REDUCING BIAS IN VALIDATING HEALTH MEASURES WITH PROPENSITY SCORE METHODS

Xian Liu, Ph.D.

Charles C. Engel, Jr., M.D., M.PH.

Kristie Gore, Ph.D.

Michael Freed, Ph.D.

Abstract

In this article, we present and develop a propensity score method to reduce bias in diagnostic tests of health measures on actual disease diagnosis. The propensity score serves as a distributional balance of covariates and is predicted by the binomial logistic regression with disease diagnosis as the response variable and a number of covariates as predictor variables. We develop two multivariate regression models to validate health measures. For health measures that are dichotomized, we establish a binomial logistic regression model to estimate sensitivity, specificity, and the likelihood ratio. For health measures involving more than two levels, we use the multinomial logit regression to estimate the probability of a true positive result, the probability of a false positive result, and the likelihood ratio at each measurement level. Our empirical examples demonstrate that without considering an individual's demographic, socioeconomic and other relevant characteristics, results from diagnostic tests of health measures can be biased and misleading.

Introduction

The propensity score method is a technique that has been widely employed in randomized controlled experiments to reduce biases in comparing effects between conditions (Brookhart et al., 2006; D'Agostino, 1998; Robins and Mark, 1992; Rosenbaum, 1983; Rosenbaum and Rubin, 1984). The score serves as a distributional balance of covariates in examining the effects of a given experimental intervention, thereby adjusting for potential confounding factors (Rosenbaum, 1983). If the propensity score is used as a covariate in a multivariate regression model, the result is a relatively unbiased association between the treatment factor and the outcome variable.

While it is commonly used in experimental studies, the propensity score method can also be effective in performing empirical analyses in other domains. In experimental studies, the confounding factors influence the relationship between treatment and outcome; consequently, the propensity score is used to adjust for such biases. Similarly, confounding factors may differentially affect results from another testing variable that is assessed with a single binary measure, thereby resulting in misleading and biased conclusions. We can borrow the propensity score method from these experimental studies and employ it to reduce biases in assessing the validity of such variables. There are many examples, among which the statistical validation of health measures is probably the most prominent.

A variety of health measures exist to assess the prevalence and severity of illnesses, and researchers commonly cross disciplines and use statistical techniques to examine the accuracy and validity of their measures. Sensitivity, specificity, positive or negative predictive values, receiver operating characteristic plots, and the likelihood ratio

(the ratio of the probability of true positives over the probability of false positives) are examples of such techniques used to validate measures (Altman and Bland, 1994a; Altman and Bland, 1994b; Altman and Bland, 1994c; Deeks and Altman, 2004; Grimes and Schulz, 2005; Simel, Samsa, and Matchar, 1991). Without careful consideration of population heterogeneity, however, testing results from a health measure can be misleading, as strong differences in individual characteristics and personal traits can confound the relationship between a specific health measurement and the true medical condition. Despite such threats for the quality of testing outcomes, it is surprising to note that little effort has been made to understand and explore potential biases in applying these statistical techniques.

In this article, we examine the relationship between a given health measure and the underlying true medical condition by constructing a propensity score model to reduce potential biases in performing diagnostic tests. We demonstrate how the propensity score mediates/moderates the detection of a specific disease using a screening measure versus the gold standard diagnostic measure. The screening measure can produce a binomial or an ordinal scaling of illness severity (a multilevel outcome). Accordingly, we derive two propensity score models using a binomial and a multilevel approach.

Review of Basic Diagnostic Testing Techniques

In this study, we concentrate on sensitivity, specificity and likelihood ratios given their extensive application and popularity. Other diagnostic testing techniques, such as positive or negative predictive value and receiver operating characteristic plots, can be viewed as their extensions.

Technically, sensitivity and specificity are just two conditional probabilities linked to a given health measure for a specific medical or mental health condition. In the context of health measures, sensitivity is the probability of positive measuring results given positive medical diagnosis, whereas specificity is defined as the probability of negative results with negative medical diagnosis. Each of the two diagnostic indicators only reflects a unilateral dimension in terms of an effective assessment of a given health measure; and the sole dependence on any one can lead to misconception and misjudgment. Statisticians have used likelihood ratios to summarize the diagnostic accuracy, serving as a more balanced tool to characterize the behavior of diagnostic tests (Deeks and Altman, 2006; Simel, Samsa and Matchar, 1991).

