OXFORD

Original Research Article

Evaluation of Complementary and Integrative Health Approaches Among US Veterans with Musculoskeletal Pain Using Propensity Score Methods

Ling Han, MD, PhD,*,† Joseph L. Goulet, PhD, MS,†,‡,§,¶ Melissa Skanderson, MSW,† Harini Bathulapalli, MPH, MSc,† Stephen L. Luther, PhD,‖ Robert D. Kerns, PhD,†,‡,§,¶ and Cynthia A. Brandt, MD, MPH†,‡,§,¶

Departments of *Internal Medicine, ‡Psychiatry, §Medicine, and ¶Emergency Medicine, Yale School of Medicine, New Haven, Connecticut; †Pain Research, Informatics, Multimorbidities and Education (PRIME) Center, VA Connecticut Healthcare System, West Haven, Connecticut; ‖CINDRR, James A. Haley Veterans Hospital, Tampa, Florida, USA

Correspondence to: Ling Han, MD, PhD, Yale School of Medicine Program on Aging, 300 George St Suite 775, New Haven, CT 06511, USA. Tel: 203-737-2162; Fax: 203-785-4823; E-mail: ling.han@yale.edu.

Abstract

Objectives. To examine the treatment effectiveness of complementary and integrative health approaches (CIH) on chronic pain using Propensity Score (PS) methods.

Design, Settings, and Participants. A retrospective cohort of 309,277 veterans with chronic musculoskeletal pain assessed over three years after initial diagnosis.

Methods. CIH exposure was defined as one or more clinical visits for massage, acupuncture, or chiropractic care. The treatment effect of CIH on self-rated pain intensity was examined using a longitudinal model. PS-matching and inverse probability of treatment weighting (IPTW) were used to account for potential selection and confounding biases.

Results. At baseline, veterans with (7,621) and without (301,656) CIH exposure differed significantly in 21 out of 35 covariates. During the follow-up period, on average CIH recipients had 0.83 (95% confidence interval [CI] = 0.77 to 0.89) points higher pain intensity ratings (range = 0–10) than nonrecipients. This apparent unfavorable effect size was reduced to 0.37 (95% CI = 0.28 to 0.45) after PS matching, 0.36 (95% CI = 0.29 to 0.44) with IPTW on the treated (IPTW-T) weighting, and diminished to null when integrating IPTW-T with PS matching (0.004, 95% CI = −0.09 to 0.10). An alternative IPTW model and conventional covariate adjustment appeared least powerful in terms of potential bias reduction. Sensitivity analyses restricting the follow-up period to one year after CIH initiation derived consistent results.

Conclusions. PS-based causal methods successfully eliminated baseline difference between exposure groups in all measured covariates, yet they

did not detect a significant difference in the self-rated pain intensity outcome between veterans who received CIHs and those who did not during the follow-up period.

Key Words. Complementary and Integrative Health Approaches (CIH); Musculoskeletal Disorders (MSD); Propensity Score (PS) Matching; Inverse Probability of Treatment Weighting (IPTW); PIRs; Average Treatment Effect; Causal Inference; Absolute Difference (AD)

Introduction

Chronic pain is a significant public health problem in the United States, affecting nearly 100 million Americans and costing $635 billion in annual treatment and lost productivity [1,2]. Among veterans receiving care in the Veterans Health Administration (VHA) system, musculoskeletal disorders (MSD) are among the most frequently diagnosed painful conditions [3–7], and the prevalence is increasing [8,9]. Presence of pain among VHA primary care patients is associated with poorer self-rated health, greater health care utilization and costs, and increased comorbidity burden with depression, post-traumatic stress, and substance use disorders [4,6–11].

The Department of Health and Human Services recently published the National Pain Strategy (NPS) [2] with specific recommendations to transform pain care in America. These recommendations highlighted the importance of access to patient-centered, evidence-based, integrated, and multimodal pain care, including nonpharmacological approaches. Consistent with the NPS, VHA policy advocated incorporation of evidence-based complementary and integrative health (CIH) approaches, including acupuncture, massage, and spinal manipulation, into the pain care plan [3,12].

Several meta-analyses of randomized and nonrandomized trials revealed small to moderate effects of acupuncture, massage, and spinal manipulation on pain intensity (0.4–0.8 standardized difference) for varying pain conditions relative to no treatment or usual care [13–16]. A more recent systematic review reported that positive trials outnumbered negative ones for acupuncture, chiropractic, and massage therapy [17]. However, most published CIH trials were conducted in small samples (<100 patients) over short follow-up durations (less than six months), and thus may not be representative of their intended target patient populations or the natural course of chronic pain [13–16].

Propensity Score (PS) denotes the conditional probability of receiving a treatment (or an exposure) given a set of baseline covariates [18–20]. Although early studies did not always find PS methods superior to traditional multivariable modeling [21–23], advances in counterfactual causation theory have revived this technique into the causal investigation of medical interventions [24–27].

These causal methods attempt to tackle potential selection and confounding biases in a similar fashion to randomized trials by resampling hypothetically more comparable treatment and control groups through matching or inverse probability of treatment weighting (IPTW) [24–31]. In recent decades, such novel methods have been used increasingly in observational studies of medical interventions [20,28,30,32]. Two recent studies also demonstrated that PS methods were able to achieve good covariate balance between chronic pain patients who either received or did not receive acupuncture [33] or chiropractic care [34]. However, the usefulness of such causal methods to real-world effectiveness studies of CIH on chronic pain has yet to be investigated [35], especially in the veteran population, in which the burden of chronic pain is high, the concern over the "opioid epidemic" is mounting, and randomized trial evidence for CIH is inadequate [2,3,14–17].

We sought to evaluate CIH effectiveness on chronic pain using PS-based causal methods among a large cohort of US veterans receiving care in VHA facilities. Our hypothesis was that treatment with CIHs should have a modest benefit on reducing intensity of chronic pain. Accordingly, we expect that on average veterans who received one or more courses of CIH therapies would report less intense pain than those who did not receive such therapies, after appropriate control of potential selection and confounding biases inherent to observational data.

