The z-difference can be used to measure covariate balance in matched propensity score analyses

Kuss O[1]

[1]Institute of Medical Epidemiology, Biostatistics, and Informatics, Faculty of Medicine, University of Halle-Wittenberg, Halle (Saale), Germany

Address for correspondence:

Oliver Kuss, Institute of Medical Epidemiology, Biostatistics, and Informatics, Medical Faculty, Martin-Luther-University of Halle-Wittenberg, Magdeburger Str. 8, 06097 Halle (Saale), Germany, Phone: +49-345-557-3582, Fax: +49-345-557-3580, e-mail: oliver.kuss@medizin.uni-halle.de

Abstract

The propensity score (PS) method is increasingly used to assess treatment effects in nonrandomized trials. While there are several methods to use the PS for analysis, matching treated and non-treated patients by the PS is recommended by some researchers, mainly because this allows assessing and comparing covariate balance before and after matching. While the standardized difference is commonly applied to compute a measure of balance, it has two deficiencies: its distribution does depend on the sample size and there is no possibility to compare standardized differences for baseline covariates on different scales. We instead propose to use the z-difference to measure covariate balance in matched propensity score analyses. It solves the two mentioned problems of the standardized difference, moreover it is simple to calculate, can also be used with second moments for continuous covariates and in most cases can also be computed from published data. The full advantage of the z-difference emerges after displaying z-differences in a Q-Q-plot which additionally allows balance comparisons of the study data to 1) a randomized trial, and 2) to a perfectly matched PS analysis in the sense of Rubin/Thomas. The method is explained by a recent matched PS analysis to compare the clampless off-pump technique to the conventional on-pump technique in coronary artery bypass grafting.

Keywords

Propensity score, Propensity score matching, Covariate balance, Standardized difference, z-difference, Q-Q-plot

Introduction

The propensity score (PS) method is increasingly used to assess treatment effects in nonrandomized trials. Compared to the apparent standard method, regression modeling of the outcome, the propensity score method has several advantages(1-4). Propensity score analyses are conducted in two steps. In the first step, the propensity score, defined as the probability of treatment conditional on the subject's covariates, is estimated. In the second step, the estimated PS for each subject can be used in four different ways to arrive at an estimate of the treatment effect(5): matching on the PS, stratification on the PS, covariate adjustment using the PS, and inverse probability of treatment weighting (IPTW). Each method has its own merits, but Austin(6) and Morgan/Harding(7) prefer matching on the PS and give some of the advantages of PS matching as compared to the other three PS methods. The most important advantage is that covariate distributions in treated and untreated subjects after matching can be made explicit, similar to the traditional baseline table (or table 1) in a randomized trial (RCT) where the distribution of covariates (or at least, some descriptive measures of those) are reported. This enables judging the success of matching, especially when the distributions of covariates before matching are also reported in the table and the pre-matching situation can easily be compared to the post-matching situation.

It is commonly agreed that the balance of covariates should not be assessed by statistical tests. Imai et al.(8) call this practice the 'balance test fallacy' and explain why this practice is considered wrong. For example, statistical tests for balance assessment lose power due to the reduced sample size after matching and thus erroneously show better balance which might actually only be due to the smaller number of observations. Even more extreme, Imai et al. show that even randomly dropping (instead of dropping the observations that cannot be matched) observations will diminish test statistics. Moreover, test statistics are not only influenced by balance and sample size, but also by other characteristics of the sample, for example the ratio of treated and untreated observations in the matched sample or (in the continuous case) the standard deviations of covariates in the two treatment groups.

To assess covariate balance it is recommended to use graphical displays, Rubin's diagnostics(9), or to compute the standardized difference. Regardless of the method applied, a statistic for assessing balance should(8;10) (1) not be affected by sample size, and (2) it should be a characteristic of the sample and not of some hypothetical population. It is common to use the standardized difference (the difference of means or proportions in both groups, divided by a common standard deviation) in this case, and a value of 10% or lower has often been proposed to define satisfactory balance. These rules of thumb, however, have also been contradicted, Imai et al.(8) suggest minimizing imbalance without limit, and Austin (10) shows that in small samples even values of 30% do not necessarily indicate a bad balance of covariates.