For dichotomized measuring results classified by medical diagnosis (a 2×2 contingency table), the conventional likelihood ratio approach is to combine information of both sensitivity and specificity for providing complete information obtained from a specific health measure. The positive likelihood ratio, denoted by LR+, represents changes in the likelihood of a specific disease given a positive measuring result, whereas the negative likelihood ratio (LR-) indicates changes in the likelihood of a specific diagnosed disease given a negative result. For a diagnosed disease, sensitivity and specificity can be expressed as the probability of true positive measuring results and the probability of true negative measuring results. Similarly, (1 – specificity) is simply the probability of false positives referring to that disease while (1 – sensitivity) is the probability of false negatives. Hence, LR+ can be expressed as the ratio of the probability of true positives over the probability of false positives, whereas LR- as the ratio of the probability of false negatives over the probability of true negatives.

When a health measure is composed of multiple scaling levels, a set of likelihood ratios need to be derived, termed the multilevel likelihood ratio approach (Deeks and Altman, 2004; Grimes and Schulz, 2005; Simel, Samsa, and Matchar, 1991). Empirically, a multilevel likelihood ratio is calculated as the proportion of diagnosed patients with positive measuring results at a given level (the probability of true positives) divided by the proportion of non-diseased patients with the same measuring result (the probability of false positives).

Creation of a Propensity Score

In conventional experimental studies, the propensity score is defined as the conditional probability of assignment to a given treatment given an individual's demographic, socioeconomic and other theoretically related characteristics. This score has been used to balance the differences in personal traits and their confounding influences on the treatment's effects while evaluating the association between conditions. Statistically, it can be estimated either by a logistic or by a probit (the standard normal cumulative distribution function) model, with the probability of being selected in the treatment group used as the response variable and the selected covariates as the independent.

As potential confounders, variables predicting the propensity score can be risk factors on both the testing result factor and the treatment variable. Empirically, the asymptotic variance of an estimator is often decreased as the number of theoretically related parameters in the prediction model is increased (Brookhart et al., 2006; Robins and Mark, 1992). When both treatment (a dichotomous factor) and the propensity score

(a continuous factor) act together as independent variables in a statistical model predicting the level of a given medical or another indicator, the potential confounding effects of individual characteristics can be coarsely controlled. Thus the generation of unbiased estimates of the treatment's effects is secured.

In the context of diagnostic tests on health measures, the propensity score is defined as the conditional probability of being diagnosed with a specific disease given values of covariates. It is used to balance differences in potential confounding factors when assessing the quality of a specific health measure. Statistically, we first define a dichotomous variable δ_i for observation i (i = 1,2,….., n), given by

$$\delta_i = \begin{cases} 1 & \text{if diagnosis for a given disease is positive} \\ 0 & \text{if diagnosis for the disease is negative,} \quad i = 1,2,....., n. \end{cases}$$

Assuming the underlying health measure variable (1 = positive, 0 = negative) to be Z_i and X_i to be the propensity score, the probability of a true positive test and the probability of a false positive test can be expressed as

$$\begin{cases} P_i^{true} = prob\left(\delta_i = 1 \middle| Z_i = 1, X_i\right) \\ P_i^{false} = prob\left(\delta_i = 0 \middle| Z_i = 1, X_i\right), \end{cases}$$

where P_i^{true} and P_i^{false} indicate, respectively, the probability of true positives and the probability of false positives for subject i. And X, the likelihood of being diagnosed with the disease, is used as the propensity score, a combination variable reflecting the information for a vector of confounders. Since the propensity score is included in this model, the effect of the diagnosis is not attributable to the measured confounders, thus adjusting for possible biases in assessing the underlying health measure.

Propensity Score Approach for 2 × 2 Tables

If the outcome of a given health measure is dichotomized, the probability of a positive test is the response variable under assessment. Accordingly, the status of disease diagnosis (yes or no) is the explanatory variable to determine whether a positive test result is true or false. We can use a link function between the two to predict the probability of true positives or of the false positives, and eventually deriving a likelihood ratio. In constructing this link function, the propensity score is used as a control variable, predicted by a set of selected individual demographic, socioeconomic and other relevant variables.

We establish a logistic regression model to estimate the probability of true positives, the probability of false positives, and the likelihood ratios plus and minus, adjusting for potential confounding factors. Letting P_{pos} be the probability of positive measuring results, we specify the following logit model,

$$Log\left(\frac{P_{pos}}{1-P_{pos}}\right) = \alpha + \beta_1(Disease) + \beta_2 X,$$

where $(1 - P_{pos})$ is the probability of negative results, α is the intercept for the log odds, and β_1 and β_2 are regression coefficients for the dichotomous variable actual disease (1 = yes, 0 = no) and the propensity score, denoted by X. We employ the maximum likelihood approach to estimate the three parameters contained in the above model.

