Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1365-2753

Graphical displays for assessing covariate balance in matching studies

Ariel Linden DrPH[1,2]

[1]President, Linden Consulting Group, LLC, Ann Arbor, MI, USA
[2]Adjunct Associate Professor, Department of Health Management and Policy, School of Public Health, University of Michigan, Ann Arbor, MI, USA

Keywords
covariate balance, matching, propensity score, standardized differences, visual displays

Correspondence
Dr Ariel Linden
Linden Consulting Group, LLC
1301 North Bay Drive
Ann Arbor, MI 48103
USA
E-mail: alinden@lindenconsulting.org

Accepted for publication: 11 November 2014

doi:10.1111/jep.12297

Abstract

Rationale, aims and objectives An essential requirement for ensuring the validity of outcomes in matching studies is that study groups are comparable on observed pre-intervention characteristics. Investigators typically use numerical diagnostics, such as *t*-tests, to assess comparability (referred to as 'balance'). However, such diagnostics only test equality along one dimension (e.g. means in the case of *t*-tests), and therefore do not adequately capture imbalances that may exist elsewhere in the distribution. Furthermore, these tests are generally sensitive to sample size, raising the concern that a reduction in power may be mistaken for an improvement in covariate balance. In this paper, we demonstrate the shortcomings of numerical diagnostics and demonstrate how visual displays provide a complete representation of the data to more robustly assess balance.
Methods We generate artificial datasets specifically designed to demonstrate how widely used equality tests capture only a single-dimension of the data and are sensitive to sample size. We then plot the covariate distributions using several graphical displays.
Results As expected, tests showing perfect covariate balance in means failed to reflect imbalances at higher moments (variances). However, these discrepancies were easily detected upon inspection of the graphic displays. Additionally, smaller sample sizes led to the appearance of covariate balance, when in fact it was a result of lower statistical power.
Conclusions Given the limitations of numerical diagnostics, we advocate using graphical displays for assessing covariate balance and encourage investigators to provide such graphs when reporting balance statistics in their matching studies.

Introduction

While the randomized controlled trial (RCT) is the gold standard for evaluating the effectiveness of health-related interventions, its implementation is often not feasible because of logistical, practical or ethical reasons. As an alternative, investigators often choose from a wide variety of matching approaches in an attempt to emulate the randomization process using observational data. The primary difference between the RCT and matching approaches is that randomization should produce balance on both observed and unobserved covariates, while matching studies can only endeavour to obtain balance on observed covariates, and must assume that the unknown characteristics will not bias the results [1]. Therefore, observed covariate balance is an essential criterion for helping to ensure that treatment effects are valid in matching studies.

Covariate balance can be assessed using both numerical and graphical diagnostics [2,3]. In practice, matching studies typically report only numerical balance statistics. This may be due to the perception that the objective criteria provided by quantifiable summaries are preferable to graphical displays, which inherently require subjective interpretation [4]. However, numerical diagnostics have some important limitations that may actually constrain their effectiveness in assessing covariate balance. For example, the most widely used metrics, such as statistical tests to measure the equality of means (e.g. *t*-tests) or variances (e.g. variance comparison tests), can only assess covariate balance in one dimension (or moment). As a result, balance (which is typically qualified by a *P*-value greater than 0.05) may be observed in one dimension, but not in others. Thus, investigators would need to run an array of numerical tests on each covariate in order to get a representation of the overall balance. Even if investigators pursued this approach, they would be faced with a second limitation – these statistical tests are sensitive to sample size, which means that the reduction in sample size that typically occurs in matching strategies (as non-matched controls are dropped from the analysis) may cause covariates to appear balanced, when in fact a *P*-value greater than 0.05 is likely to be the result of reduced power [5]. One commonly used balance measure, the standardized difference [6], circumvents the sample size sensitivity issue by dividing the difference in the two sample means by their pooled standard deviation. However, as the standardized difference does not have an associated *P*-value, investigators can only assess the improvement in balance relative to the covariate in the 'pre-matching' state. Moreover, like the other balance measures mentioned above, the standardized difference is also limited to reporting balance in a single dimension.

