Reducing Bias in Cancer Research: Application of Propensity Score Matching

Bryce B. Reeve, Ph.D., Ashley Wilder Smith, Ph.D., M.P.H., Neeraj K. Arora, Ph.D., and Ron D. Hays, Ph.D.

In cancer observational studies, differences between groups on confounding variables may have a significant effect on results when examining health outcomes. This study demonstrates the utility of propensity score matching to balance a non-cancer and cancer cohort of older adults on multiple relevant covariates. This approach matches cases to controls on a single indicator, the propensity score, rather than multiple variables. Results indicated that propensity score matching is an efficient and useful way to create a matched case-control study out of a large cohort study, and allows confidence in the strength of the observed outcomes of the study.

INTRODUCTION

Observational research is often used to understand the burden of cancer on patients and survivors, and allows scientists to examine relevant factors that may provide the framework for future intervention research. Studying the effects of cancer on health-related quality of life (HRQOL) requires an observational rather than experimental design. However, in observational studies, HRQOL comparisons to individuals without cancer are challenging because the cancer and non-cancer groups may differ in their distributions on key demographic and other factors related to HRQOL. In experimental studies, random assignment to condition ensures that on average there should be no systematic differences between groups on relevant factors (D'Agostino, 1998). Thus, the randomization process balances group differences on both measured and unobserved variables. However, in observational studies, there is no random assignment and group differences on confounding variables may potentially have significant effects on the results. It is important that the cancer and non-cancer groups are as similar as possible on the characteristics other than cancer in order to reduce the possibility of confounding when examining HRQOL differences.

Researchers have relied on statistical approaches to help reduce the bias due to group differences in observational studies. However, these approaches have limitations. For example, traditional methods to control for group differences such as matched sampling, stratification, and covariate adjustment are often limited by the number of variables that can be adjusted because of sample size concerns (D'Agostino, 1998). The focus of this article is to illustrate the strength of this bias reduction method, propensity score matching, within a cohort study of older adults with and without cancer. This approach is highlighted for its utility in selecting non-cancer controls to compare to cancer cases, matching on a number of relevant factors of interest.

Bryce B. Reeve, Ashley Wilder Smith, and Neeraj K. Arora are with the National Cancer Institute (NCI). Ron D. Hays is with the University of California, Los Angeles. Ron Hays was supported by the National Cancer Institute under the Intergovernmental Personnel Act and in part by a P01 Grant Number AG020679-01 from the National Institute on Aging and the UCLA Center for Health Improvement in Minority Elderly/Resource Centers for Minority Aging Research under Grant Number 2P30-AG-021684. The statements expressed in this article are those of the authors and do not necessarily reflect the views or policies of NCI; the University of California, Los Angeles; National Institute on Aging; or the Centers for Medicare & Medicaid Services (CMS).

Propensity score matching simply uses the traditional framework of matching two groups to make them comparable, but matches them on a single indicator, the propensity score, rather than multiple variables. When matching, controls from the non-cancer group are selected who have similar propensity scores to those in the cases (those with cancer). The goal is a dataset of cases and controls with similar characteristics on all key variables that were used to define the propensity scores. The propensity score is defined as the probability of being in the case group given the individual's level on the covariates included in the model (Rosenbaum and Rubin, 1983). The propensity score is often estimated using a logistic regression model because logistic regression makes no assumptions about the distributions of the covariates on the dichotomous outcome (D'Agostino, 1998). A single propensity score is estimated for every individual in the study, both cases and controls. This propensity score is then used to adjust for the differences between the two groups on the observed covariates in the study. Thus, the propensity score is often thought of as a balancing score allowing researchers to control for a large number of background covariates simultaneously based on a single number (Rosenbaum and Rubin, 1983).

The three most common propensity score methods to balance the groups based on the measured covariates are: (1) matching, (2) stratification, and (3) regression adjustment. Matching algorithms use the propensity score to match one or more controls to each case. Stratification (or subclassification) methods group individuals by strata based on the propensity scores. Often the boundaries of the strata are based on dividing the distribution of the entire sample into quintiles of the propensity score. Analyses are then carried out by strata. Regression adjustment techniques include the propensity score in the model as a covariate or use the propensity score as a weight. D'Agostino (1998; 2007) and Kurth, Walker, Glynn et al. (2006) provide detailed reviews of these three techniques along with examples.

Matching is an attractive choice to control for bias relative to simple adjustments of the covariates made in the regression approach because regression models generally assume there is a linear relationship between the covariates and the outcome of interest. Matching does not make this assumption. Further, unlike the regression approach, the process of matching removes individuals in the control group who are a poor match to the cases (Foster, 2003). Propensity score matching has also been found to reduce bias due to case-control differences better than stratification methods based on the propensity score (Austin, 2008).