Methods

Data Source and Study Population

This study used data from the MSD Cohort, consisting of more than 5 million veterans [6]. In brief, VHA electronic health records (EHRs) were searched for 1,685 distinct *International Classification of Diseases*, Ninth Revision, Clinical Modification (ICD-9-CM) codes consistent with MSD. Veterans were included into the cohort at the date of their first MSD diagnosis, denoted as the MSD index date, if they received such a diagnosis at two or more outpatient visits within 18 months or during one or more inpatient stays. Data on demographic and military characteristics were obtained through linkage with other VHA administrative and clinical databases. Comorbid medical and mental health conditions were identified within one year before and six months after the MSD index date. Opioid prescriptions were extracted from VHA pharmacy dispenses. Pain intensity was documented in VHA Vital Signs data using a numerical rating scale, a global pain screening tool used in routine clinical practice without regards to specific diagnoses [6]. To reflect recent VHA practice and facilitate causal inference, we restricted our study sample to 309,277 "incident" MSD cases whose MSD index date fell within a one-year time frame between October 1, 2010, and September 30, 2011. Each veteran was then followed up for a maximum of three years until death, loss to VHA care, or September 30, 2013. The study
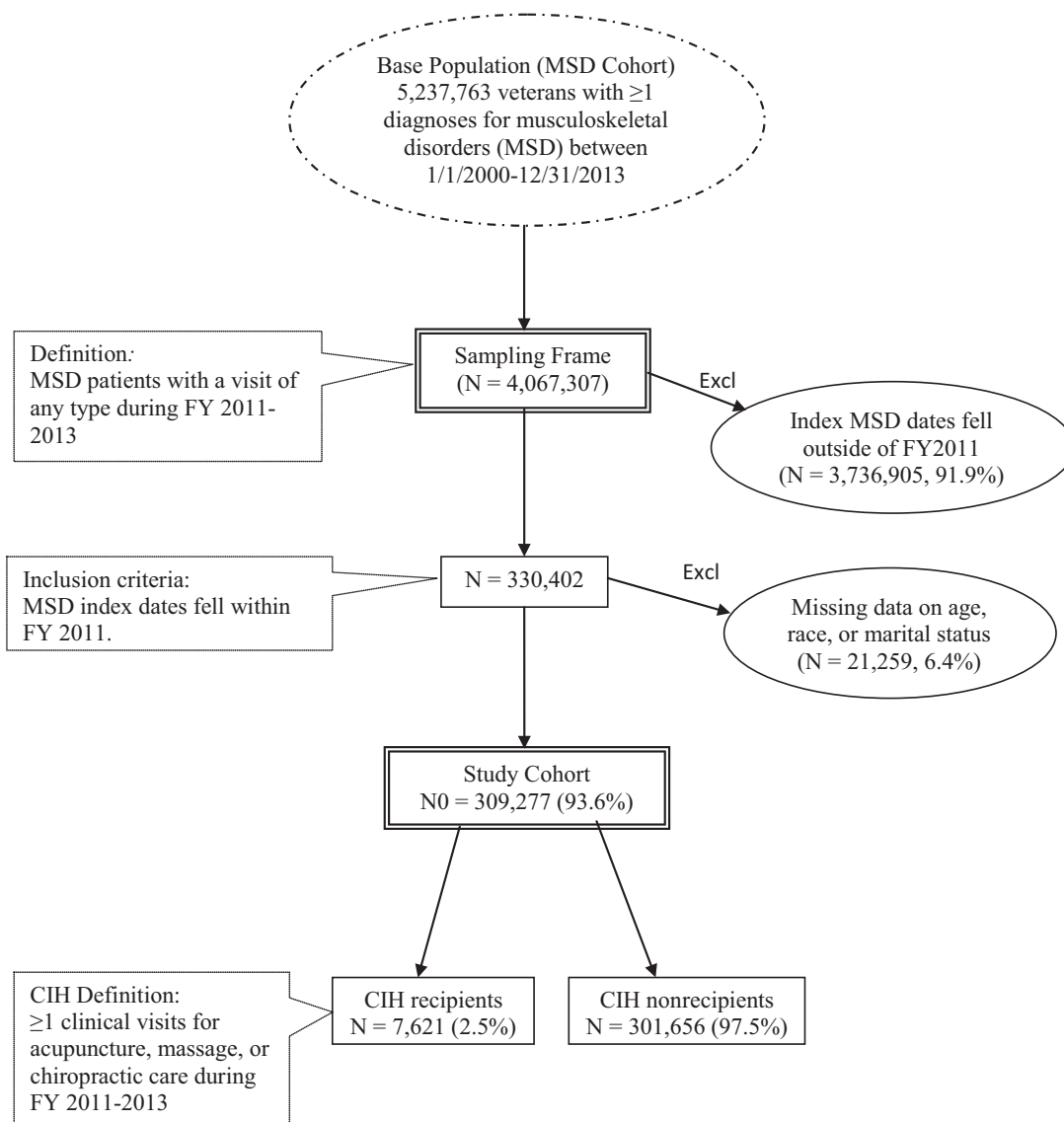
Base Population (MSD Cohort)
5,237,763 veterans with ≥1
diagnoses for musculoskeletal
disorders (MSD) between
1/1/2000-12/31/2013

Definition:
MSD patients with a visit of
any type during FY 2011-
2013

Sampling Frame
(N = 4,067,307)

Excl

Index MSD dates fell
outside of FY2011
(N = 3,736,905, 91.9%)

Inclusion criteria:
MSD index dates fell within
FY 2011.

N = 330,402

Excl

Missing data on age,
race, or marital status
(N = 21,259, 6.4%)

Study Cohort
N0 = 309,277 (93.6%)

CIH Definition:
≥1 clinical visits for
acupuncture, massage, or
chiropractic care during
FY 2011-2013

CIH recipients
N = 7,621 (2.5%)

CIH nonrecipients
N = 301,656 (97.5%)

Figure 1   A flow chart of cohort assembly among US veterans with musculoskeletal disorders.

was approved by the VA Connecticut Healthcare System Institutional Review Board.

A flowchart of the sampling process is presented in Figure 1.

Study Measurement

Exposure to CIH

We tracked veterans' EHRs for encounters of chiropractor care or acupuncture or massage procedure (Supplementary Data - Table 1). To reduce likely underreporting of such approaches by clinicians, we defined CIH exposure as having one or more clinical visits for

any of the three therapies during the three-year follow-up period. Veterans were classified as CIH recipients or nonrecipients, with the date of first clinical visits surrogating treatment initiation.

Covariates for PS Estimation

Demographic and service characteristics included age, gender, race/ethnicity, marital status, service era (recent, Vietnam or early era), body mass index (BMI), and current smoking status. MSD or other pain conditions were represented by 11 individual diagnoses with a prevalence ≥10%. Major mental health conditions included post-traumatic stress disorders (PTSDs), major

depressive disorders (MDDs), alcohol use disorders, and substance use disorders (SUDs). Other clinical conditions included diabetes, hypertension, chronic heart failure, coronary artery diseases, and a composite indicator for conditions not otherwise specified. Veterans receiving opioid treatment at baseline were identified from pharmacy dispensing records during the MSD index year. Variations of pain care over time and across VHA facilities were captured using duration under VHA care prior to study entry, calendar year of study entry, length of follow-up, facility complexity [36] for the primary facility where veterans received usual care or the index MSD diagnosis, and 21 veterans' Integrated Service Networks (VISN) regions [6,36].

