Title

Toward a better understanding of when to apply propensity scoring: a comparison with conventional regression in ethnic disparities research

Permalink

https://escholarship.org/uc/item/13h1m92f

Journal

Annals of Epidemiology, 22(10)

ISSN

1047-2797

Authors

Ye, Yu
Bond, Jason C
Schmidt, Laura A
et al.

Publication Date

2012-10-01

DOI

10.1016/j.annepidem.2012.07.008

Peer reviewed

Annals of Epidemiology 22 (2012) 691–697

Contents lists available at SciVerse ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Toward a better understanding of when to apply propensity scoring: a comparison with conventional regression in ethnic disparities research

Yu Ye MA [a,*], Jason C. Bond PhD [a], Laura A. Schmidt PhD, MSW, MPH [b], Nina Mulia DrPH [a], Tammy W. Tam PhD [a]

[a] *Alcohol Research Group, Public Health Institute, Emeryville, CA*
[b] *Philip R. Lee Institute for Health Policy Studies and Department of Anthropology, History and Social Medicine, University of California, San Francisco, San Francisco, CA*

ARTICLE INFO

Article history:
Received 20 January 2012
Accepted 20 July 2012
Available online 14 August 2012

Keywords:
Propensity scores
Multivariable regression
Marginal effect
Confounding
Effect modification
Reference sample
Disparity

ABSTRACT

Purpose: Despite growing popularity of propensity score (PS) methods used in ethnic disparities studies, many researchers lack clear understanding of when to use PS in place of conventional regression models. One such scenario is presented here: When the relationship between ethnicity and primary care utilization is confounded with and modified by socioeconomic status. Here, standard regression fails to produce an overall disparity estimate, whereas PS methods can through the choice of a reference sample (RS) to which the effect estimate is generalized.

Methods: Using data from the National Alcohol Surveys, ethnic disparities between White and Hispanics in access to primary care were estimated using PS methods (PS stratification and weighting), standard logistic regression, and the marginal effects from logistic regression models incorporating effect modification.

Results: Whites, Hispanics, and combined White/Hispanic samples were used separately as the RS. Two strategies utilizing PS generated disparities estimates different from those from standard logistic regression, but similar to marginal odd ratios from logistic regression with ethnicity by covariate interactions included in the model.

Conclusions: When effect modification is present, PS estimates are comparable with marginal estimates from regression models incorporating effect modification. The estimation process requires a priori hypotheses to guide selection of the RS.

Introduction

Epidemiologic studies investigating racial/ethnic disparities in health have grown exponentially over the past few decades [1], in conjunction with the Institute of Medicine's groundbreaking report on disparities in health care [2] and U.S. national objectives put forth in *Healthy People 2000–2020* [3–5]. Until recently, disparities research relied heavily on conventional regression modeling to document inequalities in health and health care across racial/ethnic groups [1]. One common problem with regression, however, is that the relationship between ethnicity and the health outcomes of interest are often confounded with, and modified by, socioeconomic status [6]. This study addresses why and how newer methods based on propensity scores (PS) are particularly well-suited for disparities research, although they are relevant to any area of epidemiologic research where effect modification is similarly of concern.

There is lively debate over when PS methods offer benefits over standard multivariable linear or logistic regression typically used in disparities research [7–10]. Although there is empirical support for PS's advantages [11,12], in practice, PS methods often seem to generate results quite similar to those from multivariable regression [13,14]. This paper is preoccupied with one common scenario in which PS methods should theoretically produce more interpretable effect estimates than those from multivariable regression. This occurs when the distribution of one or more confounding covariates (e.g., socioeconomic status [SES]), as well as the relationship between these confounding covariates and the health outcome, varies across the ethnic groups being compared. Entering relevant interaction terms addresses the effect modification, but the regression model can only produce disparity estimates at specific SES levels and fails to generate a single overall disparity estimate, which, in most practical cases, is desired. Without including interactions, disparity estimates will in general be biased.

Supported by grants from the U.S. National Institute on Alcohol Abuse and Alcoholism R01 AA017197 and P50 AA005595.

[*] Corresponding author. 6475 Christie Avenue, Suite 400, Emeryville, CA 94608.
E-mail address: yye@arg.org (Y. Ye).

1047-2797/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.annepidem.2012.07.008

With control variables in the model, standard regression generates *conditional* effect estimates. In contrast, PS methods produce marginal effect estimates that can be interpreted counterfactually. In the language of causal inference, which often focuses on the effect of a treatment or of exposure to a risk factor, PS methods estimate the difference in an outcome for a given population when it is treated/exposed, and the outcome for the same population when it is not treated/exposed. PS methods thus estimate the marginal effect by design, largely through specifying a reference sample (RS) to which the effect estimate is generalized. The RS often used is either the "treated" ("risk-exposed") sample or the sample combining those who were and were not treated/risk-exposed [15–17]. With regression generating conditional effect estimates and PS methods producing marginal estimates, questions have been raised regarding the comparability of the two approaches in nonlinear models [10,18]. Specifically, for a dichotomous outcome (as in the present study), it was suggested that marginal odds ratios (ORs) should be generated from both PS and logistic regression for the two methods to be comparable [19,20]. Little is known about the comparability of the two approaches, however, if effect modification is present.