The purpose of this article is to illustrate the benefits of the propensity score matching approach. We used propensity score methods in a longitudinal study that examines the burden of cancer (the cases) on patients' HRQOL relative to individuals without cancer (the controls). This is an observational study as individuals obviously could not be randomly assigned to condition (cancer versus no cancer). The cases and controls differed on a number of key demographic characteristics and comorbid conditions resulting in the reduced ability, because of potential bias, to make group comparisons to look at the unique effect of cancer diagnosis. Propensity score matching was used to control for potential confounding. For this study, the propensity score is defined as the probability of an individual having cancer, conditional on the set of measured covariates including demographics, survey characteristics, and comorbid conditions. This study uses the analytic plan described by D'Agostino

(1998) to illustrate the utility of this methodology to develop a case-control study out of a large cohort study of older adults living in the U.S.

METHODS

Sample

The current study examined the burden of a cancer diagnosis on the HRQOL in individuals age 65 or over who were participating in Medicare managed care plans. The data for this study come from collaboration between CMS and NCI. Under this collaboration, data from the CMS' Medicare Health Outcomes Survey (MHOS) (Jones, Jones, and Miller, 2004) were linked to data from the NCI Surveillance, Epidemiology, and End Results (SEER) cancer registries (Ries et al., 2007). The MHOS is a yearly survey that is administered to a random sample of 1,000 Medicare beneficiaries from each managed care plan under contract with CMS. In plans with 1,000 or fewer Medicare enrollees, all eligible members were surveyed. Each participant is asked to complete a survey at baseline and 2 years later. A detailed description of the SEER-MHOS data linkage is provided elsewhere (Ambs et al., 2008).

The linked SEER-MHOS dataset includes four MHOS cohorts (baseline and followup year): 1998 and 2000; 1999 and 2001; 2000 and 2002; and 2001 and 2003. Pooling across these four cohorts, 1,432 cancer patients and 30,964 patients without cancer were identified who have data on both the baseline and followup MHOS. The cancer patients selected for this study were those whose first cancer diagnosis occurred between the baseline and followup MHOS. The sample included 436 prostate, 320 breast, 240 colorectal, 112 non-small cell lung, 89 bladder, 80 melanoma, 56 endometrial, 53 non-Hodgkins lymphoma, and 46 kidney cancer patients. Selection of non-cancer respondents was limited to those who resided in the same SEER region and participated in the same managed care plans as the cancer patients.

Data

The primary goals for the collection of MHOS data was for CMS to evaluate the performance of the managed care plans under contract with CMS, to promote quality improvement, and to empower beneficiaries with the knowledge of program performance to make plan selections (Jones, Jones, and Miller, 2004). The data set also serves as a resource for outcomes research (Haffer and Bowen, 2004). The MHOS provides data on patient demographics, survey characteristics, chronic medical conditions, clinical symptoms, physical and mental health (including the SF-36® version 1) (Ware and Sherbourne, 1992), and smoking status.

Demographic variables in the MHOS included education, age, sex, race/ethnicity, current marital status as well as change in marital status between baseline and followup assessment. Survey characteristics included survey administration (self-administered or interviewer administered) and whether the survey was completed by the Medicare recipient person or proxy. Chronic medical conditions included hypertension or high blood pressure, coronary artery disease, congestive heart failure, myocardial infarction or heart attack, other heart conditions, stroke, chronic obstructive pulmonary disease, inflammatory bowel disease, arthritis of the hip or knee, arthritis of the hand or wrist, sciatica, and diabetes. We classified medical conditions that existed before the baseline MHOS assessment as pre-existing conditions and classified conditions that were diagnosed between baseline and

followup MHOS assessments as newly diagnosed conditions.

The SEER program includes population-based cancer registry sites throughout the U.S. The geographic areas included in the SEER program have changed over time. Currently the SEER program includes 18 population-based cancer registries that represent 26 percent of the U.S. population (Ries et al., 2007). The SEER data provide detailed clinical information including primary tumor site, tumor morphology and stage at diagnosis, time of diagnosis, first course of treatment, and followup for vital status.