While we agree that the standardized difference avoids statistical testing for baseline differences, we still see some disadvantages of it. First, while it is true that its *value* does not depend on sample size, its *distribution* does(10). Second, it is impossible to compare standardized differences for baseline covariates on different scales. For example, Austin(11) uses the phi coefficient for binary covariates and reports that a standardized difference of 10% for a continuous baseline covariate corresponds roughly to a phi coefficient of 5% for a binary one. However, there is still no solution for ordinal or nominal covariates.

In the following we introduce a measure that resolves these two problems and illustrate it with data from a PS analysis(12) that compared the clampless off-pump technique to the on-pump technique in coronary artery bypass grafting.

The z-Difference

The idea of the z-difference is to measure covariate balance by a statistic that is standard normally distributed (henceforth denoted as N(0,1)) under the null hypothesis of covariate balance. This measure has been introduced to the PS literature for continuous covariates by Hill et al.(13), however, the statistic can be traced back at least to Senn(14). To generalize this idea to another scale, we simply use a measure for imbalance on the respective scale and divide it through its standard error. For binary covariates we propose to use the risk difference, and for ordinal covariates the Wilcoxon statistic. The concrete formulas are given in the first three lines of table 1. Conveniently, the calculations need not be coded by hand, but are computed by default in most statistical software packages. For example, in SAS the procedures TTEST, FREQ, and NPAR1WAY can be used with small additional effort for data processing. A SAS macro is available from the author on request. For nominal covariates we are not aware of a measure that is N(0,1)-distributed, instead all association measures are only defined between 0 and 1, with 0 indicating no and 1 indicating perfect association. In the nominal case we thus propose to calculate the binary z-differences for all nominal categories.

At various instances (see, e.g.(15)) it is emphasized that not just means and proportions should be checked for similarity, but the entire distribution of baseline covariates in the two groups. This is straightforward for continuous covariates where the Wilcoxon statistic can be computed which compares the whole distributions in the two groups. However, in the continuous case it is also possible to calculate z-differences based on second moments. Using the standard error of the empirical variance(16) and elementary formulas on sums of variances, one finds a z-difference based on variances. In a similar fashion, a z-difference based on the coefficient of variation can be defined using the respective standard error given by Miller and Feltz(17). The respective formulas are given in the lines 4 and 5 of table 1 and also only require estimates for means and standard deviations in the two groups. For binomial covariates it is impossible to give z-differences for higher moments, because the higher moments are necessarily fixed by the first moment.

To fully recognize the advantage of the z-difference we follow Hill et al.(13), and propose to draw Q-Q-Plots where the z-differences for all covariates before and after matching are displayed. As the z-differences are N(0,1)-distributed under the null hypothesis of covariate balance, we would expect z-differences from a randomized trial (where the randomization ensures covariate balance, at least asymptotically) to be N(0,1)-

distributed and lie on the line through the origin with slope equal to 1. Interestingly, we can also give a reference line for a matched PS analysis. Rubin/Thomas(18;19) showed that under several assumptions (e.g., PS matching was performed with the logit of the PS, covariates in the PS model are normally distributed, and there is a large pool of controls that is matched with a 1:1-ratio to the treated subjects) the z-difference is N(0,1/2)-distributed. As such, we can also draw a reference line with slope $\sqrt{1/2}$ through the origin and can compare the balance after matching to the "Rubin/Thomas line".

An example

For illustration we use a matched PS analysis that compared the clampless off-pump (OPCAB) technique to the conventional on-pump technique (cCABG) in coronary artery bypass grafting(12). In this study, the PS model was estimated as a standard logistic regression model including all covariates from table 1 as main effects. We did not check if interactions would have improved covariate balance, but fitted a generalized additive model (GAM) to estimate the influence of all continuous covariates non-parametrically. However, this GAM fit did not result in better covariate balance and so we stayed with the main effects model. The fit of the PS model was additionally checked with the Hosmer-Lemeshow test and the first two diagnostics proposed by Rubin(9).

