyping only the *CYP2D6**4 allele, as well

as the positive-predictive value of reduced CYP2D6 function in controls based on genotyping only the *CYP2D6**4 allele. We set the prevalence of reduced CYP2D6 function in controls equal to the prevalence observed in the German study population based on the comprehensive genotyping data. For the probabilistic bias analysis, we assigned beta distributions to all of these classification parameters using standard methods (α =numerator+1, β =total–numerator+1) (Lash *et al.*, 2009a) (Figure 1).

Table 1: Cross-tabulation of CYP2D6 function inferred from comprehensive genotyping (rows, the assumed gold-standard) or from only *CYP2D6**4 genotype (columns, assumed susceptible to misclassification of function), as reported by Schroth *et al.* in a cohort of German breast cancer patients (Schroth *et al.*, 2010)

	≥ 1 *4 allele (reduced function)	No *4 allele (normal function)	total
reduced function	187	116	303
normal function	3	186	189
total	190	302	492

Sensitivity (s)=187/303=0.62

PPV controls (P_{co})=187/190=0.98†

Specificity (t)=186/189=0.98

NPV controls (N_{co})=186/302=0.62

Prevalence of reduced function in controls (p_{co})= 303/492=0.62

†PPV does not equal 100% because some women with the *4 allele also carry a highly active form of the *CYP2D6* enzyme, which compensates for the loss of function in the *4 allele.

The positive predictive value in cases (P_{ca}) and negative predictive value in cases (N_{ca}) are functions of sensitivity (s), specificity (t), and the prevalence of reduced CYP2D6 function in cases (p_{ca}):

(1)

$$P_{ca} = \frac{sp_{ca}}{sp_{ca} + (1 - t)(1 - p_{ca})}$$

$$N_{ca} = \frac{t(1 - p_{ca})}{p_{ca}(1 - s) + t(1 - p_{ca})}$$

The prevalence of reduced CYP2D6 function in cases (p_{ca}) is a function of the prevalence in controls and the association between reduced function and breast cancer recurrence (OR):

(2)

$$p_{ca} = \frac{e^{\ln \frac{p_{co}}{1-p_{co}} + \ln OR}}{1 + e^{\ln \frac{p_{co}}{1-p_{co}} + \ln OR}}$$

Figure 1: Beta probability density distributions assigned to the positive predictive value (P_{co}) and negative predictive value (N_{co}) of CYP2D6 function in controls based on genotyping only the *CYP2D6**4 allele, as informed by genotyping data reported by Schroth *et al.* in a cohort of German breast cancer patients (Schroth *et al.*, 2010).

OR , however, is the parameter that we wished to estimate. We substituted in its place two probability density distributions assigned to $\ln OR$ (Figure 2). The first (Distribution 1) was a beta distribution with $\alpha=1.83$ and $\beta=4.54$ scaled to the interval $\ln|1|$ to $\ln|2.77|$. This distribution has minimum $OR=1$ (null association), maximum $OR=2.77$ (the strongest association reported in the study of German breast cancer patients (Schroth *et al.*, 2010)), and mode of $OR=1.22$ (the result of

a recent meta-analysis (Seruga and Amir, 2010) of this topic among Caucasians ($OR=1.22$, 95% CI 0.88, 1.68)). The second (Distribution 2) was a lognormal distribution ($\mu=\ln|2.77|$, $\sigma=(\ln|5.89|-\ln|1.31|)/3.92$) based on the result furthest from the null ($HR=2.77$, 95% CI 1.31, 5.89) from among those reported by Schroth *et al* (Schroth *et al.*, 2010) associating reduced CYP2D6 function inferred from comprehensive CYP2D6 genotyping with breast cancer recurrence.

Note that $P_{ca} \approx P_{co} \approx 1$. Therefore, when a draw from a distribution assigned to $\ln|OR|$ yields $OR>1$ —which is required for all draws from Distribution 1 and which is expected for greater than 99.6% of the draws from Distribution 2—a case classified as having normal function based only on having no *4 allele was more likely to be reclassified as truly having reduced function than an analogous control. This relation would tend to increase the modeled prevalence of reduced function more in cases than in controls, resulting in a simulated OR greater than the OR reported above from the conventional analysis. $OR<1$ is given little weight by both distributions, since there is no biologic mechanism by which reduced CYP2D6 function could prevent breast cancer recurrence.

Figure 2: Probability density distributions assigned to Distribution 1 and Distribution 2; two distributions assigned to the association between reduced CYP2D6 function and breast cancer recurrence ($\ln OR$). See text for explanations of how the parameters of the distributions were assigned.

With values selected from each of the distributions, we completed a record-level probabilistic bias analysis following established methods (Lash *et al.*, 2009a). The outline for the statistical computing algorithm was as follows.

