Health and Quality of Life Outcomes

Research

Utilization of the propensity score method: an exploratory comparison of proxy-completed to self-completed responses in the Medicare Health Outcomes Survey

Beth Hartman Ellis*[1], Wade M Bannister[2], Jacquilyn Kay Cox[1], Brenda M Fowler[1], Erin Dowd Shannon[1], David Drachman[1], Randall W Adams[1] and Laura A Giordano[1]

Address: [1]Surveys, Research & Analysis Health Services Advisory Group, Inc. Phoenix, AZ 85020 United States of America and [2]Information Planning Data Group Mercer Human Resource Consulting Phoenix, Arizona 85016 United States of America

Email: Beth Hartman Ellis* - bellis@AZQIO.sdps.org; Wade M Bannister - wade.bannister@mercer.com; Jacquilyn Kay Cox - jcox@AZQIO.sdps.org; Brenda M Fowler - bfowler@AZQIO.sdps.org; Erin Dowd Shannon - eshannon@AZQIO.sdps.org; David Drachman - ddrachman@AZQIO.sdps.org; Randall W Adams - radams.@AZQIO.sdps.org; Laura A Giordano - lgiordano@AZQIO.sdps.org

* Corresponding author

Published: 18 September 2003

Health and Quality of Life Outcomes 2003, **1**:47

Received: 25 April 2003
Accepted: 18 September 2003

This article is available from: http://www.biomedcentral.com/1477-7525/1/47

Abstract

Background: This research examined the use of the propensity score method to compare proxy-completed responses to self-completed responses in the first three baseline cohorts of the Medicare Health Outcomes Survey, administered in 1998, 1999, and 2000, respectively. A proxy is someone other than the respondent who completes the survey for the respondent.

Methods: The propensity score method of matched sampling was used to compare proxy and self-completed responses. A propensity score is a value that equals the estimated probability of a given individual belonging to a treatment group given the observed background characteristics of that individual. Proxy and self-completed responses were compared on demographics, the SF-36, chronic conditions, activities of daily living, and depression-screening questions. For each individual survey respondent, logistic regression was used to calculate the probability that this individual belonged to the proxy respondent group (propensity score). Pre and post adjustment comparisons were tested by calculating effect sizes.

Results: Differences between self and proxy-completed responses were substantially reduced with the use of the propensity score method. However, differences were still found in the SF-36, several demographics, several impaired activities of daily living, several chronic conditions, and one depression-screening question.

Conclusion: The propensity score method helped to reduce differences between proxy-completed and self-completed survey responses, thereby providing an approximation to a randomized controlled experiment of proxy-completed versus self-completed survey responses.

Background

Surveys such as the Medicare Health Outcomes Survey (HOS) [1] are widely used to assess respondents' physical and mental health status. While survey methods are crucial to the assessment of self-reported health care conditions and outcomes, the use of proxy-completed responses in interviews and surveys may systematically affect responses (a proxy is someone other than the respondent, i.e. professional caregiver, friend, family member, or relative who completes the survey for the beneficiary). This is a particularly troublesome problem for data collected on an elderly population, since the elderly frequently must rely on a proxy. The propensity score method provides an approach for assessing bias in self-report surveys such as the Medicare HOS. The goal of using the propensity score methodology is to create balance between different groups of subjects [2]. In this research, we apply the propensity score method [3] to three cohorts of self and proxy-completed responses in the Medicare HOS to compare results for physical and mental health status.

Self-Completed and Proxy-Completed Response Differences

Literature exists documenting the differences between self and proxy-completed responses on health status surveys. For example, some research demonstrates that proxy-completed responses tend to more accurately report conditions that are less private and more observable, but tend to underestimate less observable conditions such as emotional and affective states [4,5]. Additionally, Yip, Wilber, Myrtle, and Grazman found that mean scores were significantly lower for proxy-completed responses compared to self-completed responses on the Physical Functioning, Vitality, and Mental Health scales of the 36-Item Short-Form Health Survey (SF-36)[6].

Other research also has indicated significant disagreement between proxy-completed responses and observers for instrumental activities of daily living (IADLs). In this research, proxy-completed responses underreported IADLs compared to observers who watched subjects engaged in IADLs [7]. Systematic biases were found in the National Health Interview Survey; results indicated that proxy-completed responses underreported disabilities for those aged 18 to 64 years, but overreported disabilities for those 65 and older [8]. Data from the Canadian SF-36 indicated that proxy-completed responses tended to underestimate health status, with poor to moderate agreement between proxy-completed responses and the disabled elderly [9]. In examining data from Canada's National Population Health Survey, Shields [10] found significant differences between self and proxy-completed responses. In general, proxy-completed respondents underestimated the prevalence of certain health conditions. However, disagreement between self-completed and proxy-completed respondents was less likely for conditions that the proxy respondents were more likely not to mislabel, such as diabetes, heart disease, and cancer. It is evident that there are inconsistencies in the literature regarding proxy-completed and self-completed respondents and continued research is necessary to understand these inconsistencies.

Because proxy-completed responses are often necessary in assessing health outcomes for the elderly, it is important that methods be found for examining selection bias. The propensity score is one such method used to reduce selection bias in observational studies. This paper explores the use of the propensity score method in understanding the differences between proxy and self-completed responses, by applying this method to the Medicare HOS data.

Propensity Score Methodology

Donald Rubin [3,11–13] pioneered the propensity score methodology, which has been used extensively in medical research [e.g. 14–17]. Theoretically, this method is similar to an experimental design, but it is applied to survey or observational data and has the potential to reduce selection bias. Simply put, the propensity score is the probability that an individual belongs to a naturally occurring treatment group, based on the individual's background characteristics (covariates). Since the propensity score summarizes the information on the background characteristics in a single summary score, it has a distinct advantage over standard matching techniques [18–20]. These latter techniques require the investigator to find subjects who are closely matched on each of many individual covariates, an often difficult task. Once the propensity scores have been calculated, the "treatment" and "control" groups (in this case, proxy-completed and self-completed respondents) can each be stratified into similarly matched comparison groups based upon their propensity scores. For each stratum we can then examine two groups of survey respondents, a group of proxy-respondents and a group of self-respondents that have similar propensity scores and who were "randomly" assigned to the groups in the sense of being equally likely to be a proxy or self-respondent [2].