The probability of positive results is the probability of true positives if the disease variable is 1, and the probability of negative results is the probability of true negatives if the disease variable is 0. Hence, a number of regular diagnostic testing indicators can be estimated from transforming this logit model and inserting certain values of the independent variables, given by

$$(\text{Prob of true positives}\,|\,\mathrm{X}=\overline{\mathrm{X}}) = \frac{\exp(\alpha+\beta_1+\beta_2\overline{X})}{1+\exp(\alpha+\beta_1+\beta_2\overline{X})}$$

$$(\text{Prob of true negatives}\,|\,\mathrm{X}=\overline{\mathrm{X}}) = \frac{1}{1+\exp(\alpha_1+\beta_2\overline{X})}$$

$$(\text{Prob of false positives}\,|\,\mathrm{X}=\overline{\mathrm{X}}) = \frac{\exp(\alpha_1+\beta_2\overline{X})}{1+\exp(\alpha+\beta_2\overline{X})}$$

$$(\text{Prob of false negatives}\,|\,\mathrm{X}=\overline{\mathrm{X}}) = \frac{1}{1+\exp(\alpha_1+\beta_1+\beta_2\overline{X})},$$

where $\overline{X}$ is the mean of the propensity score serving as a standardized adjustment for possible biases in the probability estimates generated from differences in observed individual characteristics. As defined, the probability of true positives and the probability of true negatives are simply sensitivity and specificity, respectively. The likelihood ratios plus and minus can then be estimated, given by

$$\mathrm{LR}+ = \frac{(\text{Prob of true positives}\,|\,\mathrm{X}=\overline{\mathrm{X}})}{(\text{Prob of false positives}\,|\,\mathrm{X}=\overline{\mathrm{X}})},$$

and

$$\mathrm{LR}- = \frac{(\text{Prob of false negatives}\,|\,\mathrm{X}=\overline{\mathrm{X}})}{(\text{Prob of true negatives}\,|\,\mathrm{X}=\overline{\mathrm{X}})}.$$

As the propensity score is set at sample mean for those diagnosed with the disease and those without, the estimates of the likelihood ratios plus and minus are coarsely independent of differences in covariates between the two groups.

Propensity Score Method for Multilevel Tables

If the test results involve R levels (R > 2), the dependent variable in the prediction model includes R competing probabilities of test results: P_0 (the probability at the lowest

level, P_1 (the probability at the second lowest level), ……, and P_{R-1} (the probability at the highest level).

We use the multinomial logit regression model to estimate the log odds of R - 1 contrasts, $\log(P_1/P_0)$, $\log(P_2/P_0)$, ……, and $\log(P_{R-1}/P_0)$, defining P_0, the probability of positive results at the lowest level, as the baseline probability. The estimation of P_0 depends on the estimates for the probabilities of the other measuring results given the condition that a given set of probabilities must sum to unity. Associated with a specific disease and the propensity score, the basic multinomial logit model in this setting is defined by

$$\begin{cases} \mathrm{Log}\left(\frac{P_1}{P_0}\right) = \alpha_1 + \beta_{11}(\mathrm{Disease}) + B_{12}X \\ \mathrm{Log}\left(\frac{P_2}{P_0}\right) = \alpha_2 + \beta_{21}(\mathrm{Disease}) + B_{22}X \\ \cdots\cdots\cdots\cdots \\ \mathrm{Log}\left(\frac{P_{R-1}}{P_0}\right) = \alpha_{R-1} + \beta_{(R-1)1}(\mathrm{Disease}) + B_{(R-1)2}X \end{cases},$$

where α's are the intercepts for (R – 1) log odds, β's are regression coefficients for the dichotomous variable actual disease (1 = yes, 0 = no) and the propensity score, X, respectively. After a series of equation transformation (the detailed procedure is available upon request), probability distribution of multiple health measures can be expressed as

$$
\begin{cases}
\hat{P}_0 = \dfrac{1}{1+\sum_{k=1}^{R-1}\exp\left[\hat{\alpha}_k + \hat{\beta}_{k1}(\text{Disease}) + \hat{B}_{k2}X\right]} \\
\hat{P}_2 = \dfrac{\exp\left[\hat{\alpha}_1 + \hat{\beta}_{11}(\text{Disease}) + \hat{B}_{12}X_2\right]}{1+\sum_{k=1}^{R-1}\exp\left[\hat{\alpha}_k + \hat{\beta}_{k1}(\text{Disease}) + \hat{B}_{k2}X\right]} \\
\cdots\cdots\cdots\cdots \\
\hat{P}_{R\text{-}1} = \dfrac{\exp\left[\hat{\alpha}_{R\text{-}1} + \hat{\beta}_{(R\text{-}1)1}(\text{Disease}) + \hat{B}_{(R\text{-}1)2}X\right]}{1+\sum_{k=1}^{R-1}\exp\left[\hat{\alpha}_k + \hat{\beta}_{k1}(\text{Disease}) + \hat{B}_{k2}X\right]}
\end{cases}.
$$

