Health Services & Outcomes Research Methodology 2:279–290, 2001

Comparing Standard Regression, Propensity Score Matching, and Instrumental Variables Methods for Determining the Influence of Mammography on Stage of Diagnosis

MICHAEL A. POSNER*
Data Coordinating Center, Boston University School of Public Health, Boston, Massachusetts
E-mail: mposner@bu.edu

ARLENE S. ASH, KAREN M. FREUND, AND MARK A. MOSKOWITZ†
Health Care Research Unit, Section of General Internal Medicine, Evans Department of Medicine, Boston Medical Center, Boston, Massachusetts
E-mail: aash@bu.edu; karen.freund@bmc.org

MICHAEL SHWARTZ
Operations Management, Boston University School of Management, Boston, Massachusetts
E-mail: mshwartz@bu.edu

Received December 20, 2000; revised March 27, 2002; accepted March 29, 2002

Abstract. In situations where randomized trials are not feasible, analysis of observational data must be used instead. However, when using observational data, there is often selection bias for which we must account in order to adjust for pre-treatment differences between groups in their baseline characteristics. As an example of this, we used the Linked Medicare-Tumor Registry Database created by the National Cancer Institute and the Centers for Medicare and Medicaid Services to look at screening with mammography in older women to determine its effectiveness in detecting cancer at an earlier stage. The standard regression method and two methods of adjusting for selection bias are compared. We start with the standard analysis, a logistic regression predicting stage at diagnosis that includes as independent variables a set of covariates to adjust for differences in baseline risk plus an indicator variable for whether the woman used screening. Next, we employ propensity score matching, which evens out the distribution of measured baseline characteristics across groups, and is more robust to model misspecification than the standard analysis. Lastly, we conduct an instrumental variable analysis, which addresses unmeasured differences between the users and non-users. This article compares these methods and discusses issues of which researchers and analysts should be aware. It is important to look beyond the standard analysis and to consider propensity score matching when there is concern about group differences in measured covariates and instrumental variable analysis when there is concern about differences in unmeasured covariates.

Keywords: selection bias, observational studies, health services research

*Corresponding author: Michael A. Posner, c/o Boston University School of Public Health, 715 Albany Street 580, Boston, MA 02118. (617) 638-8151.
E-mail: mposner@bu.edu

†Deceased.

Grant support for this research was provided by the United States Army Medical Research and Materiel Command Breast Cancer Program, Grant #DMAD-17-97-1-7121.

1. Background

Randomized trials are viewed as the "gold standard" in research. However, in many cases randomized trials cannot be conducted, due to financial restrictions or ethical constraints. Where randomized trials are not feasible, observational data provide an invaluable source of empirical information to address important clinical questions. Administrative databases are an important source of data for such observational studies, given their potential inclusion of large representative populations and the breadth of clinical and non-clinical data available. However, treatment and control groups potentially differ in their baseline characteristics, some of which might not be observed, which makes findings susceptible to bias. Failure to account for these differences in baseline characteristics can lead to biased results.

The standard method for analyzing a dichotomous outcome (like stage at diagnosis used in our analysis) is a logistic regression. An indicator for the exposure of interest is included as a predictor, with other covariates included to control for baseline differences. The odds ratio of the exposure indicator is then estimated, and tested to determine if it is statistically different from one, the value that indicates no effect. This method will work if essentially all relevant baseline risk factors have been measured, and if the relationship between risk factors and outcomes are correctly specified in the model. However, it is almost never possible to measure essentially all the important risk factors and it is often difficult to know whether relationships between measured risk factors and the outcomes have been correctly specified. This article discusses two other types of analyses drawn from different disciplines as tools for addressing these challenges.

