Biometrika (2008), **95**, 2, *pp*. 481–488
doi: 10.1093/biomet/asn004

Printed in Great Britain

The prognostic analogue of the propensity score

BY BEN B. HANSEN

Statistics Department, University of Michigan 439 West Hall, Ann Arbor, Michigan 48109, U.S.A.

ben.b.hansen@umich.edu

SUMMARY

The propensity score collapses the covariates of an observational study into a single measure summarizing their joint association with treatment conditions; prognostic scores summarize covariates' association with potential responses. As with propensity scores, stratification on prognostic scores brings to uncontrolled studies a concrete and desirable form of balance, a balance that is more familiar as an objective of experimental control. Like propensity scores, prognostic scores can reduce the dimension of the covariate, yet causal inferences conditional on them are as valid as are inferences conditional only on the unreduced covariate. As a method of adjustment unto itself, prognostic scoring has limitations not shared with propensity scoring, but it holds promise as a complement to the propensity score, particularly in certain designs for which unassisted propensity adjustment is difficult or infeasible.

Some key words: Covariate balance; Matched sampling; Matching; Observational study; Quasi-experiment; Regression discontinuity; Subclassification.

1. INTRODUCTION

Following Neyman (1990) and Rubin (1977), let us construe an intervention's effect in terms of potential outcomes, as the ensemble of differences between subjects' potential responses to treatment, y_t, and control, y_c. The basic challenge for measurement of effects is that at most one of these two outcomes is observed, according as the subject did, $z = 1$, or did not, $z = 0$, receive the treatment. This difficulty is most conclusively surmounted when the potential outcomes arise through stable, repeatable processes with known chance properties, as in laboratories with highly controlled experimental conditions, and when treatment conditions are decided by a stable, repeatable process with known chance properties, as in field or clinical studies with random assignment to treatment (Holland, 1986).

Rosenbaum & Rubins (1983) showed that modelling the observed pattern of treatment, z, as a function of covariates, $x = (x_1', \ldots, x_k')'$, collapses those covariates into a scalar, the propensity score, upon which it is beneficial to condition. The demonstration requires no ostensible assignment mechanism to exist outside the statistical model, and suggests that, even when treatment assignment models are misspecified, propensity-score stratification is likely to reduce, if not eliminate, bias. If casual models of treatment assignment favourably reduce the dimension of x, then dimension reductions of x that arise from modelling $p(y_c|x)$ should also be favourable, either as alternatives or as complements to propensity scores.

If $\Psi(X)$ is sufficient for Y_c, in the sense that $Y_c \perp X|\Psi(X)$, we call $\Psi(X)$ a prognostic score. Throughout the paper, $\perp$ denotes independence of random variables. Should Y_c follow a shift model, $p(y_c|x) \equiv f_c\{y_c - h(x)\}$ for some fixed f_c, then $h(X)$ is a prognostic score. Should $Y_c|X$

follow a generalized linear model (McCullagh & Nelder, 1989), then the linear predictor of Y_c given X is a prognostic score, as is the scalar $E(Y_c|X)$. The propensity score, $E(Z|X)$, is also a sufficient statistic, sufficient for Z rather than Y_c; whereas propensities are scalars, when Y_c is not binary there may be multi-dimensional prognostic scores. Should Y_c have a linear regression on X, for example, but with nonconstant variance depending on X, then the regression and variance functions taken together constitute a prognostic score. In prognostic scoring, one fits a model of $p(y_c|x)$ to some set of control subjects, extrapolating this fit to intervention and control subjects being compared.

In many settings, information about response in the absence of treatment, Y_c, is more available than are data on responses of treated subjects, Y_t; the definition of a prognostic score reflects an assumption that this is the case. When the reverse is true, it may be appropriate to concentrate on sufficient statistics for Y_t rather than Y_c. If so, much of what follows remains true provided that controls are relabelled as treatments, and vice versa.