We know that when the actual disease variable is 1, the P's are probabilities of true positives; similarly, when the disease variable is 0, the P's are probabilities of false positives. Hence, by inserting disease values (0 or 1) and the standardized propensity scores, we can calculate a set of probabilities for true positives or a set of probabilities for false positives adjusting for possible biases in probability estimates generated from differences in observed individual characteristics. For analytic convenience, we use sample means of the propensity scores to replace **X**, noted by $\overline{X}$ and defined as the mean conditional probability as predicted by subjects' demographic, socioeconomic and other relevant variables. Specifically, a set of probabilities for true positives are given by

$$\begin{cases} \hat{P}_0^{\text{true}} \left| \left(X = \overline{X} \right) = \dfrac{1}{1 + \sum_{k=1}^{R-1} \exp\left[\hat{\alpha}_k + \hat{\beta}_{k1} + \hat{B}_{k2} \overline{X} \right]} \right. \\ \hat{P}_1^{\text{true}} \left| \left(X = \overline{X} \right) = \dfrac{\exp\left[\hat{\alpha}_1 + \hat{\beta}_{11} + \hat{B}_{12} \overline{X} \right]}{1 + \sum_{k=1}^{R-1} \exp\left[\hat{\alpha}_k + \hat{\beta}_{k1} + \hat{B}_{k2} \overline{X} \right]} \right. \\ \ldots\ldots\ldots\ldots \\ \hat{P}_{R\text{-}1}^{\text{true}} \left| \left(X = \overline{X} \right) = \dfrac{\exp\left[\hat{\alpha}_{R\text{-}1} + \hat{\beta}_{(R\text{-}1)1} + \hat{B}_{(R\text{-}1)2} \overline{X} \right]}{1 + \sum_{k=1}^{R-1} \exp\left[\hat{\alpha}_k + \hat{\beta}_{k1} + \hat{B}_{k2} \overline{X} \right]} \right. \end{cases}.$$

Similarly, the probabilities of false positives can be estimated as

$$\begin{cases} \hat{P}_0^{\text{false}} \left| \left(X = \overline{X} \right) = \dfrac{1}{1 + \sum_{k=1}^{R-1} \exp\left[\hat{\alpha}_k + \hat{B}_{k2} \overline{X} \right]} \right. \\ \hat{P}_1^{\text{false}} \left| \left(X = \overline{X} \right) = \dfrac{\exp\left[\hat{\alpha}_1 + \hat{B}_{12} \overline{X} \right]}{1 + \sum_{k=1}^{R-1} \exp\left[\hat{\alpha}_k + \hat{B}_{k2} \overline{X} \right]} \right. \\ \ldots\ldots\ldots\ldots \\ \hat{P}_{R\text{-}1}^{\text{false}} \left| \left(X = \overline{X} \right) = \dfrac{\exp\left[\hat{\alpha}_{R\text{-}1} + \hat{B}_{(R\text{-}1)2} \overline{X} \right]}{1 + \sum_{k=1}^{R-1} \exp\left[\hat{\alpha}_k + \hat{B}_{k2} \overline{X} \right]} \right. \end{cases}.$$

Given these probability estimates, the multilevel likelihood ratios can be readily estimated by R pairs of probabilities, true positives over false positives, given by

$$\begin{cases} \text{Likelihood Ratio I}\left(\text{Health measure} = 1 \middle| \text{X} = \overline{\text{X}}\right) = \text{LR1} = \dfrac{\hat{\text{P}}_1^{\text{true}} \middle| X = \overline{X}}{\hat{\text{P}}_1^{false} \middle| X = \overline{X}} \\ \text{Likelihood Ratio II}\left(\text{Health measure} = 2 \middle| \text{X} = \overline{\text{X}}\right) = \text{LR2} = \dfrac{\hat{\text{P}}_2^{\text{true}} \middle| X = \overline{X}}{\hat{\text{P}}_2^{false} \middle| X = \overline{X}} \\ \text{.....} \\ \text{Likelihood Ratio (R - 1)}\left(\text{Health measure} = \text{R - 1} \middle| \text{X} = \overline{\text{X}}\right) = \text{LR(R - 1)} = \dfrac{\hat{\text{P}}_{\text{R-1}}^{\text{true}} \middle| X = \overline{X}}{\hat{\text{P}}_{R-1}^{false} \middle| X = \overline{X}} \end{cases}.$$