In the current study of ethnic disparities, we compare estimates from PS approaches generated using various RSs to both conditional and marginal ORs produced from ordinary logistic regression. We show how marginal ORs from logistic regression incorporating effect modification can be produced and comparable with PS estimates. Although some recent work has examined the performance of PS methods under effect modification [21–23], this study compares substantive results obtained from PS and logistic regression models; it also shows how varying the specification of the RS can influence results.

Materials and methods

Dataset and measures

Our empirical analysis focuses on the timely issue of racial/ethnic disparities between Whites and Hispanic Americans in access to primary care interventions for alcohol problems [24]. Data come from the combined 2000 and 2005 U.S. National Alcohol Surveys (NAS), two comparable probability samples of U.S. adults collected using computer-assisted telephone interviews via random digit dialing [25]. Included in this analysis are "at-risk" drinkers who meet the National Institute on Alcohol Abuse and Alcoholism drinking guidelines defined as men/women drinking more than 4/3 drinks on any day or more than 14/7 drinks per week [26]. Self-identified ethnicity, based on the current U.S. Census definition, is the key risk-exposure variable in this analysis. Only those self-identifying as White ($N = 2798$) or Hispanic ($N = 684$) were retained for analysis, because the greatest disparities have been found between these two groups [24,27,28]. The outcome of interest is the subject's report of whether she or he had one or more primary care visits in the prior year. Any visit with a private doctor, clinic, or non-emergency medical setting during the prior year counted as a primary care visit. Demographic and SES covariates were used as potential confounders and effect modifiers of interest, including gender, age, education, annual household income, and health insurance coverage.

Statistical analysis

PS stratification and weighting

A PS is defined as the estimated probability of being Hispanic versus White and is modeled as a function of gender, age, education, annual household income, and health insurance coverage using logistic regression. Two approaches utilizing the estimated PS to estimate ethnic disparities are considered in the present study, PS stratification and PS weighting.

PS-stratification [15,16] classifies subjects into five strata (quintiles) using their estimated PS. This process is repeated several times, using "Hispanics," "Whites," and the "combined White and Hispanic sample" separately as the RS. When Hispanics are treated as the RS, the Hispanic sample is divided into five equal-sized strata based on the sorted PS distribution within the Hispanic sample. Using the PS thresholds for quintile generation, the comparison White sample is then divided into five groups, presumably of unequal sizes given that the distribution of PS is different between Whites and Hispanics. This is done analogously for Whites as the RS. When the combined Whites/Hispanic sample is treated as the RS, the pooled sample is divided into five equal-sized strata based on the sorted PS distribution for the total sample. The overall response probability, for each choice of RS, is simply a weighted average of stratum-specific response probabilities. Because the RS is divided into five equal size strata in the current design, equal weights are assigned across the five strata and the overall response probability is the simple average of the stratum-specific probabilities. The marginal OR estimate is then derived using the overall response probabilities (see the Appendix for details as well as [19]).

For PS weighting [29], subjects in the two ethnic comparison groups are weighted based on their estimated PS to construct a 'pseudo-population' in which confounders are no longer associated with ethnicity. When using either the White or the Hispanic group as the RS, "standardized mortality ratio" weights are used that assign a weight of 1 to the group chosen as the RS and the propensity odds $\widehat{e}(X)/(1-\widehat{e}(X)$ (where $\widehat{e}(X)$ is the estimated PS) to the non-RS. When using the combined sample as the RS, the inverse-probability-of-treatment-weighted estimator must be used and derives the effect estimate by using the inverse of the PS $(1/\widehat{e}(X))$ as weights for one group and the inverse of 1 minus the PS $(1/(1-\widehat{e}(X))$ for the other. The marginal ORs are estimated by fitting logistic regression using the ethnic indicator as predictor, with the PS weights used as sampling weights in the model estimation [29].

Standardization and RS

Standardization is the traditional approach in epidemiology to obtain an overall effect estimate when a confounder modifies the relationship between the outcome and a risk factor. Standardization takes a weighted average of the stratum-specific rates or risks, with the strata defined by the effect modifier and with weights corresponding to the number of persons in the RS falling into each category of the effect modifier [30]. As noted, disparity estimates from the PS stratification method are created from a weighted average of stratum-specific estimates, which is essentially a standardization process where the PS serves as the effect modifier (see an illustration in [29] Appendix 1). Sato and Matsuyama [31] also show that PS weighting allows for nonparametric standardization using either the exposure group or the total combined groups as the RS.