Analyses

All analyses were performed using SAS® (version 9.13) software. Propensity scores were estimated for each person using logistic regression model regressing cancer status (0 = no cancer, 1 = cancer) on patient demographics, survey characteristics, and pre-existing and newly diagnosed chronic medical conditions other than cancer. The selection of these variables for the model was based on a desire to balance the cancer and non-cancer groups on these characteristics for the followup study that uses the propensity-matched dataset to examine differences in HRQOL between the cancer and non-cancer sample. Further, inclusion of multiple variables, even when a covariate is not different between groups, ensures the matched case-control samples will be similar (Rubin and Thomas, 1996). In SAS®, propensity scores for each person can be obtained by the predicted probabilities of the outcome specified in the output statement of the logistic procedure. Because of missing data on the covariates, 64 of the 30,964 non-cancer participants did not receive a propensity score.

The next step was to test whether balance was achieved between the cancer and non-cancer samples on the covariates included in the regression model by comparing differences between the groups before and after adjustments based on propensity scores. We separated the full sample into quintiles defined by their propensity scores. For categorical variables, we then compared frequencies/proportions and chi-square statistics before and after adjusting for propensity score quintile and non-cancer/cancer status. For continuous variables, we conducted a two-way analysis of variance (ANOVA), which included main effects for propensity score quintile and non-cancer/cancer status. Specifically, we compared the p-values for non-cancer/cancer status after adjustment for propensity score quintile with the p-values for non-cancer/cancer status before adjustment to determine whether balance on the covariates was achieved. Interactions between propensity score quintile and non-cancer/cancer status were also examined to determine whether there were significant differences between the two groups based on the level of the propensity score quintile.

After determining that balance on the covariates could be achieved, non-cancer patients were then matched to cancer cases using the propensity score. Five controls were matched to each cancer case to account for possible bias due to confounding from unmeasured variables in the study. Relative to a single match, multiple matches make more use of the available data and increase the efficiency of the estimation of group differences (Ury, 1975; Smith, 1997). Thus, more matches reduce possible bias. However, the amount of bias reduction per additional matched control lessens with each additional match, and forcing too many matches may select controls who do not have a close propensity score to the case (Smith, 1997). Matching was carried out through a SAS® macro available from the Mayo Web site that uses

the greedy matching algorithm (Bergstralh and Kosanke, 2004). In this algorithm, the control selected for each case is the one with the smallest difference between propensity scores. This is commonly referred to as nearest neighbor matching. When ties emerged, the first control participant encountered was selected. Cases and controls were randomly sorted before matching to remove any bias in the matching process. The greedy method generally produces very good matches, especially if the control pool is large relative to the number of cases. Further, this method has been shown to work relatively well in comparison with more complex matching algorithms (Rosenbaum and Rubin, 1985; Smith, 1997). The algorithm first matches one control to all cases and then proceeds to select the second set of controls and continues until five controls are identified per case. Once a control is matched, the control is not considered again as a match for any other cancer case (i.e., matching without replacement).

RESULTS

The propensity scores ranged from 0.0089 to 0.1656 across the sample. The distributions of the non-cancer and cancer samples across the quintiles are provided in Table 1. Means and standard deviations for the non-cancer and cancer samples on the covariates before stratification are provided in Table 2. For the dichotomous covariates, the means are essentially the proportions of respondents (e.g., mean for females of 0.60 indicates 60 percent of the non-cancer controls were female). Chi-square tests identified 11 covariates that were significantly different ($\alpha < 0.05$) between the two groups. The non-cancer sample relative to the cancer sample was less educated and included more females, more Asians and Hispanics, more individuals who were widowed at baseline, fewer former and current smokers, fewer proxy reporters at followup, fewer heart and chronic obstructive pulmonary disease (COPD) conditions at baseline, and fewer reports of heart conditions occurring between the baseline and followup. After stratification of non-cancer and cancer samples based on the propensity score quintiles, no significant differences were found (Table 2).

Across the tests for the 46 covariates, 3 interaction terms (propensity score quintile by non-cancer/cancer status) were significant ($\alpha < 0.01$). For example, education at baseline was found to have a significant

Table 1

Distribution of Non-Cancer and Cancer Sample, by Propensity Score Range

Quintile	Propensity Score Range	Non-Cancer[1]		Cancer[2]	
			Percent		Percent
I	0.0089 – 0.0290	6,298	20	168	12
II	0.0290 – 0.0355	6,269	20	198	14
III	0.0355 – 0.0450	6,211	20	255	18
IV	0.0450 – 0.0609	6,140	20	327	23
V	0.0609 – 0.1656	5,982	20	484	34

[1] N = 30,900.

[2] N = 1,432.

NOTE: The endpoints in the propensity scores range appear to overlap as shown here, but do not in actuality as the thresholds for deriving the quintiles was at a lower decimal place value.