One way to address potential model misspecification is through the two-stage process of propensity score matching (Joffe and Rosenbaum, 1999; Rosenbaum and Rubin, 1983). In the first stage, the propensity score, which is the likelihood (or "propensity") of a case being in the exposed group, is estimated for each case through a logistic regression model. The exposed and unexposed groups are each then sampled to identify sub-samples with similar distributions of this estimated score. As shown by Rosenbaum and Rubin, such matching results in sub-samples of the study and control group with similar distributions of observed risk factors. Cochran (1968) states that creating five strata removes 90% of the bias due to the stratifying variable or covariate, and Rosenbaum and Rubin claim that stratification on propensity score removes even more than this in each covariate used in the propensity model. While researchers typically only check to see that the matched sub-samples have similar means for all important risk factors, more sophisticated checks on the comparability of their multivariate distributions can also be done. In the second stage, standard analytic techniques are used to fit a "response model" to the matched sub-samples and, ultimately, to estimate the effect of the exposure on the outcome. The propensity score method typically has less precision, due to reduced sample sizes, but this limitation is generally of less concern than the worry over possible bias from a misspecified model. It should also be noted that propensity score matching does not address or resolve problems due to imbalances in unmeasured factors.

It is natural to consider whether the model used to estimate the propensity scores might itself be misspecified, perhaps introducing a new set of problems for the analysis. Drake (1993) used simulations to compare the consequences of misspecifications of the propensity score to those of misspecified response models. She concluded that the propensity score "seems preferable when considering model misspecifications in the response model, particularly so because an incorrect propensity score model has smaller bias" and "generally, the simulations seem to indicate that the value of the propensity score lies primarily in guarding against model misspecifications."

It is easy to see why the methodology is robust to misspecification of the response model. Suppose, for the moment, that the matching achieves exactly the same (multivariate) distribution of all risk factors in the exposed and unexposed sub-samples. In this case: 1) the response model is not needed to correct for bias because the raw difference in treatment means between the sub-samples is already an unbiased estimate of the exposure effect; 2) a response model may be used, but its sole function is to increase the precision of the estimate; and, 3) even a poor response model cannot introduce bias in the estimated exposure effect, because any error in predicting the outcome for individuals with a given covariate pattern has exactly the same effect on both the exposed and unexposed sub-samples.

Our third method is the instrumental variable approach. Such analyses are common in econometrics (Greene, 1997), dating back to the 1920s (Wright, 1928), and are becoming increasingly popular in the health services outcomes research field (Newhouse and McClellan, 1998; Robins, 1989; Sheiner and Rubin, 1995; *Health Services Research*, December 2000). This approach is also conducted in two stages. The effect of the instrument is used in the first stage of the analysis to predict exposure status. The second stage of the analysis then examines differences in outcome as a function of differences in predicted exposure, to assess the causal effect of the exposure on the outcome. For a variable, or group of variables, to be considered a valid instrument, it should be neither associated with the outcome beyond its effect on exposure, nor associated with unmeasured confounders, after adjusting for other covariates included in the model (Greenland, 2000). The instrumental variable analysis is appealing because it addresses unmeasured confounders. However, the lack of association assumed above must be conceptually credible since they cannot be verified empirically.

The use of mammography for screening women over age 70 years is one area where data are required to answer a critical clinical question, yet no randomized trial information is available or is likely to become available. While the data from most randomized controlled trials on screening mammography on women ages 50 to 70 years demonstrate a benefit of this procedure (Andersson, Aspgren and Janzon et al., 1998; Roberts, Alexander and Anderson et al., 1990; Shapiro, Venet and Strax et al., 1982; Tabar, Gad and Holmberg et al., 1985), there are no data to guide clinicians for women over age 70 years. Most of the trials included no women over age 70 years, and none of the trials reported any age-specific data within the 50–70 year old age groups to assess age-related trends. Thus, the value of continuing screening in this group is yet to be established. Breast cancer incidence continues to rise beyond age 65, and accounts for 48% of all new breast cancers (Ries, Kosary and Hankey et al., 1997). In the absence of data, clinicians are nevertheless making decisions regarding screening, and in some geographic areas, as many as one quarter of older women receive regular mammography (Burns, McCarthy and Freund et al., 1996). A methodology to understand the benefits of screening mammography in women over age 70 years is therefore critical.