Conditional and marginal ORs from logistic regression

We first fit a standard logistic regression model, which generates the conditional OR for the ethnic disparity estimate. Using the fitted model, two predicted response probabilities are generated for each individual by plugging in the actual values of their observed covariates (other than ethnicity) into the regression model. Regardless of an individual's actual ethnicity, the first predicted response probability is generated by assuming the individual is White, the second by assuming he/she is Hispanic. Using these two sets of estimated response probabilities for each individual, the separate "White"/"Hispanic" marginal response probabilities are estimated as the simple average across all respondents (thus using the total combined sample as the RS) where everyone was assumed to be

White, and (separately) Hispanic. Finally, the marginal OR is then derived using these two marginal response probabilities. Alternatively, the marginal response probabilities, as well as the corresponding marginal ORs, can also be generated by averaging the individual predicted probabilities across the White or Hispanic sample only, thus using the White or Hispanic sample as the RS. See the Appendix for details and also see [19].

Marginal ORs from logistic regression incorporating effect modification

Here, interaction terms are explicitly built into the regression model to generate ethnic-specific coefficients. Individual predicted response probabilities for the White and Hispanic effects, marginal response probabilities, and the corresponding marginal ORs are all derived in a similar manner to that described above using standard logistic regression models. More details for this procedure can be found in the Appendix. Unlike the above procedure, the presence of effect modification (e.g., SES with ethnicity) requires the appropriate covariate and coefficient values to be used in the construction of marginal probabilities. Considering a White female case, her estimated counterfactual Hispanic outcome would be obtained by plugging in her actual SES into the regression but using the estimated Hispanic SES coefficient (rather than that estimated for Whites). The marginal response probability thus derived can then be interpreted as the predicted outcome for Hispanics had they the same covariate distribution as Whites.

Results

We first examined differences in the distributions of demographic and SES covariates as well as effect modification between White and Hispanic at-risk drinkers. Table 1 shows pronounced differences in sociodemographic estimates between the two ethnic groups, with Hispanics more likely to be male, younger, have lower education and income levels, and no health insurance compared with Whites. The right-hand panel of Table 1 also shows the percentages of White and Hispanic at-risk drinkers receiving any primary care across the levels of each covariate. The magnitude of the ethnic difference in outcome rates is smaller among those with higher income and education vs. those with lower levels. The presence of effect modification was also indicated by the large differences in the magnitude of the χ^2 statistics across the two ethnic groups.

We examined the common support region and evaluated the balancing quality of the two PS approaches. The estimated PS, the probability of being Hispanics, ranged from 0.039 to 0.823 for the Hispanic sample, and from 0.024 to 0.823 for Whites. The nonoverlapping region was very small and sensitivity analysis restricting to the common support generated similar results. Overall, PS weighting performed better than PS stratification in balancing the covariates. Of the 21 dummy variables from the demographic and SES measures, and assessed separately for Hispanics, Whites, and the combined sample as the RS (63 tests altogether), no significant differences were found between Whites and Hispanics using PS weighting for any of the covariates. All comparisons had $P > .10$ and only 2 out of 63 had $P < .20$. In contrast, using PS stratification assessing balancing for all covariates separately across the five strata, about 9% to 13% of comparisons showed significant differences at $P < .05$ for the three RSs, with the 5th stratum (i.e., those with the largest PS) producing the majority of the unbalanced covariates.

Table 2 shows the effect estimates, both overall and within each stratum, using the PS stratification method. With Hispanics used as the RS, shown in the top third of Table 2, as expected, the five strata each contained a similar number of Hispanics. The middle of Table 2 shows results when Whites are used as the RS, for which the strata contain a similar proportion of Whites and analogously for the combined White/Hispanic groups used as the RS as shown in the bottom third of Table 2. In general, larger ORs, and hence larger disparity estimates, were observed for the strata with higher PS estimates for all three choices of the RS. However, with Hispanics as

Table 1
Distribution of covariates across White and Hispanic at-risk drinkers and their association with primary care use within ethnic groups

