Published in final edited form as:

J Manipulative Physiol Ther. 2015 ; 38(9): 620–628. doi:10.1016/j.jmpt.2015.10.005.

Comparing propensity score methods for creating comparable cohorts of chiropractic users and non-users in older, multiply comorbid Medicare patients with chronic low back pain

William B Weeks, MD, PhD, MBA [Professor] [Director],

The Geisel School of Medicine at Dartmouth; Health Services and Clinical Research, Palmer College of Chiropractic; The Dartmouth Institute for Health Policy and Clinical Practice Palmer Center for Chiropractic Research

Tor D Tosteson, ScD [Professor],

The Geisel School of Medicine at Dartmouth; The Dartmouth Institute for Health Policy and Clinical Practice

James M Whedon, DC, MS [Instructor] [Director],

The Geisel School of Medicine at Dartmouth; Health Services Research, Southern California University of Health Sciences, Whittier, CA; The Dartmouth Institute for Health Policy and Clinical Practice

Brent Leininger, DC [Research Fellow],

Integrative Health & Wellbeing Research Program, Center for Spirituality & Healing, University of Minnesota, Minneapolis, MN

Jon D Lurie, MD, MS [Associate Professor],

CORRESPONDING AUTHOR CONTACT INFORMATION, William B Weeks, MD, PhD, MBA, wbw@dartmouth.edu TDI, 35 Centerra Parkway, Lebanon, NH 03766, Phone number 603 208 8963.

CONFLICTS OF INTEREST

No conflicts of interest were reported for this study.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Contributorship

Concept development (provided idea for the research)

Wbw, tdt, jmw, bl, jdl, rs, cmg

Design (planned the methods to generate the results)

Wbw, tdt, jmw, bl, jdl, rs, cmg, ajo

Supervision (provided oversight, responsible for organization and implementation, writing of the manuscript)

wbw

Data collection/processing (responsible for experiments, patient management, organization, or reporting data)

wbw

Analysis/interpretation (responsible for statistical analysis, evaluation, and presentation of the results)

Wbw, tdt, jmw, bl, jdl, rs, cmg, ajo

Literature search (performed the literature search)

Wbw, tdt, jmw, bl, jdl, rs, cmg, ajo

Writing (responsible for writing a substantive part of the manuscript)

wbw

Critical review (revised manuscript for intellectual content, this does not relate to spelling and grammar checking)

Wbw, tdt, jmw, bl, jdl, rs, cmg, ajo

Other (list other specific novel contributions)

The Geisel School of Medicine at Dartmouth; The Dartmouth Institute for Health Policy and Clinical Practice

Rand Swenson, DC, MD, PhD [Professor],

The Geisel School of Medicine at Dartmouth; The Dartmouth Institute for Health Policy and Clinical Practice

Christine M Goertz, PhD, DC [Vice Chancellor], and

Palmer College of Chiropractic, Davenport, IA; Palmer Center for Chiropractic Research

Alistair J O'Malley, PhD [Professor]

The Geisel School of Medicine at Dartmouth; Department of Biomedical Data Science The Dartmouth Institute for Health Policy and Clinical Practice

Abstract

Objective—Patients who use complementary and integrative health services like chiropractic manipulative treatment (CMT) often have different characteristics than patients who do not, and these differences can confound attempts to compare outcomes across treatment groups, particularly in observational studies when selection bias may occur. The purposes of this study were to provide an overview on how propensity scoring methods can be used address selection bias by balancing treatment groups on key variables and to use Medicare data to compare different methods for doing so.

Methods—We described 2 propensity score methods (matching and weighting). Then we used Medicare data from 2006–2012 on older, multiply comorbid patients who had a chronic low back pain episode to demonstrate the impact of applying methods on the balance of demographics of patients between 2 treatment groups (those who received only CMT and those who received no CMT during their episodes).

Results—Before application of propensity score methods, patients who used only CMT had different characteristics from those who did not. Propensity score matching diminished observed differences across the treatment groups at the expense of reduced sample size. However, propensity score weighting achieved balance in patient characteristics between the groups and allowed us to keep the entire sample.

Conclusions—While propensity score matching and weighting have similar effects in terms of balancing covariates, weighting has the advantage of maintaining sample size, preserving external validity, and generalizing more naturally to comparisons of 3 or more treatment groups. Researchers should carefully consider which propensity score method to use, as using different methods can generate different results.

Keywords

Chiropractic manipulation; Medicare; Propensity Score

Introduction

Health Services research methods, such as analysis of large claims databases, are an important tool for research on effectiveness and efficiency of complementary and integrative

health services (CIHS) such as spinal manipulative therapy (SMT).¹ These methods may be particularly useful when determining whether patients who use such care for certain conditions might have lower overall care costs when compared to patients who do not,² a so called ‘medical care cost offset’ that has been demonstrated, for instance, with pharmaceutical^{3,4} and mental health^{5,6} treatment.