Outcome Measure of CIH "Treatment" Effect

We used patient-reported pain intensity ratings (PIRs) assessed monthly over three years on a numerical rating scale (range = 0–10, 0 for no pain and 10 for maximal possible pain) as a time-dependent outcome [6]. Because the numbers of ratings each veteran received varied substantially from month to month (median = 1, range = 1–60), we extracted only the highest ratings each month following the original MSD cohort convention [6].

Statistical Analyses

Baseline characteristics were summarized using frequency (percentage) and mean ($\pm$ standard deviation) or median (intraquartile range [IQR]), as appropriate, for the entire cohort and according to CIH exposure.

PS Estimation and Evaluation

We used a logistic regression model to estimate a PS for each veteran, with a binary exposure to CIH (ie, recipients vs nonrecipients) as the dependent variable and 35 selected covariates as predictors (see Table 1 for a complete list). Covariates with sparse frequency (<1%) were recoded or regrouped to avoid unstable model estimation. To derive a parsimonious model, we first included only the main effects of the 35, followed by testing potentially important two-way interactions between selected veteran characteristics and cohort features (see the footnote to the Supplementary Data - Figure 1). Distribution of the derived PS was examined using a univariate procedure, along with Box and Kernel density plots (Supplementary Data - Figure 1) to ensure adequate "common support" between the two exposure groups [19,29,30]. This process was repeated until deriving an optimal model based on the maximal likelihood ratios test between nested models (–2 times (LL_1–LL_2), $P < 0.05$), which included 35 main effects and six interactions.

Deriving PS-Matched Subcohort and Two Alternative IPTWs

We used a SAS greedy matching macro to select one non-CIH veteran for each CIH recipient by a caliper width ≤ 0.2 logit (PS) [37]. In addition, two different IPTWs were calculated, one for estimating the average treatment effect in the population or ATE, abbreviated hereafter as IPTW-P, and the other for estimating the average treatment effect on the treated or ATT, abbreviated as IPTW-T [24,25,28,31]. The formula for the IPTW-P, which was stabilized to avoid extreme weights (top 1%), was

$$IPTW - P \ (stabilized) = E * PE/PS + (1 - E) \\ * (1 - PE)/(1 - PS);$$

whereas the IPTW-T (without needing stabilization) was derived as

$$IPTW - T = E + (1 - E) * PS/(1 - PS);$$

where E denotes exposure to CIH (1 if exposed, 0 otherwise) and P denotes the overall exposure prevalence in the cohort. In nonexperimental studies, the two IPTWs are used to recover the "causal" treatment effect [24,25,31]. Studies have shown that ATT provides a more clinically interpretable and statistically consistent estimate than ATE [38], especially when the exposure (or treatment) is rare [39].

Assessing Covariate Balance Between CIH Groups

Covariate balance between the two CIH groups was assessed using standardized difference (STD) for the original cohort (N = 309,277) and the PS-matched subcohort (N = 7,621 × 2), and weighted STD with IPTW-P or IPTW-T, respectively. To facilitate intermethod comparison, for each method we calculated the overall count and percentage of STDs <0.10, a typical cutpoint for excellent balance [29–31], and an average STD per covariate (out of 35).

Examining "Causal" Treatment Effect of CIH on Chronic Pain

The treatment effectiveness of CIH was examined among 306,720 veterans (99.7% of original cohort) with one or more PIRs after the MSD index date. To avoid potential "reverse causality" bias, the analytic time zero for CIH recipients was reset to the date of initial CIH visits, which presumably flagged the start of one or more courses of CIH therapies. Accordingly, all their prior pain intensity ratings were eliminated. Details on the final causal analysis samples before and after PS matching are provided in Table 2.

Table 1 Baseline characteristics of 309,277 US veterans with musculoskeletal disorders at initial clinical visits during 2011 according to exposure to complementary and integrative health approaches, in full cohort and the Propensity Score–matched subcohort

Characteristics by Domain	Overall (N = 309,277) Mean±SD or No. (%)	CIH Recipients[†] (N = 7,621) Mean±SD or No. (%)	CIH Nonrecipients (N = 301,656) Mean±SD or No. (%)	PS-Matched Nonrecipients[‡] (N = 7,621) Mean±SD or No. (%)
Demographic, military, and behavioral factors				
Age, y	56.6±17.1	47.9±16.0	56.8±17.1***	47.8±16.7
18–49	96,779 (31.3)	3,978 (52.2)	8,148 (34.2)	3,941 (51.7)
50–64	112,793 (36.5)	2,528 (33.2)	13,208 (55.4)	2,444 (32.1)
65–104	99,705 (32.2)	1,115 (14.6)	2,477 (10.4)	1,008 (16.2)
Female	23,177 (7.5)	1,032 (13.5)	22,145 (7.3)***	1,038 (13.6)
Nonwhite race/ethnicity	79,518 (25.7)	2,411 (31.6)	77,107 (25.6)***	2,375 (31.2)
Currently married	167,128 (54.0)	3,654 (48.0)	163,474 (54.2)***	3,618 (47.5)
Serviced in OEF/OIF/OND	48,134 (15.7)	2,201 (28.9)	45,933 (15.2)***	2,240 (29.4)
Service era				
Recent	93,312 (30.2)	3,796 (49.9)	89,516 (29.7)***	3,840 (50.4)
Vietnam	163,549 (52.9)	3,253 (42.7)	160,296 (53.2)	3,240 (42.5)
Early	52,282 (16.9)	565 (7.4)	51,717 (17.2)	536 (7.0)
Current smoker	110,341 (38.1)	2,919 (39.6)	107,422 (38.1)***	2,857 (38.7)
Body mass index	29.5±5.9	29.4±5.6	29.5±5.9*	29.3±5.7
Obese (BMI ≥ 30)	124,009 (41.2)	2,986 (39.9)	121,023 (41.2)***	2,962 (39.6)
Musculoskeletal and other pain conditions				
Back pain	19,751 (6.4)	890 (11.7)	18,861 (6.3)***	857 (11.3)
Low back pain	74,761 (24.2)	3,005 (39.4)	71,756 (23.8)***	2,987 (39.2)
Neck pain	20,250 (6.6)	1,146 (15.0)	19,104 (6.3)***	1,159 (15.2)
Nontraumatic joint injuries	103,875 (33.6)	2,416 (31.7)	101,459 (33.6)***	2,463 (32.3)
Sprains/strains	9,212 (3.0)	268 (3.5)	8,944 (3.0)**	284 (3.7)
Traumatic joint/muscle/spinal cord injuries[§]	6,993 (2.3)	203 (2.7)	6,790 (2.3)*	183 (2.4)
Rheumatic arthritis/osteoarthritis	47,313 (15.3)	662 (8.7)	44,652 (15.5)***	620 (8.1)
Osteoporosis	3,112 (1.0)	27 (0.4)	3,085 (1.0)***	21 (0.3)
Gout	15,022 (4.9)	107 (1.4)	14,915 (4.9)***	92 (1.2)
Fracture	9,725 (3.1)	188 (2.5)	9537 (3.2)***	193 (2.5)
Fibromyalgia	3,444 (1.1)	146 (1.9)	3,298 (1.1)***	140 (1.8)
Mental health diagnoses				
PTSD	37,501 (12.1)	1,627 (21.4)	35,874 (11.9)***	1,618 (21.2)
Major depressive disorders	15,960 (5.2)	734 (9.6)	15,226 (5.1)***	710 (9.3)
Alcohol use disorders	26,428 (8.6)	794 (10.4)	25,634 (8.5)***	787 (10.3)
Drug use disorders	12,106 (3.9)	456 (6.0)	11,650 (3.9)***	455 (6.0)
Other clinical conditions				
Hypertension	144,904 (46.9)	2,647 (34.7)	142,257 (47.2)***	2,657 (34.9)
Diabetes	59,428 (19.2)	1,043 (13.7)	58,385 (19.4)***	1,124 (14.8)
Coronary artery diseases	39,913 (12.9)	586 (7.7)	39,327 (13.0)***	592 (7.8)
Chronic heart failure	8,759 (2.8)	108 (1.4)	8,651 (2.9)***	102 (1.3)
Conditions NOS[¶]	70,530 (22.8)	1,534 (20.1)	68,996 (22.9)***	1,510 (19.8)
Pain severity and treatment				
Self-rated pain intensity[‖]	3.4±3.2	4.1±3.1	3.4±3.3***	4.0±3.2
Intensity rating ≥4	121,350 (46.2)	3,511(56.6)	117,839 (46.0)***	3,535 (56.9)
Opioid use	66,684 (21.6)	2,056 (27.0)	64,628 (21.4)***	2,059 (27.0)