2. Data Description

The database we utilized for this cohort study is the Linked Medicare-Tumor Registry Database. The linked database was jointly created by the National Cancer Institute (NCI) and the Centers for Medicare and Medicaid Services (CMS), formerly the Health Care Financing Administration (HCFA) (Potosky, Riley and Lubitz et al., 1993). The database links Medicare data on women ages 65 and older from 1985 to 1994 with cancer registry information from the NCI's SEER program for cancers diagnosed between 1973 and 1993. The two databases overlap in three racially and socially diverse geographic areas: metropolitan Atlanta, Seattle-Puget Sound, and the state of Connecticut.

The data for these analyses were previously developed and have been described elsewhere (McCarthy, Burns and Freund et al., 2000; McCarthy, Burns and Couglin et al., 1998). Medicare Physicians' Claims files, which provide a record for every physician claim covered under Medicare Part B, provided measures of mammography utilization and primary care utilization. Medicare MedPAR (Medical Provider Analysis and Review) inpatient hospitalization records were utilized to develop measures of comorbidity. Medicare beneficiary enrollment files were used to determine race. The SEER data were used to determine age at diagnosis, stage at diagnosis using the Tumor, Nodal Status, Metastases (TNM) classification system (Beahers and Myers, 1983), and geographic location. 1990 U.S. Census data were used to obtain median household income by zip code, which was used as an ecological measure of socioeconomic status.

Our study sample consisted of all women with a first diagnosis of breast cancer in the three geographic areas whose utilization we could track for 2 years prior to the diagnosis of breast cancer (McCarthy, Burns and Freund et al., 2000). Since the Medicare utilization files provided mammography data on women aged 65 and older beginning in 1985, our sample included women aged 67 and older with their diagnosis in 1987 through 1993. We excluded women without two full years of Medicare Part B claims data prior to their diagnosis, in particular women with health maintenance organization coverage since their part B claims are not available.

We utilized the following procedure to classify mammography use of women. Women were classified as regular mammography users if they had claims for two separate bilateral mammograms (CPT code 76091 or 76092) within the two years prior to their breast cancer diagnosis, which were at least 10 months apart. Non-users were those women with no mammography claims during the two years prior to their diagnosis. The remaining women with less frequent mammography or only mammography just prior to diagnosis comprise a heterogeneous group of women receiving both screening and diagnostic studies and were excluded from further analyses because they cannot be unambiguously be classified as either users or non-users of screening. While the impact of this "uncertain" group is medically important, their exclusion does not weaken the interpretation of different modeling approaches.

Stage at diagnosis, our primary outcome variable, was classified as early (in situ and Stage I) or late (Stage II, III, and IV). Women with unstaged cancer (7.4% of the data) were excluded from further analysis. Our previous work indicates these women have both early and late stage disease, as demonstrated by a survival rate intermediate between the early and late stage groups (McCarthy, Burns and Freund et al., 2000).

The Linked database contains a number of clinical and non-clinical variables that are likely to be predictive of the outcome, and therefore were utilized as covariates in each model. Age at diagnosis was considered as both a continuous and categorical variable (67–69, 70–74, 75–69, 80–84 and 85+ years). The final analysis used the categorical version of age. Comorbidity was measured using a modified Charlson Comorbidity Index from inpatient claims' diagnosis codes (Deyo, Cherkin and Ciol, 1992). Women were categorized as having no hospital inpatient claims (i.e. comorbidity could not be assessed), or being hospitalized with either none or one or more comorbid conditions. An indicator for whether or not their race was listed as "black" was included. Median household income of zip code was considered both as a continuous variable and by quintile for each of the three geographic regions. After looking at quintiles by region, the variable used in analysis was a dichotomized split of the highest 40% versus the lower 60% of income within each region. Number of claims for office visits to primary care providers during the two years prior to breast cancer diagnosis was considered both continuously and categorically (never (0), seldom (1–3), every two to six months (4–12), and more than once every other month (13 or more)). The final analysis used the categorical version.

3. Methods

For the standard analysis, we developed a logistic regression model similar to previous work (McCarthy, Burns and Freund et al., 2000), to predict stage at diagnosis from user status, controlling for region, age, race, comorbidity, median income within zip code, and primary care visits. The c-statistic was used to measure how well the model discriminates between those diagnosed with early and late stage cancer (predictive validity).