The very high variability of healthcare expenditures for individuals⁷ might distort cost-offset analysis when sample sizes are small, as they frequently are in randomized controlled trials (RCTs) that examine the effectiveness of SMT. As a recent report on RCTs that studied the clinical effectiveness of SMT for neck and low back pain found a median sample size of 95, and an interquartile range of 47-199,⁸ research efforts searching for medical care cost offsets might need to use large databases or nationally weighted surveys⁹ to conduct observational comparative effectiveness studies with adequate sample sizes.

However, the use of observational data from large datasets or surveys to compare groups that do or do not use a particular treatment modality may be confounded by selection bias.^{10,11} For instance, studies that have compared users to non-users of a common type of SMT that is provided by doctors of chiropractic (DCs) (known as chiropractic manipulative treatment (CMT)) show that CMT users are younger, wealthier, and healthier than non-users and more likely to be insured and female.¹²⁻¹⁴ These demographic characteristics have been associated with different health and cost outcomes when comparing CMT to medical care for treatment of low back pain.¹⁵ Therefore, it is critical to recognize potential selection bias when attempting to compare treatment groups in observational studies of patients who use CMT.

While traditional risk-adjustment for demographic differences in patient populations through risk stratification and regression adjustment have narrowed cost differences between patients who obtained CMT and those who sought conventional medical care for back pain,¹⁶ newer statistical methods that calculate the propensity of patients to self-select into 1 treatment group or another, based on a variety of demographic and utilization variables, and use those calculations for further analysis, are increasingly used in light of their improved performance in estimating causal effects.^{17,18}

To date, several studies have applied propensity score methods to analysis of survey data in an effort to compare CIHS users to non-users. In a study that found that most CIHS treatment for back and neck problems was provided by DCs, Martin et al. used the Medicare Expenditure Panel Survey to compare costs of treatment for back and neck problems for patients who did and did not use CIHS: propensity scores for the 2 groups were used to match patients who had propensity scores within a ‘region of common support’ and then identify the ‘nearest neighbor’ as a method to develop 2 matched groups for comparison.¹⁹ Weigel et al. have applied inverse propensity score weighting to data from ‘Medicare Current Beneficiary Survey’²⁰ respondents and to ‘Assets and Health Dynamics among the Oldest Old’ interviews²¹ when comparing utilization and outcomes of chiropractic and conventional medical treatment of back pain in the Medicare population. However, to date, propensity score methods have not been applied directly to Medicare claims data to evaluate costs of care for patients seeking back pain treatment.

This study has 2 purposes. Firstly, the paper provides an overview and brief tutorial on how propensity score methods work and the different types of propensity score methods that are commonly used in order to help readers of the CIHS literature understand these fairly complex methods. Secondly, the study uses Medicare data from 2006-2012 on older, multiply comorbid patients who had a chronic low back pain episode to demonstrate the impact of applying different propensity score methods on the balance of demographics of patients who were in one of 2 treatment groups: those who received only CMT and those who received no CMT during their episodes.

Methods

Overview of propensity score methods

Propensity scoring methods can be used to address possible confounding in observational studies where investigators have no control over treatment assignment, such as when they retroactively analyze health care claims databases. Conditional on a set of observed covariates, the propensity score for a patient is defined as the probability of a patient with the same observed characteristics being in the treatment group.²² The application of propensity score methods is designed to construct a new analytic dataset in which treatment groups are balanced on observed confounders so that the outcomes for the different treatment groups can be directly compared.

For instance, imagine that there were 100 patients with low back pain who used CMT and 200 who did not, and the 2 groups differed in their age and sex distributions (Figure 1). When examining the actual data, it is evident that the numbers, proportions of males and females, and ages of the 2 different treatment groups differed. Propensity scores based on age and sex could be used to identify and match patients who have a similar likelihood to be in a particular treatment group even if those cases differed individually on age and sex (shown in red in the figure).²³ However, using such a matching process limits the analysis to those with similar scores, and the number of cases that can be used for analysis will decline in both treatment groups. This is generally true whether one uses a 1-to-many (where 1 case from the smaller group is matched to several from the larger group)²³ or a 1-to-1 'closest neighbor' matching mechanism.¹⁹ With the limited number of cases, the age- and gender-distributions of the larger group will begin to resemble those of the smaller group. When using this method, it must be then acknowledged that the reduced dataset used to analyze outcomes may not be representative of the original study dataset. That is, external validity may be sacrificed to improve internal validity.

