Do treatments effects differ between randomized trials and propensity score analyses:
Evidence from a meta-propensity score analysis in off-pump versus on-pump coronary artery bypass surgery

Kuss O¹, Legler T¹, Börgermann J²

¹Institute of Medical Epidemiology, Biostatistics, and Informatics, Faculty of Medicine, University of Halle-Wittenberg, Halle (Saale), Germany
²Heart and Diabetes Center North Rhine-Westphalia, Department of Thoracic- and Cardiovascular Surgery, Ruhr-University Bochum, Bad Oeynhausen, 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

Running head: A meta-propensity score analysis
Financial support: There was no financial support for this study.
Acknowledgements: We are grateful to Lena Minning, Mareike Kunze, and Jörn Klauke (Institute of Medical Epidemiology, Biostatistics, and Informatics, Faculty of Medicine, University of Halle-Wittenberg, Halle (Saale), Germany) for designing and programming the study data base.

Word count: 3178
Abstract

Background
Analyses comparing randomised to non-randomised clinical trials suffer from the fact that the study populations are usually different. Therefore, differences in treatment effects are not necessarily due to the different treatment modes but may result from differences between populations.

Methods
In a systematic review, we “meta-matched” randomised clinical trials (RCT) and propensity score analyses (PS) that compared the off- and the on-pump technique in coronary artery bypass grafting. “Meta-confounders” were used for “meta-matching” and were summarised in a “meta-propensity score”. We compared treatment effects between RCTs and PS analyses for ten previously defined binary clinical outcomes in this “meta-matched” population as differences in “meta-odds ratios”.

Results
For all clinical outcomes, the estimated differences in “meta-odds ratios” were below an absolute value of 0.15, all confidence intervals included the null. The overall difference was -0.03 [95% confidence interval: -0.12, 0.07], indicating a small, but precisely estimated difference between RCTs and PS analyses in the meta-matched sample.

Conclusions
In our example, treatment effects of off-pump versus on-pump surgery from RCTs and PS analyses were very similar in a "meta-matched" population of studies, indicating a small effect of randomisation itself.

Word count (Abstract): 188
Introduction

It is commonly agreed on that randomised clinical trials (RCTs) are the gold standard for treatment evaluation. However, RCTs have also often been criticised for limited external validity which results from the enrolment of highly selected patients groups. Study patients tend to be younger and healthier than the average patient. Non-randomised (non-RCTs) or observational studies are a possible alternative for assessing treatment effects, because those are expected to have larger external validity. Their obvious principal disadvantage is limited internal validity as non-randomised treatment allocation might bias treatment comparisons due to confounding. A wealth of methods to adjust for this confounding have been proposed, the most recent being the technique of propensity score (PS) analysis which is expected to have statistical advantages as compared to the standard methods of confounder adjustment.

Numerous investigations have been conducted to assess if treatment effects from RCTs differ systematically from those of non-RCT studies. Most of these investigations have already been collected in systematic reviews. Evidence from these reviews is still inconclusive, results from non-RCT studies differ sometimes, but not always, and not in a predictable direction from the results of RCTs.

A simple consequence of the limited external validity of RCTs is, however, the limited internal validity of all systematic comparisons of RCTs and non-RCTs: If RCTs are conducted in highly selected populations, but non-RCTs in the general population, differences between both study types are not necessarily due to the missing randomisation. They might also arise from the different study populations involved.
Ideally we would like to conduct a “meta-randomised” trial to systematically compare RCTs and non-RCTs. That is, investigators willing to conduct a study on a specific clinical question would be randomly selected to perform a RCT or a non-RCT. Proceeding this way, all characteristics of the investigator’s setting and patients would distribute evenly on the group of RCTs and non-RCTs, thus eliminating all kinds of “meta-confounding” and enabling a causal statement on the effect of randomisation. Obviously, conducting such a study would be difficult and maybe even unethical as we should not force investigators to conduct non-RCTs if they are willing to do a RCT. It is interesting to note that similar trials have already been performed, not on a study, but on a patient level\textsuperscript{14-16}. In these trials, usually called doubly randomised preference trials, patients have been randomised in a first stage to (1) a group where the actual treatment would be subsequently randomised in a second stage, or (2) to a group where they choose the treatment they prefer.