(continued)

Table 1 Continued

Characteristics by Domain	Overall (N = 309,277)	CIH Recipients† (N = 7,621)	CIH Nonrecipients (N = 301,656)	PS-Matched Nonrecipients‡ (N = 7,621)
	Mean±SD or No. (%)	Mean±SD or No. (%)	Mean±SD or No. (%)	Mean±SD or No. (%)
Cohort/study design features				
Duration of VHA care before study entry, y	3.6±3.9	3.0±3.6	3.6±3.9***	2.9±3.5
Length of follow-up, y	2.1±0.7	2.3±0.5	2.1±0.6***	2.3±0.5
Calendar year of study entry				
Oct–Dec 2010	76,317 (24.7)	2,000 (26.2)	74,317 (24.6)**	1,970 (25.9)
Jan–Sep 2011	232,960 (75.3)	5,621 (73.8)	227,339 (75.4)	5,651 (74.1)
Facility/geographical heterogeneity				
Facility complexity level‖‖‖				
1a	117,868 (38.1)	2,842 (37.3)	115,026 (38.1)***	2,973 (39.0)**
1b	43,960 (14.2)	945 (12.4)	43,015 (14.3)	1,028 (13.5)
1c	63,864 (20.7)	1,831 (24.0)	62,033 (20.6)	1,700 (22.3)
2	47,288 (15.3)	1,038 (13.6)	46,250 (15.3)	971 (12.7)
3	33,911 (11.0)	931 (12.2)	32,980 (10.9)	897 (11.8)
Not designated	2,386 (0.8)	34 (0.5)	2,352 (0.8)	52 (0.7)
VISN†† - V16 region	24,516 (7.9)	237 (3.1)	24,279 (8.1)***	234 (3.1)
Other regions	284,761 (92.1)	7,384 (96.9)	277,377 (91.9)	7,387 (96.9)

After PS-matching, no covariates were statistically significantly different between the two CIH exposure groups (*P* values ranging from 0.06 for diabetes to 0.97 for drug use disorder), except facility complexity level (*P* = 0.006). Percentages may not sum up to 100% due to missing data.

CIH=Complementary and Integrative Heath; PS=Propensity Score; PTSD=post-traumatic stress disorder; NOS=not otherwise specified; VISN=Veteran Integrated Service Networks, consisting of 21 demarcated regions.

*$P < 0.05$,

**$P < 0.01$,

***$P < 0.001$–0.0001, based on the Student *t* test for continuous variables and chi-square test for categorical variables.

†Denotes veterans with one or more visits to a VA clinic for massage, chiropractor, and acupuncture during the three-year follow-up period.

‡Selected using 1:1 greedy matching with a caliper width of 0.2 logit PS, which was estimated using a logistic regression model with CIH visit as a binary dependent variable and 35 covariates and selected interaction terms. See the *Methods* for details.

§Included one or more encounters of traumatic joint damages, traumatic muscle damages, and/or spinal cord damages.

¶Not otherwise specified.

‖Based on pain intensity ratings using a numeric rating scale, with a rating ≥4 denoting moderate to severe pain (range = 0–10).

‖‖‖Complexity levels ranged from most complex (level 1a) to least complex (level 3) based on the primary facility where veterans receive their usual health care and/or the index musculoskeletal diagnosis. This covariate was treated as an ordinal variable (0–4 corresponding to the five designated levels) for standardized difference calculation.

††Demarcates 21 geographic regions in the VA information system, which were represented by dummy indicators in the PS model and dichotomized for standardized difference calculation (all other regions combined vs the most prevalent V16 region).

We used a generalized estimating equation (GEE) to estimate the average treatment effect of CIH on the monthly varying PIRs over three years as a normal outcome [40]. The derived parameter estimate denotes absolute difference (AD) on average PIRs at any given month between the two exposure groups (i.e., CIH recipients vs nonrecipients). AD is a common measure of effect sizes used in CIH trials [13–16], with values <0 favoring the CIH treatment (i.e., reducing pain intensity). The 95% confidence intervals (CIs) of AD were obtained using a robust variance estimator. Within-subject correlations among repeated measures were accounted for using a compound symmetry covariance structure.