SOURCE: The dataset links the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) cancer registry data with Medicare beneficiaries' responses to the Centers for Medicare & Medicaid Services' Medicare Health Outcomes Survey (MHOS). The linked SEER-MHOS dataset includes four MHOS cohorts (baseline and followup year): 1998 and 2000; 1999 and 2001; 2000 and 2002; and 2001 and 2003.

Table 2

Comparison of Covariates for Non-Cancer Controls and Cancer Cases Before and After Propensity Score Stratification

Covariate	No Cancer[1] Mean	(SD)	Cancer[2] Mean	(SD)	*p*-Value before Stratification[3]	*p*-Value after Stratification[3]
Education	3.178	(1.217)	3.265	(1.249)	<0.01*	0.16
Age	73.940	(6.109)	73.856	(5.852)	0.59	0.63
Female	0.600	(0.490)	0.446	(0.497)	<0.01*	0.91
Asian	0.077	(0.267)	0.058	(0.234)	<0.01*	0.47
Black	0.055	(0.228)	0.052	(0.223)	0.70	0.57
Hispanic	0.088	(0.283)	0.059	(0.235)	<0.01*	0.67
American Indian	0.005	(0.070)	0.006	(0.079)	0.49	0.96
Other (Non-White)	0.010	(0.097)	0.007	(0.083)	0.33	0.97
Never Married	0.029	(0.168)	0.023	(0.149)	0.16	0.53
Divorced	0.097	(0.296)	0.098	(0.297)	0.98	0.83
Widow	0.284	(0.451)	0.219	(0.413)	<0.01*	0.97
Marital Status Missing	0.017	(0.128)	0.019	(0.136)	0.55	0.96
Widowed (After Baseline)	0.036	(0.185)	0.036	(0.187)	0.90	0.77
Former Smoker	0.382	(0.486)	0.448	(0.497)	<0.01*	0.86
Current Smoker	0.101	(0.301)	0.123	(0.328)	<0.01*	0.19
Smoking Missing	0.055	(0.229)	0.065	(0.247)	0.13	0.44
Assessment Mixed	0.141	(0.348)	0.133	(0.340)	0.43	0.83
Proxy (Baseline)	0.084	(0.278)	0.076	(0.265)	0.28	0.77
Proxy (Baseline) Missing	0.074	(0.262)	0.068	(0.251)	0.35	0.73
Proxy (Followup)	0.105	(0.306)	0.126	(0.332)	<0.01*	0.79
Proxy (Followup) Missing	0.092	(0.289)	0.079	(0.270)	0.09	0.45
Pre-Existing Conditions						
Hypertension/HBP	0.531	(0.499)	0.524	(0.500)	0.61	0.54
Angina/Coronary Artery Disease	0.136	(0.342)	0.150	(0.357)	0.13	0.91
Congestive Heart Failure	0.053	(0.223)	0.059	(0.235)	0.34	0.72
Myocardial Infarction/Heart Attack	0.086	(0.281)	0.097	(0.296)	0.16	0.96
Other Heart Conditions	0.194	(0.396)	0.216	(0.412)	0.04*	0.33
Stroke	0.066	(0.248)	0.061	(0.240)	0.47	0.99
Emphysema/Asthma/COPD	0.113	(0.317)	0.139	(0.346)	<0.01*	0.98
Crohn's Disease/IBD	0.043	(0.204)	0.041	(0.197)	0.62	0.95
Arthritis Hip	0.353	(0.478)	0.342	(0.474)	0.40	0.71
Arthritis Hand	0.323	(0.468)	0.300	(0.458)	0.07	0.80
Sciatica	0.211	(0.408)	0.213	(0.410)	0.88	0.74
Diabetes	0.151	(0.358)	0.156	(0.363)	0.62	0.96
Newly Diagnosed Conditions						
Hypertension/HBP	0.080	(0.271)	0.082	(0.275)	0.76	0.29
Angina/Coronary Artery Disease	0.042	(0.200)	0.045	(0.207)	0.58	0.48
Congestive Heart Failure	0.031	(0.174)	0.029	(0.167)	0.55	0.88
Myocardial Infarction/Heart Attack	0.031	(0.174)	0.027	(0.163)	0.38	0.69
Other Heart Conditions	0.071	(0.257)	0.085	(0.279)	0.05*	0.91
Stroke	0.030	(0.172)	0.033	(0.178)	0.61	0.95
Emphysema/Asthma/COPD	0.036	(0.187)	0.045	(0.207)	0.10	0.95
Crohn's Disease/IBD	0.021	(0.143)	0.026	(0.159)	0.22	0.96
Arthritis Hip	0.103	(0.303)	0.104	(0.305)	0.85	0.82
Arthritis Hand	0.101	(0.301)	0.093	(0.290)	0.34	0.89
Sciatica	0.083	(0.276)	0.069	(0.254)	0.06	0.52
Diabetes	0.037	(0.188)	0.045	(0.207)	0.12	0.75

* Significant differences (α <0 .05).