	Distribution of covariates (%)		% Receiving any primary care	
	White (N = 2798)	Hispanics (N = 684)	White (N = 2798)	Hispanics (N = 684)
Gender	χ^2 (1) = 26.9***		χ^2 (1) = 1.2†	χ^2 (1) = 7.3**
Male	52.4	63.5	33.1	15.0
Female	47.6	36.5	37.3	23.2
Age	χ^2 (3) = 108.1***		χ^2 (3) = 0.7†	χ^2 (3) = 2.3
18–29	26.7	45.9	32.7	15.9
30–49	51.4	42.8	36.4	20.1
50–64	16.5	9.5	35.5	16.9
≥65	5.4	1.8	34.4	25.0
Education	χ^2 (3) = 402.8***		χ^2 (3) = 3.9†	χ^2 (3) = 30.6***
<HS graduate	5.3	28.9	33.3	8.1
HS graduate	25.4	30.7	29.6	15.7
Some college	30.4	23.2	35.6	23.9
College graduate	39.0	17.1	38.6	30.8
Income ($)	χ^2 (6) = 233.7***		χ^2 (6) = 4.2†	χ^2 (6) = 27.3***
≤10,000	6.8	17.0	30.9	10.3
10,001–20,000	8.5	21.2	27.8	13.8
20,001–30,000	11.1	14.5	34.4	12.1
30,001–40,000	12.1	11.3	31.1	22.5
40,001–60,000	19.1	11.3	39.1	32.5
>60,000	34.6	15.8	37.7	26.9
Missing	7.9	9.1	32.7	12.9
Insurance	χ^2 (4) = 304.6***		χ^2 (4) = 5.3†	χ^2 (4) = 33.3***
None	10.0	35.7	25.7	8.2
Private	74.5	48.7	36.5	26.4
Medicaid	1.8	2.8	41.2	10.5
Medicare/federal	9.7	7.0	38.2	14.6
Other	3.4	5.8	22.5	15.0

**$P < .01$; ***$P < .001$.

† χ^2 test statistics for Whites are derived by normalizing sample size of Whites with that of Hispanics (N = 684).

Table 2
Estimated proportions and odds ratios (ORs) of any primary care use for Whites and Hispanics, for strata and combined overall across strata for the propensity score stratification method, using Hispanics, Whites, and the combined sample as reference samples

	Ethnicity	N	Mean (SE)	Difference (SE)	OR (Whites vs. Hispanics)
Propensity score stratification, Hispanics as reference sample: Stratum (from lowest to highest propensity for being Hispanic)					
1	White	1470	0.381 (0.013)	0.022 (0.042)	1.10
	Hispanics	142	0.359 (0.040)		
2	White	789	0.346 (0.017)	0.128 (0.040)**	1.90**
	Hispanics	133	0.218 (0.036)		
3	White	321	0.308 (0.026)	0.157 (0.040)***	2.51**
	Hispanics	139	0.151 (0.030)		
4	White	160	0.231 (0.033)	0.134 (0.042)**	2.80***
	Hispanics	134	0.097 (0.026)		
5	White	58	0.241 (0.056)	0.175 (0.060)**	4.49**
	Hispanics	136	0.066 (0.021)		
Overall	White	2798	0.302 (0.033)‡		
	Hispanics	684	0.178 (0.031)‡	0.123 (0.045)**	1.99 (1.58, 2.51)§
Propensity score stratification, Whites as reference sample: Stratum (from lowest to highest propensity for being Hispanic)					
1	White	572	0.406 (0.021)	0.011 (0.082)	1.05
	Hispanics	38	0.395 (0.079)		
2	White	553	0.354 (0.020)	0.054 (0.063)	1.28
	Hispanics	60	0.300 (0.059)		
3	White	562	0.386 (0.021)	0.035 (0.059)	1.16
	Hispanics	74	0.351 (0.055)		
4	White	552	0.330 (0.020)	0.118 (0.046)*	1.83*
	Hispanics	99	0.212 (0.041)		
5	White	559	0.279 (0.019)	0.175 (0.024)***	3.33***
	Hispanics	413	0.104 (0.015)		
Overall	White	2798	0.351 (0.020)‡		
	Hispanics	684	0.272 (0.054)‡	0.079 (0.058)	1.44 (1.12, 1.87)§
Propensity score stratification, combined Whites/Hispanics as reference sample: Stratum (from lowest to highest propensity for being Hispanic)					
1	White	751	0.403 (0.018)	0.101 (0.072)	1.56
	Hispanics	43	0.302 (0.070)		
2	White	571	0.354 (0.020)	0.009 (0.066)	1.04
	Hispanics	58	0.345 (0.062)		
3	White	581	0.348 (0.020)	0.069 (0.050)	1.38
	Hispanics	97	0.278 (0.043)		
4	White	553	0.340 (0.020)	0.120 (0.040)**	1.83**
	Hispanics	141	0.220 (0.035)		
5	White	342	0.257 (0.024)	0.165 (0.028)***	3.39***
	Hispanics	345	0.093 (0.015)		
Overall	White	2798	0.340 (0.020)‡		
	Hispanics	684	0.248 (0.050)‡	0.093 (0.054)†	1.57 (1.23, 2.00)§

† $P < .10$; *$P < .05$; **$P < .01$; ***$P < .001$.
‡ The overall mean is simple average of the means from the five strata and pooled SE is calculated based on [16].
§ Pooled OR is calculated from the pooled proportion from White/Hispanic groups; 95% CIs were generated using bootstrap method.

the RS, four strata had ORs significantly different from 1.00. By contrast, with Whites as the RS, only two strata had ORs significantly different from 1.00. This suggests that the overall ethnic effect would be larger with Hispanics used as the RS compared with Whites as the RS. As shown in the last rows of the three sections in Table 2, the overall marginal ORs were 1.99, 1.44, and 1.57 using Hispanics, Whites, and the combined White/Hispanic sample as the RS, respectively, in separate procedures.

