European Heart Journal (2012) **33**, 1893–1901
doi:10.1093/eurheartj/ehs114

Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes

Issa J. Dahabreh[1]*, Radley C. Sheldrick[2], Jessica K. Paulus[3,4], Mei Chung[1], Vasileia Varvarigou[5], Haseeb Jafri[6], Jeremy A. Rassen[7,8], Thomas A. Trikalinos[1], and Georgios D. Kitsios[1,9]

[1]Center for Clinical Evidence Synthesis, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, 800 Washington Street, Box No. 63, Boston, MA 02111, USA; [2]Division of Developmental-Behavioral Pediatrics, Department of Pediatrics, Floating Hospital for Children, Tufts Medical Center, Boston, MA, USA; [3]Tufts Clinical and Translational Science Institute, Tufts University, Medford, MA, USA; [4]Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA; [5]Department of Environmental and Occupational Medicine and Epidemiology, Harvard School of Public Health, Boston, MA, USA; [6]Division of Cardiology, Johns Hopkins Hospital, Johns Hopkins University School of Medicine, Baltimore, MD, USA; [7]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA; [8]Harvard Medical School, Boston, MA, USA; and [9]Division of General Internal Medicine, Lahey Clinic Medical Center, Burlington, MA, USA

Received 23 November 2011; revised 26 February 2012; accepted 16 April 2012; online publish-ahead-of-print 17 June 2012

See page 1867 for the editorial comment on this article (doi:10.1093/eurheartj/ehs186)

Aims

Randomized controlled trials (RCTs) are the gold standard for assessing the efficacy of therapeutic interventions because randomization protects from biases inherent in observational studies. Propensity score (PS) methods, proposed as a potential solution to confounding of the treatment–outcome association, are widely used in observational studies of therapeutic interventions for acute coronary syndromes (ACS). We aimed to systematically assess agreement between observational studies using PS methods and RCTs on therapeutic interventions for ACS.

Methods and results

We searched for observational studies of interventions for ACS that used PS methods to estimate treatment effects on short- or long-term mortality. Using a standardized algorithm, we matched observational studies to RCTs based on patients' characteristics, interventions, and outcomes ('topics'), and we compared estimates of treatment effect between the two designs. When multiple observational studies or RCTs were identified for the same topic, we performed a meta-analysis and used the summary relative risk for comparisons. We matched 21 observational studies investigating 17 distinct clinical topics to 63 RCTs (median = 3 RCTs per observational study) for short-term (7 topics) and long-term (10 topics) mortality. Estimates from PS analyses differed statistically significantly from randomized evidence in two instances; however, observational studies reported more extreme beneficial treatment effects compared with RCTs in 13 of 17 instances ($P = 0.049$). Sensitivity analyses limited to large RCTs, and using alternative meta-analysis models yielded similar results.

Conclusion

For the treatment of ACS, observational studies using PS methods produce treatment effect estimates that are of more extreme magnitude compared with those from RCTs, although the differences are rarely statistically significant.

Keywords

Observational studies • Randomized controlled trials • Acute coronary syndromes • Myocardial infarction • Unstable angina • Propensity score

* Corresponding author. Tel: +1 617 636 1459, Fax: +1 617 636 8628, Email: idahabreh@tuftsmedicalcenter.org

Introduction

Acute coronary syndromes (ACS), including acute myocardial infarction (AMI) and unstable angina (UA), are major causes of morbidity and mortality in the USA.[1] Many treatments for ACS have strong evidentiary support from randomized controlled trials (RCTs) and meta-analyses of RCTs. However, the conduct of RCTs is costly and often inefficient due to the large number of participants needed to estimate treatment effects with adequate precision.[2] Furthermore, conducting RCTs may not be feasible or even ethical for all clinical questions of interest, and restrictive selection criteria can limit the external validity of their results.[3] Observational studies are often a practical alternative to efficiently obtain estimates of the effectiveness of treatment in non-experimental, routine-care settings. Nonetheless, the lack of randomization and other RCT design elements renders observational studies susceptible to biases, including confounding (and particularly confounding by factors that affect treatment choice and are also causally associated with the outcome), selection, and differential ascertainment bias.[4]

Because of the efficiency gains potentially afforded by analyses of observational data, several comparisons of treatment effect estimates obtained from observational studies and RCTs have been performed, generally suggesting that the results of different designs are in moderate agreement.[5–8] Although these seminal empirical investigations provided useful insights into the frequency and magnitude of disagreements between study designs,[9] they often relied on collections of studies with heterogeneous patient populations, interventions, and analytical designs. Since the publication of these empirical comparisons, modern statistical methods allowing robust inference on treatment effects have been increasingly employed in the design and analysis of observational studies. In theory, the application of these methods should result in greater agreement between observational studies and RCTs.