There are two-sample tests available to measure the equality of overall distributions, such as the Kolmogorov–Smirnov (KS) test [7,8] and the Anderson–Darling test [9], but these tests also have limitations. For example, the KS test is known to have low sensitivity to deviations in the tails, while the Anderson–Darling test has low sensitivity in the middle of the distribution. Additionally, a brief review of the literature did not uncover any adaptations of these tests to accept probability weights or multiple samples, thereby limiting the use of these tests to 1:1 matching approaches. Finally, while a significant *P*-value on one of these tests indicates an inequality in the distributions, there is no way of knowing at which points along the continuum inequalities exist.

In this paper, we use simulated data to demonstrate the limitations of numerical diagnostics for assessing covariate balance, while highlighting the advantages of examining graphical displays of covariate distributions. We advocate the use of graphical displays as an integral component in the assessment of covariate balance and encourage investigators to provide such graphs when reporting balance statistics in their studies.

Methods

Scenario 1: sensitivity of two-sample tests to sample size

We demonstrate the sensitivity of the two-sample *t*-test to differences in sample size, holding all other statistics constant. Two variables were drawn from a multivariate normal distribution to represent the treatment and control groups' continuous baseline covariate. The treatment group's mean and standard deviation (SD) were 100 and 20, respectively, and the control group's mean and SD were 95 and 20, respectively, so that the mean difference was 5 points and the SD was neutralized. This data generating process was used to create two artificial datasets, one with a sample size of 1000 subjects per group and the second with a sample size of 100 per group. For each dataset, *P*-values were calculated from two-sample *t*-tests. Sensitivity to sample size would be demonstrated if the *P*-value was greater than 0.05 after reducing sample size.

Scenario 2: single-dimensional nature of equality tests in the two-sample setting

We demonstrate that statistical tests designed to measure the covariate equality in a particular dimension will inevitably miss inequalities in other dimensions. Two variables were drawn from a multivariate normal distribution to represent the treatment and control groups' continuous baseline covariate with both groups having sample sizes of 100 each. The treatment group's mean and SD were 100 and 20, respectively, and the control group's mean and SD were 100 and 10, respectively, so that the mean difference was 0 and the SD difference was 10. Four separate tests were then performed on these data: two-sample *t*-test, standardized difference [6], Wilcoxon rank-sum test [10] and the variance ratio test [11]. The single-dimensional nature of an equality test would be demonstrated if the *P*-value was greater than >0.05, when there is a known imbalance at other points in the distribution (demonstrated by the corresponding equality of variances test).

Scenario 3: sensitivity of the KS test

We demonstrate the low sensitivity of the KS test to deviations in the tails that appear to be associated with sample size. Two variables were drawn from a multivariate normal distribution to represent the treatment and control groups' continuous baseline covariate. The treatment group's mean and SD were 100 and 20, respectively, and the control group's mean and SD were 100 and 10, respectively, so that the mean difference was 0 and the SD difference was 10. This data generating process was used to create two artificial datasets, one with a sample size of 100 subjects per group and the second with a sample size of 50 per group. Additionally, as a sensitivity test for the reduced sample size, 50 subjects per group were randomly drawn from the first artificial dataset. For each scenario, *P*-values were calculated from two-sample KS test. Sensitivity to sample size would be demonstrated if the *P*-value was greater than 0.05 after reducing sample size.

Scenario 4: single-dimensional nature of equality tests in the multiple-sample setting

We demonstrate that, as in the two-sample case, statistical tests designed to measure the covariate equality in a particular dimension for multiple treatment groups will inevitably miss inequalities in other dimensions. Three variables were drawn from a multivariate normal distribution to represent a continuous baseline covariate for a three-level intervention. The first treatment group's mean and SD were 100 and 10, respectively, the second treatment group's mean and SD were 100 and 20, respectively, and the third treatment group's mean and SD were 100 and 30, respectively, so that the mean difference between all groups was 0 and only the SD differed between groups. *P*-values were calculated from a one-way analysis of variance (ANOVA) to test for equality in means between groups, and Levine's robust test statistic to test for equality of variances [12]. The single-dimensional nature of an equality test would be demonstrated if the *P*-value was greater than >0.05, when there is a known imbalance at other points in the distribution (demonstrated by the corresponding equality of variances test).