For example, in a 2002 study on nephrology consultation in acute renal failure, Mehta et al. [17] used the propensity score methodology to assess the timing of nephrology consultation and in-hospital mortality. In this example, the authors created the propensity score using the characteristics that differentiated the delayed and early consultation groups. These authors state, "Inclusion of the propensity score as a covariate in a multivariable regression accounts for likelihood of 'treatment' (in this case, timing of consultation), and may adjust for unobserved,

Table 1: Demographics: Unadjusted Data from Cohorts I, II, and III

Variable	Category	Proxy-Completed		Self-Completed		χ^2 Value	Effect Size
		N = 65,668		N = 457,837			
Gender	Male	31,950	(48.7%)	195,026	(42.6%)	857.8***	0.12
	Female	33,718	(51.4%)	262,811	(57.4%)		0.12
	Missing/Total	0/	65,668	0/	457,837		
Race	American Indian/Alaskan Native	674	(1.1%)	3013	(0.7%)	7055.8***	0.04
	Asian/Pacific Islander	1,991	(3.2%)	6,714	(1.5%)		0.11
	African American	7,459	(11.8%)	27,740	(6.2%)		0.20
	White	49,382	(78.4%)	399,106	(89.5%)		0.31
	Other/Multiracial	3,513	(5.6%)	9,135	(2.1%)		0.19
	Missing/Total	2,649/	63,019	12,129/	445,708		
	Hispanic or Spanish Background	7,634	(12.1%)	18,856	(4.3%)	6,730.4***	0.29
	Missing/Total	2591/	63,077	18,726/	439,111		
Age Group	Under 45	1,271	(1.9%)	3,371	(0.7%)	23,864.7***	0.11
	45 – 54	1,600	(2.4%)	6,898	(1.5%)		0.07
	55 – 64	3,126	(4.8%)	16,135	(3.5%)		0.07
	65 – 74	22,198	(33.8%)	255,642	(55.8%)		0.45
	75 – 84	23,430	(35.7%)	148,105	(32.4%)		0.07
	85 or Over	14,040	(21.4%)	27,686	(6.1%)		0.46
	Missing/Total	3/	65,665	0/	457,837		
Marital Status	Married	35,195	(54.6%)	262,519	(58.4%)	2,448.0***	0.08
	Divorced	3,600	(5.6%)	43,510	(9.7%)		0.16
	Separated	706	(1.1%)	4,304	(1.0%)		0.01
	Widowed	21,749	(33.7%)	124,536	(27.7%)		0.13
	Never Married	3,216	(5.0%)	14,271	(3.2%)		0.09
	Missing/Total	1202/	64,466	8,697/	449,140		
Educational Level	8th Grade or Less	25,470	(40.3%)	39,930	(8.9%)	53,601.4***	0.77
	Some High School	12,840	(20.3%)	78,434	(17.5%)		0.07
	High School/GED	15,925	(25.2%)	164,319	(36.6%)		0.25
	Some College	5,851	(9.3%)	100,927	(22.5%)		0.37
	College Graduate	1,839	(2.9%)	31,628	(7.0%)		0.19
	More than 4 Year Degree	1,326	(2.1%)	34,058	(7.6%)		0.27
	Missing/Total	2,417/	63,251	8,541/	449,296		
Income Level	Less than $5,000	4,999	(9.8%)	15,722	(4.2%)	8,128.5***	0.22
	$5,000 – $9,999	10,273	(20.4%)	44,570	(12.0%)		0.23
	$10,000 – $19,999	17,320	(33.9%)	113,960	(30.8%)		0.07
	$20,000 – $29,999	9,117	(17.8%)	81,182	(21.9%)		0.10
	$30,000 – $39,999	4,372	(8.5%)	48,140	(13.0%)		0.15
	$40,000 – $49,999	2,199	(4.3%)	26,959	(7.3%)		0.13
	$50,000 – $79,999	2,036	(4.0%)	26,896	(7.3%)		0.14
	$80,000 – $99,999	486	(0.9%)	5,850	(1.6%)		0.06
	$100,000 or more	460	(0.9%)	6,925	(1.9%)		0.09
	Missing /Total	14,406/	51,262	87,633/	370,204		
Homeowner Status	Owned or being bought by you	37,438	(60.4%)	342,327	(77.0%)	9,930.2***	0.36
	Owned or being bought by someone in your family other than you	9,168	(14.8%)	26,968	(6.1%)		0.29
	Rented for money	13,298	(21.5%)	68,165	(15.3%)		0.16
	Not owned and one in which you live without payment of rent	2,080	(3.4%)	7,327	(1.7%)		0.11
	Missing/Total	3,684/	61,984	13,050/	444,787		
Institutionalized	No	64,132	(98.3%)	457,136	(99.9%)	6,108.7***	0.20
	Yes	1,128	(1.7%)	246	(0.1%)	9,930.2***	0.20
	Missing/Total	408/	65,260	455/	457,382		

***$p < .0001$

confounding, and selection bias, thereby refining regression estimates".

Similarly, Kilborn et al. [14] utilized the propensity score methodology in a nonrandomized study of amiodarone and mortality among acute myocardial infarction patients with atrial fibrillation. Patient characteristics that were associated with prescriptive use of amiodarone were incorporated into a regression analysis through the use of the propensity score methodology.

Assessing the difference between proxy-completed and self-completed responses in survey research is analogous to nonrandomized treatment studies such as the two dis-

cussed above. Proxy-completed responses can be conceptualized partly as the results of selection bias, and partly the result of true differences between proxy and self-respondents. Generally, proxy respondents who answer survey items for the respondent occupy a specific role in the respondent's life such as a family member (spouse, child), friend, or professional caregiver. Essentially, these proxy respondents bring a cognitive role set with them when they complete survey items on behalf of the respondent. This role set can bias responses to survey items. For example, if an adult child completes a survey for an elderly parent, the adult child may understate the physical and mental health status of the elderly parent. Indeed, these results were found in analyses conducted on the Canadian SF-36 [9]. Hence, it becomes very important to test for differences between proxy and self-completed respondents. If differences are found, adjustment procedures such as the propensity score should be used to reduce those differences in further analyses on outcomes such as physical and mental health status.