We fit series of GEE normal models to obtain an effect estimate (i.e., AD) for CIH "treatment." To provide a reference standard for maximal possible bias, we first fit a baseline model in the full cohort, with CIH exposure as the sole predictor. Next, we introduced covariates into this baseline model, representing a conventional deconfounding approach, in contrast to causal methods that employed a PS-based "resampling" mechanism (i.e., matching or IPTW) to address confounding and selection bias. The covariates included a priori selected risk factors for chronic pain (age, service era, mental health diagnoses, number of MSDs [range = 1–11], and baseline PIRs) and factors capturing pain care

Table 2 Distribution of observed monthly pain intensity ratings over the three-year follow-up period among 309,277 US veterans with chronic musculoskeletal pain

Study Cohort		Veterans with ≥1 PIRs During Follow-up Period			Analytic Sample for Causal Evaluation of CIH Effect		
		No.	No. of PIRs Included	PIRs, Mean ± SD	No.	No. of PIRs Included	PIRs, Mean ± SD
Original cohort							
Overall		307,417*	2,205,885	3.02 ± 3.03	306,720†	2,172,286	3.00 ± 3.30
CIH exposure	No	299,829	2,131,620	2.98 ± 3.30	299,829	2,131,620	2.98 ± 3.30
	Yes	7,588	74,265	4.04 ± 3.29	6,891	40,666	3.83 ± 3.27
PS-matched subcohort							
Overall		15,114	133,043	3.78 ± 3.31	13,726‡	94,252	3.62 ± 3.31
CIH exposure	No	7,557	59,062	3.45 ± 3.31	6,863	53,738	3.46 ± 3.32
	Yes	7,557	73,981	4.04 ± 3.29	6,863	40,514	3.83 ± 3.27

CIH=Complementary and Integrative Health; PIR=pain intensity ratings.
*Excluded 1,860 (0.6%) of the 309,277 veterans in the original cohort who did not have any PIR after the index musculoskeletal diagnoses.
†Excluded 697 (0.2% of the 307,417) CIH recipients whose PIRs occurred exclusively prior to initial CIH visits.
‡Excluded 694 matched pairs (9.1% of the 7,557), in which the CIH recipients did not have any PIRs after initial CIH visits.

heterogeneity over time and across VA facilities (duration under VHA care, calendar year, length of follow-up, VHA facility complexity, and VISNs). In addition, we added opioid use during each year as a time-varying covariate to account for confounding by concomitant analgesic medications, the current firstline treatment for chronic pain [1–3].

Next, we performed causal analyses by applying different PS methods to the baseline GEE model. First, we performed GEE analyses among the PS-matched subcohort while accounting for each matched pair as a cluster. To maintain validity of the matching, 697 broken pairs (9.1% of the original matched sample) were excluded due to lack of follow-up PIRs in either party, resulting in 6,863 intact matched pairs (see Table 2 for details). Second, we fit weighted GEE models with IPTW-P or IPTW-T as a weight, respectively, and derived corresponding ATE and ATT estimates. Finally, we performed a hybrid analysis by refitting an IPTW-T weighted GEE model among the PS-matched subcohort, while accounting for the matching cluster. Recent studies have suggested that combining more than one PS method (e.g., stratification and weighting) enhances the validity and generalizability of the causal inference [38,41]. To quantify the capacity of different causal methods in removing potential selection and/or confounding bias, we calculated percent changes (reductions) in the ADs from the baseline model to each causal model, as well as to the conventional covariate-adjusted model.

To provide insights into the specificity of causal methods in terms of confounding control, we added the same set of covariates as the conventional adjusted model into each causal model, despite potential "overadjustment." Percent changes in the ADs before and after adjusting for covariates were used to judge the necessity (vs redundancy) of confounding adjustment.

Because most randomized trials examined CIH efficacy over three to 12 months, to determine whether our study may underestimate or "dilute" potential CIH benefits due to extended follow-up time [14,42], we performed sensitivity analyses by restricting the follow-up period to one year (≤365 days) after the defined "treatment" time zero. Veterans who did not have a PIR during the first 365 days after the initial CIH visit (218 CIH recipients) or the index MSD date (2,546 nonrecipients) were excluded, resulting in a one-year causal analysis sample of 6,673 CIH recipients and 297,283 nonrecipients (total N = 303,956). We expected that the predicted CIH effects over this shorter follow-up period would be closer to those reported from the randomized CIH trials.

To provide further clinical insight, we explored the potential "dose" response of chronic pain to the cumulative number of the 3 CIH modalities (i.e., acupuncture, massage, and chiropractic care) each veteran received during the follow-up period as a categorical exposure (range = 0–3), acknowledging the imprecise nature of this measurement. We calculated average PIRs over one year for each CIH exposure category (i.e., 0, 1, 2, and 3) and compared the AD estimates derived from the conventional baseline model (presumably the most biased) and the hybrid causal model (presumably the least biased) using the GEE approach.

All the statistical analyses were conducted using SAS software version 9.4 (SAS Institute, Cary, NC, USA). Model fit was examined using Quasi-Information Criterion and residual plots. The hypotheses were tested at a two-sided significance level of α = 0.05.

Results

As shown in Table 1, the study cohort had an average age of 57 years, with the majority being male (92%), white (74%), and married (54%). The most prevalent MSD diagnoses were nontraumatic joint injuries (34%) and low back pain (24%). Nearly 46% veterans had PIRs ≥4, and 26% had an active opioid prescription.

During the follow-up period, 7,621 (2.5%) veterans had one or more CIH visits, with 6,015 (1.9%) having had CIH visits in one year, 1,336 (0.4%) in two years, and 270 (0.09%) in all three years. The median (IQR) time to the initial CIH visit was 7.8 (2.3–17.1) months.

As shown in Figure 2, in the original cohort, 21 (60%) of the 35 PS covariates had an STD ≥0.10 (average STD = 0.17). After applying causal methods, all STDs were <0.10, with average STDs of 0.01 for PS matching, 0.02 for IPTW-P, and 0.003 for IPTW-T.

Table 2 presents sample sizes for causal analyses and observed PIRs. The mean (± SD) ratings of CIH recipients (3.83 ± 3.27) were higher than nonrecipients in both the full cohort (2.98 ± 3.30) and the PS-matched subcohort (3.46 ± 3.32).

Figure 3 presents AD estimates for average treatment effect of CIH from different methods over three years (mean duration = 25.6 ± 7.7 months). The baseline model estimated an AD of 0.83 (95% CI = 0.77 to 0.89), indicating that on average CIH recipients had a 0.83-point higher (worse) PIR than nonrecipients. Conventional covariate adjustment brought down this apparent unfavorable effect size to 0.46 (95% CI = 0.41 to 0.51). The AD reduced to 0.36 (95% CI = 0.29 to 0.44) after PS matching, 0.61 (95% CI = 0.51 to 0.71) with IPTW-P, and 0.35 (95% CI = 0.30 to 0.41) with IPTW-T weighting, respectively, and diminished to null when refitting the IPTW-T model among the PS-matched subcohort (0.004, 95% CI = –0.09 to 0.10, $P = 0.942$). The percent reduction in ADs from the baseline model was largest in the hybrid model (83%) and smallest in the IPTW-P model (22%).

Adjusting for covariates inflated AD estimates (i.e., more "biased") by 12% to 71 folds in the PS-matched (0.43, 95% CI = 0.36 to 0.50), the IPTW-T (0.40, 95% CI = 0 0.34 to 0.45), and the hybrid (0.25, 95% CI = 0.16 to 0.34) models, yet reduced AD (i.e, less "biased") by 7% (0.57, 95% CI = 0.49 to 0.65) in the IPTW-P model.