In table 2 we give details on the covariate distributions of the 4 continuous, 10 binary, and 1 ordinal covariate, which were included in the PS model in the original analysis. We computed z-differences before and after PS-matching to judge the success of PS-matching. To enable a comparison of the z- and the standardized difference we also computed the latter for the example data set. Following Austin(11), standardized differences for continuous covariates were computed as differences in means divided by the pooled standard deviation and for binary covariates as phi coefficients. It should be remembered, however, that with z-differences we can directly compare covariate balance between continuous and binary covariates, which is not possible for the standardized differences. Moreover, there is no standardized difference for ordinal covariates and none for continuous covariates that uses information from other sources that the means. Figures 1 and 2 give Q-Q-Plots (Figure 1) where the z-differences before and after matching are displayed together with the two reference lines for a randomized trial and a perfectly matched PS analysis in the sense of Rubin/Thomas ("Rubin/Thomas line"). Figure 1 gives the Q-Q-plot with only the z-difference based on the mean differences for the continuous covariates. Figure 2 additionally reports the z-differences based on the differences of variances, coefficients of variation, and on the Wilcoxon statistic, respectively, for the continuous covariates.

In our example we can see that there is *more* covariate imbalance *before* matching (in figure 1 the estimated mean of z-differences is actually -1.17 and the estimated variance 3.51) and *less* covariate imbalance *after* matching, both compared to a randomized trial. Moreover, the z-differences after matching lie very close to the Rubin/Thomas line, the estimated mean of the z-differences after PS-matching in figure 1 is 0.07, and the estimated variance 0.42, in close correspondence with the expected values of 0 and 0.5.

Referring to the computed standardized differences and keeping in mind the common cutpoints of 10% for continuous covariates and 0.05 for the phi coefficient, we note a very close correspondence between z- and standardized differences. Looking at, for example, the continuous covariate age before matching, we find a z-difference of -3.24 and a standardized difference of -19.3 %, both indicating a compromised balance. After matching, age is balanced with a z-difference of -0.46 and a standardized difference of -3.3 %. A similar behavior can be seen for binary covariates. Considering the history of previous myocardial infarction, we find a z-difference of -3.14 and a standardized difference of -0.085 before PS matching, both pointing to missing balance, and 0.16 and 0.006, respectively, after PS matching. As such, z- and standardized differences reach similar conclusions on balance in our example data set.

It is also obvious that there can be additional information when using all four z-differences for a continuous covariate. Referring, for example, to the covariate 'number of previous surgeries' we note inconspicuous z-differences before matching in terms of the mean (1.56), the CV (-0.19), and the Wilcoxon based z-difference (-1.48), but a large value of the z-difference based on the variance (9.80).

Discussion

We propose the z-difference to assess covariate balance in matched propensity score analyses. Compared to the present standard, the standardized difference, it allows comparison of continuous, binary, and ordinal covariates on the same scale, and also has a distribution that does not depend on sample size. As such, z-differences from samples with different sizes can easily be compared. Moreover, these advantages come not at a prize of an enhanced complexity in computation, instead the necessary calculations are comparable to the standardized difference and are implemented in statistical software. The full advantage of the z-difference emerges after displaying z-differences in a Q-Q-plot which additionally allows balance comparisons of matched PS analyses with reference to 1) a randomized trial, and 2) to a perfectly matched PS analysis in the sense of Rubin/Thomas. The z-differences for continuous, binary (but not for ordinal) covariates can also be calculated from published data, e.g. when covariate balance from published data should be judged.

Only recently, the importance of balance measures also for selecting optimal PS models (and not just for assessing balance) has been emphasized(20). Using standard meta-analytic techniques, such global balance measures can also be defined by z-differences. Mathematically convenient, the standard error of the z-difference always equals 1. Applying common ideas for combining effect sizes(21) (and realizing that an estimate with standard error 1 has also an inverse variance of 1), it turns out that the mean of k z-differences has a standard error of $^{1}\!/\!_{\sqrt{k}}$. In our example, the standard error for the 15 z-differences in each group becomes $^{1}\!/\!_{\sqrt{15}}$ = 0.258 and the mean z-difference before matching is -1.19 [95%-confidence interval: -1.69, -0.68] and 0.07 [-0.44, 0.57] after matching. Compared to the methods in Belitser et al.(20), it is a clear advantage of summarized z-differences that they can summarize covariates with different scales and also allow to explicitly assess statistical variability by confidence intervals.