[1] *N* = 30,964.

[2] *N* = 1,432.

[3] Differences between non-cancer and cancer cases on all categorical covariates were tested by a chi-square statistic. For the continuous variable, age, *F*-statistic was used.

NOTES: SD is standard deviation. HBP is high blood pressure. COPD is chronic obstructive pulmonary disease. IBD is inflammatory bowel disease.

SOURCE: The dataset links the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) cancer registry data with Medicare beneficiaries' responses to the Centers for Medicare & Medicaid Services' Medicare Health Outcomes Survey (MHOS). The linked SEER-MHOS dataset includes four MHOS cohorts (baseline and followup year): 1998 and 2000; 1999 and 2001; 2000 and 2002; and 2001 and 2003.

interaction (χ^2= 16.66, $p < 0.01$). We further examined this variable by reviewing the quintile means for the control and cancer groups as shown in Table 3. Education is categorized by levels (1 = 8th grade or less; 2 = some high school; 3 = high school graduate; 4 = some college or 2 year degree; 5 = college graduate or higher education level). As previously noted, those in the non-cancer sample had less education than those with cancer. When examining by quintile of propensity score, the cancer sample reported higher education levels on average than the non-cancer sample in the first four quintiles and lower education level on average in the fifth quintile. With the goal of propensity score matching to balance the cancer and non-cancer sample on education, the reported differences in average education levels across the five quintiles were deemed of minimal importance.

We also found significant interaction terms for former smoker (χ^2= 13.48, $p < 0.01$) and proxy response at followup MHOS assessment (χ^2= 13.80, $p < 0.01$). After reviewing each of these variables using a similar strategy to that for the education variable, we observed no meaningful differences in proportions of control and cancer cases that would indicate a need for concern about the balance of each group on the covariate.

As previously noted, five controls per cancer case were matched using the greedy algorithm, which selects controls with the minimal difference on propensity scores. Table 4 shows a comparison of covariates for non-cancer and cancer samples before and after matching. Significant changes in the make up of the non-cancer (control) sample include the percentage of females (from 60 to 45 percent), former smokers (from 38 to 44 percent), widows at baseline survey (from 28 to 22 percent), Hispanics (from 9 to 6 percent), Asians (from 8 to 6 percent), and those diagnosed with COPD at baseline (from 11 to 14 percent).

Table 3

Evaluating Non-Cancer Control and Cancer Group Differences on Education Status, by Strata Based on Propensity Score Quintiles

Distribution	Sample	*N*	Education at Baseline Mean (Standard Deviation)
Overall	Control	30,900	3.178 (1.217)
	Cancer	1,432	3.265 (1.249)
After Stratification into Quintiles Based on Propensity Scores			
Quintile 1	Control	6,298	2.743 (1.184)
	Cancer	168	2.976 (1.199)
Quintile 2	Control	6,269	3.141 (1.128)
	Cancer	198	3.187 (1.158)
Quintile 3	Control	6,211	3.226 (1.173)
	Cancer	255	3.294 (1.172)
Quintile 4	Control	6,140	3.320 (1.260)
	Cancer	327	3.379 (1.298)
Quintile 5	Control	5,982	3.476 (1.212)
	Cancer	484	3.304 (1.295)

SOURCE: The dataset links the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) cancer registry data with Medicare beneficiaries' responses to the Centers for Medicare & Medicaid Services' Medicare Health Outcomes Survey (MHOS). The linked SEER-MHOS dataset includes four MHOS cohorts (baseline and followup year): 1998 and 2000; 1999 and 2001; 2000 and 2002; and 2001 and 2003.

