Estimation of Disease-specific Health Care Costs Using Causal Inference Framework

Irina Bondarenko[1], Trivellore Raghunathan[2]

[1]Senior Statistician, Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109
[2]Chair and Professor, Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109

Abstract

Health care costs have been increasing faster than National Income per Capita during the last decade, resulting in heavy burden on patients with increasing payment for services and on taxpayers, who often fill the gap. In addition to this upward trend in Health Care Expenditures, prevalence rates of medical conditions and chronic diseases are also increasing posing a question of efficiency of health care. Partitioning of National Health Expenditures into shares attributable to various conditions is essential to understand the structure of the Health Care Expenditure and to develop strategies to improve the efficiency of health care system. One useful strategy is to estimate increases in cost for specific diseases utilizing the causal inference framework developed for observational study settings. We compare four methods--propensity score stratification, and three approaches based on multiple imputations of counterfactuals, for estimating disease-specific costs. For multiple imputation inference, we use three approaches: (1) Parametric approach using log-normal distribution; (2) Tukey's gh-distribution(GH) on the original scale; and (3) Approximate Bayesian Bootstrap (ABB). Data from the Medicare Current Beneficiary Survey (MCBS) is used to illustrate the methodology. We also evaluate the repeated sampling properties of the estimates through a simulation study.

Key Words: Cost attribution, Counterfactual, Potential outcomes, Semi-parametric imputations, Tukey GH-distributions, Approximate Bayesian Bootstrap

1. Introduction

1.1. Cost attribution

Over the last two decades, Personal Health Care Expenditures per Capita increased almost three fold from $2430 in 1990 to over $7000 in 2010 [1]. Though, the rate of growth in recent years has slowed relative to the late 1990s, it still yielded 72% from 2000 to 2010 climbing faster than national income per capita which climbed 33% over the same decade. What health conditions are linked to the increase in health care cost? Various attempts have been made to understand the reasons for the escalating cost of health care . Thorpe at el [2] showed how being overweight or obese contributed to both the increase in prevalence of many diseases and the cost of treatment. Cutler [3] and Newhouse [4] estimated percent increase in healthcare spending due to new medical technologies. Does cost attributed to chronic conditions increase proportionately to the cost attributed to the severe conditions that frequently require expensive treatment? It's crucial to attribute the cost increase to a specific factor or group of conditions. However, it's even more important to analyze the increase in cost of treatment for a specific disease

against the benefits of the new treatment [5]. Do innovations in treatment or screening technologies result in quality of life improvement and/or increases in longevity of patients?

The goal of cost attribution analysis, therefore, is to estimate a portion of health expenditure that is due to a specific disease or medical condition. A need for this analysis comes from an effort to understand trends and changes in health cost structure over the years. Developing comprehensive methods that would allow economists to attribute health care dollars to the specific diseases will lead to a better understanding of causes behind inflation of health spending, and subsequently assess the effectiveness of the health care system.

1.2 Motivating Example

Metabolic syndrome (MetS) also known as Syndrome X is a name for a group of risk factors that occur together and increase the risk of Cardiovascular Diseases. We selected Metabolic Syndrome (MetS) to study methods for cost attribution because of the association with risk of mortality [6], and declining health status in elderly [7]. We defined the metabolic syndrome variable MetS as a number of the conditions that a subject has among the four major risk factors, obesity, hypertension, diabetes, and hypercholesterolemia. The risk factors were determined based on self-report and a review of Medicare claims. We examined prevalence of metabolic syndrome for the USA population, age 65 and older through the analysis of data from the Medicare Current Beneficiary Survey (MCBS) for the years 1999-2005. Subjects who were institutionalized, or enrolled in HMO were excluded.