A sufficient statistic for Y_c is not necessarily sufficient for Y_t. The difference has to do with the possibility that a covariate or function of the covariates modifies the effect of treatment; indeed, it suggests a general perspective on effect modification. We say that there is no effect modification if any prognostic score $\Psi(X)$ is sufficient for Y_t as well as Y_c, and we call $m(X)$ an effect modifier if, for any prognostic score $\Psi(X)$, $\{\Psi(X), m(X)\}$ is sufficient for Y_t, while at least one prognostic score $\Psi(X)$ is not sufficient for Y_t. Effect modification is sometimes identified with the presence of interaction terms involving Z in regression models for $\mathrm{pr}(Y|Z, X)$; these definitions support that usage, but also apply when there is no linear predictor in view. They allow there to be effect modification even when response surfaces for Y_t and Y_c given X are parallel, if for instance the dispersion of Y_t depends on X in a way that Y_c's dispersion does not. Whenever $\mathrm{pr}(Y_t|Y_c, X) = \mathrm{pr}(Y_t|Y_c)$, on the other hand, there can be no effect modification. Section 2 tracks various implications of effect modification and its absence.

Peters (1941), Belson (1956), Cochran (1969), Rubin (1984) and Gastwirth & Greenhouse (1995) suggest extracting $\hat{E}(Y_c|X = x)$ from a parametric-model fit to the control group, and then estimating the treatment effect as the treatment group mean of $y_i - \hat{E}(Y_{ci}|X = x_i)$. Miettinen (1976) proposes regression of Y on Z and X, followed by subclassification on the part of the linear predictor that is free of Z. Zhao (2004), in a suggestion echoed by Imbens (2004), proposes matching on a weighted combination of differences in covariates, using separate control- and treatment-group regressions of Y on X to determine the weights.

In contrast with the Peters–Belson and Zhao techniques, the primary aim of prognostic scoring is to reduce the dimension of the covariate; it combines with other techniques, such as matching or propensity scoring. Miettinen's scores sometimes coincide with prognostic scores, although, with due attention to issues to be discussed in § 3, estimated prognostic scores will generally differ from Miettinen's scores.

2. PROGNOSTIC CONDITIONING: POPULATION THEORY

2·1. *Prognostic balance*

Conditioning on the propensity score, $\phi(x) \equiv E(Z|X = x)$, secures a form of covariate balance:

$$X \perp Z|\phi(X)$$

(Rosenbaum & Rubins, 1983, Theorem 1). Within level sets of the propensity score, no covariate is associated with membership in the treatment or the control group. A quintessential benefit of experimental randomization is its tendency to impose this absence of association, here called

'propensity balance'. An observational study exhibiting propensity balance on scientifically important covariates is experiment-like, in that it resembles a randomized trial in salient observed characteristics.

In a second experimental ideal, it is the process by which outcomes are generated that is repeatable, understood and carefully controlled, not the process of assigning units to treatment. Studies approaching this ideal use experimental control in the interest of removing associations between covariates and potential outcomes, not treatment assignment. If, in advance of studying a new experimental manipulation, an investigator conducts tests without the new manipulation in order better to understand the accompanying conditions and their influence on the outcome, then it is this second ideal that his or her procedure seeks to attain. Such preparations may incompletely control those nonexperimental factors with the potential to influence the outcome of the trial, but they will have succeeded if in their wake uncontrolled variation in such factors is not systematically associated with trial outcomes. Should the investigator subsequently observe a systematic association between experimental manipulations and trial outcomes, this will be evidence of a treatment effect.

This form of balance, i.e., similarity among the covariate distributions of trials or subjects with contrasting potential outcomes, $Y_c \perp X$, is quite distinct from propensity balance, $Z \perp X$. We call it prognostic balance. Principles of sufficiency and of conditional independence support a theory of prognostic balance that parallels Rosenbaum & Rubin's (1983) account of propensity balance, with a few important differences.

PROPOSITION 1. *Let Y_c be the potential response to be controlled. Then $\Psi(X)$ is a prognostic score if and only if conditioning on it induces prognostic balance within domains determined by X,*

$$Y_c \perp X | \Psi(X), X \in A, \tag{1}$$

where A may be any measurable set.