Table 4

Comparison of Covariates for Non-Cancer Controls and Cancer Patients Before and After Propensity Score Matching

Characteristic	Non-Cancer		Cancer
	Before Matching (*N* = 30,964)	After Matching (*N* = 7,160)	(*N* = 1,432)
		Percent	
Education			
8th Grade or Less	11.72	10.70	11.10
Some High School	15.04	14.08	15.08
High School Graduate or GED	33.25	32.51	29.89
Some College or 2-Year Degree	23.37	24.26	24.09
College Graduate or Higher	16.41	18.45	19.83
Age			
Mean	73.94 (6.11)	73.86 (6.05)	73.86 (5.85)
Sex			
Male	39.95	55.13	55.45
Female	60.05	44.87	44.55
Race			
White	76.41	81.69	81.63
Asian	7.69	5.88	5.80
Black	5.47	5.06	5.24
Hispanic	8.78	5.89	5.87
American Indian	0.50	0.66	0.63
Other (Non-White)	0.95	0.82	0.70
Marriage Status (Baseline)			
Married	58.00	64.01	64.87
Never Married	2.86	2.04	2.23
Divorced	9.56	9.46	9.57
Widow	27.89	22.30	21.44
Widowed from Baseline to Followup			
No	94.82	96.73	96.44
Yes	3.50	3.27	3.56
Smoking Status			
Never	46.13	35.84	36.45
Former	38.23	44.47	44.76
Current	10.10	12.77	12.29
Assessment Mode Mixed from Baseline to Followup			
No	85.92	85.89	86.66
Yes	14.08	14.11	13.34
Proxy (Baseline)			
No	84.13	85.20	85.61
Yes	8.43	7.90	7.61
Proxy (Followup)			
No	80.30	79.26	79.47
Yes	10.47	12.83	12.64
Pre-Existing Conditions			
Hypertension/HBP	52.80	52.25	51.82
Angina/Coronary Artery Disease	13.36	15.29	14.66
Congestive Heart Failure	5.19	6.33	5.73
Myocardial Infarction/Heart Attack	8.47	9.85	9.50
Other Heart Conditions	19.16	22.91	21.30
Stroke	6.51	6.31	6.01
Emphysema/Asthma/COPD	11.21	14.19	13.69
Crohn's Disease/IBD	4.27	4.15	3.98
Arthritis Hip	34.98	34.04	33.80
Arthritis Hand	31.98	30.85	29.54
Sciatica	20.89	21.75	20.95
Diabetes	15.04	15.20	15.43

Refer to footnotes at the end of the table.

Table 4—Continued

Comparison of Covariates for Non-Cancer Controls and Cancer Patients Before and After Propensity Score Matching

Characteristic	Non-Cancer		Cancer	
	Before Matching (*N* = 30,964)	After Matching (*N* = 7,160)		(*N* = 1,432)
		Percent		
Newly Diagnosed Conditions				
Hypertension/HBP	8.01	8.69		8.87
Angina/Coronary Artery Disease	4.17	5.01		4.82
Congestive Heart Failure	3.14	3.51		3.49
Myocardial Infarction/Heart Attack	3.14	3.07		3.00
Other Heart Conditions	7.11	9.13		9.08
Stroke	3.04	3.88		3.91
Emphysema/Asthma/COPD	3.64	5.00		4.75
Crohn's Disease/IBD	2.10	3.09		3.00
Arthritis Hip	10.25	11.47		11.10
Arthritis Hand	10.06	9.82		9.99
Sciatica	8.34	7.57		7.33
Diabetes	3.68	4.57		4.75
Cancer Type			*N*	
Colorectal	—	—	240	16.76
Lung (Non-Small)	—	—	112	7.82
Melanoma	—	—	80	5.59
Breast	—	—	320	22.35
Endometrial	—	—	56	3.91
Prostate	—	—	436	30.45
Bladder	—	—	89	6.22
Kidney	—	—	46	3.21
Non-Hodgkin's Lymphoma	—	—	53	3.70

NOTES: Standard deviations are shown in parentheses. GED is General Educational Development. HBP is high blood pressure. COPD is chronic obstructive pulmonary disease. IBD is inflammatory bowel disease.

SOURCE: The dataset links the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) cancer registry data with Medicare beneficiaries' responses to the Centers for Medicare & Medicaid Services' Medicare Health Outcomes Survey (MHOS). The linked SEER-MHOS dataset includes four MHOS cohorts (baseline and followup year): 1998 and 2000; 1999 and 2001; 2000 and 2002; and 2001 and 2003.

DISCUSSION

Observational research provides an important approach to help understand the impact of cancer diagnosis and treatment on HRQOL. One such observational study, the SEER-MHOS linked data (Ambs et al., 2008), allows researchers to examine longitudinal relationships between cancer and HRQOL in a large cohort of Medicare managed care recipients and compare these HRQOL changes to individuals without cancer. However, without the use of random assignment, causal inferences are impossible, and statistical approaches are needed to address some of the limitations incurred by the observational design. The current study offers one approach to reduce the potential impact of differences between cancer patients and non-cancer patients on key demographic and clinical factors when measuring change in HRQOL. Results indicated that propensity score matching is a useful way to create a case-control study out of a large cohort study, and allows more confidence in the strength of the observed outcomes. Future SEER-MHOS studies should consider this approach when examining differences both cross-sectional and longitudinal in HRQOL among cancer cases and controls.