Let

$X=1$ indicate carrying at least one reduced function *CYP2D6**4 allele,
 $X=0$ indicate carrying no reduced function *CYP2D6**4 allele,
 $X^*=1$ indicate truly reduced *CYP2D6* function,
 $X^*=0$ indicate truly full *CYP2D6* function,
 $Y=1$ indicate recurrence case status, and
 $Y=0$ indicate recurrence case status.

For $i=1$ to 100,000

Select $P_{co,i}$, $N_{co,i}$, $p_{co,i}$, s_i , and t_i from their beta distributions
 Select OR_i from its distribution and calculate $p_{ca,i}$ by equation (2)
 Calculate $P_{ca,i}$, $N_{ca,i}$ by equation (1)
 For $ii=1$ to last data record

Select u from a standard uniform
 If $X_{i,ii}=1$ and $Y_{i,ii}=1$ and $u > P_{ca,i}$ then $X^*_{i,ii}=0$, else $X^*_{i,ii}=1$
 Else if $X_{i,ii}=1$ and $Y_{i,ii}=0$ and $u > P_{co,i}$ then $X^*_{i,ii}=0$, else $X^*_{i,ii}=1$
 Else if $X_{i,ii}=0$ and $Y_{i,ii}=1$ and $u > N_{ca,i}$ then $X^*_{i,ii}=1$, else $X^*_{i,ii}=0$
 Else if $X_{i,ii}=0$ and $Y_{i,ii}=0$ and $u > N_{co,i}$ then $X^*_{i,ii}=1$, else $X^*_{i,ii}=0$

Next ii

Next i

The algorithm yields 100,000 datasets, each with theoretically different subsets of persons who have $X^* \neq X$. For each dataset, we estimate $OR_{X^*Y,i}$ using logistic regression with adjustment for the same set of covariates as in the conventional analysis. We reported the median of these 100,000 estimates as the point estimate for Distribution 1 and Distribution 2, and the 2.5th and 97.5th percentiles as the 95% simulation intervals.

Results and interpretation

The frequency of the *CYP2D6**4 minor allele was 24% among recurrent cases and 22% among controls, which compares well with the minor allele frequency reported in reference populations of Caucasians (Sachse *et al.*, 1997; Myrand *et al.*, 2008). As reported above, the conventional adjusted odds ratio associating genetically reduced function (zero or one functional *CYP2D6**4 allele) with breast cancer recurrence equaled 1.1 (95% CI 0.81, 1.4). The probabilistic bias analysis model using Distribution 1 yielded a median OR=1.3 (95% simulation interval

0.87, 1.9) and the probabilistic bias analysis model using Distribution 2 yielded a median OR=2.3 (95% simulation interval 1.2, 4.8).

The conventional result (OR=1.1, 95% CI 0.81, 1.4) suggests a null or small association between reduced CYP2D6 function and breast cancer recurrence in ER+ breast cancer patients treated with at least one year of tamoxifen. The near-null association may accurately reflect the true effect, and there are reasonable biologic explanations for why reduced CYP2D6 function may not reduce the clinical effectiveness of this endocrine therapy (Lash *et al.*, 2009b). In this analysis, the women classified as having reduced CYP2D6 function based on carrying at least one *4 allele very likely did have reduced function (*i.e.*, carrying at least one *CYP2D6**4 allele has a high positive predictive value of reduced CYP2D6 function). However, some of the women classified as having normal CYP2D6 function based on having no *4 allele very likely did have reduced function. This misclassification problem provides a second explanation for the null result obtained from the conventional analysis.

To evaluate the bias quantitatively, we developed and parameterized the above bias model. The model, however, required an assumption about the very association we were trying to measure, and at least two very reasonable distributions could be assigned to the association. The first distribution, with most of its density nearer the null than the second distribution, yielded a median estimate and simulation interval near the null, and very near the conventional result. As noted above, this result is biologically plausible and consistent with the most current meta analysis (Cronin-Fenton and Lash, 2011); but one must also remember that the studies included in the meta analyses had widely varying comprehensiveness of *CYP2D6* genotyping (Cronin-Fenton and Lash, 2011). The second distribution, with most of its density centered on a 2.8-fold risk ratio [as reported in a similar cohort of breast cancer patients (Schroth *et al.*, 2010)], yielded a median estimate and simulation interval near to a doubling of the recurrence risk. Pooled clinical trial results show that tamoxifen reduces the risk of recurrence by about half in trial participants selected without regard to genotype (2005). A two-fold reduction in the effectiveness of tamoxifen therapy among those with genetically inhibited CYP2D6 function is, therefore, also plausible (Cronin-Fenton and Lash, 2011; Lash *et al.*, 2009b). We conclude that the bias analysis cannot reliably evaluate the potential bias, conditional on the accuracy of the bias model, and that further data collection is therefore required. In particular, additional reduced-function alleles of the *CYP2D6* gene should be assayed among the original study participants to better categorize those with no *CYP2D6**4 allele as having normal or reduced CYP2D6 function.