Proposed as a potential solution to the problem of confounding of the treatment–outcome association, a propensity score (PS) expresses the probability of having been treated with an intervention based on variables measured at or before the time of treatment.[10,11] Analyses using PS methods attempt to emulate randomized comparisons because they allow contrasts between patient groups that are on average similar on all observed confounders; however, PS methods cannot adjust for unmeasured confounding.[3,12,13] The literature on ACS offers a unique opportunity for assessing agreement between RCTs and observational studies using PS methods because of the abundance of studies of both designs and the availability of multiple competing interventions for the treatment of these conditions. Clinical practice guidelines use observational studies employing PS methods as a basis for some of their recommendations, particularly when no randomized evidence is available,[14] and authors of observational studies using PS methods often perform informal comparisons of their results against RCT results;[15] however, no systematic comparison of these two lines of evidence has been performed.

In order to qualitatively and quantitatively assess the agreement between observational studies using PS methods and RCTs in the field of ACS therapeutics, we performed a systematic comparison of treatment effect estimates derived from these two designs.

Methods

Additional details of our methods for identifying observational studies and matching RCTs are presented in the Supplementary material online.

Identification and selection of observational studies

We searched Medline (through 11 February 2011) to identify studies using PS methods to obtain estimates of treatment efficacy for therapeutic interventions administered to patients with ACS. ACS was defined as AMI [ST-elevation myocardial infarction (STEMI) or non-STEMI (NSTEMI)] or UA; we accepted disease definitions as provided by each study. To increase the specificity of the search strategy, we limited our searches to the top 8 journals (by impact factor, Institute of Scientific Information, Thomson Reuters, Philadelphia, PA, USA) in the category 'Cardiac and cardiovascular systems' and to the top 4 journals in the category 'Medicine, general and internal' that publish primary clinical research studies. We screened titles and abstracts to identify studies enrolling patients with an established diagnosis of ACS that used PS to obtain estimates of the efficacy of competing therapeutic interventions.

Two reviewers (I.J.D. and G.D.K.) read potentially eligible studies in full text to determine eligibility; discrepancies were resolved by consensus. Eligible studies had to have an observational design, enrol patients with a diagnosis of ACS, and use PS methods to obtain estimates for treatment effects of therapeutic interventions on mortality. We considered only studies reporting on either short-term (typically within 30 days of ACS diagnosis) or long-term (more than 30 days following ACS diagnosis) mortality because of its clinical importance and the fact that it is less prone to misclassification compared with other outcomes. We classified the interventions investigated into pharmacological and non-pharmacological strategies.

Matching observational studies to randomized controlled trials

Two reviewers (I.J.D. and G.D.K.) independently attempted to match each observational study to at least one RCT, based on the interventions, patient populations, and type of mortality outcomes investigated, using a structured approach. Briefly, for interventions, we required that studies examined the same pharmacological or non-pharmacological interventions applied in the same clinical setting; for populations, matching was based on the examination of the same subtype of ACS (STEMI vs. NSTEMI/UA); and for mortality, matching was performed on short- and/or long-term mortality. Demographic or comorbidity characteristics of the examined populations were also considered in the matching process when they represented a selection criterion for the observational study. For example, for observational studies that examined patients with ACS and comorbid chronic kidney disease, we identified RCTs that specifically reported on the same population (ACS in patients with chronic kidney disease), and we also aimed to obtain estimates from subgroups of patients with kidney disease of the same disease stage. Similarly, for angiotensin-converting enzyme (ACE) inhibitors, because PS studies included unselected patients, we only considered RCTs that did not use enrolment criteria based on left ventricular function. Throughout the article, we refer to sets of populations, interventions/comparators, and outcomes as 'topics'.