The propensity score is a methodology that has heuristic potential for quality of life research, which generally involves self and proxy-completed responses. We apply this methodology to three cohorts of data that include self-completed and proxy-completed responses from the Medicare HOS in order to determine if differences in physical and mental health status remain after adjustment using the propensity score.

Methods

Data collection

The Medicare HOS assesses the physical and mental health status of the Medicare elderly enrolled in managed care in the United States. Beginning in 1998 and continuing annually, a baseline cohort is created from a randomly selected sample of 1,000 Medicare members from each applicable Medicare contract market area. In plans with fewer than 1,000 Medicare members, the sample includes the entire enrolled Medicare population that meets the inclusion criteria. Medicare beneficiaries who are continuously enrolled in the health plans for at least six months are eligible for sampling [21].

The data collection protocol includes a combination of mail and telephone surveys. Multiple mailings, standardized telephone interviews, interviewer training, and methods for maximizing response rates are well established in the Health Plan Employer Data and Information Set (HEDIS) specifications [22]. The Medicare HOS instrument includes the SF-36 health survey, which is a widely used multi-purpose, short-form health survey. Psychometric properties, reliability and validity studies of the SF-36 as well as normative data are available in user manuals [23,24]. The SF-36 yields an eight-scale profile of scores and is a generic measure as opposed to one that targets a specific age, disease, or treatment group. The eight scales form two distinct higher ordered clusters that are the basis for scoring the physical component summary (PCS) measure and mental component summary (MCS) measure. For this analysis, the SF-36 individual scale scores, as well as the PCS and MCS scores, have been normed to the values for the 1990 general U.S. population, so that a score of fifty represents the national average for a given scale or summary score. Higher scores on the SF-36 represent better physical and/or mental health status.

The respondents included in this study were beneficiaries in baseline cohorts I, II, and III; the data sets represented survey results for 1998, 1999, and 2000, respectively. The Medicare HOS cohort I baseline consisted of 279,135 Medicare members who were sampled from 269 Medicare+Choice organizations (M+COs) representing 287 contract market areas. Cohort II baseline consisted of 301,184 Medicare members from 283 M+COs in 312 contract market areas, and cohort III baseline consisted of 298,883 Medicare members from 275 M+COs in 306 market areas. Several criteria were met in selecting the final analytic sample. First, all duplicates were removed; i.e., only the first survey was used for any beneficiary. Second, beneficiaries in plans for Program of All-Inclusive Care for the Elderly (PACE), as well as EVERCARE (a program that provides care and care coordination to vulnerable, chronically ill beneficiaries) in cohorts II and III were removed (0 in cohort I; 4,225 PACE and 5,015 EVERCARE beneficiaries in cohort II; 3,267 beneficiaries in PACE programs in cohort III; PACE and EVERCARE beneficiaries have significantly lower PCS and MCS scores and are much more ill than non-PACE and non-EVERCARE beneficiaries in the Medicare HOS). Third, surveys must have had a response for the question, "Who completed this survey form?" and finally, responses must have had a survey for which the PCS and MCS scores were calculable. Based on these criteria, the total sample size of proxy-completed responses was 65,668 and for self-completed responses the total was 457,837.

In addition to the SF-36, demographic data, activities of daily living (ADLs), chronic conditions (angina pectoris, arthritis, cancer, congestive heart failure, Crohn's disease, diabetes, emphysema/asthma/chronic obstructive pulmonary disease [COPD], hypertension, myocardial infarction, other heart conditions, sciatica, and stroke), and three depression-screening questions were examined for differences between self-completed and proxy-completed responses.

Data Analyses

Three steps were necessary in applying the propensity score method to the Medicare HOS data. First, self-com-

Table 2: SF-36, Proxy and Self-Completed Differences in Mean Scores: Unadjusted Data

SF-36 Measure	Proxy-Completed		Self-Completed			
	Mean	(SD)	Mean	(SD)	t Value	Effect Size
Physical Component Summary Score (PCS)	33.99	(12.25)	40.97	(11.97)	129.8***	0.58
Mental Component Summary Score (MCS)	46.69	(12.56)	52.56	(10.09)	108.6***	0.56
Physical Functioning Scale	31.34	(14.44)	40.83	(12.57)	158.8***	0.74
Role-Physical Scale	37.38	(12.81)	42.92	(12.82)	101.0***	0.43
Bodily Pain Scale	40.03	(12.11)	44.61	(11.23)	90.4***	0.40
General Health Scale	37.80	(12.25)	45.64	(11.01)	153.0***	0.70
Vitality Scale	40.84	(11.82)	47.88	(10.87)	141.8***	0.64
Social Functioning Scale	39.19	(14.91)	48.07	(11.57)	145.0***	0.74
Role-Emotional Scale	43.54	(13.69)	48.41	(11.44)	84.0***	0.41
Mental Health Scale	44.62	(12.57)	51.30	(10.14)	128.0***	0.64

***$p < .001$

pleted and proxy-completed responses were examined to establish differences between the groups of respondents (unadjusted comparisons). These two groups were compared on demographic variables, the SF-36 scores, type of chronic condition, type of impaired ADL, and three depression-screening questions.

Second, the propensity score was used to create comparison samples. The propensity score matching process involved developing a stepwise logistic model [12] to determine which demographic, disease, and disability variables affected the likelihood of a proxy response. Based on the values of these predictors, each beneficiary in the data set had an estimated probability of using a proxy, which is the propensity score.