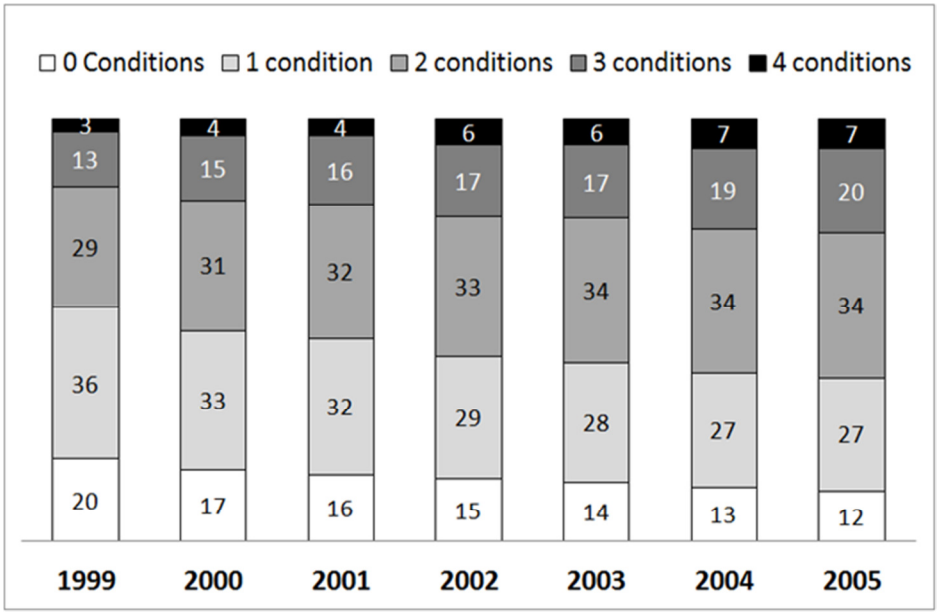

Figure1. Prevalence rates (%) of Metabolic syndrome by the number of condition in non-institutionalized Medicare population

Figure 1 shows prevalence rates of number of MetS conditions from 1999 and 2005. The prevalence rate of 0 and 1 conditions decreased from 20% to 12%, and 36% to 27% respectively. The prevalence rate of 2 conditions increased from 29% to 34%, and prevalence rates of 3 and 4 conditions jumped from 13% to 20% and 3% to 7% respectively.

Figure 2 provides estimates of average healthcare costs for subjects stratified by the number of MetS conditions. Average cost of care increased over the years for all groups. There was 25% increase in healthcare costs between 1999 and 2005 for subjects with one, two and three conditions; almost 50% increase for subjects with four conditions. Healthcare costs increased 15% for subjects with no sign of metabolic syndrome.

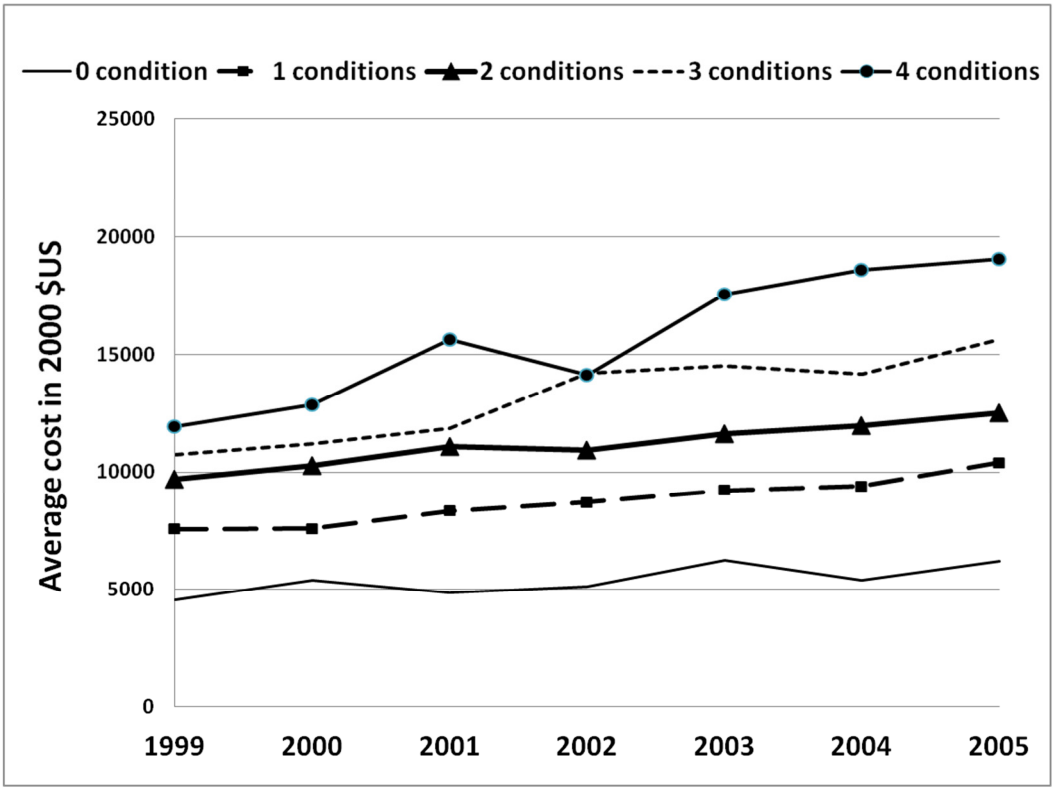

Figure 2. Average health care costs in 2000 $US by number of MetS conditions.