It is fair to discuss some limitations of the z-difference. First, the z-difference could be criticized for re-introducing statistical testing and the p-value through the back-door, because it explicitly violates the ideas of Imai et al. which were given in the introduction. Or to be concrete, the z-difference depends on sample size and will necessarily indicate improved balance in the matched sample only due to restricted sample size in the matched sample. However, in matched propensity score analyses one of the fundamental problems of testing for covariate balance in RCTs is avoided. In RCTs the underlying populations are identical due to randomization and as such it is nonsense to test for the equality of population parameters, because it is known that these population parameters are equal by construction and so the null hypothesis is true by design. This equality of underlying populations, however, does not apply to matched PS analyses. Moreover, a lot of the reported problems with p-values comes from mixing Fisher's and Neyman-Pearson's ideas of statistical testing(22). We would like the z-difference to be understood in the Fisherian sense, as a measure of the plausibility of the data under the null hypothesis of equal balance, however, without referring to an alternative hypothesis, and without referring to a particular level of significance.

A second limitation is the absence of a z-difference for a nominally scaled covariate. As noticed above, in this case we propose to calculate the binary z-differences for all nominal categories. For example, if we consider the ordinal covariate 'priority' from our example data set as a nominal one, we achieve binary z-differences of -0.26, 0.23, 0.32 and -0.57 for the four categories 'elective', 'urgent', 'emergent' and 'ultima ratio' for the matched sample. The respective figures before matching are -5.75, -5.71, -2.28 and, -0.87, and we again find the improved balance due to PS matching.

A final limitation might be that with the current proposal of the z-difference we concentrated on measuring covariate balance only in matched PS analyses. However, Austin(5) showed that standardized differences can be computed for all of the four methods that use the PS to arrive at an estimate of the treatment effect (matching on the PS, stratification on the PS, covariate adjustment using the PS, and inverse probability of treatment weighting (IPTW)). It might be interesting future work to check if some of the good properties of the z-difference would carry over at least to the IPTW approach, where standardized differences differ only in using weighting factors as compared to their counterparts from PS matched groups.

To finally conclude and despite the mentioned limitations, we feel that the advantages of the z-difference regarding simplicity, generality with respect to covariate scales, graphical accessibility, and summarizability outweighs their disadvantages and recommend it for a global assessment of covariate balance in matched PS analyses.

References

(1) Cook EF, Goldman L. Performance of Tests of Significance Based on Stratification by A Multivariate Confounder Score Or by A Propensity Score. Journal of Clinical Epidemiology 1989;42(4):317-24.

(2) Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. American Journal of Epidemiology 2003 Aug 1;158(3):280-7.

(3) Oakes JM, Johnson PJ. Propensity score matching for social epidemiology. In: Oakes JM, Kaufman JS, editors. Methods in social epidemiology.San Francisco: Jossey-Bass; 2006. p. 364-86.

(4) Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: From nave enthusiasm to intuitive understanding. Stat Methods Med Res 2011 Jan 24.

(5) Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med Decis Making 2009 Nov;29(6):661-77.

(6) Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement. Journal of Thoracic and Cardiovascular Surgery 2007 Nov;134(5):1128-U7.

(7) Morgan SL, Harding DJ. Matching estimators of causal effects - Prospects and pitfalls in theory and practice. Sociological Methods & Research 2006 Aug;35(1):3-60.

(8) Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. Journal of the Royal Statistical Society Series A-Statistics in Society 2008;171:481-502.

(9) Rubin DB. Using Propensity Scores to Help Design Observational Studies: Application to Tobacco Litigation. Health Services and Outcomes Research Methodology 2001;2(3-4):169-88.

(10) Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 2009 Nov 10;28(25):3083-107.

(11) Austin PC. Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research. Communications in Statistics-Simulation and Computation 2009;38(6):1228-34.

(12) Borgermann J, Hakim K, Renner A, Parsa A, Aboud A, Becker T, et al. Clampless off-pump versus conventional coronary artery revascularization: a propensity

score analysis of 788 patients. Circulation 2012 Sep 11;126(11 Suppl 1):S176-S182.

(13) Hill J, Rubin DB, Thomas N. The Design of the New York School Choice Scholarship Program Evaluation. In: Bickman L, editor. Research Design: Donald Campbell's Legacy.Thousand Oaks, CA, USA: Sage Publications; 2000. p. 155-80.

(14) Senn S. Testing for baseline balance in clinical trials. Stat Med 1994 Sep 15;13(17):1715-26.

