A Nonparametric Matching Method for Covariate Adjustment with Application to Economic Evaluation[*]

Jasjeet S. Sekhon[†] and Richard Grieve[‡]

Associate Professor
Travers Dept. of Political Science
UC Berkeley

Senior Lecturer
Health Services Research Unit
London School of Hygiene and Tropical Medicine

7/2/2009 (23:35)

[*] We gratefully acknowledge Sheila Harvey (LSHTM), David Harrison (ICNARC) and Kathy Rowan (ICNARC) for access to data from the PAC-Man CEA and the ICNARC CMP database; and Chris McCabe and Katherine Stevens (SCHARR) for access to the cost data. We thank Kevin Bowrin, Jim Burgess, John Cairns, James Carpenter, Carla Guerriero, Ben Hansen, Kara Hanson, Jan Abel Olsen, Rocío Titiunik and Jonathan Wand for comments. A version of this paper was presented at the International Health Economics Association, Copenhagen, July 2007; the Society for Medical Decision Making, Philadelphia, October 2008; and the Banff International Research Station conference on Causal Inference in Statistics, Banff, May 2009.

[†] Survey Research Center, 2538 Channing Way, UC Berkeley, 94720, <sekhon@berkeley.edu>, http://sekhon.berkeley.edu/

[‡] Health Services Research Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, <Richard.Grieve@lshtm.ac.uk>

Abstract

In health economic studies that use observational data, a key concern is how to adjust for imbalances in baseline covariates due to the non-random assignment of the programs under evaluation. Traditional methods of covariate adjustment such as regression, depend on correct model specification. Alternatives such as propensity score matching depend on covariate balance being achieved. We demonstrate a nonparametric matching method, Genetic Matching, which uses a search algorithm to optimize covariate balance. Genetic Matching is a generalization of propensity score and Mahalanobis distance matching. We apply Genetic Matching to an economic evaluation of a clinical intervention, Pulmonary Artery Catheterization. Our results show that Genetic Matching achieves better covariate balance than propensity score matching. Genetic Matching gives different estimates of incremental effectiveness and cost-effectiveness compared to propensity score matching. We conduct Monte Carlo simulations that show that Genetic Matching reduces bias and root mean squared error, compared to propensity score matching. We conclude that Genetic Matching improves covariate balance, and it can lead to less biased estimates than propensity score matching.

Keywords: matching methods; semiparametric and nonparametric methods; observational studies; health economic evaluation

1 Introduction

Progress has been made in statistical methods for health economic evaluations that use data from randomised controlled trials (RCTs) (Glick et al., 2007; Willan and Briggs, 2006). However, for many decision problems, relevant RCTs are unavailable or insufficient (Sculpher et al., 2006); indeed 80% of published economic evaluations rely on estimates from observational studies (OHE, 2005). Economic evaluations may use observational data for various purposes: for example to predict long-term outcomes from the intermediate measures collected in the RCT, or to estimate costs in routine clinical practice. For some decision problems the only evidence for comparing treatment alternatives[1] may come from non-randomised studies (NRS) (Deeks et al., 2003). When treatment assignment is non-random, the groups are drawn from different populations and failure to correct for the resulting baseline differences can lead to biased estimates (Basu et al., 2007, 2008; Jones, 2008).

Economic evaluations that use NRS therefore require methods that fully adjust for imbalances in baseline covariates between the treatment groups. For regression or matching methods to provide unbiased estimates, it is necessary to make the "selection on observables" assumption. If this assumption is invalid then these methods will fail to provide unbiased estimates (Basu et al., 2007). However, even if this assumption holds, using regression to adjust for baseline imbalances can be problematic: the results can be very sensitive to parametric assumptions especially if the baseline covariates are highly imbalanced (Rubin, 1997). In economic evaluations that use NRS, there may be little overlap in baseline covariates between the treatment groups so regression extrapolates outside the support of the data (Griffin et al., 2007; Grieve et al., 2008). To avoid making functional form assumptions, some researchers evaluating health care programs have moved to propensity score approaches (Merito and Pezzotti, 2006; Mitra and Indurkhya, 2005; Mojtabai and Zivin, 2003; Polsky and Basu, 2006). Matching on the estimated propensity score often results in less conditional bias than stratifying or using the propensity score as a covariate (Austin et al., 2007).

While propensity score matching makes less stringent parametric assumptions than regression,

[1] Treatment is used generally to refer to any clinical or public health intervention.

it does require that covariates are balanced between the treatment groups post matching. However, a review of propensity score matching in the medical literature found that almost all studies either failed to report covariate balance or only stated mean differences between treatment and control groups (Austin, 2008). Even if the mean of a continuous covariate is well matched, there can be differences elsewhere in the distribution leading to biased estimates (Sekhon, In Press).

This paper's aim is to demonstrate a nonparametric matching method, Genetic Matching, for addressing covariate imbalance in health economic evaluation. Genetic Matching is a generalization of propensity score and Mahalanobis distance matching (Diamond and Sekhon, 2005; Sekhon, In Press). While Genetic Matching can incorporate a propensity score, the method does not depend on knowing or estimating the propensity score. Genetic Matching performs multivariate matching using an evolutionary search algorithm to determine the weight each covariate is given with the aim of maximising the balance of observed potential confounders across treatment groups. Genetic Matching has already been shown to improve covariate imbalance in a wide range of applications (e.g., Gilligan and Sergenti, 2008; Gordon and Huber, 2007; Grieve et al., 2008; Heinrich, 2007; Herron and Wand, 2007; Korkeamäki and Uuistalo, In Press; Lenz and Ladd, 2009; Raessler and Rubin, 2005; Woo et al., In Press). However, the use of Genetic Matching in health economic evaluation has not been previously assessed.

This paper compares Genetic Matching with propensity score matching using a cost-effectiveness analysis (CEA) of Pulmonary Artery Catheterization (PAC). The results show that Genetic Matching achieves better covariate balance than propensity score matching, leading to different effectiveness and cost-effectiveness results. In Monte Carlo simulations we find that Genetic Matching markedly improves covariate balance and reduces bias and root mean squared error (RMSE) compared to propensity score matching.

The next section (2) provides an overview of causal inference, propensity score and Mahalanobis distance matching. Later in section 2, we introduce Genetic Matching (2.3), the motivating example (2.4-2.5), and the simulation study (2.6). In section 3 we present results from the case study and the simulations. In Section 4, we discuss the key findings.