For the propensity score approach, we developed a logistic regression model using age, race, comorbidity, median income within zip code, region, and primary care visits to predict the propensity to be a user. We split the data into deciles based on the propensity scores. To develop the matched samples, within each decile, a random sample of the larger group (users or non-users) was taken to get the same number in the smaller group (see Table 1). The matched sub-samples were then combined to create the analytic dataset. To examine the extent to which the above matching procedures resulted in samples of users and non-users more comparable in terms of baseline characteristics, statistical tests were used; p-values for chi-squared tests of independence were calculated for categorical risk factors and p-values for independent sample t-tests for equivalence of population means were calculated for continuous risk factors.

For the instrumental variable analysis, the first step is to determine which variable or variables are to be used as instruments. Angrist, Imbens, and Rubin (1996) describe five necessary conditions for the instrumental variable analysis. Of importance here is the association between the instrument and the exposure as well as a lack of correlation between the instrument and the unmeasured covariates that are associated with the outcome. For our application, a candidate instrument must be a predictor of user status with no residual predictive power on stage at diagnosis, after controlling for the other

Table 1. Propensity score matching results

	Pre-Matching			Post-Matching		
	Non-User	User	p-value	Non-User	User	p-value
Total Sample	2140	2516		1274	1274	
Decile 1	416	**57**		57	**57**	
Decile 2	339	**89**		89	**89**	
Decile 3	359	**136**		136	**136**	
Decile 4	239	**205**		205	**205**	
Decile 5	**193**	289		**193**	193	
Decile 6	**159**	277		**159**	159	
Decile 7	**145**	347		**145**	145	
Decile 8	**96**	327		**96**	96	
Decile 9	**113**	394		**113**	113	
Decile 10	**81**	395		**81**	81	
Age at Diagnosis	77.2	74.5	0.001	75.49	75.31	0.389
Age at Diagnosis						
67–69	14.3%	20.1%		16.2%	16.9%	
70–74	26.9%	35.6%		30.7%	29.7%	
75–79	24.3%	26.2%		27.8%	27.6%	
80–84	17.1%	13.6%		16.9%	17.1%	
85+	17.4%	4.6%	0.001	8.5%	8.7%	0.975
Charlson Comorbidities						
Not Hospitalized	27.0%	29.7%		27.4%	28.7%	
Hosp, No Comorbidies	48.0%	54.2%		51.1%	50.8%	
At Least One Comorbidity	24.9%	16.1%	0.001	21.5%	20.6%	0.726
Race						
Black	6.3%	3.1%		6.3%	5.4%	
Non-Black	93.8%	96.9%	0.001	93.7%	94.6%	0.350
Income (Median of Zip Code)	$42,030	$41,137	0.061	$41,073	$40,650	0.488
Regional Income						
Top 40%	40.2%	43.3%		41.8%	40.3%	
Lower 60%	59.8%	56.7%	0.036	58.2%	59.7%	0.444
Primary Care Visits	4.9	10.5	0.001	7.15	8.08	0.004
Primary Care Visits						
None	37.4%	6.5%		13.4%	12.8%	
1–3	22.3%	15.1%		24.3%	24.4%	
4–12	27.1%	47.3%		41.1%	39.2%	
13+	13.2%	31.1%	0.001	21.3%	23.6%	0.506
Location						
Seattle	27.0%	41.2%		32.0%	32.1%	
Atlanta	20.9%	14.6%		20.0%	20.3%	
Connecticut	52.1%	44.2%	0.001	48.0%	47.7%	0.984
Stage						
Early (TNM 0 or I)	58.9%	81.1%		58.3%	81.6%	
Late (TNM II, III, or IV)	41.1%	19.0%	0.001	41.7%	18.5%	0.001