We conclude that a causal analysis to compare costs for subjects with one, two, three and four MetS conditions with cost for similar subjects with no conditions will help in dissecting the increasing costs more accurately.

2. Methods.

2.1 Causal inference framework for Cost Attribution.

We frame the cost attribution estimation problem as causal inference in an observational study settings. Suppose, we want to estimate average cost attributed to health condition D controlling for other variables X. Then average causal effect (ACE) is equal $\Delta_D = E(Cost(D=1)|X) - E(cost(D=0)|X)$.

Established methods for Causal Inference in observational study settings are rooted in matching on the estimated propensity of being a case, or in the context of this paper probability of having a specific number of conditions D (D=1) with subjects with no conditions(D=0). Propensity scores can be easily estimated from the observed data utilizing logistic regression models, or decision trees. Cochran [8] showed that under a broad range of conditions, stratification into k=4, 5, 6 groups removes correspondingly 86% , 90%, 92% of bias yielded by crude difference in expected values of outcomes

between cases and controls. Rubin [9] proved effectiveness of sub-classification on propensity scores to remove bias.

1. The stratification approach provides an easy to implement four-step algorithm coined to attribute the health care costs to a particular conditions. Estimate probability of having condition *D*, conditional on available covariates. $\hat{p}=P(D=1|X)$.
2. Stratify data into *k* groups depending on estimated value of $\hat{p}$.
3. Within each group estimate the difference in cost between cases *(D=1)* and controls *(D=0)*.
4. Combine estimates across the strata weighting by the population size for each stratum.

Cochran [8] argued that percent reduction in bias due to propensity scores stratification depends upon the number of strata, choice of the cut off, similarity in mathematical form of regression relationship between outcome and variable used in stratification, and the overlap of distribution of this variable between cases and controls. As standardised difference in means of propensity scores increases, bias introduced by crude estimation grows and amplifies value of bias remaining after stratification. However, in observational study setting it's not uncommon for propensity scores to differ between cases and controls by more than one standard deviation.

An alternative approach, especially when the conditions laid out in Cochran[8] are not satisfied, is to multiply impute potential outcomes. There is a number of papers that draw causal inference by imputing missing values of potential outcomes. Schafer and Kang[10] provided a comprehensive review of causal inference methods in an observational study and list Multiple Imputation as an appealing solution. Dominici et al [11] proposed methods for semi-parametric imputation of counterfactuals in semi-parametric framework based on percentiles of the distributions. Other example of imputation of potential outcomes can be found in papers by Elliot, Raghunathan & Li [12], Bondarenko and Raghunathan [13].

In this article we investigate a flexible framework for multiple imputation of potential outcomes conditional on the propensity scores using parametric or non-parametric approaches that incorporates the following steps:

1. Model values of cost as a linear or non-linear function of the $\hat{p}$ separately for subjects with disease *(D =1)* and without disease *(D=0)*.
2. Impute the residuals conditional on $\hat{p}$ or $logit(\hat{p})$. Other covariates with strong prognostic score can be added if desired.
3. Combine values of residuals and predicted values to yield *cost1* and *cost0*.
4. Estimate average causal effect ACE as $\Delta_D=E(cost1-Cost0)$. Alternatively, average causal effect for treated (ACT) can be estimated on the subset of cases $\Delta'_D=E(cost1-Cost0|D=1)$.

2.2 Choice of Imputation framework

Generally speaking, the framework for imputation can be chosen based on the empirical investigation of the relationship between cost and propensity scores for cases and controls. We consider three multiple imputation approaches: parametric; semi-parametric; non-parametric. The first approach uses the Sequential Regression Multiple imputation described by Raghunathan at el [14] assuming log-normal distribution for costs which seems to be reasonable for Medicare data but the drawback is the reliance on this parametric assumption. The benefit on the other hand is easy implementations using existing software packages in R, SAS, STATA etc.