2 Methods and Motivating Example

2.1 Causal Inference in NRS

In an RCT comparing treatment and control,[2] the two groups are drawn from the same population, treatment assignment is independent of all baseline variables. Asymptotically, the multivariate distribution of both observed and unobserved variables is equal between treatment and control groups, that is the distribution is *balanced.* By contrast in a NRS, treatment and control groups are usually from different populations, and baseline covariates are unlikely to be balanced across groups. In a NRS a common quantity of interest is, therefore, the Average Treatment effect for the Treated (ATT):

$$\tau \mid (T = 1) \quad = \quad E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 1), \tag{1}$$

where T_i is a treatment indicator equal to 1 or 0 according to whether unit i is in the treatment or control group, with the potential outcomes Y_{i1} and Y_{i0}, and where the expectation is taken over the distribution of treatment assignments. Equation 1 cannot be directly estimated because Y_{i0} is not observed for the treated. Progress can be made by assuming that selection for treatment only depends on observable covariates X. Then, one can assume that conditional on X, treatment assignment is unconfounded; the conditional distributions of the potential outcomes are the same for treated and control: $\{Y_0, Y_1 \perp\!\!\!\perp T\} \mid X$.

Following Rosenbaum and Rubin (1983), treatment assignment is strongly ignorable given a vector of covariates X if unconfoundedness and common overlap hold:

$$\begin{aligned} \{Y_0, Y_1 \perp\!\!\!\perp T\} \mid X \\ 0 < Pr(T = 1 \mid X) < 1 \end{aligned}$$

for all X. To estimate ATT, as opposed to the average treatment effect, strong ignorability can be

[2] The notation throughout is given for a study comparing a treatment and control group. Extensions to the case of multiple discrete treatments are straightforward (e.g., Imbens 2000, Rosenbaum 2002).

weakened to $\{Y_0 \perp\!\!\!\perp T\} \mid X$ and $Pr(T = 1 \mid X) < 1$ (Heckman et al., 1998).

Given strong ignorability, following Rubin (1974, 1977) we obtain

$$E(Y_{ij} \mid X_i, T_i = 1) = E(Y_{ij} \mid X_i, T_i = 0). \tag{2}$$

Equation 2 is a formalization of the "as-if random" assumption made in observational studies. By conditioning on observed covariates, X_i, treatment and control groups are balanced—i.e., the distributions of the potential outcomes between treatment and control groups are the same. Then, the ATT can be estimated by calculating

$$\tau \mid (T = 1) \;\;=\;\; E\left\{E(Y_i \mid X_i, T_i = 1) - E(Y_i \mid X_i, T_i = 0) \mid T_i = 1\right\}, \tag{3}$$

where the outer expectation is taken over the distribution of $X_i \mid (T_i = 1)$, which is the distribution of X in the treated group.

2.2 Propensity Score and Mahalanobis Distance Matching

The most straightforward and nonparametric way to condition on X is to exactly match on the covariates. This approach fails in finite samples if the dimensionality of X is large or if X contains continuous covariates. Therefore, alternative methods must be used.

A common alternative way to condition on X is to match on the probability of assignment to treatment, known as the propensity score. Let $e(X_i) \equiv Pr(T_i = 1 \mid X_i) = E(T_i \mid X_i)$, defining $e(X_i)$ to be the propensity score.

Given strong ignorability, Rosenbaum and Rubin (1983) prove that the propensity score can be used to estimate ATT:

$$\tau \mid (T = 1) \;\;=\;\; E\left\{E(Y_i \mid e(X_i), T_i = 1) - E(Y_i \mid e(X_i), T_i = 0) \mid T_i = 1\right\},$$

where the outer expectation is taken over the distribution of $e(X_i) \mid (T_i = 1)$.

Propensity score matching usually involves matching each treated unit to the nearest control unit on the unidimensional metric of the propensity score vector. Since the propensity score is generally unknown, it must be estimated.

Rosenbaum and Rubin (1985) demonstrate that rather than matching solely on the propensity score, covariate balance can be improved by combining propensity score matching with matching on the individual covariates using a multivariate distance metric such as Mahalanobis distance. Individual covariates are collapsed into a single scalar metric using Mahalanobis distance, which is defined as the generalization of the standardized distance from the origin of an n-dimensional space to a point where the coordinates represent the X values for a particular observation (Cochran and Rubin, 1973; Glance et al., 2007; Rubin, 1979, 1980). The Mahalanobis distance between any two column vectors is:

$$md(X_i, X_j) \quad = \left\{(X_i - X_j)' S^{-1} (X_i - X_j)\right\}^{\frac{1}{2}}$$

where S is the sample covariance matrix of X. To estimate ATT by matching with replacement, each treated observation is matched one-to-one with the M closest control observations, defined by the distance metric, $md()$—i.e. the matches are selected to minimize the Mahalanobis distance between the matched pairs.

Whichever matching method is chosen it is important to assess balance and then modify the propensity score model or choice of distance metric with the aim of improving the balance not just of the means, but also the cross products, squared terms and variances (Rosenbaum and Rubin, 1984; Rubin, 1997). However, achieving balance on a wide range of terms is challenging; it may be unclear how best to modify the propensity score or distance metric. Rather than using propensity score matching (with or without Mahanaobis distance matching), better balance may be achieved by using a search algorithm to identify those matches that optimise covariate balance between the treatment and control groups.

2.3 Genetic Matching

The aim of Genetic Matching is to maximise the balance between treatment groups across those potential confounders that are observed. Genetic Matching achieves this aim by performing multivariate matching using an evolutionary search algorithm to determine the weight each individual covariate is given. Genetic Matching is a generalization of propensity score and Mahalanobis distance matching (Diamond and Sekhon, 2005; Sekhon, In Press). If a reasonable propensity score model is available, it should be included as one of the covariates in the Genetic Matching algorithm, but the method does not depend on knowing or estimating the propensity score. More generally, Genetic Matching searches over the space of distance metrics (including Mahalanobis distance) to find the best metric for optimizing covariate balance. Genetic Matching generalises the Mahalanobis metric by including an additional weight matrix:

$$d(X_i, X_j) \quad = \left\{(X_i - X_j)' \left(S^{-1/2}\right)' W S^{-1/2}(X_i - X_j)\right\}^{\frac{1}{2}}$$

where W is a $k \times k$ positive definite weight matrix and $S^{1/2}$ is the Cholesky decomposition of S which is the variance-covariance matrix of X. The Genetic Matching algorithm uses the distance measure $d()$ in which (by default) all elements of W are zero except down the main diagonal. The main diagonal consists of k parameters that must be chosen. If each of these parameters are set equal to 1, $d()$ is the same as Mahalanobis distance. Hence both propensity score and Mahalanobis distance matching can be considered as special, limiting cases of Genetic Matching: if the propensity score contains all of the relevant information, the other variables will be given zero weight, and Genetic Matching will converge to the Mahalanobis distance if that proves to be the appropriate distance measure.