If “meta-randomised” trials are difficult, one is still able, however, to conduct a “meta-non-randomised” trial to compare RCTs and non-RCTs. As we feel that propensity score analyses are the most valid methods for analysing non-RCTs, we conducted a “meta-propensity score analysis” to judge the differences between RCTs and Non-RCTs. That is, we “meta-matched” RCTs and non-RCTs for important and available “meta-confounders” where the latter were summarised in a “meta-propensity score”. We then compared treatment effects between RCTs and non-RCTs in this “meta-matched” population.
As a clinical example we use the comparison of off-pump versus on-pump technique in coronary artery bypass grafting, one of the most hotly debated and polarising issues in cardiac surgery. This clinical question has also some public health relevance. For example, in Germany 49.788 (isolated) bypass surgery were performed in 2007, 10.1% of those used the off-pump technique.

Methods

Study selection - Selection of non-RCTs

In the group of non-RCTs we restricted ourselves to propensity score analyses. We included all propensity-score analyses comparing off- and on-pump coronary bypass grafting from our recent systematic review, details on study selection and search strategy are given there. As the study search for this review was performed in February 2006, we again performed the described search in October 2006. PS analyses were included in the analysis presented here if they gave descriptive information on the PS study publication (e.g. average age, proportion of males etc., factors we will subsequently refer to as “meta-confounders”) and at least one of the ten short-term binary clinical outcomes death, stroke, myocardial infarction, atrial fibrillation, acute renal failure, inotropic support, RBC transfusion, wound infection, re-operation for bleeding or, IABP support.

Study selection - Selection of RCTs

Randomised trials were retrieved by including all publications that were referred to in the five largest systematic reviews. As the literature search for the most recent review included studies only up to February 2006, we additionally performed a MEDLINE
search for the keywords “randomised” and “off-pump” in October 2006. All RCT publications were gathered in full text and read independently by two reviewers (OK, TL), blinded for the results of the previous reviews. Data were abstracted into a self-developed case record form which had been tested in a small (five studies) pilot study. Results from the two reviewers were entered in a MS ACCESS data base. Disagreements between reviewers were found by automatic comparisons and resolved by (in this hierarchical order) finding a consensus, discussion with a third reviewer (JB), or referring to the previous systematic reviews. RCTs were included if they gave descriptive information on the RCT study publication and at least one of the binary clinical outcomes mentioned above. Double publications were removed, but we used all additional information on study populations and outcomes from the removed publications.

**Extracted information**

From each of the retrieved studies (PS analyses and RCTs) we extracted the observation period, location of study, number of involved centres, number of treated patients, and percentage of conversions to on-pump. We further extracted any information that described the study populations, i.e., that is, all pre-operative prognostic and risk factors provided.

To describe the estimated treatment effects from each study, we separately extracted the original four-fold tables for each of the ten binary outcomes from the RCTs and the PS analyses that used matching for treatment comparisons. In PS analyses where only effects estimates, but no four-fold tables were given (i.e., PS analyses that used PS
stratification or regression adjustment), we reconstructed four-fold tables by the Di Pietrantonj method.\textsuperscript{25}

**Meta-PS analysis**

Different statistics (mean, median, or several categories with maybe different cutpoints) were reported in the original studies to describe the distributions of continuous meta-confounders. We managed this by equating means to medians to describe the average value of the respective distribution. In case of continuous meta-confounders reported in several categories of percentages, we assumed values evenly distributed in categories and calculated the median value from the reported figures.