This study replicated the propensity score methods described by D'Agostino (1998), applying those techniques in this population-based dataset. It illustrates how propensity score methods can be used to balance two samples that differ on multiple covariates so that they have the same distribution of characteristics on all measured variables of interest in the study.

This was accomplished by matching non-cancer controls to cancer cases using the propensity score. The propensity score is a single indicator for each person that is estimated from a logistic model that includes multiple independent variables (Rosenbaum and Rubin, 1983; Rubin, 1997; Joffe and Rosenbaum, 1999). Among individuals with a given propensity score, the distribution of the covariates is on average the same among the cancer cases and controls (Kurth et al., 2006). Further, a study by Rubin (1997) has shown that when one matches on the propensity score, the group means and standard deviations on the covariates also will be equivalent (Rudner and Peyton, 2006).

Caution should be noted that this methodology is not equivalent to the comparison of two groups in an experimental design. In experimental studies, participants are randomly assigned to the treatment or control group, to adjust for differences between each group on both measured and unmeasured covariates. The same assumption cannot be made for propensity score methodology as we can only balance each group on the measured covariates used in the propensity score model (Austin, 2008). This limitation also applies to other methods used in non-randomized studies that attempt to reduce bias due to group differences using stratification techniques or adjustments in the regression model. Further, it is important to note that this method does not achieve perfect matches on all covariates; however, perfect balance is also not achieved in randomized trials (Austin, 2008).

The ability of the propensity scores to achieve balance between two groups rests on the assumption that the assignment to case (in this example, cancer) and the outcome to be analyzed in the study (HRQOL) are known to be conditionally independent given the covariates (D'Agostino, 1998; Joffe and Rosenbaum, 1999). In other words, this assumption of strong ignorability means that for individuals with the same characteristics used for matching, the regression model assumes that no additional relationship between cancer diagnosis and the measured covariates exist. This is a strong assumption that needs to be considered for each application. Addressing this assumption includes the addition of an array of characteristics in the model associated with the outcome (Foster, 2003) and the selection of multiple controls per case.

Traditional statistical methods to adjust for group differences on covariates are to include all covariates in the regression model when testing for group differences on the study outcome. For example, Smith et al. (2008) found cancer patients to have worse physical and mental health compared to non-cancer patients in a cross-sectional study using regression case-mix adjustment. Baker et al. (2003) also found poorer HRQOL in cancer relative to non-cancer respondents to the MHOS in a cross-sectional study that matched only on age categories.

A study by Rubin (1979) compared the traditional case-mix adjustment methods with methods that adjust using propensity scores and found results often lead to the same conclusions. However, methods that includes the first step of estimating propensity scores and then adjusting for group differences in the second step when analyzing the main outcome has the advantage that many more variables, even those of no relationship to the outcome of interest, can be included in the propensity score model (D'Agostino, 2007). This allows the researcher in the second step to only include the key factors of interest in the regression model; thus, simplifying the model to estimate the effect of cancer on HRQOL relative to the HRQOL of people without cancer who share similar

characteristics as the cancer group. Further, matching procedures can result in estimated effect sizes whose standard errors are smaller than those obtained using the full sample in the covariate adjusted regression model (Smith, 1997).

When examining the study outcome using the matched case-control design, it is important to account for the correlated nature of the matched data (Austin, 2008). For example, the followup study to this one will adjust for the matched non-cancer and cancer respondents when comparing their physical and mental health changes over time. Specifically, a clustering variable is added to the regression model to identify the five non-cancer controls who are matched to each cancer case. Cases and controls within the same matched cluster have similar propensity scores; thus, they are, on average, more similar than are randomly selected cancer and non-cancer respondents. Since the matched samples are not independent, statistical analyses must adjust for the matched nature of the design (Austin, 2008).

Use of propensity scores to adjust for group differences on observed covariates in observational studies is growing rapidly in the health care research field. Because of propensity score matching, what was once considered an observational study can now be considered a quasi-experimental study because of the balance achieved between the two groups on the measured covariates in the study (Austin, 2008). This is essential, particularly in the SEER-MHOS data linkage project, as covariate differences between the cancer and non-cancer samples may lead to biased estimates of the burden of cancer on HRQOL. Researchers using propensity score matching can feel confident in the strength of the observed associations between cancer diagnosis and treatment and HRQOL, which is the subject of a followup study.