(15) Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research 2011;46(3):399-424.

(16) Lehmann EL, Casella G. Theory of Point Estimation. New York: Springer-Verlag; 1998.

(17) Miller GE, Feltz CJ. Asymptotic inference for coefficients of variation. Communications in Statistics-Theory and Methods 1997;26(3):715-26.

(18) Rubin DB, Thomas N. Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions. Biometrika 1992 Dec;79(4):797-809.

(19) Rubin DB, Thomas N. Matching using estimated propensity scores: Relating theory to practice. Biometrics 1996 Mar;52(1):249-64.

(20) Belitser SV, Martens EP, Pestman WR, Groenwold RH, de Boer A, Klungel OH. Measuring balance and model selection in propensity score methods. Pharmacoepidemiol Drug Saf 2011 Nov;20(11):1115-29.

(21) Hartung J, Knapp G, Sinha BK. Statistical Meta-Analysis with Applications. Hoboken, New Jersey: 2008.

(22) Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med 1999 Jun 15;130(12):995-1004.

Table 1

Table 1: Calculation formulas for the z-difference on the three relevant scales.

Scale	Formula
Continuous (mean)	$z=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\hat{\sigma}_1^2/n_1+\hat{\sigma}_2^2/n_2}}$
Binary	$z=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1+\hat{p}_2(1-\hat{p}_2)/n_2}}$
Ordinal	$z=\frac{W-\left[n_2(n_1+n_2+1)/2\right]}{\sqrt{\frac{n_1n_2}{12}\left[n_1+n_2+1-\frac{\sum_{j=1}^{g}t_j\left(t_j^2-1\right)}{\left(n_1+n_2\right)\left(n_1+n_2+1\right)}\right]}}$
Continuous (variance)	$z=\frac{\hat{\sigma}_1^2-\hat{\sigma}_2^2}{\sqrt{\left(\hat{\sigma}_1^2\sqrt{\frac{2}{n_1-1}}\right)^2+\left(\hat{\sigma}_2^2\sqrt{\frac{2}{n_2-1}}\right)^2}}$
Continuous (coefficient of variation)	$z=\frac{\hat{\sigma}_1/\bar{x}_1-\hat{\sigma}_2/\bar{x}_2}{\sqrt{\left(\frac{1}{n_1-1}+\frac{1}{n_2-1}\right)\times\left(\hat{\sigma}/\bar{x}\right)^2\times\left(0.5+\left(\hat{\sigma}/\bar{x}\right)^2\right)}}$

$\bar{x}_1,\bar{x}_2,\hat{\sigma}_1^2,\hat{\sigma}_1^2,n_1,n_2,\hat{p}_1,\hat{p}_2$ denote estimated means, variances, sample sizes, and proportions in groups 1 and 2. W equals the sum of ranks in group 2, where ranks have been taken with values from both groups amalgamated, g is the number of different values in the data set and t_j is the number of identical values j(23).

The common coefficient of variation $\hat{\sigma}/\overline{x}$ is estimated by

$$\hat{\sigma}/\overline{x} = \frac{(n_1 - 1) \times \hat{\sigma}_1/\overline{x}_1 + (n_2 - 1) \times \hat{\sigma}_2/\overline{x}_2}{n_1 + n_2 - 2}$$

Table 2

Table 2: Distributions of baseline covariates resulting z-differences and standardized differences before and after PS-matching for the example data set. Given are mean and standard deviation for the continuous covariates, and relative frequencies for binary and ordinal covariates. The standardized differences are computed as differences in means divided by the pooled standard deviation for continuous covariates (in %) and as phi coefficients for binary covariates (11).

BMI, Body Mass Index; cCABG, conventional coronary artery bypass grafting (On-pump); COPD, Chronic Obstructive Pulmonary Disease; IABP, intra-aortic balloon pump; LVEF, left ventricular ejection fraction; MI, myocardial infarction; PAD, peripheral artery disease.