The second approach uses the Tukey's *gh* distribution as described in He and Raghunathan [15,16]. This method based on a class of G-and-H distributions proposed by

Tukey [17]. Distributions from this family are certain transformations of the standard normal variable that accommodate different skewness and elongation of tails. Let Z denote a standard normal variable then variable y is defined by the following transformation of Z

$$Y_{gh}(Z) = \mu + \sigma \frac{e^{gZ} - 1}{g} e^{hZ^2/2},$$

Where μ is location, σ is scale, g specifies skeweness and h governs elongation of tails.
Maximum likelihood estimates of these parameters are numerically intensive. On the other hand empirical quintiles can be easily estimated by method proposed by Hoaglin [18].

The third and final approach uses Approximate Bayesian Bootstrap described by Rubin and Schenker [19].

3. Application to motivating example.

We estimated the attributable cost for a specific number of MetS conditions using all four methods described in the paper. Missing values in the covariates were multiply imputed using a sequential regression framework as implemented in IVEWARE package. Due to the small sample size, we combined subjects with three and four conditions.

In each imputed dataset probability of having k conditions relative to having none was estimated using a logistic regression model.

$\hat{e}_k = logit(P(MetS=k \text{ vs } 0 \mid X) = X'\beta$. The list of covariates included age, gender, demographic and SES variables, insurance (private vs. Medicare only), self-reported cancer, arthritis, osteoporosis, hip fracture, self-reported health status, and estimated probability of death in the year of interview. We examined standardized difference in $\hat{e}_k$ for cases and controls, expressed as Cohen's D: $(\hat{e}k(D=1) - \hat{e}k(D=0))/s$, where s is standardized deviation for the data. Table 3 shows Cohen's D by the number of conditions for each year. For the combined 3+ counts of metabolic syndrome Cohen's D exceeded 1 standard deviation for all years. Difference in functional relationship between cost and propensity scores was examined and found to be was substantial.

Next, we estimated increase in health costs between subjects with no metabolic syndrome and one, two, and three or more conditions. For stratified analysis we grouped subjects into 5 strata based on the estimated value of the êk. For imputation methods, the residuals were estimated by regressing cost for cases and cost of controls on êk. For all parametric, GH, and ABB imputations we imputed five datasets for each previously imputed MCBS data, resulting in 25 imputed data sets. MI estimates and corresponding standard errors were produced according to MI combing rules. Results are presented in table 4 and are in 2000 US dollars.

Table 3. Cohen's D by year and number of conditions

MetS Conditions	1999	2000	2001	2002	2003	2004	2005
1	0.5	0.5	0.6	0.6	0.5	0.6	0.6
2	0.7	0.7	0.9	0.9	0.8	0.9	0.8
3+	1.1	1.1	1.1	1.1	1.1	1.1	1.1

Table 4. Estimated attribute cost (SE) by counts of Metabolic syndrome and year

Year	*MetS Conditions*	*Stratified Analyses*	*SRMI Imputations*	*ABB Imputations*	*GH Imputations*
1999	***1***	*1834 (563)*	*1828 (324)*	*1488 (432)*	*1689 (323)*
	2	*3596 (771)*	*3564 (337)*	*3283 (491)*	*3381 (398)*
	3+	*3867 (880)*	*4024 (458)*	*3592 (604)*	*3825 (602)*
2000	***1***	*452 (857)*	*1038 (315)*	*189 (1101)*	*534 (426)*
	2	*2412 (979)*	*3024 (365)*	*2366 (670)*	*2645 (540)*
	3+	*3237 (1221)*	*3803 (502)*	*3272 (1207)*	*3500 (541)*
2001	***1***	*1681 (650)*	*1779 (356)*	*1386 (429)*	*1657 (346)*
	2	*3853 (863)*	*3905 (397)*	*3400 (490)*	*3508 (427)*
	3+	*4806 (848)*	*4245 (452)*	*4124 (888)*	*4408 (555)*
2002	***1***	*1895 (621)*	*1765 (330)*	*1370 (513)*	*1576 (432)*
	2	*3972 (741)*	*3725 (403)*	*3260 (429)*	*3490 (435)*
	3+	*5948 (1083)*	*5640 (518)*	*5189 (826)*	*5684 (613)*
2003	***1***	*1215 (793)*	*1467 (400)*	*1270 (523)*	*1505 (485)*
	2	*2936 (779)*	*3230 (442)*	*2828 (654)*	*3287 (432)*
	3+	*4304 (1115)*	*5484 (625)*	*4455 (799)*	*4626 (666)*
2004	***1***	*2273 (731)*	*2432 (386)*	*1846 (766)*	*2225 (449)*
	2	*4092 (781)*	*4355 (390)*	*3522 (815)*	*3974 (483)*
	3+	*6740 (913)*	*5819 (477)*	*5984 (1379)*	*6429 (557)*
2005	***1***	*2480 (821)*	*2398 (463)*	*2113 (597)*	*2258 (557)*
	2	*3780 (764)*	*3716 (531)*	*3236 (576)*	*3604 (519)*
	3+	*6554 (1149)*	*6193 (623)*	*6144 (755)*	*6297 (657)*