Table 3 shows that, before adjustment, Whites had 2.47 times the odds compared with Hispanics to report a primary care visit. Next, adjusting for demographics and SES using logistic regression, the adjusted conditional OR dropped to 1.84. The marginal OR from the logistic regression adjusting for covariates was 1.81. This marginal OR was derived by averaging across the total combined sample. For illustrative purpose, we estimated the marginal OR for Whites and, separately, Hispanics as the RS as 1.82 and 1.81, respectively. Note that similar results are expected across different RSs here, because effect modification was not incorporated in the model.

Table 3
Estimated disparities in primary care use for the propensity score weighting method standardized to different reference samples, compared with raw and multivariable logistic regression, Whites versus Hispanics

	OR (95% CI)
Raw OR (95% CI)	2.47 (2.00–3.05)
Multivariable logistic regression, conditional OR (95% CI)	1.84 (1.47–2.31)
Multivariable logistic regression, marginal OR (95% CI)	1.81 (1.45–2.27)*
Propensity score weighting, marginal ORs (95% CI)	
Hispanics as reference sample	2.01 (1.58–2.57)
Whites as reference sample	1.41 (1.09–1.83)
Combined sample as reference sample	1.54 (1.19–1.99)

CI = confidence interval; OR = odds ratio.
* Marginal OR was derived for the total combined sample; 95% CI was generated using bootstrap method.

For a specific choice of RS, PS weighting methods produced results quite similar to those from PS stratification. Using PS weighting, the marginal ORs were 2.01, 1.41, and 1.54 using Hispanics, Whites, and the combined White/Hispanic sample as the RS, separately.

Table 4 shows the marginal effects from the logistic regressions incorporating effect modification by SES, including the marginal response probabilities and marginal ORs for the three RS choices. Using Hispanics, Whites, and the combined White/Hispanic sample as the RS, the marginal ORs were 2.01, 1.48, and 1.56, respectively, each of which are similar to those produced from PS methods for the corresponding RS.

Discussion

Our goal was to compare standard logistic regression and PS methods to estimate the main effect of ethnicity on a health-related

Table 4
Estimated disparities in primary care using logistic regression incorporating effect modification

	OR (95% CI)
Hispanics as reference sample	
Hispanics proportion	0.180
Counterfactual Whites proportion	0.305
Difference in proportion (Whites vs Hispanics)	0.126
Marginal OR (Whites vs Hispanics)	2.01 (1.54–2.61)*
Whites as reference sample	
Counterfactual Hispanics proportion	0.268
Whites proportion	0.351
Difference in proportion (Whites vs Hispanics)	0.083
Marginal OR (Whites vs Hispanics)	1.48 (1.18–1.86)*
Combined Whites/Hispanics as reference sample	
Counterfactual Hispanics proportion	0.251
Counterfactual Whites proportion	0.342
Difference in proportion (Whites vs Hispanics)	0.092
Marginal OR (Whites vs Hispanics)	1.56 (1.25–1.93)*

CI = confidence interval; OR = odds ratio.
* The 95% CIs were generated using the bootstrap method.

outcome in the presence of effect modification. This is a common problem in epidemiologic research, particularly in disparities research where ethnicity and SES are typically confounded. Standard regression analyses, controlling for SES and other demographics, estimated that Whites had 1.84 times the odds of receiving primary care compared with Hispanics. However, when effect modifiers are present, this conditional estimate assumes, incorrectly, the ethnicity effects are constant across SES levels. In contrast, PS methods produced marginal estimates for a given RS, using a standardization process that used weighting to reflect the covariate distribution in the RS. In this study, if our substantive research aim is the ethnic difference between Whites and Hispanics if Hispanics were like Whites in terms of measured SES characteristics (i.e., using Whites as the RS), the results from PS methods showed Whites had 1.41 or 1.44 times the odds (using PS weighting or stratification) of visiting primary care than Hispanics, markedly different from the conditional OR of 1.84 using standard logistic regression.