ACKNOWLEDGMENTS

The authors wish to thank Marie Topor and Christopher Zeruto for their help to create the dataset for this study, and Steven B. Clauser and Samuel C. (Chris) Haffer, for creating the link between the SEER and MHOS data.

REFERENCES

Ambs, A., Warren, J.L., Bellizzi, K., et al.: Overview of the SEER—Medical Health Outcomes Survey Linked Dataset. *Health Care Financing Review* 29(4): 5-22, Summer 2008.

Austin, P.C.: A Critical Appraisal of Propensity-Score Matching in the Medical Literature between 1996 and 2003. *Statistics in Medicine* 27(12):2037-2049, May 2008.

Baker, F., Haffer, S.C., and Denniston, M.: Health-Related Quality of Life of Cancer and Noncancer Patients in Medicare Managed Care. *Cancer* 97(3): 674-681, February 1, 2003.

Bergstralh, E. and Kosanke, J.: *GMATCH SAS Macro*. Mayo Clinic. April 2004. Internet address: http://mayoresearch.mayo.edu/mayo/research/biostat/sasmacros.cfm (Accessed 2008.)

D'Agostino, R.B., Jr.: Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group. *Statistics in Medicine* 17(19):2265-2281, October 1998.

D'Agostino, R.B., Jr.: Propensity Scores in Cardiovascular Research. *Circulation* 115(17):2340-2343, May 1, 2007.

Foster, E.M.: Propensity Score Matching: An Illustrative Analysis of Dose Response. *Medical Care* 41(10):1183-1192, 2003.

Haffer, S.C. and Bowen, S.E.: Measuring and Improving Health Outcomes in Medicare: The Medicare HOS Program. *Health Care Financing Review* 25(4):1-3, Summer 2004.

Joffe, M.M. and Rosenbaum, P.R.: Invited Commentary: Propensity Scores. *American Journal of Epidemiology* 150(4):327-333, August 15, 1999.

Jones, N., Jones, S.L., and Miller, N.A.: The Medicare Health Outcomes Survey Program: Overview, Context, and Near-Term Prospects. *Health and Quality of Life Outcomes* 2:33, July 2004. Internet address: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=479698 (Accessed 2008.)

Kurth, T., Walker, A.M., Glynn, R.J., et al.: Results of Multivariable Logistic Regression, Propensity

Matching, Propensity Adjustment, and Propensity-Based Weighting under Conditions of Nonuniform Effect. *American Journal of Epidemiology* 163(3): 262-270, 2006.

Ries, L.A.G., Melbert, D., Krapcho, (eds.) et al.: *SEER Cancer Statistics Review, 1975-2004.* National Cancer Institute. Internet address: http://seer.cancer.gov/csr/1975_2004 (Based on November 2006 SEER data submission, posted to the SEER Web site, 2007). (Accessed 2008.)

Rosenbaum, P.R. and Rubin, D.B.: The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70:41-55, 1983.

Rosenbaum, P.R. and Rubin, D.B.: Constructing a Control Group by Multivariate Matched Sample Methods that Incorporate the Propensity Score. *The American Statistician* 39(1):33-38, February 1985.

Rubin, D.B.: Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. *Journal of the American Statistical Association* 74(366):318-324, June 1979.

Rubin, D.B.: Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Annals of Internal Medicine* 127(8):757-763, October 1997.

Rubin, D.B. and Thomas, N.: Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics* 52(1):249-264, March 1996.

Rudner, L.M. and Peyton, J.: Consider Propensity Scores to Compare Treatments. *Practical Assessment, Research & Evaluation* 11(9):1-9, November 2006.

Smith, A.W., Reeve, B.B., Bellizzi, K., et al.: Cancer, Comorbidities and Health-Related Quality of Life of Older Adults. *Health Care Financing Review* 29(4): 41-56, Summer 2008.

Smith, H.L.: Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies. *Sociological Methodology* 27(1):325-353, 1997.

Ury, H.K.: Efficiency of Case-Control Studies with Multiple Controls Per Case: Continuous or Dichotomous Data. *Biometrics* 31(3):643-649 September 1975.

Ware, J.E. and Sherbourne, C.D.: The MOS 36-Item Short-Form Health Survey (SF-36®): I. Conceptual Framework and Item Selection. *Medical Care* 30(6): 473-83, June 1992.

Reprint Requests: Bryce B. Reeve, Ph.D., National Cancer Institute, EPN 4005, 6130 Executive Blvd., MSC 7344, Bethesda, MD 20892-7344. E-mail: reeveb@mail.nih.gov