Graphical displays

We use several graphic displays to visualize and compare the distributions of the covariate between groups [13]. For the two-sample scenario, we utilize a quantile-quantile plot (Q-Q plot), box plot, histogram and kernel density plot. The Q-Q plot [14] graphs the quantiles of the covariate for the treatment group against the quantiles of the covariate for the control group. The goal in reviewing the Q-Q plot is to determine how and where the points deviate from the diagonal line representing perfect correlation between the two distributions. Box plots [15] provide a more concise summary of each distribution for comparison than the Q-Q plot by graphing the median, the upper and lower quartiles, the upper and lower adjacent values, and outliers. In a histogram, the data are divided into non-overlapping intervals (bins), and the number of data points within each interval is counted. The graph depicts these frequency counts – the bar is centred at the midpoint of each interval – and its height reflects the average number of data points in the interval. In a kernel density plot, the range is still divided into intervals, and estimates of the density at the centre of intervals are produced; however, the intervals are allowed to overlap and are smoothed [16]. For all graphic displays, the assessment of covariate balance is assessed by the degree of overlap in the respective distributions. For example, perfect covariate balance would make one group's distribution indistinguishable from another's.

We illustrate the distributions in the multiple-group scenario using a quantile plot and box plots. The quantile plot [14] is a variant of the Q-Q plot, in which the distributions of all groups are jointly plotted against the common quantile range (0, 1), rather than contrasting one distribution against another one.

Statistical software

All analyses were conducted using Stata 13.1 (StataCorp, College Station, TX, USA). Additionally, three graphic displays were generated through user-written Stata commands: the Q-Q plot used QQPLOT3 [17], which can plot both unweighted and weighted Q-Q plots; the quantile plot used QPLOT [18]; and the two-sample histogram used BYHIST [19].

Results

Table 1 presents the results for the first three scenarios in which two-sample comparisons are made. For the first scenario, a 5-point mean difference that is statistically significant ($P < 0.001$) with 1000 subjects per group becomes statistically non-significant ($P < 0.079$) when the number of subjects drops to 100 per group. For the second scenario, when the mean difference is 0, both the parametric *t*-test and non-parametric rank-sum test provide *P*-values close to 1.0 while the standardized difference is 0, all of which indicate near perfect balance in the means. However, the variance ratio test indicates that the variances (and thus SDs) were statistically different between the two groups in these data ($P < 0.001$). For scenario 3, the KS test identifies the inequality in the overall distributions with 100 subjects per group ($P < 0.039$), but when the sample size dropped to 50 per group (either in the original sample or by random sampling, 50 from the original sample), no statistically significant inequalities were found.

Figure 1 illustrates the simulated distributions of the covariate for the two groups, in scenario 2. As is clearly evident in all graphs, the distributions of the covariate are considerably different. However, it is more apparent in the Q-Q plot than in the other graphs that the only point in the distribution where there is no divergence between groups is at the mean. This is supported by the numerical summaries presented in Table 1.

Table 2 presents the results for scenario 4 in which three treatment groups are compared. As shown, when the mean difference is 0, the ANOVA provides a *P*-value 1.0, indicating a perfect balance among the three means. However, the robust test for comparisons of variances indicates that the variances were statistically different between the three groups in these data ($P < 0.001$).

Figure 2 illustrates the simulated distributions of the covariate for the three groups generated in scenario 4. As is evident in all graphs, the distributions of the covariate are considerably different between groups, with the quantile plot making it more apparent than in the other graphs that the only point in the distribution where there is no divergence between groups is at the mean. This is supported by the numerical summaries presented in Table 2.

Discussion

The results of these simulations indicate that numerical diagnostics alone provide incomplete, and possibly misleading, summaries of covariate balance. Specifically, we demonstrate that groups can exhibit covariate balance in the means, while exhibiting fundamentally different distributions. One may argue that if the ultimate treatment effects will be measured as the difference in means (e.g. average treatment effects), then achieving balance in

			Treated		Non-treated		
Scenario no.	Test	*n* (group)	Mean	SD	Mean	SD	*P*-value*
1	Two-sample *t*-test	1000	100	20	95	20	<0.001
1	Two-sample *t*-test	100	100	20	95	20	0.079
2	Two-sample *t*-test	100	100	20	100	10	1.000
2	Standardized difference	100	100	20	100	10	0.000[†]
2	Two-sample rank-sum	100	100	20	100	10	0.832
2	Variance ratio test	100	100	20	100	10	<0.001
3	Kolmogorov–Smirnov	100	100	20	100	10	0.039
3	Kolmogorov–Smirnov	50	100	20	100	10	0.206
3	Kolmogorov–Smirnov	50	100	20	100	10	0.078[‡]

*Unless otherwise noted.