Third, due to the large sample sizes, a stratified random sample was drawn from each decile of the distribution of the propensity score from the proxy-completed and self-completed respondent groups. Once the sample was drawn, resulting in a risk adjusted sub-sample, comparisons were made to determine whether proxy-completed responses differed from self-completed responses. Given the large sample sizes, effect sizes were used to determine significance. Effect size is defined as "...the degree to which the phenomenon is present in the population...or the degree to which the null hypothesis is false." Cohen [25] operationally defines effect sizes as follows: a small effect size is one that accounts for 2% (0.02) of the variance, a medium effect size accounts for 13%, (0.13) and a large effect size accounts for 26% (0.26). Cohen's effect size for proportions (p) was used to calculate the effect sizes for Tables 1 and 2 ($h = \phi_1 - \phi_2$, where: $\phi = 2\arcsin\sqrt{p}$). Cohen's effect size for means was calculated as:

$$d = \frac{x_1 - x_2}{s^2 pooled}.$$

Results and Discussion

Unadjusted Self-Completed and Proxy-Completed Response Comparisons

Demographics

The proxy-completed responses differed from the self-completed responses on most demographic characteristics (table 1). Small effect sizes were found for all variables (with the exception of separated) and many large effects were found (white and Hispanic race; age 65–74 and 85 or over; 8th grade or less, some college, and more than a 4 year degree; homeowner status of owned and owned by someone in the family).

SF-36 Scores

Large effect sizes were found for the PCS, MCS, and all scales. Table 2 indicates that the mean PCS scores between proxy-completed and self-completed responses reflected a seven-point difference, with proxy-completed responses having lower scores than self-completed responses (33.99 and 40.97, respectively). The mean MCS score indicated a strikingly similar situation with a proxy-completed mean score of 46.69 and a self-completed mean score of 52.56.

Chronic Conditions

The proxy-completed responses differed from the self-completed responses on all chronic conditions. Additionally, proxy-completed respondents reported proportionally more of each condition; small, medium, and large effects were found for all conditions. The effect size for stroke was the largest (0.38) with about 20% of the proxy respondents who reported this condition compared to approximately 7% of the self respondents (table 3).

Impaired ADLs

Table 3 also indicates that large effects were found for all impaired ADLs. Proxy-completed responses had proportionally more impaired ADLs than self-completed

Table 3: Chronic Conditions, Activities of Daily Living, and Depression: Unadjusted Data

Chronic Condition	Proxy-Completed		Self-Completed		χ^2 Value	Effect Size
Angina Pectoris ∫	12,924	(20.6%)	70,293	(15.9%)	904.0***	0.12
Arthritis Hand/Wrist	23,973	(37.7%)	150,840	(33.8%)	377.7***	0.08
Arthritis Hip/Knee	27,684	(43.5%)	170,727	(38.2%)	655.0***	0.11
Any Cancer ‡	8,972	(14.0%)	59,747	(13.4%)	22.9***	0.02
Congestive Heart Failure	8,807	(14.0%)	29,599	(6.7%)	4,213.7***	0.24
Crohn's Disease/ Ulcerative Colitis/Inflammatory Bowel Disease	4,074	(6.5%)	23,745	(5.4%)	130.7***	0.05
Diabetes	15,410	(24.2%)	74,656	(16.7%)	2,134.6***	0.19
Emphysema/Asthma/COPD †	9,732	(15.4%)	59,100	(13.3%)	208.7***	0.06
Hypertension	36,769	(57.6%)	240,054	(53.6%)	368.1***	0.08
Myocardial Infarction	9,514	(15.1%)	46,907	(10.6%)	1,140.7***	0.13
Other heart conditions	15,845	(25.2%)	95,171	(21.5%)	448.6***	0.09
Sciatica	15,068	(24.0%)	102,293	(23.1%)	25.7***	0.02
Stroke	12,497	(19.7%)	31,086	(7.0%)	11,456.4***	0.38
Type of Impaired ADL						
Unable/Difficulty Using Toilet	16,750	(26.1%)	30,370	(6.8%)	25,169.9***	0.54
Unable/Difficulty Eating	12,431	(19.4%)	20,264	(4.5%)	20,740.4***	0.48
Unable/Difficulty Bathing	26,088	(40.5%)	54,136	(12.0%)	34,682.6***	0.67
Unable/Difficulty Transferring from Chairs	30,849	(48.0%)	113,162	(25.2%)	14,446.7***	0.48
Unable/Difficulty Dressing	23,425	(36.4%)	44,299	(9.9%)	34,647.4***	0.66
Unable/Difficulty Walking	37,992	(59.1%)	148,584	(33.1%)	16,417.2***	0.53
Depression						
Sad/Blue 2 Weeks in Past Year	25,084	(39.8%)	92,787	(20.8%)	11,131.1***	0.42
Depressed /Sad Much of the Time in the Past Year	19,274	(30.6%)	57,531	(12.9%)	13,432.3***	0.44
Depressed/Sad 2 or More Years in Life	15,778	(25.2%)	60,231	(13.6%)	5,818.1***	0.30

***$p < .0001$ ∫ or coronary artery disease ‡ other than skin cancer † Chronic obstructive pulmonary disease

responses. For example, a large difference existed in difficulty dressing. Approximately 36% of the proxy respondents reported inability or difficulty dressing, whereas only 10% of the self respondents reported a problem. Bathing was another ADL reflecting extreme differences. Approximately 41% of the proxy respondents had inability or difficulty bathing compared to only 12% of the self respondents. About 59% of the proxy respondents reported inability or difficulty walking and approximately 33% of the self respondents reported a problem walking.

Depression

Large effect sizes were also found between proxy-completed and self-completed responses for the three depression-screening questions. Proxy-completed responses had proportionally more affirmative responses to the depression-screening questions compared to self-completed responses. Approximately 40% of the proxy-completed respondents indicated feeling sad/blue for two or more weeks in the past year compared to about 21% of self-completed respondents. About 13% of the self-completed respondents reported feeling depressed or sad much of the time in the past year compared to 31% of the proxy-completed respondents. Similarly, approximately 25% of the proxy-completed respondents reported feeling depressed/sad for two or more years in their life compared to about 14% of the self-completed respondents.

These comparisons established that the proxy-completed respondents were demographically different, varied in type of chronic condition, reported more depression, and reported decreased status in physical and mental health compared to self-completed respondents.