An important issue in Genetic Matching is therefore how to choose the free elements of, the weight matrix, W. By default, Genetic Matching uses cumulative probability distribution functions of standardized statistics. The goal is to minimize the distance between the empirical distribution functions of the covariates in treatment and control. For dichotomous variables, the distance between the means of the variables in the treatment and control groups is minimized. For other types of variables, moments other then the first may also be imbalanced. So a general measure of covariate imbalance is needed. By default, Genetic Matching minimizes the Kolmogorov-Smirnov (KS) statistic. This statistic measures the distance between the empirical distribution functions of two samples.

The empirical distribution function F_n for n independent and identically distributed observations X_i is defined as:

$$F_n(x) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x},$$

where $I_{X_i \leq x}$ is the indicator function, equal to 1 if $X_i \leq x$ and equal to 0 otherwise. The KS statistic for two samples is defined to be:

$$D_{n,n'} \quad = \quad \sup_X |F_n(x) - F_{n'}(x)|\,, \tag{4}$$

where the null hypothesis is rejected at level α if

$$\sqrt{\frac{nn'}{n+n'}} D_{n,n'} > K_\alpha. \tag{5}$$

Abadie (2002) proves that bootstrapping Equation 4 provides correct p-values even when there are point-masses, a condition under which Equation 5 no longer provides correct test levels. Equation 4 corresponds to the largest distance observed in the empirical Quantile-Quantile (eQQ) plot when the distribution of a variable is plotted in two difference samples, such as treatment and control.

The KS statistic is used to provide a nonparametric distance metric (Diamond and Sekhon, 2005; Sekhon, In Press). It is not used to conduct formal hypotheses tests, because no measure of balance is a monotonic function of balance in the estimand of interest (i.e., any level of covariate imbalance is a concern no matter how small), and because multiple tests are being undertaken.

Instead of the KS statistic, alternative distance metrics can be chosen but whatever the chosen loss function, the optimisation is both difficult and irregular and is therefore conducted with a genetic algorithm implemented in the `rgenoud` software package developed by Mebane and Sekhon (1998). Details of the genetic algorithm are provided in Sekhon and Mebane (1998) and Mebane and Sekhon (In Press) with the key theorems provided by Nix and Vose (1992) and Vose (1993).

2.4 Motivating Example: CEA of Pulmonary Artery Catheterization (PAC)

We apply Genetic and propensity score matching to a NRS evaluating PAC, an invasive and controversial cardiac monitoring device, widely used in the management of critically ill patients (Dalen, 2001; Finfer and Delaney, 2006). The controversy was fueled by a NRS using propensity scores that found PAC insertion for critically ill patients was associated with increased costs and mortality (Connors et al., 1996). Other NRS came to similar conclusions leading to reduced PAC use (Chittock et al., 2004). However, an RCT (PAC-Man) found no difference in mortality between PAC and no PAC groups (Harvey et al., 2005), which substantiated the concern that the NRS results were subject to selection bias (Sakr et al., 2005).

Our CEA of PAC uses observational data from the Intensive Care National Audit Research Centre (ICNARC) Case Mix Program (CMP) database. The ICNARC database contains information on case-mix, patient outcome and resource use for 200 critical care units in the UK (Harrison et al., 2004). A total of 57 units from the CMP collected additional, prospective data on PAC use for consecutive admissions between May 2003 and December 2004.[3] The selection of cases for PAC insertion was non-random based on clinical judgment. The NRS applied the same inclusion and exclusion criteria for individual patients as the corresponding UK PAC-Man RCT, which resulted in a sample of 1,052 PAC cases and 31,447 potential controls. Information was recorded at ICU admission on covariates previously found to be associated with hospital survival, PAC use, length of stay (LOS) and vital status at hospital discharge. The covariates included age, sex, past medical history (history), LOS prior to ICU admission (LOSpre), reasons for admission (admtype); surgical status (none, elective or emergency); a measure of chemical imbalance (base excess); whether

[3] Over this time period, 10 units recorded no PAC use and were excluded from this analysis, as were units participating in the RCT (PAC-Man Study).

an infection (infect) was present, teaching hospital status, whether the centre used an alternative monitoring device (altPAC); the number of ICU beds (beds), and the rate of PAC use (rate). For each admission, detailed physiology data were collected to report the number of organ failures, the mean acute physiology score and the baseline probability of death.

2.5 Applying the Matching Methods in the Motivating Example

The propensity score for PAC insertion was estimated using logistic regression, based on the model reported in Connors et al. (1996), supplemented with covariates recently recommended for case-mix adjustment for ICU admissions (Harrison et al., 2007; Harvey, 2009). The linear predictor for the propensity score is:

$$\begin{aligned}\mu \quad = \quad & \alpha + \beta_1\text{sex} + \beta_2\text{age} + \beta_3\text{history} + \beta_4\text{LOSpre} + \beta_5\text{admtype} + \beta_6\text{surgery} + \\ & \beta_7\text{base excess} + \beta_8\text{infect} + \beta_9\text{teaching} + \beta_{10}\text{altPAC} + \beta_{11}\text{beds} + \\ & \beta_k X_{ik} + \beta_l X_{il} + s(\text{age}) + s(\text{base excess}),\end{aligned} \tag{6}$$

where X_{ik} is a vector of dummy variables for 24 different combinations of organ failure and X_{il} is a vector of 10 physiological variables that are components of the summary variable for the baseline probability of death. Prior reasoning suggested that centre-level factors such as teaching hospital status, the number of beds in the ICU, and the availability of an alternative monitoring device may be associated with PAC insertion so these variables were also included. Nonlinearities in the continuous variables were considered by fitting restricted cubic splines, the terms s(age) and s(base excess) represent splines of degree three for age and base excess. Patients in the PAC group were matched to patients in the no PAC group according to the estimated propensity score.

Genetic Matching used the same individual covariates included in the propensity score. The algorithm minimized the differences in the means and the maximum differences in the two-sample eQQ-plots (see Equation 4). All the options for Genetic Matching were set to their default values apart from population size which was set to 5000 (see Sekhon, In Press, for details). Both methods were implemented by matching one-to-one (with replacement), as this approach minimizes the imbalance in the observed covariates (Abadie and Imbens, 2006).