[†] Standardized difference value (lower score is better with 0 being the floor).

[‡] Derived by sampling 50 subjects per group (of the original 100).

Table 1 Comparisons of covariate balance between simulated two-sample treatment and control groups under various scenarios

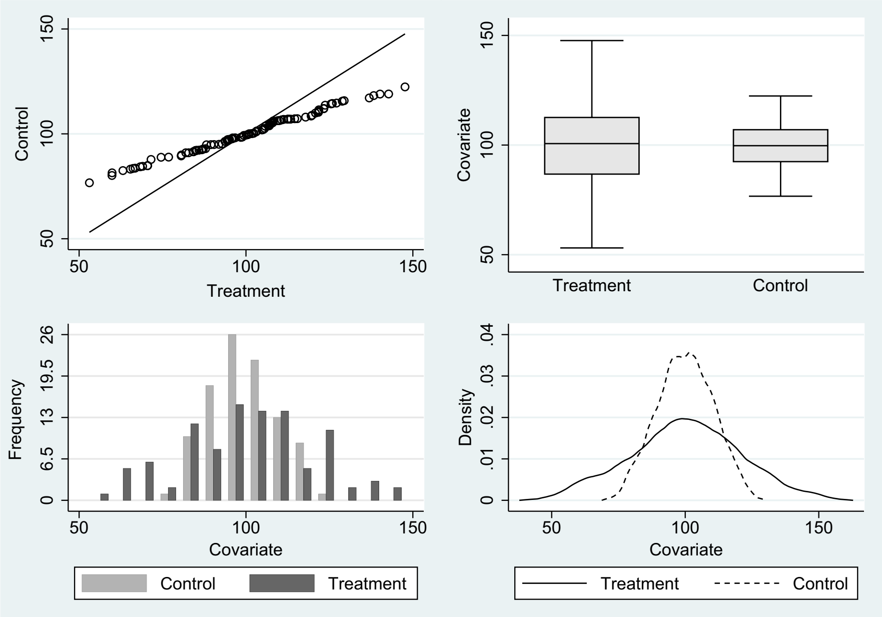

Figure 1 Graphic displays of the two-sample simulated data (treatment: *n* = 100, mean = 100, SD = 20; control: *n* = 100, mean = 100, SD = 10). Graph types are (clockwise from upper left): Q-Q plot, box plot, kernel density plot, histogram.

Table 2 Comparison of covariate balance between simulated multiple treatment groups

		n	Group 1		Group 2		Group 3		
Scenario no.	Test	(group)	Mean	SD	Mean	SD	Mean	SD	*P*-value
4	Analysis of variance	100	100	10	100	20	100	30	1.000
4	Robust variance test	100	100	10	100	20	100	30	<0.001

the means is all that is necessary. However, we also show that what may appear as covariate balance in the means is actually a result of decreased sample size that leads to reduced power. Moreover, decreasing sample size also appears to impact the results of the KS test, which is designed to compare the equality of distributions (as opposed to just a single dimension of the data). As most matching approaches will likely result in smaller samples (as non-matched individuals will be discarded from the analysis), non-significant *P*-values due to insufficient power will falsely appear as covariate balance. This, of course, raises concerns that if the groups are not truly comparable on observed baseline covariates, then the outcomes will be biased.

We also have demonstrated how graphs help visually inspect the equality of the distributions. In the simulated data, the graphic displays clearly illustrate that the groups are balanced in the means, but not at other points in the distribution, particularly in the tails where, by design, there was no overlap at all. This issue carries important meaning for health care studies in which imbalances in confounding variables may very likely result in biased outcomes. For example, assume that the covariate we generated in our second simulation represents a disease severity score. While, on average, the treatment and control groups have comparable severity scores, there is no overlap in the tails of the distribution, with the treatment group having both much lower and much higher scores than the control group. Thus, if a hypothetical evaluation indicated that the treatment group had better outcomes than controls, we would not know if it was due to a larger influence of the patients in the treatment group with lower disease severity, which, naturally, would be expected to have better outcomes than those with higher disease severity. Given that there are no comparable controls at the tails of the severity score, we cannot perform a direct comparison to address this question empirically.