An alternative method that allows researchers to retain all of the data and maintain generalizability to the larger groups is propensity score weighting²⁴ (Figure 2). Here, cases are weighted according to the inverse of the propensity of a patient to self-select into a particular treatment group, based on demographics and other variables. In effect, this method relatively over-weights results for patients who are least likely to have self-selected into their respective treatment group, given their characteristics. In the example, one can see that older female patients are less likely to be in the CMT group; by over-weighting the older patients (and the females) and under-weighting the younger patients (and the males),

one can generate a distribution that approximates the demographics of the non-CMT treatment group.

The variables that should be used to calculate propensity scores differ from study to study, and even the number of variables used is open to consideration. While some authors advocate broad use of all variables, others suggest that experts use the literature and the data to determine which variables are likely to represent important differences that might affect the outcome between treatment groups. Recent studies indicate that a combination of those methods is probably the most effective way of determining propensity score variables.²⁵ However, the trimmed propensity score methodology, wherein only statistically significant differences in the characteristics of the comparison groups are included in the propensity score calculation, has generated the best predictive models when examining rare events.²⁶

Analysis becomes much more challenging when comparing multiple treatment groups. In this case, multinomial regression or alternative approaches to analysis of ordinal or nominal data are required to generate propensity scores that allow comparison of multiple treatment groups. Matching would then require application of algorithms to identify closest matches across the several treatment groups.²⁷ However, propensity score weighting methods generalized more easily to comparisons of 3 or more treatment groups than methods for propensity score weighting methods, but either method can be used to address selection bias with,²⁸ or without,²⁹ stratification.

Impact of applying propensity score methods on the balance of demographics of older, multiply comorbid patients with a chronic back pain episode who were in one of two treatment groups

A. Identification of patients and assignment to treatment groups—Using Medicare fee-for-service files from 2006-2012, we identified older patients who had an episode of chronic low back pain (cLBP) during which they also had an additional musculoskeletal disorder diagnosis and a mental health disorder diagnosis. We focused on patients with multiple comorbidities because they consume a disproportionate amount of healthcare resources.³⁰ We then identified patients who exclusively used CMT and who did not use CMT during their cLBP episodes. The sections below provide the methodology that we used to identify the multiply comorbid patients with a cLBP episode that we examined, to determine whether each of these used or did not use CMT, to calculate propensity scores, and to apply them using the different methods described above. Dartmouth College's Committee for the Protection of Human Subjects approved the study.

Following Cherkin's methodology,³¹ we eliminated patients with ICD-9-CM codes that are unlikely to be responsive to non-surgical treatment: neoplasms (CPT-9-CM codes 140-239.9), intra-spinal abscess (324.1), osteomyelitis (730-730.99), vertebral fractures with spinal cord injury (806.0-806.9), open vertebral fractures without spinal cord injury (805.1, 805.3, 805.5, 805.7, 805.9), vertebral dislocations (839-839.59), chordotomy (03.2-03.29), and cervical and dorsal fusions (81.01-81.03). Because DCs frequently use ICD-9-CM diagnostic codes 839-839.59, we did not exclude patients who had these codes if the code was recorded during a chiropractic visit.

Guided by Martin's analysis of the Medical Expenditures Panel Survey (MEPS) prevalence of diagnostic categories of back problems,³² we identified back pain patients as those who had a primary ICD-9-CM diagnosis code in the 724 series: 724.2 (lumbago), 724.3 (sciatica), 724.4 (thoracic or lumbosacral neuritis or radiculitis, unspecified), 724.5 (backache, unspecified), 724.8 (other symptoms referable to back), and 724.9 (other unspecified back disorders). These conditions account for approximate 55% of spine disorders.³² However, because DCs frequently use additional ICD-9-CM codes, we included diagnoses in the 722 (intervertebral disc disorders), 739 (non-allopathic subluxation), and 839 (other multiple and ill-defined dislocations) series if such diagnoses were made during a chiropractic visit.

Consistent with the literature,³³ we defined a cLBP episode as beginning with the recording of a low back pain diagnosis at least twice over at least a 90-day period following a period of 6 months during which no such diagnosis was recorded; the episode was defined to end on the day that the last such diagnosis was followed by at least 180 days without any such diagnosis. Although we had access to 7 years of data, our inclusion criteria required a 1-year look back and a 1-year follow up period. Thus, our analyses included episodes of chronic low back pain that occurred during 5 years, between 2007-2011.