Example 2: Vitamin K Antagonist therapy and site-specific cancer incidence, using heart valve replacement as an instrumental variable

Motivation

Vitamin K antagonists (VKAs)—such as dicoumarol, warfarin, and phenprocoumon—are oral anticoagulants used since the 1950s (Goodman and Gilman A., 2006) to prevent stroke in patients with replacement heart valves (Goodman and Gilman A., 2006; Katzung, 2004) and for other indications. Studies have suggested that VKA users have a reduced risk of some cancers of the urogenital tract (Schulman and Lindmarker, 2000; Blumentals *et al.*, 2004; Tagalakakis *et al.*, 2007; Pengo *et al.*, 2011), but others have reported near null associations overall (Taliani *et al.*, 2003) and at other sites of the urogenital track (Tagalakakis *et al.*, 2007). A review concluded that current evidence for an association between VKA therapy and cancer is inconclusive (Pengo *et al.*, 2010).

We therefore conducted a nationwide Danish cohort study of the associations between VKA therapy and the incidence of 24 different cancers. The Danish National Registry of Patients (DNRP) has electronically recorded all surgical procedures since 1977. The Danish National Registry of Medicinal Products, which can provide data directly on VKA prescriptions, only dates to 1995. We capitalized on the longer data collection period of the DNRP to extend follow-up by using heart valve replacement surgery as an instrumental variable (IV) for VKA treatment.

Use of this variable also permits estimation of unconfounded associations between VKA therapy and cancer incidence, assuming the IV conditions are satisfied (Hernan and Robins, 2006). To evaluate these conditions, we constructed a directed acyclic graph (DAG; Figure 3), which depicts the hypothesized relations between heart valve replacement, VKA prescription, and cancer incidence. Assuming that the DAG faithfully depicts the relations, the conditions required for an IV are met because (a) there is no evidence of a causal relation between heart valve replacement and the incidence of any cancer (i.e., no arrow ‘b’ or its equivalent), (b) heart valve replacement requires lifelong VKA therapy (Aurigemma and Gaasch, 2010), and the prevalence of VKA treatment is low in the general population (i.e., there exists a strong causal association depicted by arrow ‘a’), and (c) after matching valve recipients to non-recipients on age and sex, we expect no important uncontrolled common causes of valve replacement and cancer incidence (i.e., no other variables that satisfy the U2 node for any of the site-specific cancers examined (Greenland *et al.*, 1999)). If the IV conditions are met, then bias due to residual confounding of the VKA/cancer associations

(node U_1 in Figure 3) is negated, at the expense of non-differential misclassification of VKA exposure by the instrument (Hernan and Robins, 2006; Greenland, 2000).

Figure 3: Directed Acyclic Graph (DAG) depicting the conditions necessary for heart valve replacement to serve as an instrumental variable (Z) for the associations between VKA therapy (X) and site-specific cancer incidence (Y).

We enrolled 9,727 Danish residents who received a replacement heart valve between 1989 and 2006 and matched 95,481 individuals without a history of valve replacement to them on age and sex. We used the heart valve replacement instrument to estimate rate ratios (RR) associating VKA therapy with incidence of the 24 site-specific cancers using Poisson regression models. These instrumental variable associations were plotted according to the inverse normal of rank percentile, and the pattern of associations was consistent with a null-centered Gaussian distribution as depicted by the fact that the fitted line, weighted by the inverse-variance of each association, very nearly intersects the origin (Figure 4). These results do not support the hypothesis that VKA therapy is associated with any reduced site-specific cancer risk. A more detailed description of the methods can be found in the Appendix. A complete description of the methods and results, along with site-specific associations and their intervals, are published elsewhere (Ahern *et al.*, 2011).

Figure 4: Plot of site-specific cancer incidence rate ratios, estimated by using heart valve replacement as an instrumental variable for VKA therapy, according to inverse normal of rank percentile.

It is possible that this result reflects a true null association between VKA therapy and site-specific cancer incidence. It is also possible, though, that the inherent non-differential misclassification of VKA therapy resulting from use of heart valve replacement as an IV has biased the associations toward the null. The question is whether this bias is sufficient to explain our study's null results, and thereby to have masked a true underlying preventive effect.