covariates in the model. Considering these conditions in the context of our example, we chose region as our instrument, defined as a trichotomous variable for the three regions covered in our data file (Atlanta, Seattle, and Connecticut). First, we must show that there is an association between region and use of mammography. Studies have found variation in regional practice patterns (Burns,. McCarthy and Freund et al., 1996). Second, we claim a lack of correlation between the region and unobserved covariates associated with the outcome (after adjusting for observed covariates in the model). This assertion cannot be tested statistically, although this assumption appears to be reasonable in the context of our example: it seems reasonable that the outcome for someone using mammography in one region should not differ from the outcome for someone of similar characteristics who uses mammography in another. For example, we would expect that a woman with certain characteristics (primary care visits, age, race, etc.) receiving regular screening in Seattle would have the same likelihood of early stage disease diagnosed from mammography had she lived in Atlanta or Connecticut. If this assumption were not met, it would imply that, after conditioning on observed covariates, follow-up after a positive mammogram in one region is different than follow-up in another region.

The instrumental variable approach involves a two-stage model. In the first stage, covariates plus the instruments are used to predict user status. The predicted probability of being a regular user is then used in lieu of user status as an independent variable in the second stage to predict stage at diagnosis, along with other measured covariates. Variables used as instruments in the first stage model are excluded, since these variables are assumed to effect the outcome only through their association with user status. The coefficient associated with predicted user status is used to measure the impact of use. This method is known to produce an increased estimate of the standard error for the estimated treatment effect when compared to the standard analysis that does not replace the exposure variable with predicted exposure, especially when the instrument is not highly predictive of exposure, leading to reduction in precision. Murphy-Topel variance-covariance corrections were used to account for this correlation between the first and second stage models (Murphy and Topel, 1985).

To examine the strength of the instrumental variables, we used results from the first stage model. The odds ratio as well as the log-likelihood test was used. Staiger and Stock (1997) suggest that a partial F-statistic of about 10 indicates a sufficiently strong instrument. Instead we use the difference in the -2 log likelihoods between the full and reduced (excluding the instrument) models to predict user status. They state that this F-statistic can be approximated by the chi-squared divided by its degrees of freedom. Thus, a chi-squared with two degrees of freedom, like we have by using our trichotornous region variable, would need to be at least 20 to satisfy their condition. In addition, we look at the the odds ratios to compare the propensity of screening for each region to further investigate this association.

4. Results

The study sample size was 4667 women. Of these, 11 were excluded from analyses due to missing zip codes, which made the zip code-income matching impossible. Of the

Table 2. Odds ratios (and 95% Confidence Intervals) from the three models to predict early stage at diagnosis

Variable	Standard Model	Propensity Score Matching	Instrumental Variable Analysis
Age, 67–69	1.00 (ref)	1.00 (ref)	1.00 (ref)
Age, 70–74	0.92 (0.75, 1.13)	1.07 (0.81, 1.40)	0.92 (0.75, 1.13)
Age, 75–79	0.93 (0.75, 1.15)	0.97 (0.74, 1.28)	0.93 (0.75, 1.16)
Age, 80–84	0.83 (0.66, 1.05)	0.92 (0.68, 1.25)	0.82 (0.63, 1.07)
Age, 85+	1.02 (0.79, 1.32)	1.31 (0.89, 1.93)	1.02 (0.70, 1.49)
Black	0.67 (0.50, 0.92)	0.70 (0.48, 1.02)	0.70 (0.51, 0.98)
Not Hospitalized	1.00 (ref)	1.00 (ref)	1.00 (ref)
No Comorbid	0.54 (0.45, 0.63)	0.56 (0.45, 0.69)	0.55 (0.47, 0.65)
Some Comorbid	0.48 (0.40, 0.59)	0.54 (0.41, 0.70)	0.50 (0.40, 0.62)
High Reg Inc	1.23 (1.08, 1.41)	0.36 (1.13, 1.64)	1.22 (1.07, 1.41)
Connecticut	1.00 (ref)	1.00 (ref)	–
Seattle	1.23 (1.06, 1.43)	1.29 (1.05, 1.59)	–
Atlanta	1.17 (0.97, 1.41)	1.07 (0.84, 1.37)	–
No PC Visits	1.00 (ref)	1.00 (ref)	1.00 (ref)
Seldom (1–3 vis)	0.77 (0.62, 0.94)	0.82 (0.61, 1.12)	0.76 (0.55, 1.04)
Often (4–12 vis)	0.97 (0.80, 1.17)	0.98 (0.74, 1.31)	0.95 (0.58, 1.58)
Very Often (13+)	0.79 (0.64, 0.98)	0.93 (0.67, 1.27)	0.80 (0.45, 1.45)
User of Mammo.	**2.97 (2.56, 3.45)**	**3.24 (2.69, 3.88)**	**3.01 (1.09, 8.34)**