Of course, not all studies reported values for all potential meta-confounders. We included prognostic factors as meta-confounders in our meta-PS model if information on them was given in at least two thirds of all publications (PS analyses and RCTs). We used the following strategy for dealing with missing values. Missing values for categorical meta-confounders were collected and fitted in a separate category. Missing values for continuous meta-confounders were adjusted for by multiple imputation. This was accomplished by generating 1000 imputed data sets and fitting the meta-PS model to each of those. The median estimated propensity score across these 1000 fitted data sets was calculated for each study and used in subsequent meta-matching. The functional form of all continuous meta-confounders was checked by estimating a semi-parametric generalised additive meta-PS model\textsuperscript{26} with continuous meta-confounders fitted by spline functions and visual inspection of fitted curves. The final meta-PS model was then estimated by standard logistic regression with the functional form of continuous...
meta-confounders as assessed from the semi-parametric model. No interactions of meta-confounders were included into the meta-PS model. We computed the c-statistic (median c-statistic across the 1000 imputed data sets) to assess predictive ability, and the Hosmer-Lemeshow test (median p-value across the 1000 imputed data sets) to judge goodness-of-fit of the meta-PS model. For meta-matching we applied an optimal matching algorithm\(^27\). The number of RCTs that were meta-matched to a PS analysis was allowed to vary between 1 and 4\(^28\). We used the values of the linear predictor of the meta-PS model for matching\(^29\), the caliper width was set to 0.2 times the standard deviation of the linear predictor\(^30\). We assessed balance of meta-confounders by statistical tests, and, as those have been criticised for their dependency on sample sizes\(^31\) also with calculating standardised differences (at least for continuous meta-confounders). It has been suggested that a standardised difference of below 10\(^%\)\(^32\) supports the assumption of balance between groups, however, with small numbers of observations as in our case, larger differences to not necessarily point to an important imbalance\(^33\).

**Statistical methods**

Estimated treatment effects from RCTs and PS analyses were compared in the meta-matched sample as differences in meta-odds ratios (with 95%-confidence intervals) with a 3-level (patients, studies, matching strata) random effects logistic regression model. Parameter were estimated by penalised quasi-likelihood (PQL)\(^34\), and confidence intervals computed by the multivariate delta method\(^35\). Additionally, an overall difference in meta-odds ratios across all clinical outcomes was computed. This was calculated by incorporating a fourth hierarchical level (outcomes) in the model.
To check the robustness of our results we performed two types of sensitivity analyses. First, we estimated differences between meta-odds ratios also by regression adjustment for the meta-PS in the whole group of studies and by using an IPTW estimator. Second, we also allowed the meta-matching ratio to vary between 1:1 and 1:6.

We emphasise that all specifications and assumptions regarding building of the meta-PS-model and the statistical analysis were determined a priori and independent of outcome data. All analyses were conducted with SAS Version 9.1, the %match macro of Bergstralh et al.\textsuperscript{36} was used for performing the meta-matching.

\section*{Results}

\subsection*{Included PS analyses}
Updating the search from our previous systematic review yielded another 6 PS analyses. Of this additional 6 PS analyses, one was excluded because of double publication, one because it did not compare an off-pump group to an on-pump group, and one because it gave no information on the pre-specified outcomes. Of the original 35 PS analyses, 10 PS analyses were excluded as no information on the PS population was given. A total of 28 PS analyses (with 97,478 individual patients) were finally included into the analysis reported here.

\subsection*{Included RCTs}
Abstracting the RCTs from the five systematic reviews resulted in 80 publications; our additional search gave another 26 publications. We excluded (each study may meet more than one exclusion criterion) 36 publications because of double publication, 10 because none of our pre-specified outcomes was given, 8 because they did not compare an off-pump to an on-pump group, 8 because they reported on non-randomised trials, 1 because it reported a systematic review, and 1 because no information on the study publication was given. As we used all additional available information on meta-confounders and outcomes, information from 57 different reports on 51 RCTs with 4,958 individual patients was eventually included into the analysis.

**Included meta-confounders**

Seven meta-confounders (table 1) met the inclusion criterion of having available information in at least two thirds of the PS analyses as well as of the RCTs. Between study groups, meta-confounders differed considerably. PS analyses are significantly more often conducted in Northern America, in multiple centres, and in older populations with a worse left ventricular ejection fraction. Moreover, all meta-confounders have a standardised difference of more than 10%, i.e., they are not well balanced.

**The Meta-PS model**

Visual inspection of fitted curves from the semi-parametric model suggested a cubic fit of the average LVEF (%) and the average proportion of diabetic patients in the meta-PS model, and these two confounders were thus fitted as cubic terms in the final meta-PS model. The remaining continuous meta-confounders average age, average proportion of males, and average proportion of prior MI, and the two categorical confounders were
included linearly in the meta-PS model. The c-statistic of the resulting meta-PS model was estimated to be 89.6%, the p-value of the Hosmer-Lemeshow test was p=0.345, indicating very good predictive ability and a good fit of the meta-PS model.