Based on all four methods applied, cost attributed to the one and two conditions of metabolic syndrome do not show an upward trend over the span of seven years of data analyzed here. However, increase in healthcare cost attributed to three and more conditions climbed noticeable over the same years increasing by more than 2000 dollars per subject between 1999 and 2005. Trends across the years for one, two, and three or more conditions of the syndrome are depicted in Figure 3.

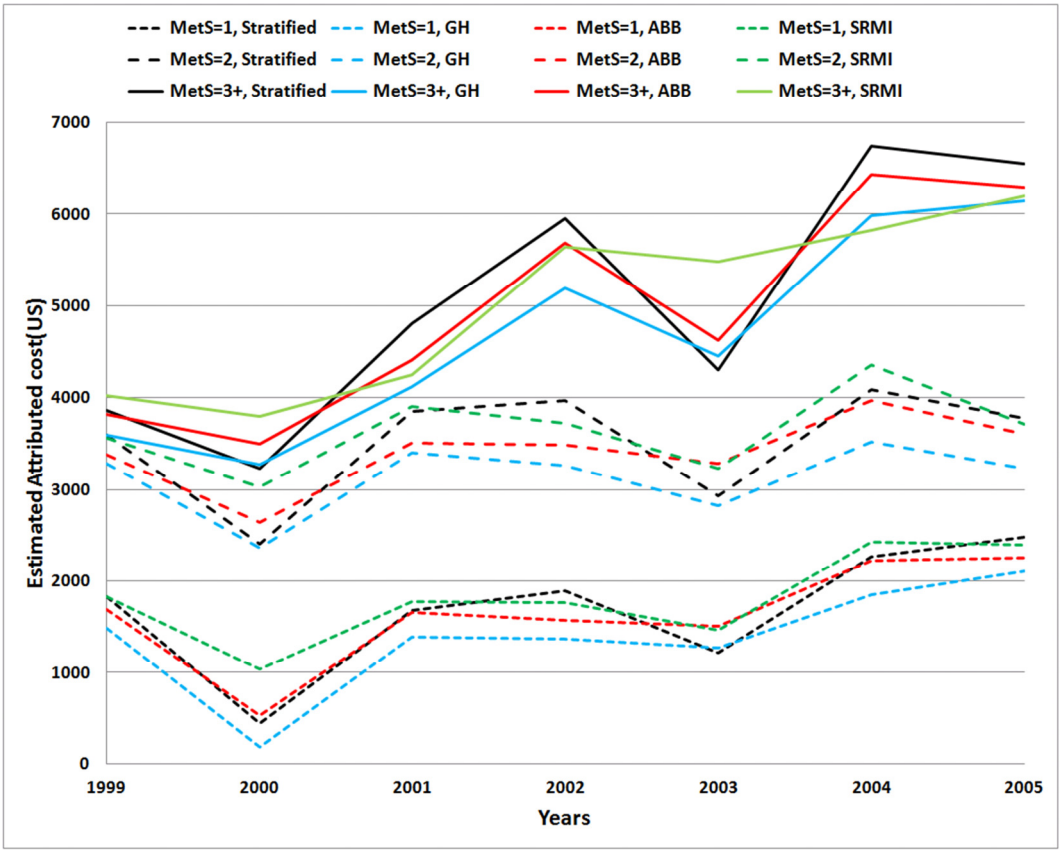

Figure 3. Estimates of cost attributed to metabolic syndrome by the number of conditions yielded by four proposed methods.