If $m(X)$ is an effect modifier, then in addition to (1) one has $Y_t \perp X | \Psi(X), m(X), g(X)$ for all measurable $g(\cdot)$.

Proof. The 'if' implication is immediate. For the other direction, given A let $g(x) = 1$ if $x \in A$ and $g(x) = 0$ otherwise. Since $(\Psi(X), g(X))$ is sufficient for Y_c if $\Psi(X)$ is, (1) follows from the definition of a prognostic score. □

Proposition 1 supports the checking of prognostic scores on samples from which treatment has been entirely withheld. Since Y_c is observed only when $Z = 0$, it does not suggest any practicable tests for samples containing both control and treatment subjects. For those settings another principle is needed, one that is valid only in the absence of hidden bias, or confounding due to omitted variables: i.e., only when

$$Y_c \perp Z | X. \tag{2}$$

This added condition marks one noteworthy difference between prognostic and propensity diagnostics: propensity balance can validly be assessed even when important confounders have been omitted. However, note that in that case propensity balance is no longer sufficient for causal inference.

PROPOSITION 2. *In the absence of hidden bias (2), $\Psi(X)$ is a prognostic score if and only if conditioning on it induces prognostic balance over domains determined jointly by X and Z: for any measurable A,*

$$Y_c \perp X | \Psi(X), (X, Z) \in A.$$

Proof. The 'if' direction is immediate. For 'only if', $Y_c \perp Z|X$ says that $\text{pr}(Y_c \in \cdot|Z, X) \equiv \text{pr}(Y_c \in \cdot|X)$, so that $\text{pr}\{Y_c \in \cdot|Z, X, \Psi(X)\} \equiv \text{pr}\{Y_c \in \cdot|X, \Psi(X)\}$. Since $\Psi(X)$ is a prognostic score, $\text{pr}\{Y_c \in \cdot|Z, X, \Psi(X)\} \equiv \text{pr}\{Y_c \in \cdot|\Psi(X)\}$ follows. In particular, $\text{pr}\{Y_c \in \cdot|X, (Z, X) \in A, \Psi(X)\} \equiv \text{pr}\{Y_c \in \cdot|\Psi(X)\} \equiv \text{pr}\{Y_c \in \cdot|(Z, X) \in A, \Psi(X)\}$. □

2·2. *Absence of confounding within prognostically balanced strata*

PROPOSITION 3. *If there is no hidden bias (2), conditioning on a prognostic score deconfounds potential responses from treatment assignment*:

$$Y_c \perp Z|\Psi(X), X \in A,$$

for any A. If also $Y_t \perp Z|X$, *and there is no effect modification, then, for any A,*

$$Y_t \perp Z|\Psi(X), X \in A.$$

Proof. Combining (2) with the defining property of prognostic scores, we have

$$\text{pr}(Y_c, Z|X) = \text{pr}(Y_c|X)\text{pr}(Z|X) = \text{pr}\{Y_c|\Psi(X)\}\text{pr}(Z|X).$$

The joint distribution of Y_c and Z given $\Psi(X)$ is, then, expressible as the product of the distribution of Y_c given $b(X)$ and a distribution produced by conditioning the propensity score, $E(Z|X)$, on $\Psi(X)$. For the claim about Y_t, no effect modification entails this demonstration's validity when Y_t is substituted for Y_c throughout. □

2·3. *Direct adjustment with prognostic scores*

In the absence of hidden bias, by prognostic score subclassification one can estimate a treatment's effects upon treatment-group subjects, provided that there is no level of the prognostic subclassification at which subjects receive the treatment with certainty. This parallels a principle of propensity subclassification, with the difference that in propensity subclassification it is required that there be no level of the unreduced covariate at which subjects receive the treatment with certainty (Rosenbaum & Rubins, 1983; Heckman et al., 1998). The propensity condition may fail while the weaker condition, on prognostic scores, holds.