Applying the optimal matching algorithm resulted in 10 included PS analyses (25,552 included patients, 7,242 treated in the off-pump group), and 29 RCTs (2,723 included patients, 1,342 treated in the off-pump group) in the meta-matched population of studies, accounting for 28,275 patients altogether, 8,584 of which were treated off-pump. In these 39 studies there was information on 186 effect estimates. The most often reported estimates were those for postoperative death (38 studies), stroke (28), and myocardial infarction (27).

In table 2 we give the distribution of the meta-confounders after meta-matching. There is no significant difference in any of the meta-confounders. Referring to the standardised difference, three continuous meta-confounders have values below 10%. There is still some moderate imbalance in the meta-confounders average proportion of males and average proportion of prior MI.

**Comparison of treatment effects in the meta-matched sample**

In Figure 1 we give the differences of the meta-odds ratios (PS-RCT) in the meta-matched sample. For the ten binary outcomes all estimated differences are below an absolute value of 0.15, all confidence intervals include the null. That is, for none of the clinical outcomes we find a statistically different difference between meta-odds ratios from PS analyses and RCTs. The overall difference amounts to -0.03 with a 95%
confidence interval of [-0.12, 0.07], indicating a small, but precisely estimated difference between RCTs and PS analyses in the meta-matched sample. Results from the described sensitivity analyses gave very similar results and are omitted here.

Discussion

In our meta-matched population of studies, treatment effects of off-pump versus on-pump surgery from RCTs and PS analyses were very similar, indicating a small effect of randomisation itself. This finding is in line with many recent results from systematic comparisons of RCTs and non-RCTs\textsuperscript{14,16,37-44}. This is exemplified by the analysis of Furlan et al.\textsuperscript{41} who found that “homogeneity in terms of settings, population, interventions, and outcomes predicts the agreement between an NRS (non-randomised study) and an RCT of the same intervention”. King et al.\textsuperscript{37} concluded from a systematic review of preference trials that “differences in outcome across the trials between randomised and preference groups were generally small, particularly in large trials and after accounting for baseline measures of outcome. Therefore, there was little evidence that preferences substantially interfere with the internal validity of randomised trials”. Even the most prominent difference between randomised and observational evidence in recent years, the disagreement between the WHI randomised trial and the Nurses’ Health Study on CHD risk in postmenopausal hormone users can possibly be resolved. According to Hernán et al.\textsuperscript{40}, who carefully emulated the design and intention-to-treat (ITT) analysis of a randomised trial with data from the Nurses’ Health Study, this disagreeing results can be explained “largely by differences in the distribution of time since menopause and length of follow-up”. Recent reviews actually finding differences
between RCTs and non-RCTs\textsuperscript{45,46} did not adjust for the potentially different populations involved, that is, there was no meta-matching.

Of course, our study has several limitations. The most important threat to internal validity here is meta-confounding\textsuperscript{12}. As explicated in the introduction we would ideally want to conduct a meta-randomised trial to compare RCTs and non-RCTs with optimal internal validity. By using a meta-propensity score analysis we can not rule out systematic differences between RCTs and non-RCTs that are not accounted for by our meta-confounders. However, besides meta-matching, our analysis has a number of design features that facilitate a valid comparison\textsuperscript{47}. By using the coronary artery bypass grafting example we were able to analyse identical interventions with identical doses in both RCTs and non-RCTs. We used the same valid outcomes with the same length of follow-up in both groups. Though not perfect, there was a considerable overlap in observation times of RCTs and non-RCTs. We further note that an unknown meta-confounder to confound our findings would have to be largely independent of all meta-confounders in the meta-PS model. Otherwise, the meta-confounders already included would account for this unknown meta-confounder\textsuperscript{48}.

We realise that the number of meta-confounders (n=7) in our meta-PS model is rather low. Shadish et al.\textsuperscript{16} only achieved a satisfactory similarity between their randomised and non-randomised group after having adjusted for a large set of confounders. Having used only a small “convenience” sample of confounders resulted in large differences between their randomised and non-randomised group. The small number of meta-confounders used in our analysis is due to the fact that we only allowed “meta-
confounders” to be included if they provided information in at least 2/3 in both RCTs and PS analyses. We refrained from contacting the authors of the original studies to yield additional information on meta-confounders, mostly because of the unfortunate experience we had when we contacted authors for our previous review.