remaining 4656, 1354 were diagnosed with late stage cancer and 3302 with early stage cancer; 2516 met the criteria for users and 2140 for non-users. Note that 6354 women were excluded from the analysis because their user status was uncertain.

The standard model controlling for age at diagnosis, race, comorbidities, regional income, region, and number of primary care had a c-statistic of 0.68 (results in Table 2, second column). The conclusion, based on this model, is that regular users have 2.97 times the odds of being diagnosed at an early stage relative to non-users (95% CI: 2.56, 3.45). These results are quite similar to those reported from a similar model applied to the same data (McCarthy, Burns and Freund et al., 2000).

In the propensity score analysis, the model to predict user status using age, race, comorbidities, regional income, region, and primary care visits as independent variables had a c-statistic of 0.68. Table 1 shows the pre- and post-matching number of cases by decile and compares baseline characteristics of users and non-users. The most extreme propensity scores were examined and found to be close to the others, suggesting that none of these women were so different that they should be excluded.

The matching resulted in much more balanced groups in terms of the measured covariates. Each of these variables showed a statistically significant difference by user status before matching (or almost significant in the case of median income of zip code) and no difference afterward (see Table 1). The exceptions to this are the continuous measure of primary care visits, which was not used in that form in the final analyses, and stage at diagnosis, our outcome variable, which was not included in the propensity score model (as outcomes should not be).

When the logistic regression was run on the matched sample, the odds ratio associated with user status was 3.24 (95% CI: 2.69, 3.88), a result similar to the pre-matching result from the standard logistic regression analysis. The c-statistic was 0.69. Table 2 shows the results from this analysis.

The result from the instrumental variable approach using region as the instrument produced an odds ratio of 3.01 (95% CI: 1.09, 8.34). The point estimate is similar to the standard analysis, while the precision decreased drastically, as expected, due to the predicted use probability varying from 0 to 1, while the observed dichotomous user status only takes on the two extreme values. Table 2 shows the results from this analysis.

The chi-squared statistic for using region as a predictor of user status was 53.0 ($p < .001$, $df = 2$, and greater than our cutoff of 20), indicating that region is a strong candidate as an instrumental variable. The odds ratio for predicting user status from region is 1.34 for Seattle and 0.64 for Atlanta, both using Connecticut as the reference group. The chi-squared value presents a summary measure of this variable with which to make assessments of the strength of the instrument, while the odds ratios presented allow us to better see the individual relationships between the regions.

5. Discussion

The standard analysis gives us results to which we can compare the others to examine the effect of uneven distribution of measured covariates or model misspecification (when results are compared to the propensity score results) and of unmeasured confounders (when the results are compared to the instrumental variable analysis). It is important to note that the odds ratio of 2.97 from the standard analysis estimates the average effect (of being a user) on the entire population.

The propensity score matching method produced an odds ratio of 3.24. While this effect is larger than that from the standard analysis, the confidence intervals overlap over most of their confidence regions. The propensity score matching approach estimates the impact of being a user of mammography for the population whose measured covariates conform to the matched sample included in the propensity score analysis (for example, the age distribution in the reduced sample differs from the age distribution in the general population). This result being so close to that of the standard model provides some reassurance that the standard model has adjusted correctly for any differences in measured covariates between the user and non-user groups.