	Before PS-matching (n = 1.282)				**After PS-matching (n = 788)**			
Covariates	**Clampless OPCAB (n = 395)**	**cCABG (n = 887)**	**z-difference**	**Standard. difference**	**Clampless OPCAB (n = 394)**	**cCABG (n = 394)**	**z-difference**	**Standard. difference**
Continuous scale (based on the difference of means)								
Age [years]	69.3 (9.1)	67.5 (9.4)	-3.24	-19.3	69.3 (9.1)	69.0 (8.9)	-0.46	-3.3
BMI [kg/m²]	27.8 (4.2)	28.3 (4.5)	1.83	10.8	27.8 (4.2)	28.0 (4.2)	0.60	4.2
Previous surgeries [n]	0.05 (0.26)	0.07 (0.39)	1.56	8.1	0.05 (0.26)	0.06 (0.27)	0.80	5.7
LVEF [%]	56.7 (12.2)	55.4 (14.1)	-1.64	-9.4	56.6 (12.2)	56.9 (13.3)	0.28	2.0
Binary scale								
Gender [% female]	21.8	22.1	0.13	0.004	21.8	22.1	0.09	0.003
Previous MI [%]	27.1	35.7	-3.14	-0.085	27.2	26.7	0.16	0.006
Diabetes [%]	22.8	31.7	-3.39	-0.091	22.8	19.8	1.04	0.037
Hypertension [%]	82.3	84.1	-0.80	-0.023	82.2	82.2	0.00	0.000
Previous stroke [%]	1.0	2.4	-1.89	-0.045	1.0	1.8	-0.91	-0.032

COPD [%]	5.8	7.1	-0.88	-0.024	5.8	6.1	-0.15	-0.005
Renal insufficiency [%]	0.8	1.2	-0.84	-0.021	0.8	0.3	1.00	0.036
Main stem stenosis [%]	25.3	25.5	-0.06	-0.002	25.1	24.9	0.08	0.003
PAD [%]	11.9	11.4	0.26	0.007	11.7	14.7	-1.26	-0.045
Pre-OP IABP [%]	1.0	1.5	-0.70	-0.018	1.0	1.0	0.00	0.000
Ordinal scale								
Priority [%]								
elective	91.9	81.0			91.9	92.4		
urgent	2.5	9.8	-4.82	--	2.5	2.3	-0.25	--
emergent	5.3	8.7			5.3	4.8		
ultima ratio	0.3	0.6			0.3	0.5		
Continuous scale (based on the difference of variances)								
Age [years]	69.3 (9.1)	67.5 (9.4)	0.90	--	69.3 (9.1)	69.0 (8.9)	-0.42	--
BMI [kg/m²]	27.8 (4.2)	28.3 (4.5)	1.43	--	27.8 (4.2)	28.0 (4.2)	-0.31	--
Previous surgeries [n]	0.05 (0.26)	0.07 (0.39)	9.80	--	0.05 (0.26)	0.06 (0.27)	1.19	--
LVEF [%]	56.7 (12.2)	55.4 (14.1)	3.35	--	56.6 (12.2)	56.9 (13.3)	1.62	--
Continuous scale (based on the difference of coefficients of variation)								
Age [years]	69.3 (9.1)	67.5 (9.4)	1.46	--	69.3 (9.1)	69.0 (8.9)	-0.33	--
BMI [kg/m²]	27.8 (4.2)	28.3 (4.5)	0.97	--	27.8 (4.2)	28.0 (4.2)	-0.43	--
Previous surgeries [n]	0.05 (0.26)	0.07 (0.39)	-0.19	--	0.05 (0.26)	0.06 (0.27)	-0.61	--
LVEF [%]	56.7 (12.2)	55.4 (14.1)	3.44	--	56.6 (12.2)	56.9 (13.3)	1.47	--
Continuous scale (based on the Wilcoxon statistic)								
Age [years]	69.3 (9.1)	67.5 (9.4)	2.98	--	69.3 (9.1)	69.0 (8.9)	-0.41	--
BMI [kg/m²]	27.8 (4.2)	28.3 (4.5)	-1.59	--	27.8 (4.2)	28.0 (4.2)	0.56	--
Previous surgeries [n]	0.05 (0.26)	0.07 (0.39)	-1.48	--	0.05 (0.26)	0.06 (0.27)	1.00	--
LVEF [%]	56.7 (12.2)	55.4 (14.1)	1.53	--	56.6 (12.2)	56.9 (13.3)	0.09	--

Figure 1: Q-Q-Plot for the z-differences (only those based on first moments) from the example data set before and after PS-matching. The broken gray line corresponds to the expected distribution of z-differences from a randomized trial. The solid gray line corresponds to the Rubin/Thomas line for z-differences in a matched PS analysis.