Quantitative bias analysis

To begin, we conducted a validation substudy among subjects whose residence and index dates fell within the coverage period of county-specific prescription registries. Patients prescribed a VKA must initially pay the entire medication cost. Once paid, the medication is dispensed and the transaction is recorded in the prescription registry. The Danish government subsequently reimburses the patient for a portion of the cost. These registries allowed us to determine whether members of the validation subset ($n=24,647$) received a VKA prescription during follow-up. This validation study yielded positive predictive values of $P_{ca}=0.96$

and $P_{co}=0.97$ among persons with and without an incident cancer, respectively, and negative predictive values of $N_{ca}=0.88$ and $N_{co}=0.92$ (Table 2).

Table 2: Stratified cross-tabulation of VKA therapy recorded in the county prescription registries (rows, the assumed gold-standard) versus history of heart valve replacement (columns, the IV assumed susceptible to misclassification of function), in the validation subset (n=24,647).

Any cancer	heart valve replacement	no heart valve replacement
VKA prescription	220	223
no VKA prescription	10	1634
total	230	1857
PPV cases (P_{ca})=220/230=0.96 NPV cases(N_{ca})=1634/1857=0.88		
No cancer	heart valve replacement	no heart valve replacement
VKA prescription	2240	1704
no VKA prescription	65	18,551
total	2305	20,255
PPV non-cases (P_{co})=2240/2305=0.97 NPV non-cases(N_{co})=18,551/20,255=0.92		

IV analyses commonly use two-stage linear regression to estimate associations. In the first stage, the outcome is regressed on the instrument—as we have done—which yields an unconfounded estimate of the exposure-outcome association that is distorted by misclassification (Brookhart *et al.*, 2006b; Brookhart *et al.*, 2006a). In the second stage, the instrument-outcome association is scaled by the instrument-exposure association, yielding an unconfounded estimate of the exposure-outcome risk difference, adjusted for misclassification of the target exposure (Brookhart *et al.*, 2006b). This approach requires that validation data are available for all subjects, but in our study only about 25% of subjects had data on VKA prescription. We could not, therefore, complete the usual second stage regression to correct for the inherent exposure misclassification with the available validation data.

It would be possible to collect the VKA prescription history by medical record review for many of the 70,834 participants without prescription data recorded in the registry. This task would require visits to general practitioner offices as well as specialty clinics where valve replacement recipients might have been followed and prescribed VKA. Some of these practices and clinics would no longer be in operation. The costs to complete the new data collection would be very high, especially given the doubtful feasibility of collecting adequate data for at least some of subjects by manual chart review. To evaluate whether the effort was worthwhile, we undertook a probabilistic bias analysis.

Estimation of the predictive values (Table 2) enabled us to adjust for misclassification of exposure in the non-validated subset of the cohort, yielding VKA/cancer associations rooted in a larger body of data. To begin, we allowed for uncertainty in the predictive values by assigning beta distributions to each, following the algorithm described above for determining α and β . Tables for the site-specific cancer associations were arranged according to the example in the top of Table 3. The subscript ‘val’ denotes that the observed cell frequencies arise from the validation subset, in which we assumed that VKA exposure status was correctly classified. Site-specific cancer association data in the non-validation subset were arranged according to the example in the bottom of Table 3, where subscript ‘mis’ denotes that the observed cell frequencies arise from the non-validation subset and are subject to misclassification of VKA exposure by the heart valve replacement proxy variable.

Table 3: Data arrangement for the j th site-specific cancer association with known VKA exposure data in the prescription validation subset.

(Validated subset)	VKA +	VKA –
Cancer cases	$a_{val,j}$	$b_{val,j}$
Person-years	$PT1_{val,j}$	$PT0_{val,j}$
(Non-validated subset)	Heart valve +	Heart valve –
Cancer cases	$a_{mis,j}$	$b_{mis,j}$
Person-years	$PT1_{mis,j}$	$PT0_{mis,j}$

The outline for the statistical computing algorithm was then as follows. Let

- $a_{val,j}$ = number of cancer cases at site j among those with documented VKA prescription,
- $b_{val,j}$ = number of cancer cases at site j among those with no VKA prescription documented in the prescription registry,
- $PT1_{val,j}$ = the site-specific person-time at risk for site j contributed by those exposed to VKA according to the prescription registry,
- $PT0_{val,j}$ = the site-specific person-time at risk for site j contributed by those unexposed to VKA according to the prescription registry,
- $a_{mis,j}$ = number of cancer cases at site j among those with heart valve replacement surgery and not in the validation subset,
- $b_{mis,j}$ = number of cancer cases at site j among those with no heart valve replacement surgery and not in the validation subset,
- $PT1_{mis,j}$ = the site-specific person-time at risk for site j contributed by those with heart valve replacement surgery and not in the validation subset,
- $PT0_{mis,j}$ = the site-specific person-time at risk for site j contributed by those with no heart valve replacement surgery and not in the validation subset,
- $SE_{RR,j}$ = the standard error of the natural logarithm of the conventional site-specific incidence rate ratio at site j