Stepwise Regression

Since significant differences were found between self and proxy-completed responses on the above stated characteristics, these variables were entered in a stepwise logistic regression as independent variables, with the dependent variable coded as 1 for proxy-completed responses and 0 for self-completed responses. The independent variables were: age under 45, age 45 – 54, age 55 – 64, age 75 – 84, age 85 and over (reference group was 65 – 74); Black/African American race, Asian race, other race (reference group was white); Hispanic ethnicity; widowed, never married, divorced or separated (reference group was married); educational attainment of eighth grade or less, educational attainment of some high school, education beyond high school (reference group was high school graduate or GED); homeownership; female; Medicaid enrolled; insti-

Table 4: Significant Variables for Propensity Score Adjustment

Effect	Odds Ratio	95% CI	χ^2 Value	*p* Value
Age *Reference Group: 65–74*				
<45	2.04	1.86, 2.24	222.51	.0001
45–54	0.98	0.93, 1.04	26.01	.0001
55–64	0.98	0.93, 1.04	0.50	.479
75–84	1.60	1.56, 1.64	1,282.39	.0001
>84	3.86	3.72, 4.00	5,149.46	.0001
Race *Reference Group: White*				
Black	1.62	1.56, 1.69	594.21	.0001
Asian	3.00	2.81, 3.21	1,057.17	.0001
Other	1.36	1.29, 1.44	126.49	.0001
Hispanic	1.93	1.84, 2.01	854.71	.0001
Marital Status *Reference Group: Married*				
Never Married	0.96	0.91, 1.02	1.64	.200
Widowed	0.73	0.71, 0.75	472.63	.0001
Divorced/Separated	0.45	0.43, 0.47	1,162.60	.0001
Educational Attainment *Reference Group: HS diploma or GED*				
8th Grade or Less	5.41	5.26, 5.57	12,746.04	.0001
Some High School	1.59	1.54, 1.64	866.01	.0001
> High School	0.55	0.53, 0.57	1,342.89	.0001
Own Home	1.56	1.52, 1.60	1,178.30	.0001
Female	0.76	0.75, 0.78	464.84	.0001
Medicaid Benefits	1.53	1.45, 1.60	277.41	.0001
Institutionalized	5.33	4.23, 6.71	202.38	.0001
Activities of Daily Living				
Toileting	1.25	1.20, 1.30	121.24	.0001
Eating	1.53	1.47, 1.60	438.03	.0001
Dressing	1.68	1.62, 1.75	553.91	.0001
Bathing	1.53	1.47, 1.60	420.76	.0001
Transferring from Chairs	1.04	1.01, 1.07	5.20	.023
Walking	1.28	1.24, 1.32	239.80	.0001
Chronic Conditions				
Angina Pectoris ∫	1.02	0.99, 1.06	1.24	.266
Arthritis of Hand/Wrist	0.92	0.90, 0.95	36.61	.0001
Arthritis of Hip/Knee	0.89	0.87, 0.92	70.28	.0001
Any Cancer ‡	0.99	0.97, 1.03	.012	.913
Congestive Heart Failure	1.26	1.21, 1.31	119.98	.0001
Crohn's Disease/Ulcerative Colitis/inflammatory Bowel Disease	0.88	0.84, 0.92	28.19	.0001
Diabetes	1.13	1.10, 1.16	74.91	.0001
Emphysema/Asthma/COPD†	0.91	0.88, 0.94	34.87	.0001
Hypertension	0.95	0.93, 0.97	21.66	.0001
Myocardial Infarction	0.92	0.93, 1.01	2.02	.155
Other Heart Conditions	0.90	0.88, 0.93	48.59	.0001
Sciatica	0.74	0.72, 0.76	437.63	.0001
Stroke	1.97	1.91, 2.04	1,597.27	.0001
Depression				
Sad/Blue 2 Weeks in Past Year	1.35	1.30, 1.39	319.35	.0001
Depressed /Sad Much of the Time in the Past Year	1.33	1.28, 1.38	210.86	.0001
Depressed/Sad 2 or More Years in Life	0.93	0.90, 0.97	15.69	.0001

∫ or coronary artery disease ‡ other than skin cancer † Chronic obstructive pulmonary disease

tutionalized; activities of daily living; chronic conditions; and three depression-screening questions (table 4).

Based on the results of the regression analyses, proxy-completed respondents were about twice as likely to be under 45 years old and approximately four times as likely to be over the age of 84. They were about five times as

Table 5: Distribution of Propensity Scores Prior to Matched Sampling and Random Samples from Each Decile

		Number and Percentage Prior to Sampling				Random Samples from Each Decile	
Range of Propensity Score	**Decile of Propensity Score Distribution**	**Self-Completed Group**		**Proxy-Completed Group**		**Self-Completed Group**	**Proxy-Completed Group**
0.00 – 0.09	1	278,690	(76.5%)	11,266	(23.6%)	400	400
0.10 – 0.19	2	45,960	(12.6%)	8,739	(18.3%)	400	400
0.20 – 0.29	3	18,771	(5.2%)	6,887	(14.4%)	400	400
0.30 – 0.39	4	8,866	(2.4%)	5,345	(11.2%)	400	400
0.40 – 0.49	5	5,172	(1.4%)	4,279	(9.0%)	400	400
0.50 – 0.59	6	3,145	(.9%)	3,397	(7.1%)	400	400
0.60 – 0.69	7	1,827	(.5%)	2,600	(5.5%)	400	400
0.70 – 0.79	8	1,122	(.3%)	2,234	(4.7%)	400	400
0.80 – 0.89	9	631	(.2%)	1,949	(4.1%)	200	200
0.90 – 1.00	10	152	(<.1%)	1,002	(2.1%)	50	50
	Total	364,336		47,698		3,450	3,450

likely to have an 8th grade education or less and to be institutionalized. They were about one and a quarter times more likely to have congestive heart failure and to be depressed two or more weeks in the past year.

The next step in the propensity score process involved creating a distribution from which to randomly sample respondents.

Sampling. Based on the values of the variables listed above, each beneficiary in the three cohorts had an associated probability of having a proxy-completed survey (i.e. the propensity score). The distribution of the propensity score was divided into deciles, and stratified random samples of 400 from the first through the eighth deciles were drawn from both the proxy-completed responses and the self-completed responses (table 5). Due to the large size of decile one, the stratified random sample was selected from the midpoint of that decile. Due to the small size of deciles nine and ten, 200 were drawn from the ninth decile and 50 were drawn from the tenth decile for both groups (see table 5). This methodology had the effect of providing a sample that was reasonably well distributed between the groups with respect to the characteristics that helped to determine proxy-completed responses. Thus, respondents in the proxy-completed and self-completed groups with equal (or nearly equal) propensity scores should have the same (or nearly the same) distributions on the variables included in the logistic regression model [22].