To identify a multiply comorbid cohort, we limited the study population to patients who had an ICD-9-CM code indicating at least 1 of the following additional musculoskeletal comorbidities during the episode or in the 12 months preceding episode onset: other back and neck pain (307.89, 722.3, 722.32, 722.8, 722.83, 846 series, 847.2, 847.3, 847.9, 996.41), non-back rheumatism (729.0), arthritis and other arthropathies (716.19, 716.68, 716.95) osteoarthritis (715.00, 715.09, 715.18, 715.21, 715.22, 715.28, 715.98), and diffuse diseases of the connective tissue (728.89, 728.9, 729.1, 729.3, 729.4). We further limited the study population to patients who had an ICD-9-CM code indicating at least 1 of the following psychological comorbidities during the episode or in the 12 months preceding episode onset: depression (296.3, 300.4, 309.0, 309.2, 309.8, 311), anxiety (293.84, 300.00, 300.09), or insomnia (327.01, 327.02).

Patients who used CMT were identified as having as at least 1 visit to a DC during the year (code 35 in the provider specialty field) and generating at least 1 CPT based charge for chiropractic manipulative treatment of the spine (code 98940 (1 or 2 regions), 98941 (3 to 4 regions), or 98942 (5 regions)) with a concurrent diagnosis of low back pain, defined above.

Therefore, we were able to identify patients who did and did not use CMT. Because the timing of CMT in the context of concurrently received medical care matters,^{16,34} we subdivided patients who used CMT into those who used only CMT, those who used CMT followed by medical care, and those who used medical care followed by CMT during their chronic low back pain episode. While we generated propensity scores for all treatment groups, for the purposes of this example, we limited analyses to patients who used only CMT and those who did not use CMT.

B. Calculation of propensity scores—Because we anticipated that we would find that the different groups studied would have different patient characteristics,^{19,35,36} we used

multinomial stepwise logistic regression³⁷ (with criteria to enter the stepwise regression set at $p < 0.05$ and criteria to exit set at $p > 0.10$, and with statistically significant interaction terms retained) to generate the propensity for each patient to be in 1 of the 4 particular treatment groups, based on the following variables and their interactions, all of which were included in calculation of propensity scores:

- a. From Medicare files:
 - i. Socio-demographic variables (patient sex, patient race, whether the patient was concurrently enrolled in Medicaid, whether the patient was originally enrolled in Medicare by virtue of disability or age, and whether the beneficiary has a low-income subsidy for Medicare Part D (a measure of patient poverty)); and
 - ii. Diagnostic codes which we used to identify patients who had the aforementioned categorically defined comorbidities that defined our study population and from which we calculated modified Deyo-Charlson³⁸ scores and Iezzoni³⁹ scores.
 - iii. The per-capita supply of DCs in the Dartmouth Atlas Defined Hospital Referral Region. We included this in the propensity score because the supply of DCs has been associated with CMT utilization in the Medicare population;^{40,41} therefore, the supply of DCs represents a confounder that should be included in propensity score generation as it might be associated with both treatment exposure and outcomes.
 - iv. The regional carrier that was used by Medicare to process CMT claims in the region where the patient lived as a variable in generating propensity scores.
- b. From ZIP Code-linked datasets:
 - i. An estimate of patient income (ZIP Code specific median annual household income in 2010) as well as the proportion of the population that lived under the federal poverty level.

C. Application of propensity scores and comparisons of the two methods—

For the 2 treatment groups we examine here – patients who used only CMT and patients who used no CMT during their cLBP episode – we compared results from 2 different propensity score application methods: propensity score matching and inverse propensity score weighting. Further, we examined 2 methods of propensity score matching: 1 in which we used the best single match,¹⁹ and a 1-to-many methodology in which we used the closest 4 matches from the larger group.²³ For propensity score weighting, we inversely weighted cases by a patient's propensity to self-select into their respective treatment group.

Results

The top lines of Table 1 show demographics, measures of poverty and income, and the per-capita supply of DCs in the hospital referral region for patients in the 2 treatment groups

examined, before application of the different propensity scoring methods. As expected, when compared to those who did not use CMT during their back pain episode, patients who used only CMT were younger, more likely to be white, more likely to have a greater than high school education, less likely to identified as being impoverished through Medicare Part D, and less likely to be Medicaid enrolled. In this subpopulation of multiply comorbid patients that we examined, we were surprised to find that patients who used only CMT were somewhat more likely to be male and to live in ZIP Codes with lower incomes than patients did not use CMT. As predicted from studies that have shown a relationship between chiropractic care use and the supply of DCs in the community,^{40,41} patients in our cohort who used CMT lived in hospital referral regions that had a higher per-capita supply of DCs.