Our findings highlight the importance of careful selection of the RS to generate appropriate marginal effects when potential confounders are also effect modifiers. In the disparities research, the decisions about the RS should be informed by the substantive research question at hand and should also consider policy implications of such a choice. Findings from disparities research are often politically charged and rather technical choices concerning the selection of RS can substantially impact the "take home" message of one's research. As illustrated, ethnic disparity estimate using Whites as the RS (OR of 1.41 or 1.44 using two PS approaches) was substantially different from those using Hispanics as the RS (OR of 1.99 or 2.01). In this study, as in many other disparities investigations, a comparison of greater interest is the Hispanics–White disparity, where the minority population (Hispanics) had the same demographic and SES characteristics as Whites, when Whites, an advantaged group, serve as the RS.

Given the importance of the selection of RS, it should be carefully thought through by researchers designing PS analyses. One relevant issue is incomplete matching [32], when it is not possible to find matches for each case in the RS. PS methods, particularly the PS matching approach, often produce a matched subsample to achieve well-balanced groups. When the number of unmatched individuals is large, there is concern whether the matched individuals are representative of the population being studied [33]. This is particularly problematic when effect modification is present and the covariate distribution in the matched subsample is different from that in the total RS. Recent work has shown inconsistent estimates when different matched subsamples were chosen [21,23]. In this study, to use PS matching with Whites as the RS, one would need to identify enough individuals from a much smaller sample (Hispanics) to match with all subjects from the much larger sample (Whites). Because of this potential problem, the PS matching method was not implemented here and left for future research.

PS is a complex approach and researchers implementing it are confronted with technical choices that can potentially impact the results. Other than the choice of RS, there are questions regarding choice of specific PS approach to apply (matching, stratification, or weighting), evaluation of common support, and assessment of balancing quality, and so on [33]. In this analysis, we found that the PS stratification approach might not fully balance the covariates owing to uncontrolled, intra-stratum confounding. One remedy is to further adjust for the SES covariates within each PS stratum. For example, using the combined sample as the RS, the ORs adjusting for covariates are 1.62, 0.96, 1.42, 1.81, and 3.01 for strata one through five, respectively, compared with 1.56, 1.04, 1.38, 1.83, and 3.39 without covariate adjustment as in the bottom section of Table 2. Comparing the two sets of results suggests the overall OR using the new approach might be slightly lower than we previously found (1.57). A problem of this adjusting approach is that, given the non-collapsibility property of ORs, the overall OR cannot be derived simply by averaging the stratum-specific ORs (=1.76) as was done for stratum-specific response probabilities, which are collapsible.

Finally, we found that the marginal effects derived from logistic regression that included interaction terms seemed quite similar to those produced from the PS methods. We stress the comparability between the PS estimates and regression estimates for marginal effects only. Conceptually, PS methods produce counterfactual outcome estimates for a specific RS and generate a marginal effect. Quantitatively, the marginal and conditional ORs from logistic regressions do not always converge owing to the non-collapsibility property of ORs [34,35]. Furthermore, when interaction terms are entered into the regression model, marginal ORs for a specific RS can be produced in a process involving the averaging across a given population. This generalized regression approach incorporating effect modification has its roots both in the Rubin casual model as well as the economics literature. Rubin [36] developed a formula estimating the treatment effect using linear regression for the combined treated and nontreated samples, basically equivalent to the marginal effects we use here for the combined sample as the RS. Separately, labor economists devised the so-called Oaxaca–Blinder decomposition [37,38] to estimate gender and racial disparities due to discrimination. Its generalized form for dichotomous outcomes [39] is equivalent to what we present here using Whites or Hispanics as the RS. Last, although a model specification including interaction terms is more flexible than one only having main effects, all regressions are parametric models, which assume some underlying data generation processes and might not be correct and may have potential problems stemming from untrustworthy extrapolation [7]. In this regard, PS methods as a quasi-parametric approach may be more robust.

References

[1] Adler NE, Rehkopf DH. U.S. disparities in health: descriptions, causes, and mechanisms. Annu Rev Public Health 2008;29:235–52.

[2] Smedley BD, Stith AY, Nelson AR, editors. Unequal treatment. Confronting racial and ethnic disparities in health care. Washington, DC: Board on Health Sciences Policy, Institute of Medicine, National Academies Press; 2002.

[3] U.S. Department of Health and Human Services. Healthy people 2000: National health promotion and disease prevention objectives. Washington, DC: U.S. Department of Health and Human Services; 1991.

[4] U.S. Department of Health and Human Services. Healthy people 2010: Understanding and improving health. 2nd ed.). Washington, DC: U.S. Department of Health and Human Services; 2000.

[5] U.S. Department of Health and Human Services. Healthy people 2020: Disparities. Archived by WebCite at http://www.webcitation.org/63iByEQI3.

Washington, DC: U.S. Department of Health and Human Services Office of Disease Prevention and Health Promotion [accessed 12.05.11].

[6] Williams DR, Collins C. US Socioeconomic and racial differences in health: patterns and explanations. Annu Rev Sociol 1995;21:349–86.