Our instrument was region. As stated in the methods section, this variable meets the condition of association with the exposure and is assumed to meet the condition of lack of correlation with unmeasured covariates. In addition, the significance of region on predicting user status was quite strong (Chi-squared with 2 df of 53, $p < .001$), indicating that our instrument is sufficiently strong. For example, for identical women with average characteristics, the effect of being in Seattle rather than Connecticut increases the odds of mammogram use by 34%. This is consistent with known geographic variation in practice patterns, and in particular for breast cancer care (Nattinger, Gottlieb and Veum et al., 1992; Nattinger, Gottlieb and Hoffman et al., 1996). If our instrument were weak, we could not conclude that similarity of instrumental variable estimates and standard analysis estimates implied that there were not unmeasured confounders (Staiger and Stock, 1997).

The instrumental variable analysis, using region as the instrument, produced an odds ratio of 3.01. Again, this result is similar to that of the standard analysis, implying that there is no large bias coming from unmeasured variables that are unassociated with region. This value measures the impact among those whose behavior (mammography use) changes depending on their region, with all else being equal. The confidence interval for this value (1.09, 8.34) is much wider than for the odds ratio for the standard model (2.56, 3.45). This is expected, because the instrumental variable approach depends on much less variation in user status.

In summary, all three analyses—the standard regression, the propensity score matching, and the instrumental variable analysis using region as the instrument—produced very similar results. The similarity of these results helps strengthen the credibility of the standard regression analysis. There is little model mis-specification, either from measured variables, as seen via the propensity score matching, nor from unmeasured variables (that meet the instrumental variable criteria), as seen via the instrumental variable analysis.

We recommend that investigators analyzing administrative databases or other observational studies consider the sources of bias that may affect their results. The standard analysis should not be blindly accepted before considering what types of bias may contribute to the results. Propensity score matching is a good method for addressing differences observed in baseline characteristics or other differences between observed variables. Instrumental variable analysis should be considered when selection bias due to unmeasured covariates is of concern, and plausible instrumental variables are available. These methods can help improve the validity of the analysis results and mitigate potential sources of bias.

Acknowledgments

This study used the linked Medicare-Tumor Registry database. The interpretation and reporting of these data are the sole responsibility of the authors. The authors acknowledge the efforts of several groups responsible for the creation and dissemination of the linked database, including the Applied Research Branch, Division of Cancer Prevention and Population Science, National Cancer Institute; the Office of Information Services and the Office of Strategic Planning, Centers for Medicare and Medicaid Services; Information Management Services, Inc.; and the Surveillance, Epidemiology, and End Results (SEER) Program tumor registries. We acknowledge the prior contributions of Ellen McCarthy in dataset development and analysis on which this work is based. In addition, the authors would like to thank Mingshan Lu for coding and Michael Ash for his economics expertise.

References

I. Andersson, K. Aspgren and L. Janzon, et al., "Mammographic screening and mortality from breast cancer: the malmo mammographic screening trial," *BMJ*, 297, pp. 943–948, 1998.

J. D. Angrist, G. W. Imbens and D. B. Rubin, "Identification of causal effects using instrumental variables," *JASA*, 96, pp. 444–472, 1996.

O. H. Beahers and M. H. Myers, *Manual for Staging of Cancer*, 2^{nd} ed. JB Lipincott, Philadelphia, 1983.

J. Bound, D. A. Jaeger and R. M. Baker, "Problems with instrumental variables estimation when the correction between the instruments and the endogenous explanatory variable is weak," *JASA*, 90, pp. 443–450, 1995.

R. B. Burns, McCarthy and K. M. Freund, et al., "Variability in mammography use among older women," *JAGS*, 44, pp. 922–926, 1996.

W. G. Cochran, "The effectiveness of adjustment by subclassification in removing bias in observational studies," *Biometrics*, 24, pp. 205–213, 1968.R. A. Deyo, D. C. Cherkin and M. A. Ciol, "Adapting a clinical comoridity index for use with IDC-9-CM Administrative database," *J Clin Epidemiol*, 45, pp. 613–619, 1992.

R. A. Deyo, D. C. Cherkin and M. A. Ciol, "Adapting a clinical comoridity index for use with IDC-9-CM Administrative database," *J Clin Epidemiol*, 45, pp. 613–619, 1992.