The bottom lines of Table 1 show that, after application of the different propensity score methods, the 2 comparison treatment groups looked more similar. As anticipated, when propensity score matching was used, the total numbers of cases dropped (dramatically so when only the single best match method was used). When 1-to-1 matching was used, demographic values of the 2 groups were more similar to those in the treatment group with the smaller number of patients (in this case, the group that only used CMT). When 4-to-1 matching was used, results that were intermediate between the 2 groups were found. In contrast, when inverse propensity score weighting was used, all of the cases remained in the analysis, and the demographic values were more similar to the treatment group with the larger number of patients (in this case, the group that did not use CMT).

Table 2 shows the prevalence of the diagnostic groups and indicators of illness burden that were used to generate our propensity scores before and after application of the different propensity score methods. As expected, before application of propensity scores, patients who used only CMT were less likely to be obese, had lower illness burdens (as measured by Charlson and Iezzoni scores), were less likely to be disabled, and how lower Medicare spending levels in the prior year than patients who used no chiropractic services during their back pain episodes. Patients who sought only CMT during their episodes had a lower prevalence of osteoarthritis, non-back pain, other arthritis, depression, and anxiety but a higher prevalence of other back and neck pain and diffuse connective tissue disease. Again, all uses of propensity scores reduced differences between groups. Also, when 1-to-1 matching was used, the values for both comparison groups were more similar to those for the group that used only CMT; when 4-to-1 matching was used, values were between the 2 groups, seemingly not representative of either. Propensity score weighting resulted in values that were more similar to the group that did not use any CMT and diminished differences across treatment groups.

Discussion

Propensity score methods have proven to be useful in observational studies that compare treatment groups when there is the possibility of selection bias. We described the application of 3 propensity score methods to generate balanced key variables across 2 cLBP treatment exposures in the older, multiply comorbid Medicare population. We showed that application of propensity score methods markedly diminished differences in baseline characteristics across the 2 treatment groups examined, thereby permitting balanced comparison of

outcome variables between these groups. However, we found that, as a method, inverse propensity score weighing had 2 advantages: it retained all of the cases and it generated more balanced demographic values that were more like the treatment group with a larger number of patients, thereby preserving external validity.

Our findings suggest that propensity score calculation and application should be considered when conducting observational trials of CIHS service utilization and outcomes any time investigators have no control over treatment assignment that may be susceptible to selection bias. While findings may differ in other studies, we found that application of inverse propensity score weighting created comparison groups that were more similar to the larger group and generated findings that were more generalizable than application of propensity score matching algorithms. Further, our results indicate that different propensity score methods may generate different results.

Limitations and suggestions for future studies

Our study has several limitations. First, as is always the case when conducting retroactive analyses of claims databases, we were limited by data availability. There remain potentially confounding variables for which we were unable to correct. Second, our inclusion criteria were dependent on the accuracy and completeness of data reporting; we did not have the ability to independently verify diagnostic accuracy, for instance. To the extent that diagnostic inaccuracies were present in our data, our results are inaccurate. Third, we studied a very narrow subset of the Medicare population; findings may vary when examining broader cohorts of patients with back pain episodes.

Future studies might reexamine observational studies that did not use propensity scoring methods by generating and applying propensity scores and then comparing new to original findings. If original findings held, they might be considered much more robust and policy relevant. Future observational studies that examine the clinical or economic impact of CIHS should use propensity scoring methods and should provide justification for the particular method used.

Conclusions

This study shows that propensity score methods are useful in making comparisons across treatment groups in observational studies. Such analytic methods can make treatment groups more comparable and, therefore, improve the accuracy and generalizability of studies utilizing databases to compare outcomes across groups of patients. Further, application of propensity score methodologies approaches the main goal of randomization in controlled studies: to assure the comparability of groups for which outcomes are being assessed.

Acknowledgments

FUNDING SOURCES

The National Institutes of Health National Center for Complementary and Integrative Health (1R21AT008287-01) National Center for Advancing Translational Sciences (1UL1TR001086); and the National Institute of Arthritis and Musculoskeletal Skin Diseases (1P60AR062799) funded this work.