We compared the two matching methods using appropriate measures of balance: eQQ-plots for continuous variables and the differences in means for binary variables. The CEA based on the matched datasets followed the same approach as that accompanying the PAC-Man RCT (Stevens et al., 2005) and reported hospital mortality, incremental lifetime QALYs, costs and net monetary benefits (INBs). For each method we considered: whether the odds ratio of mortality for PAC versus no PAC differed from one, and whether the INBs were positive. No exact variance estimator is currently available for data matched using an estimated propensity score (Imbens and Wooldridge, 2009), so confidence intervals were estimated using the nonparametric bootstrap conditional on the matched dataset (Hill and Reiter, 2006).[4]

2.6 Simulation Study Methods

The relative performance of Genetic and propensity score matching was assessed using Monte Carlo simulations. It was assumed that while the strong ignorability was satisfied, the functional form of the true propensity score was unknown. The distributions of covariates and outcomes taken from the PAC-Man RCT (n=1,014) were typical of CEA more generally (Harvey et al., 2008; Stevens et al., 2005); the covariates were dichotomous, or continuous with non-normal distributions; the lifetime QALYs had a typically irregular distribution with a spike at zero (hospital decedents) and a heavy right tail (young survivors), and the costs were skewed. To simulate NRS, individuals from the RCT were reassigned to either treatment group according to a known assignment mechanism given by Equation 7. This data generating process allowed for a nonlinear relationship between a covariate, in this case age, and treatment assignment, other covariates were assumed to have a linear relationship with treatment. These assumed nonlinearities for age were based on those observed in the NRS case study; for example between baseline probability of death

[4] The bootstrap unconditional on the matched dataset does not provide correct variance estimates (Abadie and Imbens, 2008).

and treatment.

$$\begin{aligned} P(Tr) = {} & -3 + 0.5\,\text{BaselineProbDeath} - 0.5\,\text{missing BaselineProbDeath} + \\ & 0.5\log(\text{age}_{<57.25}) + 0.3\log(\text{age}_{>57.25,<67.44}) - 0.1\log(\text{age}_{>67.44,<75.18}) + \\ & 0.3\log(\text{age}_{>75.18}) + 0.5\,\text{history} + 0.1\,\text{elecsurg} \times \log(\text{age}) \\ & 0.2\,\text{emersurg} \times \log(\text{age}) + 0.5\,\text{teaching} + 0.2\,\text{rate} \end{aligned} \tag{7}$$

Treatment assignment was assumed conditional only on these observed characteristics and differed randomly across replications according to this true propensity score.

Data generating processes for both costs and QALYs were obtained by applying models to the RCT data that recognised the nonlinear response surfaces. Costs were generated by a Generalised Linear Model (GLM) that assumed a Gamma distribution with a log link, where the linear predictor was:

$$\begin{aligned} \text{cost} = {} & \text{BaselineProbDeath} + \text{age}_{<50} + \text{age}_{>65,<75} + \text{age}_{>75} + \\ & \text{elecsurg} + \text{emersurg} + \text{teaching} + \text{rate} + \text{history} \end{aligned} \tag{8}$$

QALYs were modelled with a two-part model (Manning and Mullahy, 2001) where the first part was a logistic regression and the second part, a GLM assuming a Gamma distribution with an identity link. The linear predictor for both parts of the model was:

$$\text{QALYS} = \text{BaselineProbDeath} + \text{age}_{<40} + \text{age}_{>60} \tag{9}$$

The true mean incremental costs, QALYs and INBs (λ=£30,000 per QALY) for PAC versus no PAC were all set to zero. This data generating process was repeated across 1000 replications. Consistent with the Neyman-Rubin model, the only randomness across replications was the treatment allocation: costs and outcomes were fixed.

For each replication the propensity score was re-estimated by a logistic regression with an incorrect but richly specified model:

$$\begin{aligned}\widehat{\text{pscore}} \;=\; & \text{BaselineProbDeath} + \text{BaselineProbDeath}^2 + \text{missing BaselineProbDeath} + \\ & \text{age} \times \text{BaselineProbDeath} + \text{age}^2 \times \text{BaselineProbDeath} + \\ & \text{age} + \text{age}^2 + \text{age}^3 + \text{age}^4 + \text{elecsurg} + \text{emersurg} + \\ & \text{teaching} + \text{rate} + \text{history}\end{aligned}$$

Notwithstanding the complexity of the estimated propensity score model, it did not capture the nonlinearities in the true propensity score defined in Equation 7.

The Genetic Matching algorithm used a population size of 5000, and matched on the same covariates included in the misspecified propensity score. The algorithm was not given any information about the functional form of the true propensity score.

Performance was assessed according to covariate balance post matching, together with the bias and RMSE of the estimates. For each covariate we calculated the mean difference between treatment and controls in each Monte Carlo replication, and we report the medians of these differences across Monte Carlo draws.

3 Results

3.1 Motivating Example: Covariate Balance

Before matching, the treatment groups were unbalanced; the case-mix was much more severe in the PAC group (Table I). Both matching methods identified a control that matched each treated observation giving 1,052 matched pairs. Following propensity score matching, the balance of the baseline measures improved, and the means were similar across the groups. However, the importance of considering balance across the distribution of each baseline covariate is well illustrated by the eQQ-plots comparing the standardized distributions of the baseline probability of death across groups (Figure 1). Following propensity score matching, the means for this important confounding variable are well balanced. However, there is a large gap in the eQQ-plot at the lower end of the distribution; here the baseline probability of death is higher for the PAC group (Figure 1a).

Genetic Matching achieved better balance for each covariate than propensity score matching (Table I, Figure 1). The eQQ-plot shows that Genetic Matching achieves excellent balance for the key summary measure, the baseline probability of death, not just for the means but right across the distribution (Figure 1b). For each continuous variable the maximum gap in the eQQ-plot is smaller following Genetic Matching than for propensity score matching and this improved balance is summarized by the KS measure (columns 4 and 5, Table I). For example, following propensity score matching the mean acute physiology scores were similar across the groups, however the KS balance statistics show that there were still imbalances elsewhere in the distribution (D=0.05, p=0.06). While Genetic Matching also leads to similar mean physiology scores between the groups, in contrast to propensity score matching it improved balance right across the distribution (D=0.02, p=0.77). For all the categorical variables, including centre-level factors such as admission to teaching hospitals, balance is reported as difference in means, and here Genetic Matching again improved balance compared to propensity score matching (Table I).

3.2 Motivating Example: Outcomes

Before matching, the odds ratio for hospital mortality in the PAC versus no PAC group was 3.51(95% CI from 3.09 to 3.97). Following propensity score matching, the corresponding odds ratio was 1.22 (p=0.03) whereas following Genetic Matching the odds ratio was 1.09 (p=0.35).

The reduction in the mean lifetime QALYs associated with PAC was large following propensity score matching, whereas after Genetic Matching the reduction in QALYs was relatively small. Following both matching methods, the PAC group had higher mean LOS and hospitalization costs (Table II).

Following propensity score matching, the mean INBs for PAC compared to no PAC were highly negative at all levels of λ, and at λ=£30,000 the 95% CIs excluded zero (Table II). Following

Genetic Matching the mean INBs were higher, and at λ=£30,000 per QALY the 95% CIs included zero.

The sensitivity analysis found that using alternative matching strategies (matching with replacement or matching two controls for each treated observation) led to the same substantive conclusions but resulted in worse covariate balance for both methods. The results were also similar when the Genetic Matching algorithm was rerun including the propensity score (Equation 6) as well as the underlying covariates.