Adjusted Self-Completed and Proxy-Completed Response Comparisons

Demographics

Despite the propensity score adjustment, small effects in demographics existed within the adjusted self-completed and proxy-completed comparisons, as shown in table 6. Small effect sizes were found for both male and female gender; greater proportions of self-completed responses were male and higher proportions of proxy-completed responses were female. Small effects were found for American Indian/Alaskan Native, with proportionally less proxy-completed responses in this category. A small effect was also found for white race (more white proxy-completed responses). Small effects were found for ages 55–64 (more self-completed responses), 65–74 (more self-completed responses), and 85 or over (more proxy-completed responses). Small effects were found for all categories of marital status and a medium effect was found for divorced (more self-completed responses). Small effects were found for all educational levels. More self-completed respondents reported an 8th grade or less education and some high school; more proxy respondents reported an educational level of high school/GED, some college and college graduate; however, more self-completed respondents had more than a four year degree. and for all income levels, with the exception of $5,000 – $9,999 and $80,000 – $99,999. Small effects were found for all categories of homeowner status and for institutionalization.

SF-36

Small effects were found for PCS and MCS scores (table 7). Medium effects were found for the Physical Functioning scale, the Vitality scale, the Social Functioning scale, and the Role-Emotional scale. Small effect sizes were found for all other scales.

Table 6: Demographics: Adjusted Data from Cohorts I, II, and III

Variable	Category		Proxy-Completed		Self-Completed		χ^2 Value	Effect Size
			N = 3,450		N = 3,450			
Gender	Male		1,557	(45.1%)	1,654	(47.9%)	5.5**	0.06
	Female		1,893	(54.9%)	1,796	(52.1%)		0.06
		Missing/Total	0/	3,450	0/	3,450		
Race	American Indian/Alaskan Native		49	(1.4%)	78	(2.3%)	9.3*	0.07
	Asian/Pacific Islander		120	(3.5%)	129	(3.7%)		0.01
	African American		423	(12.3%)	458	(13.3%)		0.01
	White		2,619	(75.9%)	2,549	(73.9%)		0.05
	Other/Multiracial		239	(6.9%)	236	(6.8%)		0.00
		Missing/Total	0/	3,450	0/	3,450		
	Hispanic or Spanish Background		454	(13.2%)	466	(13.5%)	0.18	0.01
		Missing/Total	0/	3,450	0/	3,450		
Age Group	Under 45		77	(2.2%)	77	(2.2%)	6.7	0.00
	45 – 54		82	(2.4%)	80	(2.3%)		0.01
	55 – 64		177	(5.1%)	191	(5.5%)		0.02
	65 – 74		1,126	(32.6%)	1,180	(34.2%)		0.03
	75 – 84		1,107	(32.1%)	1,129	(32.7%)		0.01
	85 or Over		881	(25.5%)	793	(23.0%)		0.06
		Missing/Total	0/	3,450	0/	3,450		
Marital Status	Married		1,734	(50.3%)	1,649	(47.8%)	92.3***	0.05
	Divorced		217	(6.3%)	445	(12.9%)		0.23
	Separated		30	(0.9%)	44	(1.3%)		0.04
	Widowed		1,271	(36.8%)	1,142	(33.1%)		0.08
	Never Married		198	(5.7%)	170	(4.9%)		0.04
		Missing/Total	0/	3,450	0/	3,450		
Educational Level	8th Grade or Less		1,814	(52.6%)	1,880	(54.5%)	30.7***	0.04
	Some High School		526	(15.3%)	576	(16.7%)		0.04
	High School/GED		537	(15.6%)	432	(12.5%)		0.09
	Some College		375	(10.9%)	345	(10.0%)		0.03
	College Graduate		109	(3.2%)	80	(2.3%)		0.06
	More than 4 Year Degree		89	(2.6%)	137	(4.0%)		0.08
		Missing/Total	0/	3,450	0/	3,450		
Income Level	Less than $5,000		259	(9.2%)	269	(9.8%)	48.3***	0.02
	$5,000 – $9,999		602	(21.3%)	588	(21.5%)		0.00
	$10,000 – $19,999		968	(34.2%)	999	(36.4%)		0.05
	$20,000 – $29,999		449	(15.9%)	498	(18.2%)		0.06
	$30,000 – $39,999		218	(7.7%)	187	(6.8%)		0.03
	$40,000 – $49,999		123	(4.3%)	83	(3.0%)		0.07
	$50,000 – $79,999		155	(5.5%)	82	(3.0%)		0.13
	$80,000 – $99,999		23	(0.8%)	20	(0.7%)		0.01
	$100,000 or more		31	(1.1%)	15	(0.5%)		0.07
	Missing /Total		622/	2,828	709/	2,741		
Homeowner Status	Owned or being bought by you		1,916	(55.5%)	2,036	(59.0%)	23.9***	0.07
	Owned or being bought by someone in your family other than you		574	(16.6%)	442	(12.8%)		0.11
	Rented for money		825	(23.9%)	861	(25.0%)		0.03
	Not owned and one in which you live without payment of rent		135	(3.9%)	111	(3.2%)		0.04
		Missing/Total	0/	3,450	0/	3,450		
Institutionalized	No		3,408	(98.8%)	3,438	(99.7%)	16.8***	0.11
	Yes		42	(1.2%)	12	(0.4%)		0.09
		Missing/Total	0/	3,450	0/	3,450		

Chronic Conditions

Table 8 indicates that small effect sizes were found for all chronic conditions except any cancer, congestive heart failure, and stroke.

Impaired ADLs

Small effects were found for all impaired ADLs; inability/difficulty toileting, eating, bathing, and dressing. However, inability/difficulty getting in or out of chairs and walking did not meet the effect size criterion (table 8).