PROPOSITION 4. *Suppose that X deconfounds* Y_c *and Z, i.e.* $Y_c \perp Z|X$, *and that, with probability one,* $\text{pr}\{Z = 1|\Psi(X)\} < 1$. *Then*

$$E(Y_t - Y_c|Z = 1) = E[E\{Y|Z = 1, \Psi(X)\} - E\{Y|Z = 0, \Psi(X)\}|Z = 1].$$

Proof. Certainly $E\{Y|Z = 1, \Psi(X)\} = E\{Y_t|Z = 1, \Psi(X)\}$ and $E\{Y|Z = 0, \Psi(X)\} = E\{Y_c|Z = 0, \Psi(X)\}$, while Proposition 3 entails that $E\{Y_c|Z = 0, \Psi(X)\} = E\{Y_c|Z = 1, \Psi(X)\}$: in short, $E\{Y|Z = 1, \Psi(X)\} - E\{Y|Z = 0, \Psi(X)\} = E\{Y_t - Y_c|Z = 1, \Psi(X)\}$. □

Prognostic stratification, then, permits estimation of $E(Y_t - Y_c|Z = 1)$ under a weaker condition than does stratification on the propensity score. To estimate $E(Y_t - Y_c)$, the two approaches strengthen the conditions of Proposition 4 in parallel ways; but then prognostic stratification sometimes imposes an additional requirement. If there is effect modification, then valid estimation of overall treatment effects requires that it be captured in the conditioning statement.

PROPOSITION 5. *In the setting of Proposition 4, suppose in addition that* $\text{pr}\{Z = 1|\Psi(X)\} > 0$ *with probability 1, and* $Y_t \perp Z|X$. *If* $m(X)$ *modifies the effect of treatment, then*

$$E(Y_t - Y_c) = E[E\{Y|Z = 1, \Psi(X), m(X)\} - E\{Y|Z = 0, \Psi(X), m(X)\}]; \quad (3)$$

while if there is no effect modification then

$$E(Y_t - Y_c) = E[E\{Y|Z = 1, \Psi(X)\} - E\{Y|Z = 0, \Psi(X)\}]. \quad (4)$$

Proof. For (3), observe that the initial assumptions entail $E\{Y|Z = 0, \Psi(X), m(X)\} = E\{Y_c|\Psi(X), m(X)\}$, whereas the additional assumptions give $E\{Y|Z = 1, \Psi(X), m(X)\} = E\{Y_c|\Psi(X), m(X)\}$. When there is no effect modification, $m(X)$ in (3) can be taken to be degenerate, and (4) follows. □

3. Estimating Prognostic Scores: Two Caveats

3·1. *From theory to practice*

As in the basic theory of propensity scores, given by Rosenbaum & Rubins (1983), Propositions 1–5 refer literally only to unlikely cases in which the form of the score is known. In practice, both propensity and prognostic scores must be approximated, typically through the specification and fitting of a model. One hopes that estimated scores sufficiently like a known score in terms of balance, here Propositions 1 and 2, will share in known scores' capacity to deconfound treatment effects, Propositions 3–5. As it pertains to propensity scores, the hypothesis has been corroborated in a variety of studies (Drake, 1993; Dehejia & Wahba, 1999; Rubin & Thomas, 2000; Kurth et al., 2006; Rubin & Stuart, 2006); however, propensity balance can be checked for a whole sample, whereas prognostic balance can ordinarily be checked only in the control group. What does this suggest about adjustment with prognostic scores?

3·2. *The difficulty with same-sample estimation*

Overfitting affects both prognostic score and propensity estimation, but the fact that only controls contribute to the estimation of prognostic scores makes overfitting more acute for them, and potentially more consequential. To fix ideas, let there be treatment and control groups of size $n = 500$, in both of which Y_c and $X_1, \ldots, X_{10}$ are standard normal and mutually independent, so that the true propensity and prognostic scores are degenerate. Unaware of this, the statistician estimates a nondegenerate propensity score, fitting a logistic regression of Z on X to the sample as a whole, and a nondegenerate prognostic score, fitting a linear regression of Y_c on X to the control group only. Suppose that no additional control observations are available for fitting the prognostic score: this is same-sample estimation. Despite the absence of structural propensity or prognostic relationships, and the fact that neither of these regressions is likely to be declared significant by ordinary F tests with appropriate degrees of freedom, simulation readily verifies that, with high probability, sample deciles of the estimated propensity appear significantly to predict membership in the treatment group, and sample deciles of the estimated prognostic score appear significantly to predict controls' y_c-values.