References

1. Coulter ID, Khorsan R. Is health services research the Holy Grail of complementary and alternative medicine research? *Altern Ther Health Med*. 2008; 14:40–5. [PubMed: 18616068]
2. Mootz RD, Hansen DT, Breen A, Killinger LZ, Nelson C. Health services research related to chiropractic: review and recommendations for research prioritization by the chiropractic profession. *J Manipulative Physiol Ther*. 2006; 29:707–25. [PubMed: 17142165]
3. Lichtenberg FR. Are the benefits of newer drugs worth their cost? Evidence from the 1996 MEPS. *Health Aff (Millwood)*. 2001; 20:241–51. [PubMed: 11558710]
4. Briesacher BA, Stuart B, Ren X, Doshi JA, Wrobel MV. Medicare beneficiaries and the impact of gaining prescription drug coverage on inpatient and physician spending. *Health Serv Res*. 2005; 40:1279–96. [PubMed: 16174134]
5. Olfson M, Sing M, Schlesinger HJ. Mental health/medical care cost offsets: opportunities for managed care. *Health Aff (Millwood)*. 1999; 18:79–90. [PubMed: 10091434]
6. Polen MR, Freeborn DK, Lynch FL, Mullooly JP, Dickinson DM. Medical cost-offset following treatment referral for alcohol and other drug use disorders in a group model HMO. *J Behav Health Serv Res*. 2006; 33:335–46. [PubMed: 16752110]
7. Selden TM. The within-year concentration of medical care: implications for family out-of-pocket expenditure burdens. *Health Serv Res*. 2009; 44:1029–51. [PubMed: 19674431]
8. Rubinstein SM, van Eekelen R, Oosterhuis T, de Boer MR, Ostelo RW, van Tulder MW. The risk of bias and sample size of trials of spinal manipulative therapy for low back and neck pain: analysis and recommendations. *J Manipulative Physiol Ther*. 2014; 37:523–41. [PubMed: 25194968]
9. Smith M, Carber L. A compilation of chiropractic and complementary/alternative medicine (CAM) data from public-use national surveys: report on a health-services research resource for the chiropractic and CAM scientific community. *J Manipulative Physiol Ther*. 2002; 25:573–8. [PubMed: 12466775]
10. Swanson SA, Robins JM, Miller M, Hernan MA. Selecting on treatment: a pervasive form of bias in instrumental variable analyses. *Am J Epidemiol*. 2015; 181:191–7. [PubMed: 25609096]
11. Luedtke R, Weber U, Fischer I, Fries KH, Moeller H. An example on the value of non-randomisation in clinical trials in complementary medicine. *Forsch Komplementarmed Klass Naturheilkd*. 2002; 9:105–9. [PubMed: 12006728]
12. Ndetan HT, Bae S, Evans MW Jr, Rupert RL, Singh KP. Characterization of health status and modifiable risk behavior among United States adults using chiropractic care as compared with general medical care. *J Manipulative Physiol Ther*. 2009; 32:414–22. [PubMed: 19712783]
13. Stano M, Smith M. Chiropractic and medical costs of low back care. *Med Care*. 1996; 34:191–204. [PubMed: 8628040]
14. Haas M, Sharma R, Stano M. Cost-effectiveness of medical and chiropractic care for acute and chronic low back pain. *J Manipulative Physiol Ther*. 2005; 28:555–63. [PubMed: 16226622]
15. Sharma R, Haas M, Stano M, Spegman A, Gehring R. Determinants of costs and pain improvement for medical and chiropractic care of low back pain. *J Manipulative Physiol Ther*. 2009; 32:252–61. [PubMed: 19447261]
16. Liliedahl RL, Finch MD, Axene DV, Goertz CM. Cost of care for common back pain conditions initiated with chiropractic doctor vs medical doctor/doctor of osteopathy as first physician: experience of one Tennessee-based general health insurer. *J Manipulative Physiol Ther*. 2010; 33:640–3. [PubMed: 21109053]
17. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*. 2003; 158:280–7. [PubMed: 12882951]
18. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. 1997; 127:757–63. [PubMed: 9382394]
19. Martin BI, Gerkovich MM, Deyo RA, Sherman KJ, Cherkin DC, Lind BK, Goertz CM, Lafferty WE. The association of Complementary and Alternative Medicine use and health care expenditures for back and neck problems. *Medical Care*. 2012; 50:1029–36. [PubMed: 23132198]