C. Drake, "Effects of misspecification of the propensity score on estimators of treatment effect," *Biometrics*, 49, pp. 1231–1236, 1993.

W. H. Greene, *Econometric Analysis*, Prentice Hall, New Jersey, 288–295, 1997.

S. Greenland, "An introduction to instrumental variables for epidemiologists," *Int J Epidemiol*, 29, pp. 722–729, 2000.

M. M. Joffe and P. R. Rosenbaum, "Invited commentary: propensity scores," *Am J Epidemiol*, 150, pp. 327–333, 1999.

E. P. McCarthy, R. B. Burns, K. M. Freund, A. S. Ash, M. Shwartz, S. L. Marwill and M. A. Moskowitz, "Mammography use, breast cancer stage at diagnosis, and survival among older women," *J Am Geriatr Soc*, 48, pp. 1226–1233, 2000.

E. P. McCarthy, R. B. Burns, S. S. Couglin, K. M. Freund, J. Rick, S. L. Marwill, A. Ash, M. Shwartz and M. A. Moskowitz, " Mammography use helps to explain differences in breast cancer stage at diagnosis between older black and white women," *Ann Intern Med*, 128, pp. 729–736, 1998.

K. M. Murphy, R. H. Topel, "Estimation and inference in two-step econometic models," *Journal of Business and Economic Statistics*, 3(4), pp. 370–379, 1985.

A. B. Nattinger, M. S. Gottlieb, J. Veum, D. L. Yagnke and J. S. Goodwin, "Goegraphic variation in the use of brest conserving treatment for breast cancer," *N Engl J Med*, 326, pp. 1102–1107, 1992.

A. B. Nattinger, M. S. Gottlieb, R. G. Hoffman, A. P. Walker and J. S. Goodwin, "Minimal increase in use of breast-conserving surgery from 1986 to 1990," *Med Care*, 24, pp. 479–489, 1996.

J. P. Newhouse and M. McClellan, "Econometrics in outcomes research: the use of instrumental variables," *Ann Rev Public Health*, 19, pp. 17–34, 1998.

A. L. Potosky, G. F. Riley, J. D. Lubitz, R. M. Mentnech and L. G. Kessler, "Potential for cancer realted health services research using a linked medicare-tumor registry database," *Med Care*, 31, pp. 732–748, 1993.

L. A. G. Ries, C. L. Kosary and B. F. Hankey, et al., *SEER Cancer Statistics Review*, 1973–1994. National Cancer Institute, Bethesda, MD, 1997. NIH Publication, 97-2789.

M. Roberts, F. E. Alexander and T. J. Anderson, et al., "Edinburgh trial of screening for breast cancer mortality of seven years," *Lancet*, 335, pp. 241–246, 1990.

J. M. Robins, "The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies," in *Health Service Research Methodology: A Focus on AIDS* (L. Sechrest, H. Freeman, and A. Mulley, eds.), US Public Health Service, Washington, DC, pp. 113–159, 1989.

P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, pp. 41–55, 1983.

S. Shapiro, W. Venet, P. Strax, et al., "Ten to fourteen year effect of screening on breast cancer mortality," *J Natl Cancer Inst*, 69, pp. 349–355, 1982.

L. B. Sheiner, D. B. Rubin, "Intention-to-treat analysis and the goals of clinical trials," *Clin Pharm Therap*, 57, pp. 6–15, 1995.

D. Staiger and J. H. Stock, "Instrumental variables regression with weak instruments," *Econometrica*, 63(3), pp. 557–586, 1997.

L. Tabar, A. Gad, L. H. Holmberg, et al., “Reduction in mortality from breast bancer after mass screening with mammography. Randomized trial from the breast cancer working group of Swedish national board of health and welfare,” *Lancet*, 1, pp. 829–832, 1985.

S. Wright, “Appendix,” in *The Tariff on Animal and Vegetable Oils*, (P. G. Wright, ed.), Macmillan, New York, 1928.

Health Services Research, Vol 35, No. 5. Part II. December 2000.