4. Simulation Study

To study performance of the proposed methods we conducted a simulation study. We drew 200 samples of size *n=1000* from a joint distribution of four variables *(y0, y1, x1, x2)*. Variables *y1*, and *y0* were designed to represent two potential outcomes. The other two variables *x1* and *x2* play roles of covariates.

Next, we introduced a treatment assignment mechanism *D={0,1}*, defined by the logistic model $e=logit(P(D=1))=b0+b1*x1+b2*x2$. Next, we combined values of *y1* and *y0* into 'observed' outcome y, according to variable *D*. We considered different scenarios for data generation as well as for treatment assignment mechanism.

Scenarios used in data generation process included:

1. Two distributional families for potential outcomes – normal and gamma.
2. We explored a variety of correlation matrices. We limited results presented here to a case when correlation between potential outcomes is equal 0.5.

The treatment assignment mechanism incorporated the following features:

1. P(D=1)=0.5.
2. Absolute values of Cohen's D between $\hat{e}=logit(P(D=1|X))$ for *D=1* and *D=0* were fixed at 0.5, 1 and 1.3 to emulate the MCBS data example.
3. We varied correlation between $\hat{e}=logit(P(D=1|X))$ and observed *y*. The results are shown for correlation $\rho(y,\hat{e})$ *being equal to 0.15, 0.3 and 0.5* correlations.

Our goal was to estimate effect of *D* on *y* given *x1* and *x2*. We report five estimates.

1. CRUDE is unadjusted difference in expected values of *y* $E(y(D=1)-y(D=0))$;
2. SM is estimated by stratification into 5 equal size groups based on $\hat{e}$.

Imputation based methods included:

3. SRMI is parametric imputations conditional on $\hat{e}$.
4. GH stands for gh-imputations conditional on ê.
5. ABB abbreviates Approximate Bayesian Bootstrap Multiple Imputation conditional on $\hat{e}$.

We studied bias reduction properties for each of these methods. Figures 4 and 5 show percent bias for crude, standard and imputation-based methods for normally and gamma distributed outcomes, respectively. As expected, for both distributional families crude estimates were severely biased. As standardized difference between distributions of $\hat{e}$ *for D=1 and D=0*, measured by Cohen's D widens, bias increases. Magnitude of correlation between $\hat{e}$ and y affects amount of bias. Direction of bias is determined by the sign of Cohen's D.

In agreement with Cochran's studies of bias reduction, stratification into quintiles of the propensity scores removes around 90% of bias for normally distributed outcomes, and almost 85% for gamma distributed outcomes. As Figure 4 and 5 show, that the residual bias was negligible for Cohen's D equal 0.5. However, the residual bias in estimates based on stratification methods were 10% for normal and 12% for gamma distribution for Cohen's D equal 1.3 and strong correlation between y and e .

For normally distributed outcomes the imputation-based methods performed equally well, yielding less than 1% of bias and were not affected by Cohen's D or amount of correlation between outcome and the propensity scores.