For $i=1$ to 100,000

Select u_1 and u_2 from the standard uniform with correlation $r=0.8$ (as recommended and discussed elsewhere (Greenland and Lash, 2008))

Select u_3 and u_4 from the standard uniform with correlation $r=0.8$

Select $P_{ca,i}$, $P_{co,i}$, $N_{ca,i}$, $N_{co,i}$ from their beta distributions using u_1 to u_4 , respectively

For $j=1$ to 24

Select $z_{i,j}$ from the standard normal

Calculate site-specific rate ratios as

$$a_{i,j} = P_{ca,i}a_{mis,j} + (1 - N_{ca,i})b_{mis,j} + a_{val,j}$$

$$b_{i,j} = N_{ca,i}b_{mis,j} + (1 - P_{ca,i})a_{mis,j} + b_{val,j}$$

$$PT1_{i,j} = P_{co,i}PT1_{mis,j} + (1 - N_{co,i})PT0_{mis,j} + PT1_{val,j}$$

$$PT0_{i,j} = N_{co,i}PT0_{mis,j} + (1 - P_{co,i})PT1_{mis,j} + PT0_{val,j}$$

$$RR_{i,j} = e^{\ln\left(\frac{a_{i,j}/PT1_{i,j}}{b_{i,j}/PT0_{i,j}}\right) - z_{i,j}SE_{RR,j}}$$

Next j

Next i

The RR_{ij} value from each iteration was stored, and the bias analysis routine was iterated until 100,000 estimates were accumulated. Thus, for each site-specific cancer outcome, we generated a distribution of 100,000 misclassification-corrected estimates. We reported the median of each distribution as the point estimate for each association and 2.5th and 97.5th percentiles as the 95% simulation interval.

Results and interpretation

Figure 5 shows the plot of the median IRRs and 95% simulation intervals for the site-specific cancer associations according to their inverse normal of rank percentile, after probabilistic misclassification adjustment. The effect of misclassification adjustment was to positively displace the distribution of IRR estimates, as depicted by the fact that the fitted line, weighted by the inverse-variance of each association, intersects the y-axis above the origin. In addition, the misclassification adjustment modestly widened the intervals about the estimated associations. The positive displacement occurred primarily because of the difference in negative predictive value among cases ($N_{ca}=0.88$) versus non-cases ($N_{co}=0.92$), which results in cases without history of heart valve replacement being reclassified as exposed to VKA more often than the person-time of persons without history of heart valve replacement being reclassified as exposed to VKA. This difference in predictive values may be artifact, but might also arise from VKA-therapy being prescribed to persons with venous thromboembolism, a condition that often precedes a cancer diagnosis (Prandoni and Piccioli, 2006).

The positive displacement of the distribution of associations following probabilistic bias analysis suggests that the IV results were not biased towards the null from a true underlying protective distribution. Given the generally very high predictive values in the validation study set, and the displacement in a causal rather than protective association observed in the probabilistic bias analysis, we concluded that collecting the medical record review required to increase the completeness of validation data would be a poor use of limited research resources. The distribution of associations between VKA therapy and cancer incidence is likely to be null, and very unlikely to be protective.

Figure 5: Plot of site-specific cancer incidence rate ratios, estimated by probabilistic bias analysis methods.

Discussion

Epidemiologic data collection and analysis often require substantial resources, including financial resources to support data collection, potentially long durations over which the data are collected, and an investment of intellectual effort in the design of the data collection methods and analysis of the collected data. These resources should be expended with care, particularly when the gain in information obtained by expending additional resources is expected to be marginal.

With a particular result in hand, bias analysis allows evaluation of the plausibility of competing explanations for the observed association. One explanation is usually the causal hypothesis, and competing explanations are often sources of systematic error. In both of our examples, non-differential misclassification of the exposure was the systematic error of primary concern. Other types of bias analysis, however, allow evaluation of the potential for any source of systematic error to explain the association as well as of the causal hypothesis, conditional on the validity of the bias model (Greenland and Lash, 2008). When systematic errors provide an equally plausible explanation as the causal hypothesis, new data collection may be well-justified to resolve the competition. When systematic errors do not provide an equally plausible explanation as the causal hypothesis, new data collection may well be avoided, and scarce research resources can then be expended on more productive endeavors.