20. Weigel PA, Hockenberry JM, Wolinsky FD. Chiropractic use in the Medicare population: prevalence, patterns, and associations with 1-year changes in health and satisfaction with care. *J Manipulative Physiol Ther.* 2014; 37:542–51. [PubMed: 25233887]
21. Weigel PA, Hockenberry JM, Bentler SE, Kaskie B, Wolinsky FD. Chiropractic episodes and the co-occurrence of chiropractic and health services use among older Medicare beneficiaries. *J Manipulative Physiol Ther.* 2012; 35:168–75. [PubMed: 22386915]
22. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med.* 1998; 17:2265–81. [PubMed: 9802183]
23. Rassen JA, Shelat AA, Myers J, Glynn RJ, Rothman KJ, Schneeweiss S. One-to-many propensity score matching in cohort studies. *Pharmacoepidemiol Drug Saf.* 2012; 21(Suppl 2):69–80. [PubMed: 22552982]
24. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res.* 2011; 46:399–424. [PubMed: 21818162]
25. Patorno E, Glynn RJ, Hernandez-Diaz S, Liu J, Schneeweiss S. Studies with many covariates and few outcomes: selecting covariates and implementing propensity-score-based confounding adjustments. *Epidemiology.* 2014; 25:268–78. [PubMed: 24487209]
26. Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA, Hartzema AG. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Stat Med.* 2012; 31:4401–15. [PubMed: 23015364]
27. Rassen JA, Shelat AA, Franklin JM, Glynn RJ, Solomon DH, Schneeweiss S. Matching by propensity score in cohort studies with three treatment groups. *Epidemiology.* 2013; 24:401–9. [PubMed: 23532053]
28. Linden A. Combining propensity score-based stratification and weighting to improve causal inference in the evaluation of health care interventions. *J Eval Clin Pract.* 2014; 20:1065–71. [PubMed: 25266868]
29. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med.* 2013; 32:3388–414. [PubMed: 23508673]
30. Parekh AK, Barton MB. The challenge of multiple comorbidity for the US health care system. *JAMA.* 2010; 303:1303–4. [PubMed: 20371790]
31. Cherkin DC, Deyo RA, Volinn E, Loeser JD. Use of the International Classification of Diseases (ICD-9-CM) to identify hospitalization for mechanical low back problems in administrative databases. *Spine.* 1992; 17:817–25. [PubMed: 1386943]
32. Martin BI, Deyo RA, Mirza SK, Turner JA, Comstock BA, Hollingsworth W, Sullivan SD. Expenditures and health status among adults with back and neck problems. *JAMA.* 2008; 299:656–64. [PubMed: 18270354]
33. Rozenberg S. Chronic low back pain: definition and treatment. *Rev Prat.* 2008; 58:265–72. [PubMed: 18536200]
34. Weigel PAM, Hockenberry JM, Bentler SE, Kaskie B, Wolinsky FD. Chiropractic episodes and the co-occurrence of chiropractic and health services use among older Medicare beneficiaries. *J Manipulative Physiol Ther.* 2012; 35:168–75. [PubMed: 22386915]
35. Weeks WB, M WJ, Toler A, Goertz CM. Medicare's Demonstration of Expanded Coverage for Chiropractic Services: limitations of the Demonstration and an alternative direct cost estimate. *J Manipulative Physiol Ther.* 2013; 36:468–81. [PubMed: 23993755]
36. Coulter ID, Hurwitz EL, Adams AH, Genovese BJ, Hays R, Shekelle PG. Patients using chiropractors in North America: who are they, and why are they in chiropractic care? *Spine (Phila Pa 1976).* 2002; 27:291–6. discussion 7-8. [PubMed: 11805694]
37. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *Am J Epidemiol.* 2011; 173:1404–13. [PubMed: 21602301]
38. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol.* 1992; 45:613–9. [PubMed: 1607900]
39. Iezzoni, L. Risk Adjustment for Measuring Healthcare Outcomes. 4th ed. Health Administration Press; Chicago, IL: 2012.

40. Whedon JM, Song Y. Geographic variations in availability and use of chiropractic under Medicare. *J Manipulative Physiol Ther.* 2012; 35:101–9. [PubMed: 22257945]
41. Whedon JM, Song Y, Davis MA, Lurie JD. Use of chiropractic spinal manipulation in older adults is strongly correlated with supply. *Spine (Phila Pa 1976).* 2012; 37:1771–7. [PubMed: 22487711]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

- Patients who use CMT often have different characteristics than patients who do not, and these differences can confound attempts to compare outcomes across treatment groups.
- Propensity score methods reduce differences across treatment groups in observational studies and may lead to different policy ramifications than alternative approaches.
- Propensity score calculation and application should be considered when conducting observational analyses of service utilization and outcomes when investigators have no control over treatment assignment,

Figure 1.

Example of propensity score matching. Open circles represent females, targeted circles represent males, and red indicates cases where propensity scores are similar. CMT means chiropractic manipulative treatment.

Figure 2.

Example of propensity score weighting. Open circles represent females, targeted circles represent males, and the size of the circle indicates its weight, with larger sizes indicating greater weight. CMT means chiropractic manipulative treatment.

Mean values for demographic, poverty and income, and supply variables that were used to calculate the propensity score, before and after application of propensity score methods. Standard errors are given in parentheses. All measures are from individual level data except those marked with an asterisk, which are derived from the ZIP Code in which the patient lived. CMT means chiropractic manipulative treatment.