3.3 Simulation Study Results

The Monte Carlo results find that Genetic Matching improves balance for each covariate compared with propensity score matching. Over 1000 simulations, the median difference in the covariate means between treatment and control groups were small for both matching methods except for the age variable. For age, the median gap for propensity score matching (0.837) is ten times that for Genetic Matching (0.081).

As differences in means are only partially informative for continuous covariates, the gaps across the distributions are also examined using the D-statistics (Equation 4). Figure 2 presents the densities of the D-statistics from the eQQ-plots across the 1000 Monte Carlo draws. Using this measure, Genetic Matching still dominates propensity score matching, although the latter achieves reasonable balance for every covariate (for example baseline probability of death) aside from age. For age, propensity score matching produces far worse balance than Genetic Matching; this is to be expected given that in the true propensity score (Equation 7) all the covariates are linearly mapped to treatment assignment apart from age, where the relationship is highly nonlinear.[5]

Table III shows that the bias and RMSE is much lower following Genetic versus propensity score Matching. For the mean incremental QALYs, which are assumed to be more strongly as-

[5] The median differences in the D-statistics for the remaining variables for propensity score versus genetic matching are: baseline probability of death 0.00591 vs. 0.00240; emerg -0.00226 vs. 0.000; nonsurg -0.00490 vs. 0.000; univ 0.000 vs. 0.000; rate 0.00524 vs. 0.00503; and history 0.000 vs. 0.000. Genetic Matching (weakly) improves balance in every case.

sociated with age than costs, the relative bias following propensity score matching is especially high. Using this misspecified propensity score the bias for the mean INB is nearly 10 times that for Genetic Matching.[6]

4 Discussion

This paper demonstrates a nonparametric approach, Genetic Matching, for addressing covariate imbalance in health economic evaluations based on NRS. The paper extends previous studies that introduced propensity score methods for economic evaluation (Basu et al., 2008; Mitra and Indurkhya, 2005; Polsky and Basu, 2006). We find in a case study and Monte Carlo simulations that Genetic Matching dominates propensity score matching on measures of covariate balance. The simulations show that when the true propensity score is unknown, Genetic Matching can lead to less bias and lower RMSE.

In our case study, following propensity score matching, the intervention (PAC) was associated with increased hospital mortality (odds ratio 1.22, 95% CI from 1.03 to 1.45) whereas Genetic Matching reported that PAC had no effect on mortality (odds ratio 1.09, 95% CI from 0.91 to 1.29). The PAC-Man RCT also reported that PAC had no effect on mortality, either overall (odds ratio of 1.13, 95% CI from 0.87 to 1.47) or for specific subgroups (Harvey et al., 2008).[7] Our CEAs based on both matching methods and the RCT, reported that PAC had negative mean INBs. However, the mean INBs following propensity score matching were relatively low, with 95% that excluded zero, whereas the corresponding INBs following Genetic Matching and the RCT were higher with 95% confidence intervals that included zero. In general comparing results between NRS and RCTs is problematic because of methodological differences. However, these PAC studies both recruited from a similar population (UK ICUs during 2000-2004), used the same methods to measure costs

[6] If the true propensity score is used instead of the misspecified one, propensity score matching has lower bias than Genetic Matching, but Genetic Matching still has lower RMSE. These results are not presented for reasons of space, but are available upon request.

[7] A meta-analysis also concluded that PAC was not associated with increased mortality (Shah et al., 2005). Furthermore, randomization inference of the RCT cannot reject the sharp-null of no treatment effect.

and outcomes, and applied the same exclusion criteria; hence it is of interest that applying Genetic Matching to the NRS gave similar cost-effectiveness results to using the RCT data.

The case study compared Genetic Matching to matching on a previously published propensity score. According to conventional measures of balance such as difference in means, the groups were well balanced following propensity score matching, certainly compared to the few studies in the medical literature that have reported balance following matching (Austin, 2008). Therefore, a key lesson from our case study is that it is important to improve balance as much as possible: not just the differences in means across treatment groups for each covariate, but also more general nonparametric measures of balance such as those provided by eQQ-plots. We extended propensity score matching beyond the conventional approaches taken in the medical literature. We attempted to improve the propensity score model by including higher order terms, and interaction terms and then re-checking balance. However, like previous researchers (Basu et al., 2008), we were unable to specify a propensity score model that could achieve excellent balance in the matched data.

The relative advantage of Genetic Matching is that it uses an automated process to search and find the best matches in the data; achieving excellent levels of balance does not rely on the analyst correctly estimating the propensity score. The approach follows general recommendations for NRS and emphasises the importance of balancing baseline covariates without considering outcome data (Rubin, 2001, 2007). Genetic Matching allows the analyst to draw on prior knowledge about the relative importance of balancing different covariates: the user can stipulate the relative weight given to reducing imbalance for different covariates.

When there are wide imbalances between covariates at baseline, matching methods can improve on regression (e.g., Rubin, 1997). However, regression methods should be viewed as complementary to matching; indeed regression methods may be used to reduce residual biases once matching has removed most of the covariate imbalances—i.e., one can conduct post-matching bias adjustment (Abadie and Imbens, 2006). Hence, parametric and semiparametric adjustments proposed for health economic evaluations based on RCTs and unmatched NRS should be considered for matched data (Basu and Rathouz, 2005; Hoch et al., 2002; Manning and Mullahy, 2001; Nixon and Thompson, 2005; Willan et al., 2004). In our case study we found that the Genetic Matching results were robust after applying various parametric and nonparametric models to the matched data.

This paper has shown Genetic Matching can reduce conditional bias from differences in observable characteristics using an evaluation of a clinical intervention. The method has the potential to reduce bias in health economic evaluations more generally. However, it must be recognised that Genetic Matching is neither a panacea for eliminating bias in NRS nor a substitute for RCTs. Genetic Matching, like the other methods discussed, relies on the selection on observables assumption. The plausibility of this assumption cannot be tested statistically; it must be carefully scrutinised in each application using evidence beyond the statistical method (Freedman, 1991).

In conclusion, we find that, compared to propensity score matching, Genetic Matching improves balance on observed characteristics. Genetic Matching may therefore lead to less bias in common circumstances that face CEA: when the treatment assignment mechanism is unknown, the covariates have non-normal distributions and nonlinear relationships with outcomes. Where RCT data are unavailable or insufficient, Genetic Matching may reduce the bias that is due to observables, and allow CEAs to provide a firmer basis for policy-making than conventional methods.[8]

[8] Software for Genetic Matching and a variety of other matching algorithms is available in the Matching package for R by Sekhon. See `http://sekhon.berkeley.edu/matching/`.