The likelihood ratio of the baseline measure, LR0, serves as the ratio of two residual probabilities, given by

$$LR0 = \frac{1 - \left(\hat{P}_1^{true} \middle| X = \overline{X}\right) - \left(\hat{P}_2^{true} \middle| X = \overline{X}\right) - ..... - \left(\hat{P}_{R-1}^{true} \middle| X = \overline{X}\right)}{1 - \left(\hat{P}_1^{false} \middle| X = \overline{X}\right) - \left(\hat{P}_2^{false} \middle| X = \overline{X}\right) - ..... - \left(\hat{P}_{R-1}^{false} \middle| X = \overline{X}\right)}$$

It is suggested that a subset of the covariates predicting the propensity score should also be used in the regression adjustment (D'Agostino, 1998). Prior studies have found that in specification of a propensity score model one should include all variables theoretically related to the outcome, regardless of whether they are related to the exposure or treatment (in this case, health measures) (Brookhart et al., 2006; Rubin, 2004). Therefore, some covariates significantly predicting the test results should be considered in executing the aforementioned propensity score models. Since addition or exclusion of predictor variables can cause severe variation in covariance, caution should apply in selecting control variables in the multivariate analytic model.

Multivariate Standard Errors and Confidence Intervals

The precise derivation of standard errors and confidence intervals for the binary or multivariate likelihood ratios, obtained by the logistic regression modeling, is very complex and tedious because variations in a set of probabilities are a function of multiple

factors. A coarse approximation is to randomly draw, with replacement, 30 or more sub-samples (400 or 500 cases, depending on the total sample size of a given dataset) from a large sample, and then calculate sample estimates of standard errors and confidence intervals from approximately 30 sample estimates.

Application

We present an empirical example to demonstrate the statistical techniques developed in the present research. In particular, we seek to show how the propensity score can influence results of diagnostic tests on the prediction power of five bodily pains in terms of somatic disorders among American adults.

Sample, Data and Methods

We use data from the Epidemiologic Catchment Areas Study (ECA), a nationally representative investigation conducted by National Institute of Mental Health in the mid-1980s. The ECA study and its methods are extensively described in other publications (Eaton et al., 1984; Regier et al., 1984; Robins and Regier, 1991). In brief, the ECA study was a collaborative research effort to determine the epidemiology of specific mental disorders in the United States and associated utilization of health services. The study was conducted in five geographic regions of the United States, including the New Haven (Conn.), Baltimore, St. Louis, Durham (N.C.), and Los Angeles areas. About 3,000 households and 300 institutionalized individuals were targeted for interviews at each site. The present research uses the cross-sectional data of the original 18,571 respondents at the baseline survey, excluding institutionalized individuals.

We measure somatic disorders by a dichotomous variable ("yes" = 1; "no" = 0), using the criterion definition developed by Escobar and colleagues (1989) that is characterized by four or more unexplained physical symptoms among men or six or more symptoms by women. Unexplained physical symptoms were determined using a highly structured interview that polled patients on 37 different physical symptoms encompassing a wide range of body regions and systems (Robins et al., 1981; Swartz et al., 1991). Table 1 presents the distribution of ECA Wave I respondents by the number of bodily pains, classified by status of somatic disorders.

<Table 1 about here>

We use eight covariates to generate an individual's propensity score in the multivariate analysis for balancing differences in individual characteristics and their potential confounding effects. Specifically, we consider seven demographic and socioeconomic variables (age, gender, educational attainment, ethnicity, marital status, veteran status, and socioeconomic score), and one health factor (disability payment). Table 2 presents group comparison of means (or proportions), standard deviations, and two-sample t-statistics for nine predictor variables (ethnicity is specified by two dichotomous variables), classified by status of somatic disorders (yes or no).