Table 1

	Number	Demographics				Measures of poverty and income			The per-capita supply of DCs in the Hospital Referral Region (N/10,000 beneficiaries)
		Mean age (yrs)	Male (%)	White race (%)	Less than high school education* (%)	Part D poverty flag (%)	Medicaid enrolled (%)	Median income* (\$) (N/10,000 beneficiaries)	
Before application of propensity scores	Only used CMT	73.7 (0.12)	24.6 (0.69)	96.1 (0.31)	5.61 (0.10)	11.8 (0.52)	8.4 (0.44)	53,793 (335)	2.68 (0.016)
	Did not use CMT	76.9 (0.03)	17.4 (0.16)	86.7 (0.14)	7.08 (0.03)	29.9 (0.19)	24.0 (0.17)	54,585 (87)	2.24 (0.004)
Propensity score method used	1:1 nearest neighbor match	73.7 (0.11)	24.5 (0.69)	96.2 (0.31)	5.61 (0.08)	11.7 (0.52)	8.3 (0.44)	53,805 (301)	2.68 (0.018)
		73.7 (0.11)	24.0 (0.69)	95.0 (0.35)	5.63 (0.08)	12.0 (0.52)	8.4 (0.45)	53,583 (301)	2.66 (0.018)
	4:1 'greedy' match	74.5 (0.13)	17.5 (0.14)	95.0 (0.21)	5.93 (0.10)	14.1 (0.63)	10.2 (0.55)	54,849 (356)	2.46 (0.018)
		74.6 (0.06)	23.9 (0.33)	95.3 (0.38)	6.05 (0.05)	14.6 (0.32)	10.7 (0.28)	54,422 (177)	2.44 (0.009)
	Inverse weighting	76.4 (0.06)	18.2 (0.14)	88.0 (0.12)	6.91 (0.05)	27.0 (0.16)	21.5 (0.15)	54,817 (158)	2.28 (0.007)
		76.4 (0.06)	18.5 (0.14)	88.1 (0.12)	6.87 (0.05)	26.9 (0.17)	21.4 (0.15)	54,520 (159)	2.30 (0.007)

Table 2

Mean proportions of patients who had the following diagnoses that were used to calculate the propensity score, before and after propensity score application. Standard errors are given in parentheses. CMT means chiropractic manipulative treatment.

		Obesity	Musculoskeletal conditions					Psychiatric conditions		Indicators of illness burden			
			Osteo-arthritis	Other back and neck pain	Non-back pain	Other arthritis	Diffuse connective tissue disease	Depression	Anxiety	Charlson score	Iezzoni score	Original reason for Medicare eligibility was disability (%)	Total Medicare spending in year prior to episode (\$)
Before application of propensity scores	Only used CMT	8.6 (0.45)	22.0 (0.66)	41.3 (0.79)	0.84 (0.15)	0.90 (0.15)	50.6 (0.80)	67.4 (0.75)	48.4 (0.80)	1.30 (0.03)	0.90 (0.02)	7.8 (0.43)	4,718 (168)
	Did not use CMT	13.2 (0.14)	53.2 (0.20)	23.5 (0.17)	1.38 (0.05)	3.40 (0.07)	38.2 (0.20)	72.1 (0.18)	51.3 (0.20)	2.13 (0.01)	1.38 (0.01)	12.9 (0.14)	6,586 (51)
Propensity score method used	1:1 nearest neighbor match	8.7 (0.45)	22.1 (0.67)	41.4 (0.79)	0.9 (0.15)	0.9 (0.15)	50.5 (0.80)	67.4 (0.75)	48.3 (0.80)	1.30 (0.03)	0.90 (0.02)	7.8 (0.43)	4,707 (164)
	4:1 'greedy' match	9.6 (0.53)	27.7 (0.81)	36.0 (0.87)	1.1 (0.19)	1.1 (0.19)	51.1 (0.90)	67.9 (0.84)	50.0 (0.90)	1.50 (0.03)	1.02 (0.02)	9.3 (0.52)	5,321 (198)
	Only used CMT	9.8 (0.27)	27.4 (0.40)	35.9 (0.43)	1.0 (0.09)	0.9 (0.09)	50.5 (0.45)	67.8 (0.42)	49.9 (0.45)	1.51 (0.02)	1.03 (0.01)	9.4 (0.26)	5,310 (99)
	Did not use CMT	11.6 (0.12)	50.3 (0.19)	25.6 (0.16)	2.12 (0.053)	2.62 (0.059)	41.1 (0.18)	71.2 (0.17)	51.3 (0.19)	1.93 (0.02)	1.30 (0.01)	12.6 (0.12)	6,277 (91)
Inverse weighting	Only used CMT	12.7 (0.12)	49.0 (0.19)	26.4 (0.16)	1.31 (0.043)	3.12 (0.065)	40.3 (0.18)	71.4 (0.17)	51.2 (0.19)	1.99 (0.02)	1.31 (0.01)	12.1 (0.12)	6,280 (92)
	Did not use CMT												