References

Abadie A. 2002. Bootstrap tests for distributional treatment effect in instrumental variable models. *Journal of the American Statistical Association* 97: 284–292.

Abadie A, Imbens G. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74: 235–267.

Abadie A, Imbens G. 2008. On the failure of the bootstrap for matching estimators. *Econometrica* 76: 1537–1557.

Austin P. 2008. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine* 27: 2037–2049.

Austin P, Grootendorst P, Anderson G. 2007. A comparison of the ability of different propensity score models to balance measured variables between the treated and untreated subjects: a monte carlo study. *Health Economics* 26: 734–753.

Basu A, Heckman J, Navarro-Lozano S, Urzua S. 2007. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. *Health Economics* 16: 1153–1157.

Basu A, Polsky D, Manning W. 2008. Use of propensity scores in non-linear response models: the case for health care expenditures. NBER Working Paper No. 14086. [accessed April 1, 2009]. `http://www.nber.org/papers/w1408602138`

Basu A, Rathouz P. 2005. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* 6: 93–109.

Chittock D, Dhingra V, Ronco J, Russell J, Forrest D, Tweeddale M, Fenwick J. 2004. Severity of illness and risk of death associated with pulmonary artery catheter use. *Critical Care Medicine* 32: 911–915.

Cochran W, Rubin D. 1973. Controlling bias in observational studies: A review. *Sankhya,* Ser. A 35: 417–446.

Connors A, Speroff T, Dawson N, Thomas C, Harrell F, Wagner D, Desbiens N, Goldman L, Wu A, Califf R, Fulkerson W, Vidaillet H, Broste S, Bellamy P, Lynn J, Knaus W. 1996. The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association* 276: 889–897.

Dalen J. 2001. The pulmonary artery catheter—friend, foe, or accomplice? *Journal of the American Medical Association* 286: 348–350.

Deeks J, Dinnes J, D'Amico R, Sowden A, Sakarovitch C, Song F, Petticrew M, Altman D. 2003. Evaluating non-randomised intervention studies. *Health Technology Assessment* 7: 1–173.

Diamond A, Sekhon J. 2005. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. Working Paper.

Finfer S, Delaney A. 2006. Pulmonary artery catheters as currently used, do not benefit patients. *British Medical Journal* 333: 930–1.

Freedman D. A. 1991. Statistical models and shoe leather. *Sociological Methodology* 21: 291–313.

Gilligan M. J, Sergenti E. J. 2008. Do un interventions cause peace? using matching to improve causal inference. *Quarterly Journal of Political Science* 3: 89–122.

Glance L, Osler T, Mukamel D, Dick A. 2007. Use of a matching algorithm to evaluate hospital coronary artery bypass grafting performance as an alternative to conventional risk adjustment. *Medical Care* 45: 292–9.

Glick H, Doshi J, Sonnad S, Polsky D. 2007. *Economic Evaluation in Clinical Trials*. Oxford: Oxford University Press. Series editors Gray A, Briggs A.

Gordon S, Huber G. 2007. The effect of electoral competitiveness on incumbent behavior. *Quarterly Journal of Political Science* 2: 107–138.

Grieve R, Sekhon J, Hu T, Bloom J. 2008. Evaluating health care programs by combining cost with quality of life measures: A case study comparing capitation and fee for service. *Health Services Research* 43: 1204–1222.

Griffin S, Barber J, Manca A, Sculpher M, Thompson S, Buxton M, Hemingway H. 2007. Cost effectiveness of clinically appropriate decisions on alternative treatments for angina. *British Medical Journal* 334: 624–8.

Harrison D, Brady A, Rowan K. 2004. Case mix, outcome and length of stay for admissions to adult, general critical care units in england, wales and northern ireland: the intensive care national audit & research centre case mix programme database. *Critical Care* 8: R99–111.

Harrison D, Parry G, Carpenter J, Short A, Rowan K. 2007. A new risk prediction model for critical care: the intensive care national audit & research centre (icnarc) model. *Critical Care Medicine* 36: 1091–8.

Harvey S. 2009. A comparison of randomised and non-randomised study designs to evaluate health care interventions. Ph.D. Thesis, University of London.

Harvey S, Harrison D, Singer M, Ashcroft J, Jones C, Elbourne D, Brampton W, Williams D, Young D, Rowan K. 2005. An assessment of the clinical effectiveness of pulmonary artery catherters in patient management in intensive care (pac-man): a randomized controlled trial. *Lancet* 366: 472–77.

Harvey S, Welch C, Harrison D, Rowan K, Singer M. 2008. Post hoc insights from pac-man—the uk pulmonary artery catheter trial. *Critical Care Medicine* 36: 1714–21.

Heckman J. J, Ichimura H, Smith J, Todd P. 1998. Characterizing selection bias using experimental data. *Econometrica* 66: 1017–1098.

Heinrich C. J. 2007. Demand and supply-side determinants of conditional cash transfer program effectiveness. *World Development* 35: 121–143.

Herron M, Wand J. 2007. Assessing partisan bias in voting technology: The case of the 2004 new hampshire recount. *Electoral Studies* 26: 247–261.

Hill J, Reiter J. 2006. Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine* 25: 2230–56.

Hoch J, Briggs A, Willan A. 2002. Something old, something new, something borrowed, something blue: A framework for the marriage of econometrics and cost-effectiveness analysis. *Health Economics* 11: 415–430.

Imbens G. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87: 706–710.

Imbens G, Wooldridge J. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47: 5–86.

Jones A. 2008. Identification of treatment effects in health economics. *Health Economics* 16: 1127–31.

Korkeamäki O, Uuistalo R. In Press. Employment and wage effects of a payroll-tax cut—evidence from a regional experiment. *International Tax and Public Finance*.

Lenz G. S, Ladd J. M. 2009. Exploiting a rare communication shift to document the persuasive power of the news media. *American Journal of Political Science* 53: 394–410.

Manning W, Mullahy J. 2001. Estimating log models, to transform or not to transform? *Journal of Health Economics* 20: 461–494.

Mebane, Jr. W, Sekhon J. 1998. Genetic optimization using derivatives (genoud). Software Package. `http://sekhon.berkeley.edu/rgenoud/`.

Mebane, Jr. W, Sekhon J. In Press. Genetic optimization using derivatives: The `rgenoud` package for `R`. *Journal of Statistical Software*.

Merito M, Pezzotti P. 2006. Comparing costs and effectiveness of different starting points for highly active antiretroviral therapy in hiv-positive patients. evidence from the icona cohort. *European Journal of Health Economics* 7: 30–6.

Mitra N, Indurkhya A. 2005. A propensity score approach to estimating the cost-effectiveness of medical therapies from observational data. *Health Economics* 14: 805–815.