<Table 2 about here>

We apply both the conventional approach and the propensity score method to derive two sets of diagnostic tests on the number of five bodily pains predicting diagnosed somatic disorder. Specifically, we first dichotomize five bodily pains into two categories, zero pain versus any pain, from which we calculate sensitivity, specificity, likelihood ratio plus and likelihood ratio minus from both the descriptive approach and

the binomial logistic regression. Then we group the patient's bodily pains into three categories, 0 – 1 pain, 2 – 3 pains, and 4 – 5 pains to estimate likelihood ratios at three levels. In addition to the conventional multilevel approach, we employ the multinomial logic model to estimate the propensity score multilevel model, with the probability of having 0-1 pain being zero serving as the baseline probability. Specifically, defining P_1, P_2 and P_3 as the probabilities of three bodily pain levels, we examine $\log(P_2/P_1)$ and $\log(P_3/P_1)$ as linear functions of somatic disorder, the propensity score and some selected covariates. From the eight covariates predicting the propensity score, we select several control variables according to our statistical assessment, including age, gender, ethnicity and disability payment in the propensity model, together with status of somatic disorder and the propensity score.

We use the formulas recommended by Simel and associates (1991) to calculate the standard errors and confidence intervals for the conventional diagnostic tests. In terms of the propensity score models involving multiple explanatory variables, we perform the bootstrap re-sampling procedure (SAS/STAT 9.1, 2004) to draw 30 simple random samples with replacement, each containing 600 cases, from which we estimate the approximate standard errors and confidence intervals.

Analytic Results

Table 3 presents the results of the binomial logit model predicting the propensity score. Educational attainment is significantly and inversely associated with the probability of having somatic disorder, whereas women and blacks are significantly more likely to have somatization, other things equal. The effects of other covariates are not

statistically significant. The estimate of mean propensity score is 0.0481, indicating that about 5 percent of American adults is diagnosed with somatic disorder. Using the results presented in Table 3, we create the propensity score for each individual according to his or her values of the eleven covariates and then use it as a new covariate in the propensity score models.

<Table 3 about here>

Table 4 demonstrates the results for two sets of diagnostic tests on dichotomized bodily pains. The upper panel of the table showed sensitivity, specificity, likelihood ratio plus and likelihood minus calculated by the conventional approach, while the second panel demonstrates results of the same diagnostic indicators derived from the propensity score model. There are no distinct differences between the two sets of estimates, with or without the effects of individuals' covariates. For example, by including the propensity score and other control variables in the model, the value of specificity (the probability of true negatives) is 0.5447 compared to 0.5466 estimated from the conventional approach. As a result, the likelihood ratios plus and minus also remain relatively unchanged. As the likelihood ratio plus is below two with both approaches, the dichotomization of bodily pains in measuring an adult's status of somatic disorder is not efficient.

<Table 4 about here>

Table 5 presents two sets of likelihood ratios for three bodily pain levels, derived from, respectively, the conventional descriptive approach and the multinomial logit regression. In estimating the three population-based likelihood ratios, we fix values of the propensity score and other control variables as their sample means, so that the likelihood ratios can be derived adjusting for differences in covariates between those

diagnosed with somatization and those without. The three likelihood ratios, estimated from the multinomial logit model, are 3.0101, 1.5592 and 0.2486, somewhat deviant from the sample figures, which are, respectively, 2.7466, 1.5169 and 0.2548. As the variable somatic disorder has very strong impact on the two logits (2.2039 and 1.3803; statistically significant at $\alpha = 0.05$), not presented here, the probability of true positives and the probability of false positives are significantly different, thus highlighting the statistical reliability of the likelihood ratios.

Although we consider multiple factors in estimating the multivariate likelihood ratios, the confidence intervals for such ratios are not necessarily wider than those obtained from the conventional approaches, as evidenced in the present research.

Discussion

This study introduces a multivariate propensity score method to reduce biases in performing diagnostic tests on health measures. Specifically, we adjust differences in individual characteristics and personal traits by standardizing the propensity score in the multivariate regression model, so that sub-classification or matching on the propensity score is not required thereby avoiding the potential "residual confounding" (Robins and Mark, 1992). As a consequence, the diagnostic tests on health measures, as associated with a given disease, can be performed with much adequacy and statistical confidence.

One might question the use of a single combination factor to control the confounding effects of multiple factors. The concern may come from the fact that because the propensity score is predicted by a number of covariates, researchers can directly use those explanatory variables in the model as controls thereby deriving

unbiased diagnostic tests. We argue, however, that in presence of a large number of covariates in a statistical model, the existence of endogeneity can result in statistical inefficiency in deriving the effects of covariates, including that of treatment, exposure or, in this context, diagnosis factor. Regression models involving a large number of potential confounding factors as independent variables often derive imprecise estimates of the main effect of the variable in assessment, resulting in wide confidence intervals and high p values for the main effect (Sonis, 2006). Although it often absorbs more information for the prediction of the dependent variable, a model consisting of all covariates cannot reflect the true set of the influences generated by covariates on the stochastic processes (Liu, 2000).