[7] Imbens GW, Wooldridge JM. Recent developments in the econometrics of program evaluation. Journal of Economic Literature 2009;47(1):5–86.

[8] D'Agostino Jr RB. 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.

[9] Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997;127(8, Pt. 2):757–63.

[10] Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. Int J Epidemiol 2008;37(5):1142–7.

[11] Dehejia RH, Wahba S. Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. J Am Stat Assoc 1999; 94(448):1053–62.

[12] Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. Rev Econ Stat 2002;84(1):151–61.

[13] 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.

[14] Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. 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):437–47.

[15] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.

[16] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984;79(387): 516–24.

[17] Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 1985;39(1):33–8.

[18] Kaufman JS. Marginalia: comparing adjusted effect measures. Epidemiology 2010;21(4):490–3.

[19] Stampf S, Graf E, Schmoor C, Schumacher M. Estimators and confidence intervals for the marginal odds ratio using logic regression and propensity score stratification. Stat Med 2010;29(7-8):760–9.

[20] Austin PC. The performance of different propensity score methods for estimating odds ratios. Stat Med 2007;26(16):3078–94.

[21] Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol 2006;163(3):262–70.

[22] Stürmer T, Rothman KJ, Glynn RJ. Insights into different results from different causal contrasts in the presence of effect-measure modification. Pharmacoepidemiol Drug Saf 2006;15(10):698–709.

[23] Lunt M, Solomon D, Rothman K, Glynn R, Hyrich K, Symmons DP, et al. Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. Am J Epidemiol 2009;169(7):909–17.

[24] Schmidt LA, Ye Y, Greenfield TK, Bond J. Ethnic disparities in clinical severity and services for alcohol problems: results from the National Alcohol Survey. Alcohol Clin Exp Res 2007;31(1):48–56.

[25] Kerr WC, Greenfield TK, Bond J, Ye Y, Rehm J. Age-period-cohort modeling of alcohol volume and heavy drinking days in the US National Alcohol Surveys: divergence in younger and older adult trends. Addiction 2009; 104(1):27–37.

[26] National Institute on Alcohol Abuse and Alcoholism. Helping patients who drink too much: a clinician's guide. Updated 2005 Edition (NIH Publication No. 07-3769) http://pubs.niaaa.nih.gov/publications/Practitioner/CliniciansGuide2005/guide.pdf; 2005. [accessed 12.11.09]. Rockville, MD: National Institute on Alcohol Abuse and Alcoholism.

[27] Chartier K, Caetano R. Ethnicity and health disparities in alcohol research. Alcohol Res Health 2010;33(1-2):152–60.

[28] Mulia N, Schmidt LA, Ye Y, Greenfield TK. Preventing disparities in alcohol screening and brief intervention: the need to move beyond primary care. Alcohol Clin Exp Res 2011;35(9):1557–60.

[29] Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11(5):550–60.

[30] Rothman KJ, Greenland S. Modern epidemiology. 2nd edition).. Philadelphia: Lippincott-Raven Publishers; 1998.

[31] Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. Epidemiology 2003;14(6):680–6.

[32] Rosenbaum PR, Rubin DB. The bias due to incomplete matching. Biometrics 1985;41(1):103–16.

[33] Caliendo M, Kopeinig S. Propensity score matching, implementation, evaluation, sensitivity analysis. J Econ Surv 2008;22(1):31–72.

[34] Gail MH, Wieand HS, Piantadosi S. Biased estimates of treatment effect in randomized experiments with non-linear regressions and omitted covariates. Biometrika 1984;71(3):431–44.

[35] Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. Stat Sci 1999;14(1):29–46.

[36] Rubin DB. Assignment to treatment group on the basis of a covariate. J Educ Stat 1977;2(1):1–26.

[37] Oaxaca RL. Male-female wage differentials in urban labor markets. Int Econ Rev 1973;14(3):693–709.

[38] Blinder AS. Wage discrimination: reduced form and structural estimates. J Hum Resour 1973;8(4):436–55.

[39] Bauer TK, Sinning M. An extension of the Blinder-Oaxaca decomposition to nonlinear models. Adv Stat Anal 2008;92(2):197–206.

Appendix

1. The calculation of propensity score stratification marginal response probability and marginal ORs

For each of the three reference sample (RS) selections (Hispanics, Whites, and the combined sample), the overall (marginal) response probability for the White and Hispanic effect is estimated by weighting the stratum-specific response probability

$$\widehat{p}_{1,Resp} = \sum_{s=1}^{5} w_s \widehat{p}_{1s} \; \& \; \widehat{p}_{0,Resp} = \sum_{s=1}^{5} w_s \widehat{p}_{0s} \qquad (1)$$