Mojtabai R, Zivin J. 2003. Effectiveness and cost-effectiveness of four treatment modalities for substance disorders: a propensity score analysis. *Health Services Research* 38: 233–59.

Nix A, Vose M. 1992. Modeling genetic algorithms with markov chains. *Annals of Mathematics and Artificial Intelligence* 5: 79–88.

Nixon R, Thompson S. 2005. Incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics* 14: 1217—1229.

Office of Health Economics (OHE). 2005. *OHE—Health Economic Evaluation Database*. London: OHE.

Polsky D, Basu A. 2006. Selection bias in observational data. In Jones A, editor, *The Elgar Companion to Health Economics* Cheltenham, UK: Edward Elgar. page Chapter 13.

Raessler S, Rubin D. 2005. Complications when using nonrandomized job training data to draw causal inferences. *Proceedings of the International Statistical Institute*.

Rosenbaum P. 2002. *Observational Studies*. New York: Springer-Verlag 2nd edition.

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

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

Rosenbaum P, Rubin D. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39: 33–38.

Rubin D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66: 688–701.

Rubin D. 1977. Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics* 2: 1–26.

Rubin D. 1979. Using multivariate sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* 74: 318–328.

Rubin D. 1980. Bias reduction using mahalanobis-metric matching. *Biometrics* 36: 293–298.

Rubin D. 1997. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* 127: 757–763.

Rubin D. 2001. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology* 2: 169–188.

Rubin D. 2007. The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine* 26: 20–30.

Sakr Y, Vincent J, Reinhart K, Payen D, Wiedermann C, Zandstra D, Sprung C. 2005. Sepsis occurrence in acutely ill patients investigators. use of the pulmonary artery catheter is not associated with worse outcome in the ICU. *Chest* 128: 2722–31.

Sculpher M, Claxton K, M D, McCabe C. 2006. Whither trial-based economic evaluation for trial-based decision-making? *Health Economics* 15: 677–688.

Sekhon J. In Press. Matching: Multivariate and propensity score matching with automated balance search. *Journal of Statistical Software*. Computer program available at `http://sekhon.berkeley.edu/matching/`.

Sekhon J, Mebane, Jr. W. 1998. Genetic optimization using derivatives: Theory and application to nonlinear models. *Political Analysis* 7: 189–203.

Shah M, Hasselblad V, Stevenson L, Binanay C, O'Connor C, Sopko G, Califf R. 2005. Impact of the pulmonary artery catheter in critically ill patients: meta-analysis of randomized clinical trials. *Journal of the American Statistical Association* 294: 1664–70.

Stevens K, McCabe C, Jones C, Ashcroft J, Harvey S, Rowan K. 2005. The incremental cost effectiveness of withdrawing pulmonary artery catheters from routine use in critical care. *Applied Health Economics and Health Policy* 4: 257–64.

Vose M. 1993. Modeling simple genetic algorithms. In Whitley L. D, editor, *Foundations of Genetic Algorithms 2* San Mateo, CA: Morgan Kaufmann.

Willan A, Briggs A. 2006. *Statistical Analysis of Cost-Effectiveness Data*. Wiley.

Willan A, Briggs A, Hoch J. 2004. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics* 13: 461–475.

Woo M, Reiter J, Karr A. In Press. Estimation of propensity scores using generalized additive models. *Statistics in Medicine*.

Table I: Balance Before Matching and After Propensity Score and Genetic Matching

	PAC	No PAC	D-statistic	p-value
Mean baseline probability of death				
before matching	56.2	30.8	0.40	< 0.001
after Propensity Score		55.1	0.06	0.03
after Genetic Matching		56.1	0.03	0.87
Mean acute physiology score				
before matching	27.7	17.9	0.43	< 0.00
after Propensity Score		28.0	0.05	0.06
after Genetic Matching		27.6	0.02	0.77
Mean age				
before matching	61.9	60.1	0.06	< 0.001
after Propensity Score		61.2	0.04	0.41
after Genetic Matching		62.0	0.03	0.61
Mean LOS prior to ICU (days)				
before matching	5.73	4.56	0.08	< 0.001
after Propensity Score		5.58	0.02	0.91
after Genetic Matching		5.42	0.02	0.79
% admitted for elective surgery				
before matching	9.3	26.1		< 0.001
after Propensity Score		10.1		0.56
after Genetic Matching		9.1		0.41
% admitted for emergency surgery				
before matching	23.1	20.2		0.03
after Propensity Score		21.3		0.32
after Genetic Matching		23.7		0.66
% admitted to teaching hospital				
before matching	42.6	37.8		0.002
after Propensity Score		44.5		0.38
after Genetic Matching		42.6		1

Both matching methods are estimating ATT; so, the statistics for the PAC group are the same as unadjusted. The D-statistic is the maximum difference in the empirical QQ-plot—e.g., see Figure 1. For the continuous variables the D-statistic and the p-value are from bootstrapped KS test. For dichotomous variables the p-values are from paired $t-$tests.

Table II: Hospital Mortality, QALYS, Cost, and Incremental Net Benefit (INB) of PAC vs No PAC According to Matching Method