Appendix (reprinted with permission from the *American Journal of Epidemiology*)

Study population and data collection

The source population for this study was the population of Denmark between 1 January 1989 and 31 December 2006. We identified all heart valve replacement surgeries in Denmark during the study period by searching the DNRP for relevant procedure codes. To allow a reasonable induction period, we began follow-up on an index date, defined as the date of valve replacement surgery plus one year. Replacements were for the mitral, aortic, tricuspid, or pulmonic valve, and included biological and artificial valves. We defined each subject's index date as the date of valve replacement surgery, and excluded any subject with a cancer history before, or within one year after, their index date. This roster comprised our exposed cohort. For each exposed subject we matched up to 10 unexposed subjects from the general population, sampled randomly and without replacement, within strata of birth year and sex. Unexposed subjects were assigned the same index date as their exposed counterparts, and were required to have no history of heart valve replacement or cancer on, or within one year after, that date.

Using each subject's unique Civil Personal Registry (CPR) number, we searched the Danish Cancer Registry (DCR) for diagnoses of 49 site-specific cancers. We used the DNRP to determine prevalent medical conditions on the index date, corresponding to the Charlson Comorbidity Index (CCI) (Charlson *et al.*, 1987), and to document history of two major indications for VKA therapy, venous thromboembolism and atrial fibrillation.

Definitions of analytic variables

Cancers diagnosed during follow-up were identified in the DCR using International Classification of Diseases, 10th edition (ICD-10) codes. The DCR has translated past records, entered under earlier ICD editions, into ICD-10 to standardize case ascertainment. All cancers with five or more cases among valve recipients were considered as separate outcomes, which reduced the number of cancer sites included in the analyses from 49 to 24.

For subjects without a cancer diagnosis, we characterized end of follow-up by linkage with the Danish Civil Registry, which updates residential address and vital status for all Danish residents on a daily basis. Each subject contributed person-time from one year after their index date until the first of (a) cancer diagnosis, (b) emigration from Denmark, (c) death from any cause, or (d) 31 December 2006.

Age was defined as the number of complete years between the birth date and index date. Based on diagnosis history as of the index date, we calculated the Charlson Comorbidity Index according to the published method (Charlson *et al.*, 1987). In addition to the diagnoses included in the CCI, we assessed history of atrial fibrillation, superficial and deep venous thrombosis, and pulmonary embolism using ICD-8 and ICD-10 codes to search the DNRP. We defined a positive history of venous thromboembolism as having been diagnosed with superficial or deep venous thrombosis and/or pulmonary embolism before the index date.

Instrumental variable analysis

In our age- and sex-matched cohort, heart valve replacement appears to satisfy the criteria to be an IV for the associations between VKA therapy and site-specific cancer incidence, as described in the manuscript text. IV analyses commonly use two-stage linear regression to estimate associations. In the primary stage, the outcome is regressed on the instrument, yielding an estimate of the exposure-outcome association that is unconfounded but distorted by misclassification (Brookhart *et al.*, 2006a; Brookhart *et al.*, 2006b). In the secondary stage, the instrument-outcome association is scaled by the instrument-exposure association, yielding an unconfounded estimate of the exposure-outcome risk difference, adjusted for misclassification of the target exposure (Brookhart *et al.*, 2006b). This approach requires complete data for both the instrument and the target exposure. In our study, all subjects had data on the instrument (heart valve replacement), but only about 25% of those subjects had data on the target exposure (VKA prescription). Under a traditional two-stage IV analysis framework, we could use our data to estimate theoretically unconfounded instrument-outcome associations. However, because not all subjects had VKA prescription data, we could not implement the secondary regression to correct the inherent exposure misclassification, which provided the motivation for the bias analysis described in the manuscript.

We estimated IRRs and 95% confidence intervals associating heart valve replacement or measured VKA prescription with site-specific cancer incidence in separate Poisson regression models. VKA status was the sole independent variable in each of these models, and the logarithm of person-years at risk served as the offset variable. To accommodate over-dispersion of Poisson parameters, we used generalized estimating equations (GEE), with covariance matrices initially presumed to be exchangeable.