These spurious rejections do not in themselves speak against using either technique to test for treatment effects, and indeed in this specific scenario such tests produce false positives no more than they should. By an artefact of regression, at higher deciles of the estimated prognosis controls have atypically high responses, while at lower deciles controls' responses are particularly low. However, the two biases tend to cancel; see the simulation results in Table 1, column 1. When the treatment and control groups are separated, however, so that comparisons of treatment and control subjects are concentrated at one or the other end of the scale, the two biases need no longer compensate for one another. Columns 2 and 3 of Table 1 give results from a simulation study in which the Ys and Xs are again unrelated but the comparison groups differ, in varying degrees, on X. To be specific, in each of 1000 replications there are $n = 500$ controls and $n = 500$ treatments, all with Y_c and $X_1, \ldots, X_{10}$ drawn from

Table 1. *Type* I *error rates after stratification on deciles of prognostic scores, as estimated from the same and from a separate sample of controls. The results suggest that same-sample estimation of prognostic scores may make inference less reliable, particularly when treatment and control groups are separated*

	$\bar{\phi}_t - \bar{\phi}_c$		
Basis for stratification	0	1	5
Propensity score, $\hat{\phi}(X)$	0·05	0·05	0·05
Prognostic score, $\hat{\Psi}(X)$			
Separate-sample	0·05	0·05	0·05
Same-sample	0·04	0·05	0·18

independent normal distributions, making $\Psi(X)$ degenerate. Covariates $X_2, \ldots, X_{10}$ are $\mathcal{N}(0, 1)$ in both groups, and in the control group X_1 is also $\mathcal{N}(0, 1)$; but the treatment group has either $\mathcal{N}(1, 1)$, in column 2, or $\mathcal{N}(5, 1)$, in column 3. This makes the propensity score nondegenerate: identifying $\phi(x)$ with $\text{logit}\{E(Z|X = x)\}$, one has $\phi(X) = X_1$. Throughout, propensity scores come from logistic regression of Z on X, prognostic scores are the regression predictions of Y_c extrapolated from a linear model fit either to the control group or to an independent sample of $n = 500$ controls, Y_t is set equal to Y_c, and the hypothesis that $Y_t \equiv Y_c$ is tested using a 0·05-level aligned rank test (Hodges & Lehmann, 1963). The results show that, when comparison groups differ substantially on X, adjustment based on same-sample estimation of prognostic scores can be much worse than no adjustment at all.

The potential for inference to be undermined in this way affects the Peters–Belson and Miettinen approaches also, and has been discussed in some detail by Barsky et al. (2002, § 2), who link it to issues of model misspecification. The simulation also shows that the difficulty was mitigated by estimating the scores on a separate sample of controls. In practice, a sample of historical controls might play this role; there is some precedent for this in case-control matching (Silber et al., 2001).

3·3. *Should the treatment group contribute to the prognosis for controls?*

If difficulties arise when a prognostic score must be extrapolated from a control to a treatment group, a possible solution is to estimate the scores using a model fitted to both groups. However, this solution has problems of its own, except in the uncommon event that much is known a priori about $\text{pr}(Y_t|Y_c, X)$. Consider settings such that neither $E(Y_c|X)$ nor $E(Z|X)$ is degenerate, but the true prognostic and propensity scores are unassociated. If treatment increases Y, but the analyst fits a regression ignoring the distinction between treatments and controls, then the estimated prognosis will tend to be a mixture of the true propensity and prognostic scores. Adjusting for it will tend to compare high-propensity, low-prognosis treatments to low-propensity, high-prognosis controls, downwardly biasing estimators of the treatment effect.