Depression

A small effect size was found for the depression screening question, "depressed /sad 2 or more years in life." However, the other two depression screening questions did not meet the small effect size criterion.

Table 7: SF-36, Proxy, and Self-Completed Differences in Mean Normed Scores: Adjusted Data

SF-36 Measure	Proxy-Completed		Self-Completed			
	Mean	(SD)	Mean	(SD)	t Value	Effect Size
Physical Component Summary Score (PCS)	32.96	(12.17)	34.23	(11.97)	4.2***	0.11
Mental Component Summary Score (MCS)	45.97	(12.91)	46.22	(12.32)	0.8	0.02
Physical Functioning Scale	30.08	(14.43)	32.47	(13.89)	7.0***	0.17
Role-Physical Scale	36.88	(12.71)	36.16	(12.36)	-2.4*	0.06
Bodily Pain Scale	39.31	(12.14)	39.09	(12.21)	-0.7	0.02
General Health Scale	36.97	(12.38)	38.34	(12.02)	4.6***	0.11
Vitality Scale	39.85	(11.98)	41.86	(11.70)	7.0***	0.17
Social Functioning Scale	37.99	(15.17)	40.15	(14.05)	6.1***	0.15
Role-Emotional Scale	43.29	(13.85)	41.33	(13.53)	-5.8***	0.14
Mental Health Scale	43.77	(12.92)	44.84	(12.86)	3.4**	0.08

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 8: Chronic Conditions, Activities of Daily Living, Depression: Adjusted Data from Cohorts I, II, and III

Chronic Condition	Proxy-Completed		Self-Completed		χ^2 Value	Effect Size
Angina Pectoris ∫	669	(20.3%)	669	(19.4%)	0.8	0.02
Arthritis Hand/Wrist	1,353	(39.2%)	1,394	(40.4%)	1.0	0.02
Arthritis Hip/Knee	1,509	(43.7%)	1,590	(46.1%)	3.8*	0.05
Any Cancer ‡	495	(14.4%)	482	(14.0%)	0.2	0.01
Congestive Heart Failure	496	(14.4%)	508	(14.7%)	0.2	0.01
Crohn's Disease/ Ulcerative Colitis/Inflammatory Bowel Disease	238	(6.9%)	225	(6.5%)	0.4	0.02
Diabetes	805	(23.3%)	887	(25.7%)	5.3**	0.06
Emphysema/Asthma/COPD †	526	(15.3%)	607	(17.6%)	6.9**	0.06
Hypertension	2,007	(58.2%)	1,962	(56.9%)	1.2	0.03
Myocardial Infarction	501	(14.5%)	525	(15.2%)	0.7	0.02
Other heart conditions	883	(25.6%)	914	(26.5%)	0.7	0.02
Sciatica	846	(24.5%)	1,002	(29.0%)	18.0***	0.10
Stroke	757	(21.9%)	768	(22.3%)	0.1	0.01
Type of Impaired ADL						
Unable/Difficulty Using Toilet	1,079	(31.3%)	991	(28.7%)	5.3	0.06
Unable/Difficulty Eating	844	(24.5%)	753	(21.8%)	6.7**	0.06
Unable/Difficulty Bathing	1,679	(48.7%)	1,599	(46.4%)	3.7*	0.05
Unable/Difficulty Getting in or out of Chairs	1,820	(52.8%)	1,804	(52.3%)	0.1	0.01
Unable/Difficulty Dressing	1,527	(44.3%)	1,408	(40.8%)	8.4*	0.07
Unable/Difficulty Walking	2,177	(63.1%)	2,188	63.4%)	0.1	0.01
Depression						
Sad/Blue 2 Weeks in Past Year	1,497	(43.4%)	1,472	(42.7%)	0.4	0.01
Depressed /Sad Much of the Time in the Past Year	1,183	(34.3%)	1,197	(34.7%)	0.1	0.01
Depressed/Sad 2 or More Years in Life	952	(27.6%)	986	(28.6%)	0.8	0.02

* $p < .05$, ** $p < .01$, *** $p < .0001$ ∫ or coronary artery disease ‡ other than skin cancer † Chronic obstructive pulmonary disease

Conclusions

The results of this exploratory use of the propensity score method to compare proxy-completed and self-completed responses indicate that differences between the two samples were substantially reduced, although some differences remained after utilizing the propensity score methodology. We believe that three conclusions can be drawn from this research. First, the use of the propensity score method may be quite useful in reducing selection bias between self and proxy respondents in survey research. This methodology provides a unique tool and innovative approach for reducing this bias.

Second, though some differences between self and proxy-completed responses in this research remained after applying the propensity score methodology, the consistent use of this methodology in the literature should result in increased understanding regarding the differences between self and proxy-completed responses. For example, future research should consider the role of the proxy respondent to the self-respondent. Relatives and professional caregivers may systematically overstate or understate a respondent's physical and/or mental health status. Information on the nature of the role relationship between the proxy respondent and the self-respondent may be important to assess in health status surveys and may be a crucial factor in understanding selection bias.

While more research is needed on applying the propensity score method to self and proxy-completed responses, the use of this method in these populations can help researchers understand the differences in self and proxy-completed responses, and to reduce these differences.

Finally, the propensity score method should be examined in the context of the literature on cognitive psychology. Response bias is a phenomenon entirely consistent with the social and cognitive psychological literature on attributional biases. Overall, the findings from dozens of empirical studies indicate that humans are relatively poor processors of information and form biases and inferences that systematically distort perception [26]. Using the National Health Interview Survey on Disability, recent research indicates that conditional likelihood judgments (for example, the likelihood that an individual has a disability given another disability) predicted the number of disabilities for proxy-completed responses but not for self-completed responses [27]. The continuing search for methods to understand how to reduce proxy bias in quality of life research is important since the implications for policy direction may depend on such research.

Authors' Contributions

BHE participated in the interpretation of statistical analyses and co-wrote the manuscript.

WMB conceptualized applying the propensity score method to the proxy and self-respondent Medicare HOS data, wrote the majority of the SAS code, conducted much of the statistical analyses, participated in the interpretation of the statistical analyses, and co-wrote the manuscript.