As in every meta-analysis we also face the problem of publication bias, i.e., the selective reporting of trials depending on study results. As the comparison between off- and on-pump in CABG is such a hotly debated issue\textsuperscript{17}, we expect most (or hopefully all) of the studies to be submitted and published, as also predicted by Sedrakyan et al.\textsuperscript{23}

In this analysis, we refrained from using study quality as a potential meta-confounder. The reason for this was the absence of a valid and reliable score for assessing the quality of propensity score analyses\textsuperscript{49}. Moreover, even if such a score would exist, it is not clear how to compare scoring results from RCTs and PS analyses. Using some ad hoc measures (data not shown) we found that in the group of RCTs the subgroup of meta-matched RCTs have a similar study quality, whereas in the group of PS analyses the meta-matched PS analyses have a higher quality.

We finally point to two statistical limitations: First, by using several clinical outcomes from each study, we can not rule out the influence of competing risks: A person that dies during surgery is no longer at risk for a stroke in the follow-up. It is not clear how this might influence treatment effect estimates, and, more important in our case, if this issue would have different implications in RCTs and PS analyses. Second, it is also possible that individual outcomes within a single study are associated: A person that suffered a
myocardial infarction might also die in the follow-up, thus making effect estimates for MI and post-operative death correlated. However, also for this second point it is impossible to judge if this will bias the comparison between RCTs and PS analyses.

An interesting point is that of the generalisability of our findings. We found remarkable similar treatment effects in our meta-matched sample. Due to meta-matching, this meta-matched sample consists mainly of the younger and healthier study populations that are found in RCTs. Does that mean that RCTs and PS analyses would also have similar results in typical study populations from PS analyses? If yes, this would mean that RCTs in these populations would be unnecessary. However, this interpretation goes way too far, we still feel that RCTs should stay the gold standard for treatment evaluation in medical sciences.

We finally conclude that our study is another example showing that results from randomised and non-randomised studies give very similar results when performed in similar populations. In the future we plan an independent replication of our study in a different clinical setting.
Table 1: Distributions of meta-confounders in two study groups *before* meta-matching for *all* studies.

<table>
<thead>
<tr>
<th>Meta-confounder</th>
<th>PS analyses (N=28)</th>
<th>RCTs (N=51)</th>
<th>p-value</th>
<th>stand. diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Categorical meta-confounders</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study region</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>17 (61%)</td>
<td>36 (71%)</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>North America</td>
<td>10 (36%)</td>
<td>5 (10%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>1 (3%)</td>
<td>10 (19%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of centres</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>18 (65%)</td>
<td>47 (92%)</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>&gt;1</td>
<td>9 (32%)</td>
<td>3 (6%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>1 (3%)</td>
<td>1 (2%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Continuous meta-confounders</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average age (y)</td>
<td>65.8 (58.5 - 73.0, 27)</td>
<td>63.1 (48.3 - 75.5, 50)</td>
<td>0.002</td>
<td>75.1%</td>
</tr>
<tr>
<td>Average proportion of males (%)</td>
<td>72.1 (0.0 - 90.4, 25)</td>
<td>77.1 (25.0 - 89.2, 48)</td>
<td>0.138</td>
<td>37.0%</td>
</tr>
<tr>
<td>Average proportion of prior MI (&gt;1 month, %)</td>
<td>44.5 (11.6 - 68.0, 19)</td>
<td>41.6 (8.3 - 73.3, 29)</td>
<td>0.480</td>
<td>21.0%</td>
</tr>
<tr>
<td>Average LVEF (%)</td>
<td>58.8 (36.4 - 67.3, 23)</td>
<td>62.7 (44.5 - 75.0, 44)</td>
<td>0.033</td>
<td>55.9%</td>
</tr>
<tr>
<td>Average proportion of diabetic patients</td>
<td>26.2 (15.2 - 46.8, 25)</td>
<td>24.4 (0.0 - 52.7, 36)</td>
<td>0.595</td>
<td>13.9%</td>
</tr>
</tbody>
</table>
Table 2: Distributions of meta-confounders in two study groups after meta-matching for the matched studies.