A simplified model using a propensity score and a subset of selected covariates, combines strengths inherent in both the prediction model and the model excluding confounding factors. Moreover, sample sizes for experimental data are usually fairly small, and using a large number of covariates in a statistical model is not always realistic. Very often researchers are not interested in causal linkages of confounding factors, and a combination factor like the propensity score can well serve their purposes as controls. In spite of these advantages, we should bear in mind that inclusion of the propensity score in diagnostic tests can substantially reduce biases in diagnostic tests, but by no means will it completely eliminate deviations from true stochastic processes of a health event.

References

Alman, D.G. and Bland, J.M. 1994. Diagnostic test 1: sensitivity and specificity. *BMJ* 308:1552.

Alman, D.G. and Bland, J.M. 1994. Diagnostic test 2: predictive values. *BMJ* 309:102.

Alman, D.G. and Bland, J.M. 1994. Diagnostic test 3: receiver operating characteristic plots. *BMJ* 309:188.

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., and Sturmer, T. 2006. Variable selection for propensity score models. *American Journal of Epidemiology* 163:1149-1156.

Buchsbaum, D.G., Buchanan, R.G., Centor, R.M., Schnoll, S.H., Lawton, M.J. 1991. Screening for alcohol abuse using CAGE scores and likelihood ratios. *Annals of Internal Medicine*115:774-777.

D'Agostino, Jr., R.B. 1998. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 17:2265-2281.

Deeks, J.J. and Altman, D.G. 2004. Diagnostic test 4: likelihood ratios. *BMJ* 329:168-169.

Eaton, W.W., Holzer, C.E. III, Von Korff, M., Anthony, J.C., Helzer, J.E., George, L., Burnam, A., Boyd, J.H., Kessler, L.G., and Locke, B.Z. 1984. The design of the Epidemiologic Catchment Area survey: the control and measurement of error. *Archives of General Psychiatry* 41:942-948.

Escobar, J.I., Rubio-Stipec, M., Canino, G.J., and Karno, M. 1989. Somatic symptom index (SSI): a new and abridged somatization construct. *Journal of Nervous and Mental Disease* 177:140-146

Grimes, D.A. and Schulz, K.F. 2005. Refining clinical diagnosis with likelihood ratios. *Lancet* 365:1500-1505.

Liu, X. 2000. Development of a structural hazard rate model in sociological research. *Sociological Methods & Research* 29:77-117.

Regier, D.A., Myers, J.K., Kramer, M., Robins, L.N., Blazer, D.G., Hough, R.L., Eaton, W.W., and Locke, B.Z. 1984. The NIMH Epidemiologic Catchment Area program: historical context, major objectives, and study population characteristics. *Archives of General Psychiatry* 41:934-941.

Robins, J.M. and Mark, S.D. 1992. Estimating exposure effects by modeling the expectation of exposure conditional on confounders. *Biometrics* 48:479-495.

Robins, L.N., Helzer, J.E., Croughan, J., and Ratcliff, K.S. 1981. The National Institute of Mental Health Diagnostic Interview Schedule: its history, characteristics, and validity. *Archives of General Psychiatry* 38:381-389.

Robins, L.N. and Regier, D.A. (eds). 1991. *Psychiatric Disorders in America: The Epidemiologic Catchment Area Study*. New York: Free Press.

Rosenbaum, P.R. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41-55.

Rosenbaum, P.R. and Rubin, D.B. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79:516-524.

Rubin, D.B. 2004. On principles for modeling propensity scores in medical research (Editorial). *Pharmacoepidemiology and Drug Safety* 13:855-857.

Simel, D.L., Samsa, G.P. and Matchar, D.B. 1991. Likelihood ratios with confidence: Sample size estimation for diagnostic test studies. *Journal of Clinical Epidemiology* 44:763-770.

Sonis, J. 2006. Newer approaches to controlling confounding: Propensity scores. *Traumatic Stresspoints*, Spring:7.

Swartz, M., Landerman, R., George, L.K., Blazer, D.G., and Escobar, J.I. 1991. Somatization disorder, in *Psychiatric Disorders in America: the Epidemiologic Catchment Area Study*. Edited by Robins, L.N. and Regier, D.A. New York: Free Press.