References

- (2005) Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet* 365 (9472): 1687-1717.
- Abraham JE, Maranian MJ, Driver KE, Platte R, Kalmyrzaev B, Baynes C, Luccarini C, Shah M, Ingle S, Greenberg D, Earl HM, Dunning AM, Pharoah PD, Caldas C (2010) CYP2D6 gene variants: association with breast cancer specific survival in a cohort of breast cancer patients from the United Kingdom treated with adjuvant tamoxifen. *Breast Cancer Res* 12 (4): R64.
- Ahern TP, Pedersen L, Svaerke C, Rothman KJ, Sorensen HT, Lash TL (2011) The association between vitamin K antagonist therapy and site-specific cancer incidence, estimated by using heart valve replacement as an instrumental variable. *Am J Epidemiol* In Press.
- Aurigemma GP, Gaasch WH (2010) Antithrombotic therapy in patients with prosthetic heart valves. In UpToDate, Otto CM (ed) UpToDate: Waltham, MA
- Blichert-Toft M, Christiansen P, Mouridsen HT (2008) Danish Breast Cancer Cooperative Group--DBCG: History, organization, and status of scientific achievements at 30-year anniversary. *Acta Oncol* 47 (4): 497-505.
- Blumentals WA, Foulis PR, Schwartz SW, Mason TJ (2004) Does warfarin therapy influence the risk of bladder cancer? *Thromb Haemost* 91 (4): 801-805.
- Brookhart MA, Wang PS, Solomon DH, Schneeweiss S (2006a) Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* 17 (3): 268-275.
- Brookhart MA, Wang PS, Solomon DH, Schneeweiss S (2006b) Instrumental variable analysis of secondary pharmacoepidemiologic data. *Epidemiology* 17 (4): 373-374.
- Bross I (1954) Misclassification in 2x2 tables. *Biometrics* 478-489.
- Bross ID (1967) Pertinency of an extraneous variable. *J Chronic Dis* 20 (7): 487-495.

- Charlson ME, Pompei P, Ales KL, MacKenzie CR (1987) A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 40 (5).
- Coller JK, Krebsfaenger N, Klein K, Endrizzi K, Wolbold R, Lang T, Nussler A, Neuhaus P, Zanger UM, Eichelbaum M, Murdter TE (2002) The influence of CYP2B6, CYP2C9 and CYP2D6 genotypes on the formation of the potent antioestrogen Z-4-hydroxy-tamoxifen in human liver. *Br J Clin Pharmacol* 54 (2): 157-167.
- Cronin-Fenton DP, Lash TL (2011) Clinical epidemiology and pharmacology of CYP2D6 inhibition related to breast cancer outcomes. *Expert Rev Clin Pharmacol* 4 (3): 363-377.
- Desta Z, Flockhart DA (2007) Germline pharmacogenetics of tamoxifen response: have we learned enough? *J Clin Oncol* 25 (33): 5147-5149.
- Desta Z, Ward BA, Soukhova NV, Flockhart DA (2004) Comprehensive evaluation of tamoxifen sequential biotransformation by the human cytochrome P450 system in vitro: prominent roles for CYP3A and CYP2D6. *J Pharmacol Exp Ther* 310 (3): 1062-1075.
- Gjerde J, Hauglid M, Breilid H, Lundgren S, Varhaug JE, Kisanga ER, Mellgren G, Steen VM, Lien EA (2008) Effects of CYP2D6 and SULT1A1 genotypes including SULT1A1 gene copy number on tamoxifen metabolism. *Ann Oncol* 19 (1): 56-61.
- Goodman LS, Gilman A. (2006) Goodman & Gilman's the Pharmacological Basis of Therapeutics. McGraw-Hill: New York
- Greenland S, Lash TL (2008) Bias Analysis. In *Modern Epidemiology*, Rothman KJ, Greenland S, Lash TL (eds) pp 345-380. Lippincott Williams & Wilkins: Philadelphia
- Greenland S (1996) Basic methods for sensitivity analysis of biases. *Int J Epidemiol* 25 (6): 1107-1116.
- Greenland S (2000) An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 29 (4): 722-729.
- Greenland S, Pearl J, Robins JM (1999) Causal diagrams for epidemiologic research. *Epidemiology* 10 (1).

- Hernan MA, Robins JM (2006) Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 17 (4): 360-372.
- Jin Y, Desta Z, Stearns V, Ward B, Ho H, Lee KH, Skaar T, Storniolo AM, Li L, Araba A, Blanchard R, Nguyen A, Ullmer L, Hayden J, Lemler S, Weinshilboum RM, Rae JM, Hayes DF, Flockhart DA (2005) CYP2D6 genotype, antidepressant use, and tamoxifen metabolism during adjuvant breast cancer treatment. *J Natl Cancer Inst* 97 (1): 30-39.
- Jordan VC (2007) New insights into the metabolism of tamoxifen and its role in the treatment and prevention of breast cancer. *Steroids* 72 (13): 829-842.
- Katzung BG (2004) *Basic & Clinical Pharmacology*. Lange Medical Books/McGraw Hill: New York
- Lash TL (2007) Heuristic thinking and inference from observational epidemiology. *Epidemiology* 18 (1): 67-72.
- Lash TL, Cronin-Fenton D, Ahern TP, Rosenberg CL, Lunetta KL, Silliman RA, Garne JP, Sorensen HT, Hellberg Y, Christensen M, Pedersen L, Hamilton-Dutoit S (2011) CYP2D6 Inhibition and Breast Cancer Recurrence in a Population-Based Study in Denmark. *J Natl Cancer Inst* 103 (6): 489-500.
- Lash TL, Fox MP, Fink AK (2009a) *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer: New York
- Lash TL, Lien EA, Sorensen HT, Hamilton-Dutoit S (2009b) Genotype-guided tamoxifen therapy: time to pause for reflection? *Lancet Oncol* 10 (8): 825-833.
- Lim YC, Desta Z, Flockhart DA, Skaar TC (2005) Endoxifen (4-hydroxy-N-desmethyl-tamoxifen) has anti-estrogenic effects in breast cancer cells with potency similar to 4-hydroxy-tamoxifen. *Cancer Chemother Pharmacol* 55 (5): 471-478.