<table>
<thead>
<tr>
<th>Meta-confounder</th>
<th>PS analyses (N=10)</th>
<th>RCTs (N=29)</th>
<th>p-value</th>
<th>stand. diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Categorical meta-confounders</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study region</td>
<td></td>
<td></td>
<td></td>
<td>0.999</td>
</tr>
<tr>
<td>Europe</td>
<td>8 (80%)</td>
<td>23 (80%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>North America</td>
<td>1 (10%)</td>
<td>3 (10%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>1 (10%)</td>
<td>3 (10%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of centres</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>8 (80%)</td>
<td>25 (86%)</td>
<td>0.631</td>
<td></td>
</tr>
<tr>
<td>&gt;1</td>
<td>2 (20%)</td>
<td>3 (10%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>0 (0%)</td>
<td>1 (4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Continuous meta-confounders</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average age (y)</td>
<td>64.1 (58.5 – 70.1, 10)</td>
<td>63.9 (59.3 - 75.5, 28)</td>
<td>0.916</td>
<td>3.9%</td>
</tr>
<tr>
<td>Average proportion of males (%)</td>
<td>80.5 (69.5 – 90.4, 9)</td>
<td>76.9 (25.0 - 89.2, 28)</td>
<td>0.431</td>
<td>30.5%</td>
</tr>
<tr>
<td>Average proportion of prior MI (&gt;1 month, %)</td>
<td>44.0 (17.5 – 68.0, 8)</td>
<td>39.9 (8.3 - 73.3, 16)</td>
<td>0.530</td>
<td>27.6%</td>
</tr>
<tr>
<td>Average LVEF (%)</td>
<td>61.1 (50.6 – 67.3, 9)</td>
<td>60.7 (47.8 - 68.5, 25)</td>
<td>0.861</td>
<td>6.9%</td>
</tr>
<tr>
<td>Average proportion of diabetic patients</td>
<td>24.8 (15.2 – 40.0, 10)</td>
<td>25.2 (12.0 - 52.7, 18)</td>
<td>0.925</td>
<td>3.7%</td>
</tr>
</tbody>
</table>
Figure 1:

Differences of ORs (PS-RCT) (with 95%-CI)

- RBC transfusion: 0.11 (-1.29, 1.50)
- Inotrope support: 0.12 (-0.36, 0.60)
- Stroke: 0.05 (-0.45, 0.54)
- Wound infection: 0.15 (-0.14, 0.43)
- Death: 0.14 (-0.29, 0.57)
- Myocardial infarct: 0.00 (-0.24, 0.24)
- Re-OP for bleeding: -0.05 (-0.20, 0.10)
- Renal failure: 0.12 (-0.36, 0.60)
- IABP support: 0.15 (-0.14, 0.43)
- Atrial fibrillation: 0.14 (-0.09, 0.38)
- Myocardial infarct.: 0.14 (-0.29, 0.57)
- Stroke: -0.10 (-0.57, 0.38)
- Death: -0.05 (-0.57, 0.47)
- Overall: -0.03 (-0.12, 0.07)
Figure and table legends

Table 1: Distributions of meta-confounders in two study groups before meta-matching for all studies. For categorical meta-confounders we give absolute and relative frequencies, for continuous meta-confounders we give the mean, the range and the number of observations in the respective group. P-values are calculated from standard $\chi^2$-tests (categorical meta-confounders) and t-tests (continuous meta-confounders). For continuous meta-confounders also the standardised difference is given.

Table 2: Distributions of meta-confounders in two study groups after meta-matching for the matched studies. For categorical meta-confounders we give absolute and relative frequencies, for continuous meta-confounders we give the mean, the range and the number of observations in the respective group. P-values are calculated from standard $\chi^2$-tests (categorical meta-confounders) and t-tests (continuous meta-confounders). For continuous meta-confounders also the standardised difference is given.

Figure 1: Differences of the meta-odds ratios (PS-RCT) in the meta-matched sample for the ten binary outcomes and the overall estimate.
Reference List

(1) Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet*. 2005;365(9453):82-93.


(9) 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;158(3):280-287.