Table 1. Percentage Distribution of ECA Respondents by Number of Pains,

Classified by Status of Somatic Disorder (Sample Size = 10,155)

Number of Pains	Somatic Disorder	
	No	Yes
0	26.95%	2.65%
1	27.71	11.28
2	20.16	21.58
3	11.11	25.86
4	10.20	24.81
5	3.86	13.83
Chi-square	526.70**	
Sample Size	9,707	448

** $P < 0.01$

Table 2. Group Comparison of Nine Covariates used for Generating Propensity Scores:

Epidemiologic Catchment Area Survey, Wave I Sample (Sample Size = 10,155)

Variables used for	No Somatic Disorder (N = 9,707)		Somatic Disoder (N = 448)		Two-sample
Propensity Scores	Mean	SD	Mean	SD	t-statistic
Age	40.04	18.15	41.15	16.62	-1.45
Female	0.46	0.57	0.57	0.53	-4.57**
Educational attainment	12.32	3.58	11.45	3.78	5.72**
White	0.71	0.52	0.63	0.51	3.58**
Black	0.17	0.43	0.26	0.47	-5.08**
Currently married	0.59	0.56	0.56	0.53	1.11
Veteran status	0.21	0.47	0.17	0.40	2.14*
Scioeconomic status score	54.72	25.64	48.33	25.00	5.90**
Disability payment	0.03	0.20	0.05	0.24	-2.39*

* $0.01 < p < 0.05$, two-tailed; ** $p < 0.01$, two-tailed.

Table 3. Results of Logistic Regression Model on Propensity Score:

ECA Wave I Survey (N = 10,155)

Independent Variables	Regression Coefficient	Standard Error	P Value	Odds Ratio
Intercept	-2.5816	0.2476	<0.0001	–
Age	-0.0015	0.0029	0.6088	0.9985
Female	0.4134**	0.1011	< 0.0001	1.5119
Educational attainment	-0.0667**	0.0240	0.0055	0.9355
White	0.1727	0.1495	0.2482	1.1885
Black	0.6022**	0.1613	0.0002	1.8261
Currently married	0.0318	0.0902	0.7245	1.0323
Veteran status	0.0217	0.1358	0.8730	1.0219
Socioeconomic status score	-0.0026	0.0034	0.4451	0.9974
Disability payment	0.3521	0.1983	0.0757	1.4221
Model Chi-sqaure	94.3581**			

** $p < 0.01$, two-tailed.

Table 4. Likelihood Ratios Plus and Minus for Dichotomized Level of Pains: Sample Figures and Population Estimates

Derived from Logistic Regression Adjusting for Propensity Scores (N = 10,155)

Diagnostic Statistics	Diagnostic Score	Standard Error for Test Score	Lower 95% Confidence Interval	Upper 95% Confidence Interval
	Sample figures without adjusting for propensity scores			
Sensitivity (prob. of true positives)	0.8607	–	–	–
Specificity (prob. of true negatives)	0.5466	–	–	–
Likelihood ratio plus	1.8984	0.0193	1.8280	1.9715
Likelihood ratio minus	0.2548	0.1031	0.2082	0.3119
	Population estimates adjusting for propensity scores			
Sensitivity (prob. of true positives)	0.8656	–	–	–
Specificity (prob. of true negatives)	0.5447	–	–	–
Likelihood ratio plus	1.9010	0.0372	1.8282	1.9738
Likelihood ratio minus	0.2468	0.0261	0.1956	0.2980

Table 5. Likelihood Ratios for Three Body Pain Levels: Sample Figures and Population Estimates Derived from Multinomial Logistic Regression Adjusting for Propensity Scores (N = 10,155)

Body Pain Levels	Likelihood Ratio	Standard Error for Likelihood Ratio	Lower 95% CI for Likelihood Ratio	Upper 95% CI for Likelihood Ratio
	Sample figures without adjusting for propensity scores			
Body Pain Level 1 (Pains = 4 – 5)	2.7466	0.0565	2.4586	3.0684
Body Pain Level 2 (Pains = 2 – 3)	1.5169	0.0455	1.3876	1.6582
Body Pain Level 3 (Pains = 0 – 1)	0.2548	0.1031	0.2082	0.3119
	Population estimates adjusting for propensity scores			
Body Pain Level 1 (Pains = 4 – 5)	3.0101	0.1941	2.6296	3.3906
Body Pain Level 2 (Pains = 2 – 3)	1.5592	0.0644	1.4330	1.6854
Body Pain Level 3 (Pains = 0 – 1)	0.2486	0.0217	0.2061	0.2911