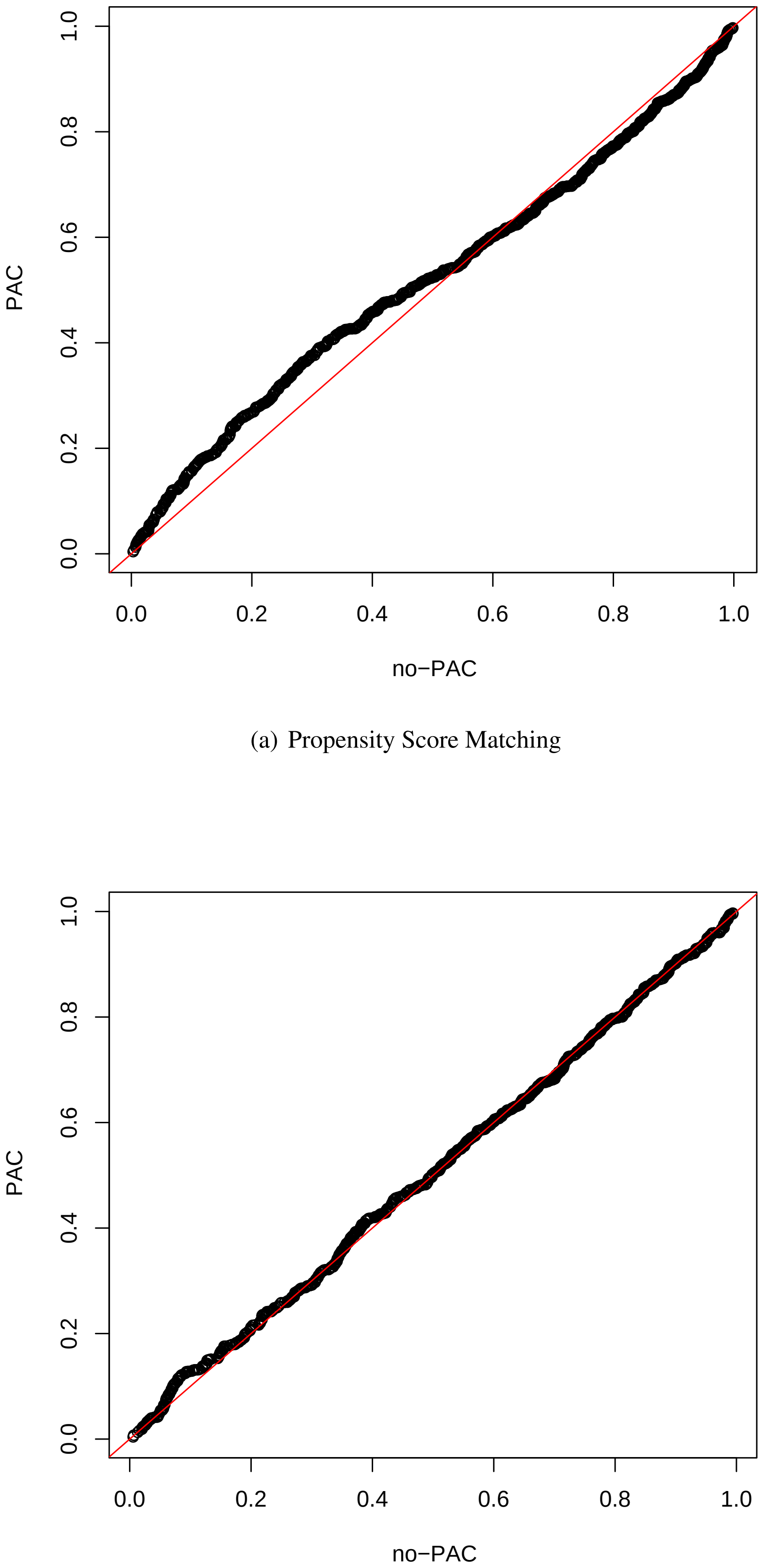
	Estimator	PAC	No PAC	PAC vs. No PAC
Deaths in Hospital		n (%)	n (%)	odds ratio (95% CI)
	Propensity Score	624 (59%)	572 (54%)	1.22 (1.03 to 1.45)
	Genetic Matching	624 (59%)	603 (57%)	1.09 (0.91 to 1.29)
Mean QALY		mean (sd)	mean (sd)	mean difference (95% CI)
	Propensity Score	3.47 (4.96)	4.22 (5.48)	-0.75 (-1.19 to -0.31)
	Genetic Matching	3.47 (4.96)	3.69 (5.06)	-0.23 (-0.65 to 0.22)
Mean Cost (£)				
	Propensity Score	19,560 (24,373)	14,921 (18,876)	4,639 (2,736 to 6,471)
	Genetic Matching	19,560 (24,373)	14,531 (18,683)	5,029 (3,180 to 6,891)
Mean INB (λ=£30,000/QALY)				
	Propensity Score			-27,215 (-39,864 to -14,154)
	Genetic Matching			-11,830 (-24,960 to 834)

Table III: Monte Carlo, Estimates

Estimator	Bias	RMSE	$\frac{\text{Bias}}{\text{Bias GenMatch}}$	$\frac{\text{RMSE}}{\text{RMSE GenMatch}}$
Mean INB:				
Genetic Matching	980.39	4778.16		
Propensity Score	−9769.41	13328.23	9.96	2.78
Mean QALY:				
Genetic Matching	0.0248	0.159		
Propensity Score	−0.3462	0.460	13.96	2.89
Mean Cost:				
Genetic Matching	−237.15	263.77		
Propensity Score	−616.69	665.56	2.60	2.523

The true treatment effects are all zero. 1000 Monte Carlos. Estimating ATT with 1-to-1 matching with replacement. Genetic Matching results use a population size of 5000.

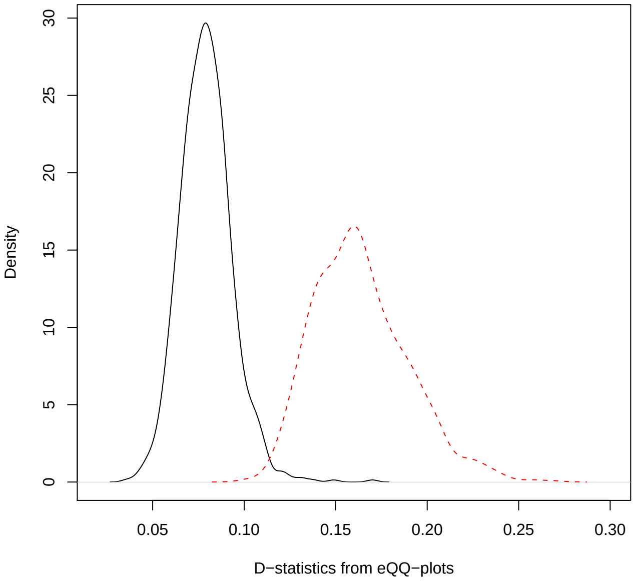
Figure 1: Balance of Baseline Probability of Death



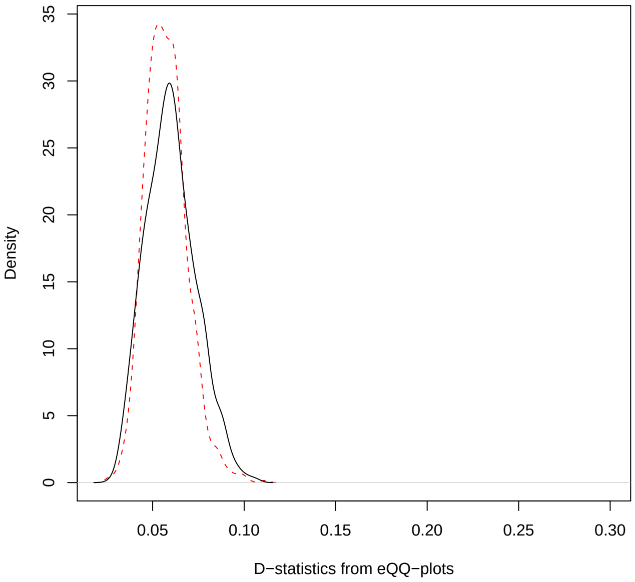
(a) Propensity Score Matching

(b) Genetic Matching

Figure 2: Covariate Balance Post-Matching in the Monte Carlo Simulations



(a) Age



(b) Baseline Probability of Death

Note: Solid black density is for Genetic Matching and the dashed red density is for Propensity Score matching.