Sensitivity analyses restricting the "treatment" course to one year (mean duration = 7.6 ± 3.9 months) derived consistent results. The baseline AD estimate (0.78, 95% CI = 072 to 0.84) decreased to 0.31 (95% CI = 0.05 to 0.22) after PS matching, 0.56 (95% CI = 0.06 to 0.45) and 0.28 (95% CI = 0.03 to 0.22) with IPTW-P and IPTW-T weighting, respectively, and diminished to null (–0.04, 95% CI = –0.16 to 0.06, $P = 0.557$) in the hybrid model (Supplementary Data - Figure 2). Covariate adjustment suggested a similar "counterproductive" impact, with 3% to 59-fold AD inflation (i.e., more "biased") across different causal models, but it had almost no impact on the conventional method (0.78, 95% CI = 071 to 0.84).

The use of three modalities as a proxy to CIH "dose" suggested that the average PIRs tended to increase as the CIH "dose" increased. In comparison with the baseline model, the hybrid causal analysis successfully diminished the unfavorable AD estimates of all the three exposure categories to null, and even reversed the direction to protective for those who received only one CIH modality (–0.06, 95% CI = –0.18 to 0.07) (Supplementary Data - Table 2).

Discussion

Our study suggests that in this large cohort of US veterans the course of chronic pain appears no better for those who received one or more courses of acupuncture, massage, or chiropractor care than those who never received such modalities during the three-year follow-up period. Application of causal methods successfully balanced the two exposure groups in terms of baseline confounding, yet it did not detect a significant difference in the self-rated pain intensity outcome between the two "treatment" groups, as we hypothesized based on randomized trial evidence [14–17]. There might be several reasons for this.

First, given their "nonmainstream" nature, CIH therapies may have been received by some veterans but never documented by their clinicians. Consequently, the average pain intensity of the non-CIH group would be contaminated with potential treatment benefit from those misclassified CIH recipients. Second, the composite measure of the three different CIH approaches may hide the potential benefits of individual therapy [17]. According to a frequently cited meta-analysis, the evidence for acupuncture and massage was relatively weaker than that for cognitive-behavioral therapy, exercise, and spinal manipulation [16]. Alternatively, the pain symptoms of the CIH recipients may be too severe or enduring, which would have driven these patients to seek CIH in the first place after failing pharmacological treatments. As a result, the estimated CIH effect would reflect a "reverse causality" bias, which may explain, at least in part, the apparent unfavorable effect estimates from each individual causal model, as well as the seemingly counterintuitive (though not statistically significant) effect estimates for receiving two or three CIH modalities. Third, this large veteran cohort may be characteristically different from the small and selective samples of randomized trials that have reported CIH benefits. Finally, it is also possible that the "indifferential" effect
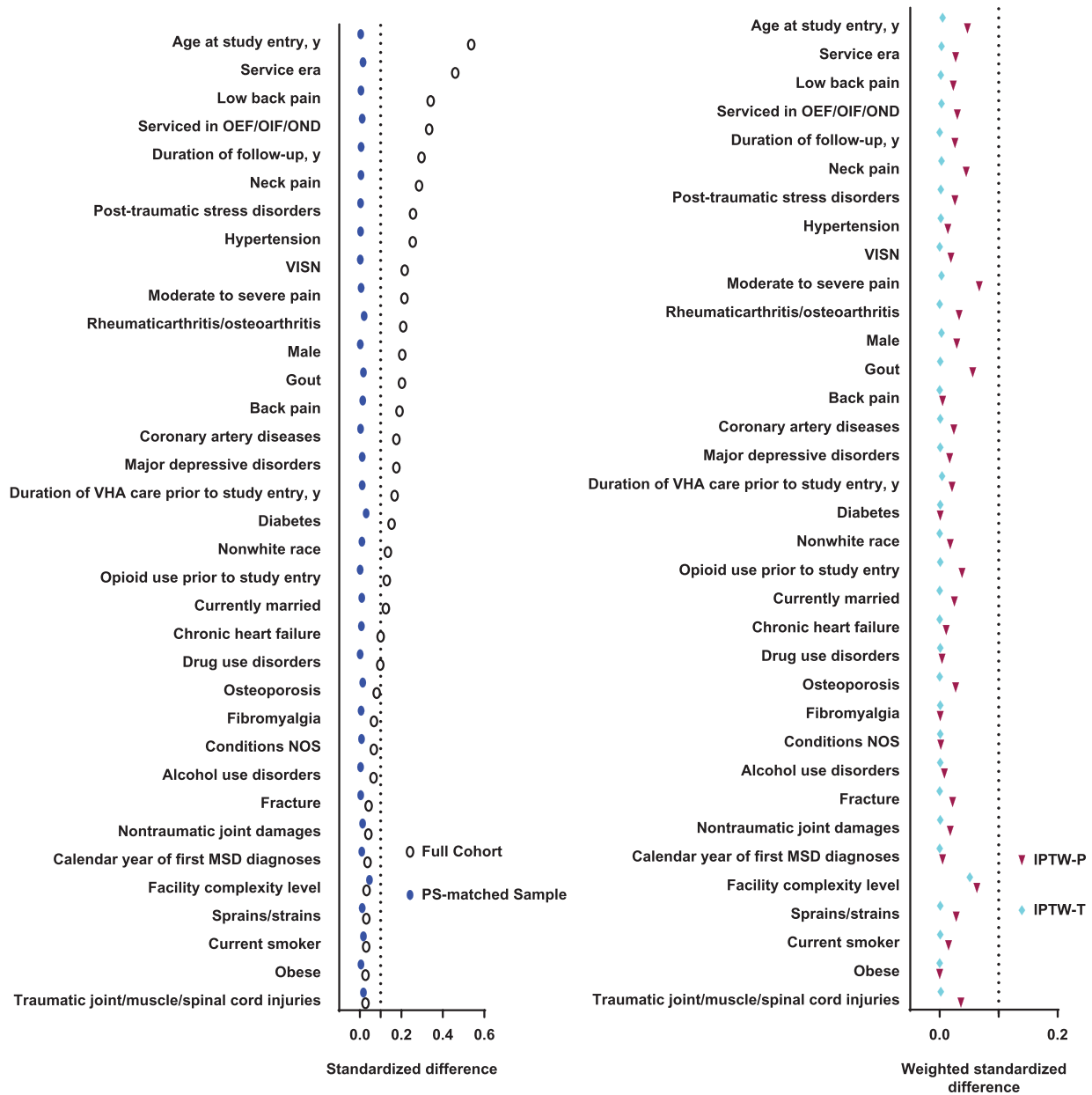
Figure 2 Standardized difference (STD) of 35 measured covariates between exposure groups to complementary health (CIH) approaches according to different causal methods. (A) STD in original full cohort and the Propensity Score–matched subcohort, respectively (B). Weighted STD by inverse probability of treatment weights in the population (IPTW-P) and the treated (IPTW-T), respectively. (A) Twenty-one (60.0% of 35) covariates were unbalanced (STD > 0.10) in the original full cohort, with an average STD of 0.16. After Propensity Score matching, all the 35 covariates were balanced, with an average STD of 0.01. (B) All the 35 covariates were balanced after IPTW-P or IPTW-T weighting, with an average STD of 0.02 for IPTW-P and 0.003 for IPTW-T.

estimate observed in this study may actually reveal the clinical truth that such CIH therapies may not have an overall benefit as expected, at least as delivered in current practice. Indeed, the effect sizes reported from randomized trials are generally small to moderate, with multiple negative findings [13–17]. Additional studies are needed to determine whether CIH approaches may be effective for certain types of chronic pain or for subgroups of patients with unique clinical characteristics.