In general, the problem is not fixed simply by adding Z to the regression and taking as prognostic score the part of the linear predictor that is free of Z. To take a scenario of particular concern in economics (Heckman, 1997), add to the previous assumptions that $E(Y_t - Y_c|X)$ increases with the propensity to receive treatment. If the investigator fits an outcome regression with only a linear contribution from Z, then again the estimated prognosis will be a mixture of the true prognostic and propensity scores, because high-propensity treatments will tend to have larger y-values than low-propensity treatment units irrespective of their prognostic scores; again the treatment effect will be obscured. Were they available, checks of prognostic balance over the whole sample would be likely to reveal such problems. Checks that can be made using only

the control group would not necessarily reveal them, although potentially a check for prognostic balance along the estimated propensity score could.

To put the issue in more general terms, true prognostic scores and scores formed by fitting a correct or incorrect model only to controls do not carry information about $E\{Y_t - Y_c|\Psi(X)\}$, $E\{Y_t - Y_c|\hat{\Psi}(X)\}$ or $E(Y_t - Y_c|Z = 1)$ and the like, this fitting being beyond the influence of any y_t's. In other words, $\mathrm{pr}(Y_c, Z, X)$ may be seen as a nuisance parameter, in which case $\Psi(X)$ and $\hat{\Psi}(X)$ are partial ancillaries for the interest parameter $\mathrm{pr}(Y_t|Y_c, X)$. A conditionality principle supports conditioning on such statistics; see for example Cox & Hinkley (1974, § 2·2·8) or Pace & Salvan (1997, § 4·2). True or estimated propensity scores are generally ancillary in this sense, even if the estimator is based on a misspecified $\mathrm{pr}(Z|X)$, but prognostic scores fitted to both groups may fail to be ancillaries.

4. Discussion

Whenever conditioning on a prognostic score, $\Phi(X)$, purges treatment–control comparisons of confounding due to X, so too does conditioning on that score and any other function of the covariate. An attractive possibility is to match or subclassify on both prognostic and propensity scores; in situations where treatment and control samples are separated on the covariate, this may reduce extrapolation, minimizing the impact of errors of estimation in the prognostic score and of the impossibility of checking prognostic balance in the treatment group.

As a result of this impossibility, and because of complications associated with effect modification, adjustment based on prognostic scores will often require stronger assumptions than propensity adjustment. However, § 2·3 noted one respect in which prognostic adjustment makes weaker assumptions than propensity adjustment: inference after propensity adjustment requires that there be no level of X at which treatment is received with certainty, whereas prognostic adjustment requires only that there be no level of $\Psi(X)$ at which treatment is certain. Regression discontinuity designs (Campbell & Stanley, 1966; Berk & de Leeuw, 1999; Hahn et al., 2001) are characterized by the presence of a threshold in a covariate that determines eligibility for the treatment, or perhaps compulsion to receive it. This means that $\mathrm{pr}\{\mathrm{pr}(Z = 1|X) = 1\} > 0$, or even that $\mathrm{pr}(Z = 1|X = x) = 1$ for most or all x's represented in the treatment group, so propensity adjustment is not possible. However, it does not entail that the treatment group is characterized by $\mathrm{pr}\{Z = 1|\Psi(X) = \Psi(x)\} = 1$. It may be that, after prognostic scoring, some regression discontinuity designs can be deconfounded using propensity techniques. One might begin by constructing a multi-dimensional score from the fitting of several candidate models to controls, perhaps historical controls as suggested in § 3·2, diagnosing its suitability in the manner indicated above. This process might be repeated for each of several outcomes of interest, with the results joined into one transformation $\hat{\Psi}(X)$ of the covariate. If observations on either side of the covariate threshold are comparable in respects that matter for the outcomes, then after subclassifying on $\hat{\Psi}(X)$ there should be controls interspersed among the treatment subjects. If so, the analysis could proceed as if $\tilde{X} = \hat{\Psi}(X)$, not X, had been the covariate; one might next match or subclassify on a 'prognostic propensity', $\phi(\tilde{x})$, alone or in combination with other functions of $\tilde{x}$.