- Myrand S, Sekiguchi K, Man M, Lin X, Tzeng RY, Teng CH, Hee B, Garrett M, Kikkawa H, Lin CY, Eddy S, Dostalík J, Mount J, Azuma J, Fujio Y, Jang IJ, Shin SG, Bleavins M, Williams J, Paulauskis J, Wilner K (2008) Pharmacokinetics/Genotype Associations for Major Cytochrome P450 Enzymes in Native and First- and Third-generation Japanese Populations: Comparison With Korean, Chinese, and Caucasian Populations. *Clin Pharmacol Ther* 84 (3): 347-361.
- Pengo V, Denas G, Jose SP, Pengo MF (2010) Cancer prevention and vitamin K antagonists: an overview. *Thromb Res* 125 Suppl 2 S103-S105.
- Pengo V, Noventa F, Denas G, Pengo MF, Gallo U, Grion AM, Ilceto S, Prandoni P (2011) Long-term use of vitamin K antagonists and incidence of cancer: a population-based study. *Blood* 117 (5): 1707-1709.
- Prandoni P, Piccioli A (2006) Thrombosis as a harbinger of cancer. *Curr Opin Hematol* 13 (5): 362-365.
- Riggins RB, Schrecengost RS, Guerrero MS, Bouton AH (2007) Pathways to tamoxifen resistance. *Cancer Lett* 256 (1): 1-24.
- Sachse C, Brockmoller J, Bauer S, Roots I (1997) Cytochrome P450 2D6 variants in a Caucasian population: allele frequencies and phenotypic consequences. *Am J Hum Genet* 60 (2): 284-295.
- Schroth W, Hamann U, Fasching PA, Dauser S, Winter S, Eichelbaum M, Schwab M, Brauch H (2010) CYP2D6 polymorphisms as predictors of outcome in breast cancer patients treated with tamoxifen: expanded polymorphism coverage improves risk stratification. *Clin Cancer Res* 16 (17): 4468-4477.
- Schulman S, Lindmarker P (2000) Incidence of cancer after prophylaxis with warfarin against recurrent venous thromboembolism. Duration of Anticoagulation Trial. *N Engl J Med* 342 (26): 1953-1958.
- Seruga B, Amir E (2010) Cytochrome P450 2D6 and outcomes of adjuvant tamoxifen therapy: results of a meta-analysis. *Breast Cancer Res Treat* 122 (3): 609-617.

- Stearns V, Johnson MD, Rae JM, Morocho A, Novielli A, Bhargava P, Hayes DF, Desta Z, Flockhart DA (2003) Active tamoxifen metabolite plasma concentrations after coadministration of tamoxifen and the selective serotonin reuptake inhibitor paroxetine. *J Natl Cancer Inst* 95 (23): 1758-1764.
- Tagalakakis V, Tamim H, Blostein M, Collet JP, Hanley JA, Kahn SR (2007) Use of warfarin and risk of urogenital cancer: a population-based, nested case-control study. *Lancet Oncol* 8 (5): 395-402.
- Taliani MR, Agnelli G, Prandoni P, Becattini C, Moia M, Bazzan M, Ageno W, Tomasi C, Guazzaloca G, Ambrosio GB, Bertoldi A, Salvi R, Poggio R, Silingardi M (2003) Incidence of cancer after a first episode of idiopathic venous thromboembolism treated with 3 months or 1 year of oral anticoagulation. *J Thromb Haemost* 1 (8): 1730-1733.
- Thompson AM, Johnson A, Quinlan P, Hillman G, Fontecha M, Bray SE, Purdie CA, Jordan LB, Ferraldeschi R, Latif A, Hadfield KD, Clarke RB, Ashcroft L, Evans DG, Howell A, Nikoloff M, Lawrence J, Newman WG (2011) Comprehensive CYP2D6 genotype and adherence affect outcome in breast cancer patients treated with tamoxifen monotherapy. *Breast Cancer Res Treat* 125 (1): 279-287.