From a methodological perspective, our results agree with the existing literature in several aspects. We
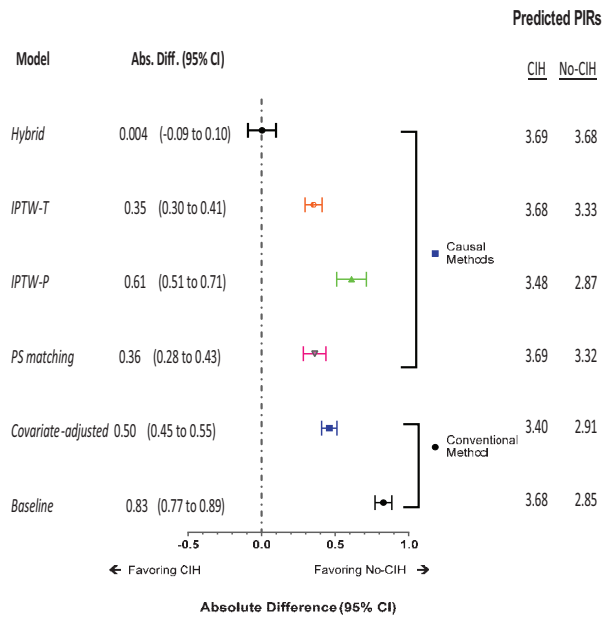
Figure 3 Estimated treatment effect of receiving complementary and integrative health approaches (CIH) on monthly pain intensity ratings (PIRs) over three years among 306,720 US veterans with chronic musculoskeletal pain. The absolute difference (AD; 95% confidence interval [CI]) represents the absolute difference (95% CIs) in average PIRs between the two CIH exposure groups, estimated using a generalized estimating equation (GEE) normal model. Bars represent upper and lower boundaries of the 95% confidence interval of AD. An AD >0 indicates an unfavorable CIH effect (i.e., increasing pain intensity), whereas an AD <0 indicates a favorable or beneficial CIH effect (i.e., decreasing pain intensity). The conventional method was applied to the original full cohort and included a baseline model (with CIH exposure as the sole predictor) and a covariate-adjusted model (for an array of selected covariates, as described in the *Methods*). Under the causal method, the Propensity Score (PS)–matching analysis refit the baseline model among the PS-matched subcohort while counting each matched pair as a cluster. The two inverse probability of treatment weighting (IPTW) analyses fit a weighted baseline model in the full cohort, with IPTW in the population (IPTW-P) or IPTW in the treated (IPTW-T) as a weight, respectively. The hybrid analysis fit a weighted baseline model in the PS-matched subcohort with the IPTW-T as a weight while accounting for matching.

demonstrated that both the PS matching and IPTW methods have the desired properties to balance exposure groups in terms of measured baseline confounding that jointly determine the likelihood of receiving CIH therapies [30–34,39,41]. In addition, we found the ATT to be a less biased and more robust causal estimator than

the ATE, confirming previous observations and recommendations [39]. Furthermore, we demonstrated that although direct confounding adjustment helped remove confounding bias in conventional outcome modeling, it may be redundant or counterproductive to the PS-based causal methods that had already "removed" baseline confounding for treatment selection. This observation somehow echoes a randomized trial that covariate adjustment in the outcome model becomes essentially unnecessary after randomization of treatment arms. Finally, our hybrid approach integrating IPTW with PS matching extends the current empirical literature and supports the recommendation that combining more than one propensity method may help maximize bias reduction and facilitate causal inference [38,41].

The strengths of our study include a large study population and an extended follow-up period for pain outcome assessments, which represents the natural history of chronic pain more closely while augmenting statistical power and efficiency for examining a rare exposure under alternative hypotheses. In addition, the use of a longitudinal statistical model to estimate an average treatment effect across repeated outcome assessments, instead of a single post-treatment end point, adds fresh evidence to the empirical causal investigation literature. Finally, by assembling a virtual inception cohort of "incident" MSD cases and accounting for a large array of potential confounding factors in the analyses, the validity and clinical implication of the study findings are enhanced.

The limitations of the study include potential misclassification of CIH exposure and lack of data on the doses, frequency, and duration of the treatment. As a result, our causal analyses were carried out under an assumption that veterans who were exposed to CIH would have complied to doctors' prescriptions and actually treated for one or more courses of such modalities over a period of time. Second, our PS model was estimated using an administrative database that did not include all potentially important covariates, such as indications for CIH treatment, history of previous analgesic treatment, and psychological and behavioral factors. This unmeasured confounding problem is inherent to all non-experimental studies and may compromise the validity and precision of PS-based causal inference [20–23,26,27]. We also did not track subsequent episodes of a new MSD diagnosis. Finally, our estimated causal treatment effect was based on comparison between two "resampled" treatment groups that were hypothetically most representative and comparable under counterfactual causation assumptions, for which criticisms and debates remain [43]. Therefore, the results reported here should be interpreted cautiously in a broader clinical and research context [43,44] and replicated using different analytical methods and different study populations.

To conclude, we evaluated CIH approaches among >300,000 veterans with musculoskeletal pain over three

years and did not find a significant difference in self-rated pain intensity between those receiving and not receiving the approaches. PS-based causal methods were able to balance the measured baseline confounding between the exposure groups, yet they failed to "recover" the potential CIH benefit reported from some randomized trials. Given the inadequate evidence for CIH effectiveness for chronic pain and the mounting concern over side effects of chronic opioid treatment, future studies using innovative and rigorous research methodologies and more comprehensive and precise data on CIH "dose," both randomized and observational, are warranted.

Acknowledgments

Supplementary Data

Supplementary Data may be found online at http://pain-medicine.oxfordjournals.org.

References

1 Institute of Medicine. Relieving Pain in America: A Blueprint for Transforming Pain Prevention, Care, Education and Research. Washington, DC: The National Academies Press; 2011.