Identifying where in the distribution imbalances lie can help investigators determine the best approach to pursue for adjustment. For example, if a propensity score [20] was used as the basis for matching, and visual displays identified distributional imbalances away from the mean, the investigator could re-estimate the propensity score with the inclusion of polynomials (i.e. squares and cubes) of the imbalanced variable, or perhaps utilize boosted logistic regression in lieu of standard logistic regression for estimating the propensity score. Regression boosting is a general, automated, data-adaptive modelling algorithm that can estimate the non-linear relationship between the outcome variable (in this case, treatment assignment) and a large number of covariates including multiple-level interaction terms resulting in greater accuracy over standard linear models [21]. This approach should result in better balance in the distribution of a given covariate as well as in the distribution of interacted variables.

Graphical displays of distributions have their limitations. Most notably, the extent of imbalance is not always clear from visual inspection. While the data generated for the current examples were specifically designed to illustrate obvious imbalances, in many datasets, the imbalances are not that apparent. For example, covariates may overlap for the greater part of their distributions

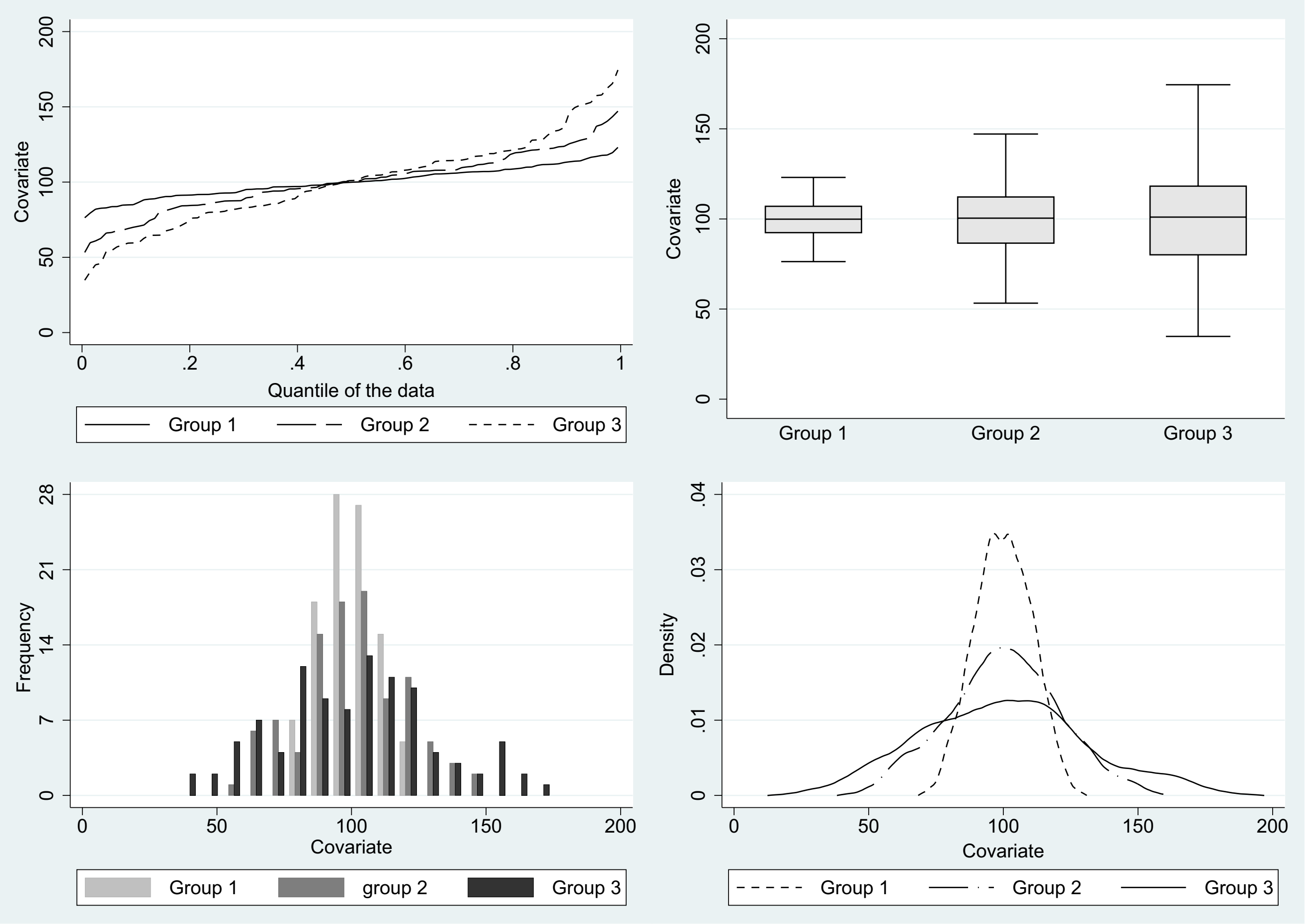

Figure 2 Graphic displays of the multiple-sample simulated data (group 1: *n* = 100, mean = 100, SD = 10; group 2: *n* = 100, mean = 100, SD = 20; group 3: *n* = 100, mean = 100, SD = 30). Graph types are (clockwise from upper left): quantile plot, box plot, kernel density plot, histogram.

with only small divergences, making balance difficult to assess. In such situations, supplementing graphic displays with complementary numerical diagnostics may be helpful – provided that the numerical diagnostics chosen capture multiple dimensions of the data.

In summary, visual displays allow us to gain a real insight into the underlying nature of the data while numerical diagnostics provide only a limited, and in some cases misleading, representation of those same data. We are not suggesting that investigators cease to use numerical diagnostics for testing and reporting covariate balance; rather, we are advocating that graphical displays of the distributions be considered an additional integral component in the assessment of covariate balance. We therefore encourage investigators to provide such graphs when reporting balance statistics in their studies.

Acknowledgements

I would like to thank Julia Adler-Milstein for reviewing the manuscript and providing helpful comments.