Acknowledgement

The author acknowledges helpful discussions with Jake Bowers, Jennifer Hill, Gary King, Paul Rosenbaum, Edward Rothman, Donald Rubin and Jeffrey Smith, and the helpful comments of Professor D. M. Titterington and three reviewers. Rosenbaum brought to his attention the issues

discussed in § 3·2. Research support came from the U.S. National Institutes of Child Health and Human Development and the U.S. National Science Foundation.

REFERENCES

BARSKY, R., BOUND, J., CHARLES, K. K. & LUPTON, J. P. (2002). Accounting for the black-white wealth gap: a nonparametric approach. *J. Am. Statist. Assoc.* **97**, 663–74.

BELSON, W. A. (1956). A technique for studying the effects of a television broadcast. *Appl. Statist.* **5**, 195–202.

BERK, R. A. & DE LEEUW, J. (1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design. *J. Am. Statist. Assoc.* **94**, 1045–52.

CAMPBELL, D. & STANLEY, J. (1966). *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin.

COCHRAN, W. G. (1969). The use of covariance in observational studies. *Appl. Statist.* **18**, 270–5.

COX, D. R. & HINKLEY, D. V. (1974). *Theoretical Statistics*. London: Chapman & Hall.

DEHEJIA, R. & WAHBA, S. (1999). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *J. Am. Statist. Assoc.* **94**, 1053–62.

DRAKE, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* **49**, 1231–6.

GASTWIRTH, J. & GREENHOUSE, S. (1995). Biostatistical concepts and methods in the legal setting. *Statist. Med.* **14**, 1641–53.

HAHN, J., TODD, P. & VAN DER KLAAUW, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69**, 201–9.

HECKMAN, J. (1997). Instrumental variables: a study of implicit behavioral assumptions in one widely used estimator. *J. Hum. Resour.* **32**, 441–62.

HECKMAN, J. J., ICHIMURA, H. & TODD, P. E. (1998). Matching as an econometric evaluation estimator. *Rev. Econ. Studies* **65**, 261–94.

HODGES, J. L. & LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34**, 598–611.

HOLLAND, P. W. (1986). Statistics and causal inference (with Discussion). *J. Am. Statist. Assoc.* **81**, 945–70.

IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev. Econ. Statist.* **86**, 4–29.

KURTH, T., WALKER, A., GLYNN, R., CHAN, K., GAZIANO, J., BERGER, K. & ROBINS, J. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am. J. Epidemiol.* **163**, 262–70.

MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models,* 2nd ed. London: Chapman and Hall.

MIETTINEN, O. S. (1976). Stratification by a multivariate confounder score. *Am. J. Epidemiol.* **104**, 609–20.

NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated by D. M. Dabrowska and T. P. Speed. *Statist. Sci.* **5**, 463–80.

PACE, L. & SALVAN, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. Singapore: World Scientific.

PETERS, C. C. (1941). A method of matching groups for experiment with no loss of population. *J. Educ. Res.* **34**, 606–12.

ROSENBAUM, P. R. & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

RUBIN, D. B. (1977). Assignment to treatment group on the basis of a covariate. *J. Educ. Statist.* **2**, 1–26. (Correction (1978), **3**, 384).

RUBIN, D. B. (1984). William G. Cochran's contributions to the design, analysis, and evaluation of observational studies. In *W. G. Cochran's Impact on Statistics*, Ed. P. S. Rao and J. Sedransk, pp. 37–69. New York: Wiley.

RUBIN, D. B. & STUART, E. A. (2006). Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *Ann. Statist.* **34**, 1814–26.

RUBIN, D. B. & THOMAS, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *J. Am. Statist. Assoc.* **95**, 573–85.

SILBER, J., ROSENBAUM, P., TRUDEAU, M., EVEN-SHOSHAN, O., CHEN, W., ZHANG, X. & MOSHER, R. (2001). Multivariate matching and bias reduction in the surgical outcomes study. *Med. Care* **39**, 1048–64.

ZHAO, Z. (2004). Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. *Rev. Econ. Statist.* **86**, 91–107.

[*Received June* 2006. *Revised November* 2007]