JKC participated in the statistical analyses and interpretation.

BMF wrote portions of the SAS code for analyses.

EDS wrote portions of SAS code, participated in the interpretation of statistical analyses, and edited the manuscript.

DD participated in the interpretation of statistical analyses and offered valuable critical thought on previous versions of this manuscript.

RWA secured funding.

LAG secured funding.

All authors read and approved the final manuscript.

Acknowledgments

The authors acknowledge Samuel C. Haffer, PhD, Patricia Wright-Gaines, and Sonya Bowen, MSW of the Centers for Medicare & Medicaid for their support of the Medicare Health Outcomes Survey and research emanating from the survey. The authors thank Wendy A. Richard, MA for proofing, and also acknowledge Susan Grace, BSN; Max Johnson, MPH; Efthimios Laios, MPH; Barbara Mayl, RN; Rajesh Shrestha, MPH for data cleaning and dissemination of reports to health plans.

References

1. Cooper JK, Kohlmann T, Michael JA, Haffer SC and Stevic M: **Health outcomes: new quality measure for Medicare.** *Int J Qual Health Care* 2001, **13:**9-16.
2. D'Agostino RB: **Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group.** *Stat Med* 1998, **17:**2265-2281.
3. Rubin DB: **Matching to remove bias in observational studies.** *Biometrics* 1973, **29:**159-183.
4. Neumann PJ, Araki SS and Gutterman EM: **The use of proxy-completed responses in studies of older adults: lessons, challenges, and opportunities.** *J Am Geriatr Soc* 2000, **48:**1646-1654.
5. Dorman PJ, Waddel F, Slattery J, Dennis M and Sandercock P: **Are proxy assessments of health status after stroke with the EuroQol questionnaire feasible, accurate and unbiased?** *Stroke* 1997, **28:**1883-1887.
6. Yip JY, Wilber KH, Myrtle RC and Grazman DN: **Comparison of older adult subject and proxy responses on the SF-36 health-related quality of life instrument.** *Aging Ment Health* 2001, **5:**136-142.
7. Magaziner J, Zimmerman SI, Gruber-Baldini AL, Hebel JR and Fox KM: **Proxy reporting in five areas of functional status: comparison with self reports and observations of performance.** *Am J Epidemiol* 1997, **146:**418-428.
8. Tordorov A and Kirchner C: **Bias in proxies' reports of disability: data from the national health interview survey on disability.** *Am J Public Health* 2000, **90:**1248-1253.
9. Pierre U, Wood-Dauphinee S, Korner-Bitensky N, Gayton D and Hanley J: **Proxy use of the Canadian SF-36 in rating health status of the disabled elderly.** *J Clin Epidemiol* 1998, **51:**983-990.
10. Shields M: **Proxy reporting in the national population health survey.** *Health Rep* 2000, **12:**21-39.
11. Rosenbaum PR and Rubin DB: **The central role of the propensity score in observational studies for causal effects.** *Biometrika* 1983, **70:**41-55.
12. Rubin DB and Thomas N: **Characterizing the effect of matching using linear propensity score methods with normal distributions.** *Biometrika* 1992, **79:**797-809.
13. Rubin DB and Thomas N: **Combining propensity score matching with additional adjustments for prognostic covariates.** *J Am Stat Assoc* 2000, **95:**573-585.
14. Kilborn MJ, Rathore SS, Gersh BJ, Oetgen WJ and Solomon AJ: **Amiodarone and mortality among elderly patients with acute myocardial infarction with atrial fibrillation.** *Am Heart J* 2002, **144:**1095-1101.

15. Teufelsbauer H, Prusa AM, Wolff K, Polterauer P, Nanobashvili J, Prager M, Holzenbein T, Thurnher S, Lammer J, Schemper M, Kretschmer G and Huk I: **Endovascular stent grafting versus open surgical operation in patients with infrarenal aortic aneurysms: a propensity score-adjusted analysis.** *Circulation* 2002, **106:**782-787.
16. Neugut AI, Fleischauer AT, Sundararajan V, Mitra N, Heitjan DF, Jacobson JS and Grann VR: **Use of adjuvant chemotherapy and radiation therapy for rectal cancer among the elderly: a population-based study.** *J Clin Oncol* 2002, **20:**2643-2650.
17. Mehta RL, McDonald B, Gabbai F, Pahl M, Farkas A, Pascual MTA, Zhuang S, Kaplan RM and Chertow GM: **Nephrology consultation in acute renal failure: does timing matter?** *Am J Med* 2002, **113:**456-528.
18. Drake C: **Effects of misspecification of the propensity score on estimators of treatment effect.** *Biometrics* 1993, **49:**1231-1236.
19. Gu XS and Rosenbaum PR: **Comparison of multivariate matching methods: structures, distances, and algorithms.** *J Comp Graph Stat* 1993, **2:**405-420.
20. Dehejia RF and Wahba S: **Causal effects in nonexperimental studies: reevaluating the evaluation of training programs.** *J Am Stat Assoc* 1999, **94:**1053-1062.
21. **Medicare Health Outcomes Survey** [http://cms.hhs.gov/surveys/hos]
22. National Committee for Quality Assurance: **HEDIS® 3.0 Volume 6 *Health of Seniors Survey Manual.*** *Washington DC* 1998.
23. Ware JE, Snow KK, Kosinski M and Gandek B: *SF-36® Health Status Survey Manual and Interpretation Guide New Haven: The Health Institute, New England Medical Center*; 1993.
24. Ware JE and Kosinski M: *SF-36® Physical and Mental Health Summary Scales: A Manual for Users of Version 1* Secondth edition. *Lincoln: QualityMetric, Inc*; 2001.
25. Cohen J: *Statistical power analysis for the behavioral sciences Hillsdale: Lawrence Erlbaum*; 1988.
26. Markus H and Zajonc RB: **The cognitive perspective in social psychology.** In *The handbook of social psychology* 3rd edition. Edited by: *Lindzey G, Aronson E. Hillsdale: Erlbaum*; 1985:137-230.
27. Todorov A: **Cognitive procedures for correcting proxy-response bias in surveys.** *Appl Cog Psychol* 2002, **17:**215-224.