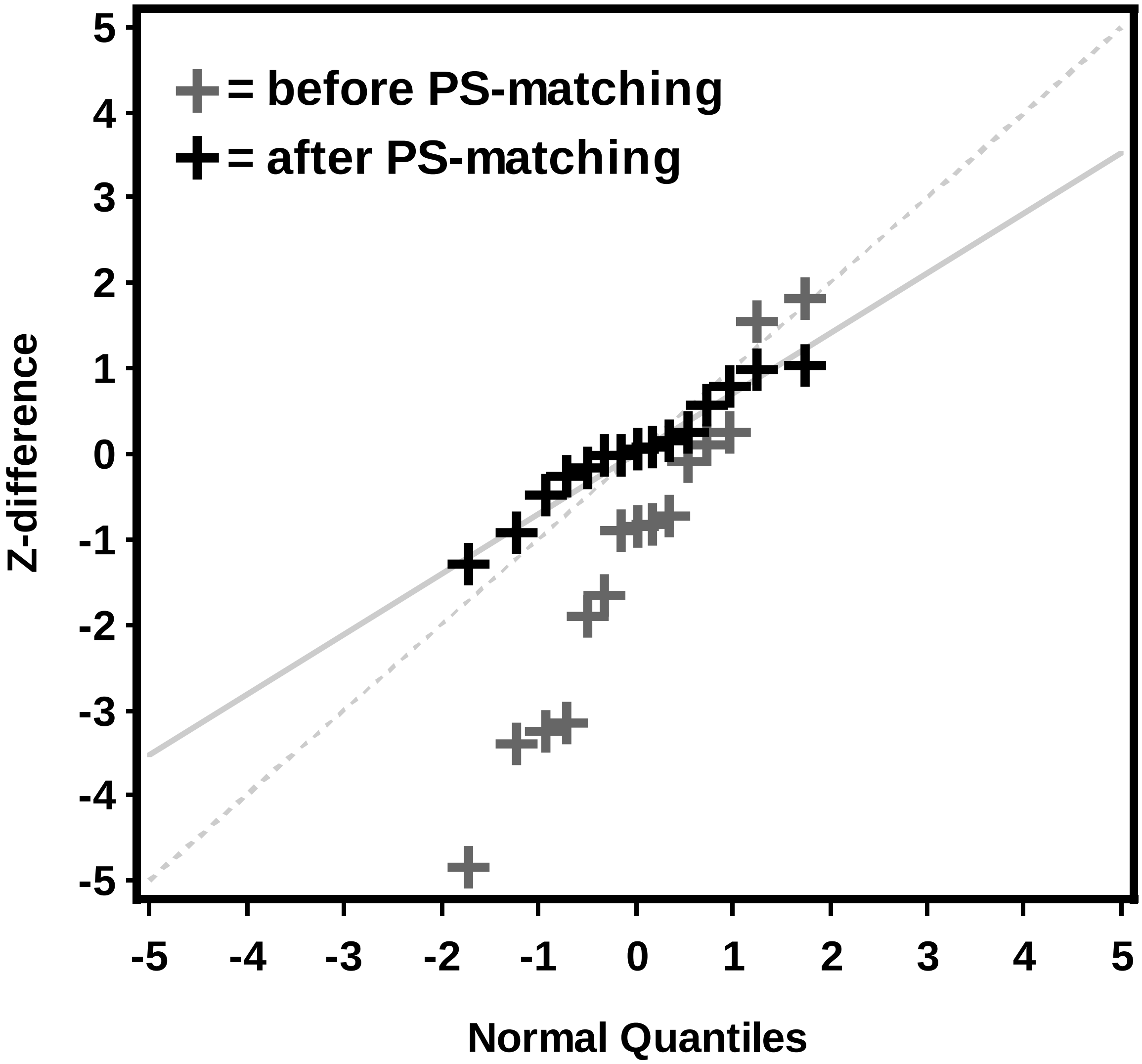
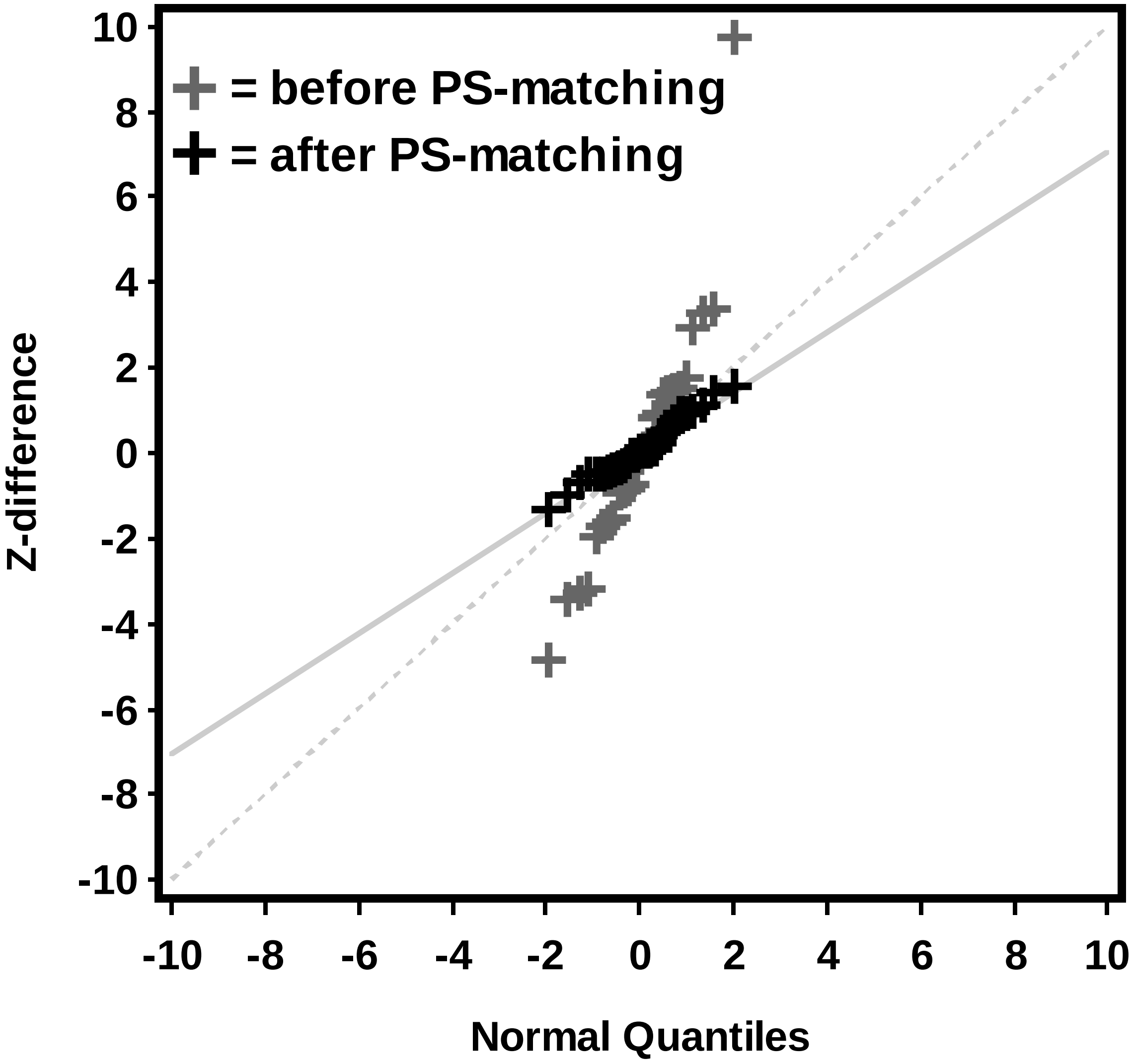


Figure 2: Q-Q-Plot for the z-differences from the example data set before and after PS-matching. For continuous covariates four z-differences are reported: The z-differences based on the differences of means, variances, coefficients of variation, and on the Wilcoxon statistic. The broken gray line corresponds to the expected distribution of z-differences from a randomized trial. The solid gray line corresponds to the Rubin/Thomas line for z-differences in a matched PS analysis.


+ = before PS-matching
+ = after PS-matching
Z-difference
Normal Quantiles