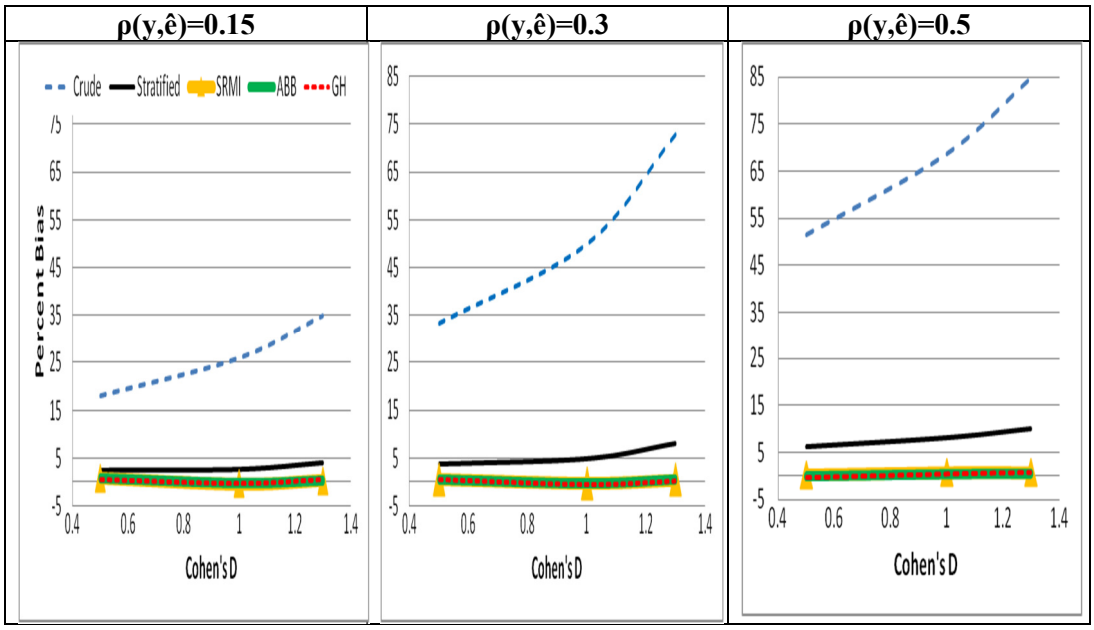

Figure 4. Percent bias for *normally*-distributed outcomes as a function of correlation and Cohen's D.

For gamma distributed outcomes ABB and GH imputations removed 99% of bias. For the SRMI method gamma-distributed outcomes were transformed to satisfy residual assumptions implied by parametric imputation model. For this purpose, we explored two transformations for outcomes: log-transformation and cubic-root transformation. Residuals were imputed on altered scale and then back-transformed and combined with predicted values of *y1* and *y0*. SRMI of log-transformed outcome shows substantial bias. Whereas, the cubic-root transformation produced results comparable to

ABB and GH imputations. We conclude that in general 1) imputation methods remove bias more efficiently than stratification on propensity scores, 2) in absence of ideal transformation to achieve normality of the residuals GH and ABB imputations perform better than SRMI.

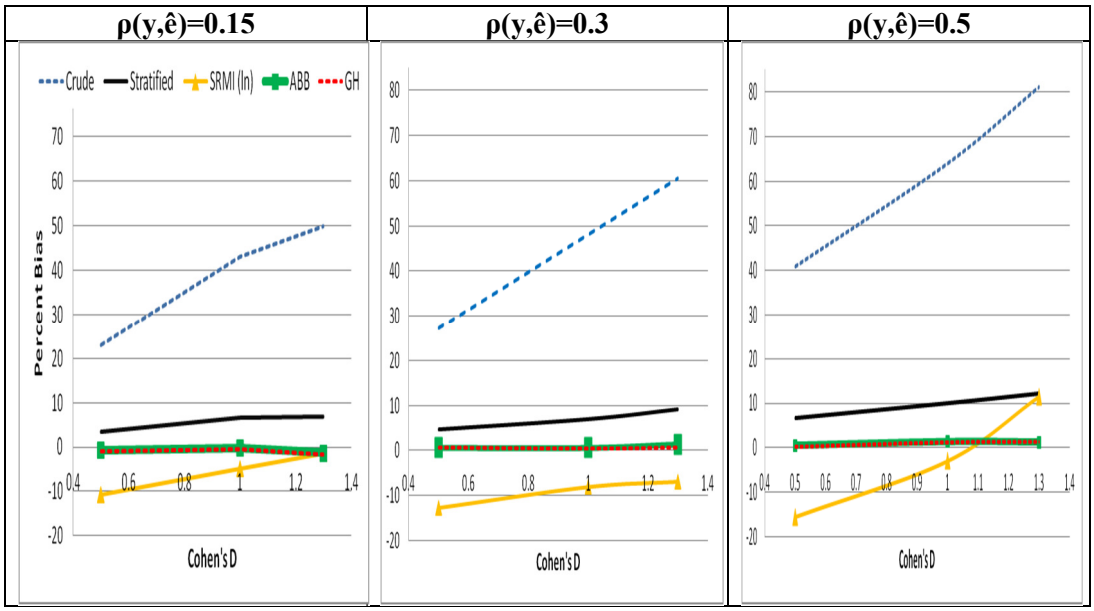

Figure 5. Percent bias for gamma-distributed outcomes as a function of correlation and Cohen's D.

5. Conclusions

We have explored methods estimation methods for disease-specific health care costs to understand factors behind increase in health care spending. We developed four methods and illustrated them using the MCBS data. This study was motivated by an example of estimating the cost associated with metabolic syndrome. The analysis of data for the years 1999 to 2005 show considerable increase in costs associated with subjects who have three or more of the conditions among Obesity, Hypertension, Diabetes and Hypercholesterolemia. Increases in both prevalence rates of these conditions and costs per subject are useful summaries to analyze efficacy of the health care system.