References

1. Rubin, D. B. (1973) Matching to remove bias in observational studies. *Biometrics*, 29, 159–184.
2. Austin, P. C. (2009) Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28, 3083–3107.
3. Stuart, E. A. (2010) Matching methods for causal inference: a review and a look forward. *Statistical Science*, 25 (1), 1–21.
4. Linden, A. & Samuels, S. J. (2013) Using balance statistics to determine the optimal number of controls in matching studies. *Journal of Evaluation in Clinical Practice*, 19 (5), 968–975.
5. Linden, A. (2008) Sample size in disease management program evaluation: the challenge of demonstrating a statistically significant reduction in admissions. *Disease Management*, 11 (2), 95–101.
6. Flury, B. K. & Reidwyl, H. (1986) Standard distance in univariate and multivariate analysis. *The American Statistician*, 40, 249–251.
7. Kolmogorov, A. N. (1933) Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4, 83–91.
8. Smirnov, N. V. (1933) Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow University*, 2, 3–16.

9. Anderson, T. W. & Darling, D. A. (1952) Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23 (2), 193–212.
10. Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics*, 1, 80–83.
11. Bland, M. (2000) An Introduction to Medical Statistics, 3rd edn. Oxford: Oxford University Press.
12. Levene, H. (1960) Robust tests for equality of variances. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling (eds I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow & H. B. Mann), pp. 278–292. Menlo Park, CA: Stanford University Press.
13. Chambers, J. M., Cleveland, W. S., Kleiner, B. & Tukey, P. A. (1983) Graphical Methods for Data Analysis. Belmont, CA: Wadsworth.
14. Wilk, M. B. & Gnanadesikan, R. (1968) Probability plotting methods for the analysis of data. *Biometrika*, 55, 1–17.
15. Tukey, J. W. (1977) Exploratory Data Analysis. Reading, MA: Addison–Wesley.
16. StataCorp (2013) Stata 13 Base Reference Manual. College Station, TX: Stata Press.
17. Linden, A. (2014) qqplot3: Stata module for plotting unweighted and weighted Q-Q plots. Available at: http://ideas.repec.org/c/boc/bocode/s457856.html (last accessed 1 November 2014).
18. Cox, N. J. (1999) gr42: quantile plots, generalized. *Stata Technical Bulletin*, 51, 16–18.
19. Nichols, A. (2010) byhist: Stata module to graph interlaced histograms: for comparing histograms by a categorical variable. Available at: http://ideas.repec.org/c/boc/bocode/s456982.html (last accessed 1 November 2014).
20. Rosenbaum, P. R. & Rubin, D. B. (1983) The central role of propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
21. McCaffrey, D., Ridgeway, G. & Morral, A. (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9 (4), 403–425.