Where $\widehat{p}_{1,Resp}$ and $\widehat{p}_{0,Resp}$ are the overall response probabilities for White effect and Hispanic effect separately, and $\widehat{p}_{1s}$ and $\widehat{p}_{0s}$ are the stratum-specific response probabilities for Whites and Hispanics separately, where s = 1 to 5. Because the RS is divided into five equal size strata in the current design, equal weights are assigned across the five strata and the overall response probability is just the simple average of the stratum-specific probability. The overall (marginal) OR is defined for Whites versus Hispanics as:

$$OR = \frac{\widehat{p}_{1,Resp}/\left(1-\widehat{p}_{1,Resp}\right)}{\widehat{p}_{0,Resp}/\left(1-\widehat{p}_{0,Resp}\right)} \qquad (2)$$

2. The calculation of marginal ORs from standard logistic regression

Consider a standard logistic regression model

$$logit\{P(Y=1|Z,X)\} = \beta_0 + Z\beta_Z + X\beta_x \qquad (3)$$

where Y is the dichotomous observed response, Z is indicator for ethnicity (coded as 1 if White and 0 if Hispanic), and X are covariates. For this model, the conditional OR is estimated as $\exp(\widehat{\beta}_Z)$. Using the fitted model (3) above, for each subject i, two predicted response probabilities are generated given individual observed covariates x_i. The first is generated by assuming the respondent is White (i.e., using $Z = 1$); the second by assuming the respondent is Hispanic (i.e., $Z = 0$), regardless of the respondent's actual ethnicity. The response probability can be represented as

$$\widehat{p}_l^{\,i} = \frac{1}{\left(1+\exp\left\{-\left(\widehat{\beta}_0+\widehat{\beta}_Z l + x_i\widehat{\beta}_x\right)\right\}\right)} \qquad (4)$$

where $l = 1$, then $= 0$, respectively, for each respondent for the White and Hispanic effects. The marginal response probability for the total combined sample is just the average across the whole sample of individual predicted probability for both White and Hispanic effects ($l = 1$ and 0, respectively):

$$\widehat{p}_l = \frac{1}{n}\sum_{i=1}^{n} \widehat{p}_l^{\,i} \qquad (5)$$

Finally the marginal OR for the total combined sample is as below.

$$OR = \frac{\widehat{p}_1/(1-\widehat{p}_1)}{\widehat{p}_0/(1-\widehat{p}_0)} \tag{6}$$

Alternatively, the marginal response probabilities, as well as the corresponding marginal ORs, can also be estimated for the White or Hispanic sample only as the RS (Z = 1 or 0, respectively), for both the White and Hispanic effects ($l = 1$ and 0, respectively):

$$\widehat{p}_l^Z = \frac{1}{n_z}\sum_{i=1}^{n_z}\widehat{p}_l^{zi} \tag{7}$$

3. The calculation of marginal ORs from standard incorporating effect modification

The logistic regression (3) above is generalized to include ethnic-specific coefficients in the model

$$logit\{P(Y=1|Z_1,Z_0,X)\} = \beta_{01}Z_1 + \beta_{00}Z_0 + XZ_1\beta_{x1} + XZ_0\beta_{x0} \tag{8}$$

where Z_1 is the indicator for Whites and Z_0 is the indicator for Hispanics, and the intercept is removed from the model. Similar to (4) above, for each subject i, two predicted response probabilities can be generated that, respectively, assuming the respondent is White or Hispanic (regardless of their true ethnicity) to generate the White and Hispanic effects ($l = 1$ and 0, respectively). For example, for a White individual, the predicted response probability for the White effect ($l = 1$) is

$$\widehat{p}_{l=1}^{wi} = \frac{1}{\left(1 + exp\left\{-\left(\widehat{\beta}_{01} + x_{wi}\widehat{\beta}_{x1}\right)\right\}\right)} \tag{9}$$

For this same White individual, the predicted response probability for the Hispanic effect ($l = 0$) is

$$\widehat{p}_{l=0}^{wi} = \frac{1}{\left(1 + exp\left\{-\left(\widehat{\beta}_{00} + x_{wi}\widehat{\beta}_{x0}\right)\right\}\right)} \tag{10}$$

Note in (10) that, although this White individual's own covariates x_{wi} are used, the coefficients for Hispanics ($\widehat{\beta}_{00}$ and $\widehat{\beta}_{x0}$) are plugged into the model to generate the counterfactual predicted outcome if the respondent were Hispanic. The predicted response probability in (10) can also be interpreted as the predicted outcome for Hispanics should they have the same covariate distribution as Whites. Similarly, the predicted response outcome for each Hispanic individual can be estimated to generate the White and Hispanic effects ($l = 1$ and 0, respectively). The marginal response probability can then be calculated by taking the average of the individual predicted probabilities as shown in (5) for the combined overall sample as the RS or (7) for Whites or Hispanics as the RS. Finally, the marginal OR for each RS is derived as shown in (6), using the corresponding marginal response probabilities.