We also evaluated proposed methods using a simulation study. Since in observational studies, the extent of differences between the covariates for cases and controls can be large even after propensity score stratification, imputation based methods may be better in terms of reducing the residual bias. The simulation study also shows that a carefully selected parametric model may be useful, but semi-parametric or non-parametric approaches may be more suitable for routine applications in estimating the attributable costs.

Acknowledgments

This research was supported by National Institute of Aging (NIA) award AG031098, "Expanding the National Health Accounts (NHA)".

References

1. Centers for Medicare and Medicaid Services, Office of the Actuary, National Health Statistics Group, National Health Care Expenditures Data, January 2012.

2. Thorpe KE, Florence CS, Howard DH, Joski P. (2004) "Trends: The impact of Obesity on Rising Medical Spending". Health Affairs, web Exclusive, October 20, 2004.
3. Cutler DM. "Technology, Health Costs, and NIH" (1995). Paper prepared for the national Institutes of Health Economic Roundtable on Biomedical Research.
4. Newhouse JP. (1992) "Medical care Costs: How Much Welfare Loss" Journal of economic Perspectives, vol 6, no. 3.
5. Cutler DM, McClellan M, (2001)"Is Technological Change In Medicine Worth It?" Health Affairs, vol. 20 no. 5. pp. 11-29.
6. Tasnime N. Akbaraly, Mika Kivimaki,a Marie-Laure Ancelin,a Pascale Barberger-Gateau, Thibault Mura, Christophe Tzourio, Jacques Touchon, Karen Ritchie, and Claudine Berr. (2010) "Metabolic Syndrome, Its Components, and Mortality in the Elderly" J Clin Endocrinol Metab. November; 95(11): E327–E332.
7. Otto Lindberg, Reijo S. Tilvis, Timo E. Strandberg, Jaakko Valvanne, Sirpa Sairanen, Christian Ehnholm and Jaakko Tuomilehto (1997) "Impacts of Components of the Metabolic Syndrome on Health Status and Survival in an Aged Population" European Journal of Epidemiology , Vol. 13, No. 4 , pp. 429-434.
8. Cochran W. G. (1968) "The effectiveness of adjustment by subclassification in removing bias in observational studies." Biometrics, 24, pp. 295–313.
9. Paul R. Rosenbaum and Donald B. Rubin (1984) "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score" Journal of the American Statistical Association , Vol. 79, No. 387, pp. 516-524.
10. Schafer JL, Kang J. (2008) "Average causal effects from nonrandomized studies: a practical guide and simulated example." Psychol Methods;13(4), pp.279-313.
11. Dominici F, Zeger SL, Parmigiani G, Katz J, Christian P (2006) "Estimating percentile-specific treatment effects in counterfactual models: a case-study of micronutrient supplementation, birth weight and infant mortality." Applied Statistics 55(2), pp.261-280.
12. Elliott MR, TE Raghunathan TE, Li Y(2010) " Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes" Biostatistics 11 (2), pp. 353-372.
13. Bondarenko I, Raghunathan TE "Multiple Imputation for Causal Inference" (2010) JSM proceedings, Section on Survey Research Methods, pp. 3934-3944.
14. Raghunathan TE, Lepkowski JM, VanHoewyk J, Solenberger P.(2001) "A multivariate technique for multiply imputing missing values using a sequence of regression models." Survey Methodol. 27, pp. 85–95.
15. He, Y., and Raghunathan T. (2006). "Tukey's gh distribution for multiple imputation." American Statistician, 60(3): pp. 251-256.
16. He Y., Raghunathan T.(2012) "Multiple imputation using multivariate gh transformations", Journal of Applied Statistics 39 (10), pp. 2177-2198.
17. Tukey, J.W. (1977) "Modern Techniques in Data Analysis", NSF-sponsored regional research conference at Southeastern Massachusetts University, North Dartmouth, MA..
18. Hoaglin, D.C. (1985), "Summarizing Shape Numerically: The g-and-h Distributions" in Exploring Data Tables, Trends, and Shapes,eds D.C. Hoaglin, F. Mosteller, and J.W. Tukey, New York: Willey, pp.461-513.
19. Rubin, D., Schenker N., (1986) "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse." Journal of the American Statistical Association, 81, pp. 366-374.