2 Office of the Assistant Secretary for Health, US Department of Health and Human Services. National Pain Strategy; 2016.

3 Veterans Health Administration. VA Pain Management Directive (2009-053). Washington, DC: Department of Veterans Affairs; 2009.

4 Lew HL, Otis JD, Tun C, et al. Prevalence of chronic pain, posttraumatic stress disorder, and persistent postconcussive symptoms in OIF/OEF veterans: Polytrauma clinical triad. J Rehabilit Res Develop 2009;46(6):697–702.

5 Kerns RD, Philip EJ, Lee AW, Rosenberger PH. Implementation of the Veterans Health Administration national pain management strategy. Translat Behav Med 2011;1(4):635–43.

6 Goulet JL, Kerns RD, Bair M, et al. The musculoskeletal diagnosis cohort: Examining pain and pain care among veterans. Pain 2016;157(8):1696–703.

7 Sinnott P, Wagner TH. Low back pain in VA users. Arch Intern Med 2009;169(14):1338–9.

8 Veterans Health Administration. Analysis of VA Health Care Utilization Among Operation Enduring Freedom (OEF), Operation Iraqi Freedom (OIF), and Operation New Dawn (OND) Veterans. Washington, DC: Department of Veterans Affairs; 2013.

9 Kerns RD, Otis J, Rosenberg R, Reid MC. Veterans' reports of pain and associations with ratings of health, health-risk behaviors, affective distress, and use of the healthcare system. J Rehabilit Res Develop 2003;40(5):371–9.

10 Edlund MJ, Sullivan M, Steffick D, Harris KM, Wells KB. Do users of regularly prescribed opioids have higher rates of substance abuse problems than non-users? Pain Med 2007;8(8):647–56. [PubMed: 18028043]

11 Otis JD, Keane TM, Kerns RD, Monson C, Scioli E. The development of an integrated treatment for veterans with comorbid chronic pain and posttraumatic stress disorder. Pain Med 2009;10 (7):1300–11.

12 Von Korff M, Moore JC. Stepped care for back pain: Activating approaches for primary care. Ann Intern Med 2001;134(9 Pt 2):911–7.

13 Patel M, Gutzwiller F, Paccaud F, Marazzi A. A meta-analysis of acupuncture for chronic pain. Int J Epidemiol 1989;18(4):900–6.

14 Furlan AD, Yazdi F, Tsertsvadze A, et al. A systematic review and meta-analysis of efficacy, cost-effectiveness, and safety of selected complementary and alternative medicine for neck and low-back pain. Evid Based Complement Alternat Med 2012; 2012:1.

15 Vickers AJ, Cronin AM, Maschino AC, et al; Acupuncture Trialists' Collaboration. Acupuncture for chronic pain: Individual patient data meta-analysis. Arch Intern Med 2012;172(19):1444–53.

16 Chou R, Huffman LH; American Pain Society; American College of Physicians. Nonpharmacologic therapies for acute and chronic low back pain: A review of the evidence for an American Pain Society/American College of Physicians clinical practice guideline. Ann Intern Med 2007;147 (7):492–504.

17 Nahin RL, Boineau R, Khalsa PS, Stussman BJ, Weber WJ. Evidence-based evaluation of complementary health approaches for pain management in the United States. Mayo Clin Proceed 2016;91 (9):1292–306.

18 Rosenbaum PR, Rubin DB. The central role of the Propensity Score in observational studies for causal effects. Biometrika 1983;70(1):41–55.

19 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 (19):2265–81.

20 Glynn RJ, Schneeweiss S, Sturmer T. Indications for Propensity Scores and review of their use in pharmacoepidemiology. Basic Clin Pharmacol Toxicol 2006;98(3):253–9.

21 Sturmer T, Joshi M, Glynn RJ, et al. A review of the application of Propensity Score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol 2006;59(5):437.e1–47.

22 Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: A systematic review. J Clin Epidemiol 2005;58 (6):550–9.

23 Han L, Kim N, Brandt C, Allore A. Antidepressant use and cognitive deficits in older men: Addressing confounding by indications using different methods. Ann Epidemiol 2012;22(1):9–16.

24 Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated Propensity Score. Econometrica 2003;71 (4):1161–89.

25 Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: A review. Review Econom Stat 2004;86(1):4–29.

26 Rubin DB. The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. Stat Med 2007;26 (1):20–36.

27 Rubin DB. For objective causal inference, design trumps analysis. Ann Appl Stat 2008;2(3):808–40.

28 Sturmer T, Wyss R, Glynn RJ, Brookhart MA. Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. J Intern Med 2014; 275(6):570–80.

29 Harder VS1, Stuart EA, Anthony JC. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychol Methods 2010;15 (3):234–49.

30 Austin PC. The relative ability of different Propensity Score methods to balance measured covariates

between treated and untreated subjects in observational studies. Med Decis Making 2009;29 (6):661–77.

31 Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the Propensity Score to estimate causal treatment effects in observational studies. Stat Med 2015;34(28):3661–79.

32 Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. Stat Med 2010;29 (20):2137–48.

33 Johnson ES, Dickerson JF, Vollmer WM, et al. The feasibility of matching on a Propensity Score for acupuncture in a prospective cohort study of patients with chronic pain. BMC Med Res Methodol 2017;17(1):42.

34 Weeks WB, Tosteson TD, Whedon JM, et al. 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. J Manipulative Physiol Ther 2015;38(9):620–8.

35 Garrido MM. Propensity scores: A practical method for assessing treatment effects in pain and symptom management research. J Pain Symptom Manage 2014;48(4):711–8.

36 Carney BT, West P, Neily J, Mills PD, Bagian JP. The effect of facility complexity on perceptions of safety climate in the operating room: size matters. Am J Med Qual 2010;25(6):457–61.

37 Faries DE, Leon AC, Haro JM, Obenchain RL. Analysis of Observational Health Care Data Using SAS. Cary, NC, USA: SAS Institute; 2010.

38 Colson KE, Rudolph KE, Zimmerman SC, et al. Optimizing matching and analysis combinations for estimating causal effects. Sci Rep 2016;6(1):23222.

39 Hajage D, Tubach F, Steg PG, Bhatt DL, De Rycke Y. On the use of Propensity Scores in case of rare exposure. BMC Med Res Methodol 2016;16(1):38.

40 McCulloch CE, Searle SR. Generalized, Linear and Mixed Models. New York: John Wiley & Sons; 2001.

41 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(6):1065–71.

42 Clarke R, Shipley M, Lewington S, et al. Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. Am J Epidemiol 1999;150(4):341–53.

43 Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: The need for a pluralistic approach. Int J Epidemiol 2016;45(6):1776–86.

44 Petersen ML, van der Laan MJ. Causal models and learning from data integrating causal modeling and statistical estimation. Epidemiology 2014;25 (3):418–26.