Evidence from RCTs was identified through the following sources using a stepwise approach: (i) the Cochrane Database of Systematic Reviews; (ii) Medline-indexed meta-analyses; (iii) evidence-based guidelines from the American Heart Association/American College of Cardiologists; (iv) a compendium of medical therapeutics;[16] (v) focused Medline searches to identify eligible primary publications of RCTs; (vi) subgroups of interest from individual patient data meta-analyses of RCTs; (vii) subgroups of interest from single RCTs enrolling at least 1000 patients; and (viii) reference lists of the observational studies to identify any RCT that the study authors had considered comparable with their investigation. We searched these sources successively: we proceeded to a step only if at least one matching RCT was not identified at the previous step. When a relevant meta-analysis was identified, all included trials were retrieved and examined in full text for potential matches; when searches for individual trials were conducted, all trials identified through our searches were considered in full text. Two physicians with training in quantitative methods (I.J.D. and G.D.K.) verified all matches independently and a practicing cardiologist (H.J.) evaluated the final matched set.

Data extraction

For each matched set of observational studies and RCTs, we extracted the following information: study design aspects of the observational studies, sample size, duration of follow-up, statistical analysis methods including the specific approach to using the PS for estimating treatment effects (i.e. whether matching, stratification, regression, and inverse probability weighting were used), and treatment effect sizes for short-term and long-term mortality. Because only three studies reported treatment effect estimates from both regression-based and matched or stratified analyses utilizing the PS, we used the regression estimates for our primary analysis (when available, to ensure consistency) and we performed sensitivity analysis by considering the matched or stratified analysis results.

Assessment of the validity of propensity score-based analyses

Based on previously published surveys of the methodological features of studies using PS methods,[17–19] we identified a set of items as potentially indicative of the validity of PS-based analyses (provided in Supplementary material online). For each study considered in this review, a single reviewer (R.C.S., J.K.P., M.C., or V.V.) extracted these items, and extraction was verified by a second reviewer (I.J.D. or G.D.K.).

Statistical analyses

To the extent possible, we used the same metric of treatment effectiveness in RCTs and in observational studies. For example, if PS analyses reported odds ratios (ORs) for treatment effects (e.g. from logistic regression models), we extracted or calculated ORs from the RCTs as well; if they reported hazard ratios (HRs, from time-to-event models), we preferred HR estimates from time-to-event analyses of RCTs, when available. For all comparisons, we coined treatment effect metrics (ORs, HRs, or risk ratios), so that estimates lower than 1 indicate benefit (reduction in mortality) for the experimental treatment. For parsimony, in the Results section, we opt to refer to all relative effects metrics as 'relative risks' (RRs); this choice does not affect our results or their interpretation.

When multiple observational studies or RCTs were available for a topic, we performed meta-analyses using random-effects models (DerSimonian–Laird) to obtain a summary estimate of treatment efficacy and then used the summary estimate in all comparisons.[20]

We used a test for interaction to compare RR estimates from observational studies using PS methods and RCTs. This test compares whether the relative effect size (ratio of the observational study effects to the RCT effects) is significantly different from 1. Relative effect sizes (i.e. relative RRs) lower than 1 indicate that PS-based analyses produced results that were more favourable for the experimental treatment compared with RCTs. We compared how often the direction of the treatment effect estimated from the observational and the randomized study is the same, and we used a binomial (sign) test to evaluate whether a particular design tended to produce favourable results for the experimental treatment more often than would be expected by chance. We also described how often the ratio of the treatment effects in the PS studies over the RCTs was lower than the arbitrary threshold of 0.70 (or larger than the reciprocal, 1.43) indicating large differences in the magnitude of effect estimates, irrespective of statistical significance.

We performed the following sensitivity analyses: (i) we repeated all comparisons by using the single largest study (observational or randomized) instead of meta-analysis estimates for topics where more than one observational or randomized study was available; (ii) we performed a comparison limited to RCTs enrolling at least 1000 participants (mega-trials); and (iii) we repeated the meta-analysis under a fixed-effect model.

All analyses were conducted using Stata version SE/11.2 (Stata Corp., College Station, TX, USA). Statistical significance was defined as a two-tailed *P*-value less than 0.05 for all comparisons.

Results

Eligible observational studies and matched randomized controlled trials

Our searches for observational studies identified 599 citations, of which 70 were considered to be potentially eligible and were retrieved in full text. *Figure 1* presents the search strategy flow along with reasons for exclusion for studies reviewed in full text. Forty-nine observational studies using PS methods were considered eligible for inclusion, of which 21 were successfully matched to 63 RCTs and were considered further.

A median of three RCTs was considered for each topic (from one up to nine). In three topics, more than one observational study using PS was deemed to have investigated the same population and interventions and was matched to the same RCTs (two observational studies for each of the two topics and three for a third topic). Overall, we considered 17 topics in which at least one observational study using PS methods was matched to at least one RCT. *Table 1* summarizes the 17 topics considered in this review, and Supplementary material online, *Table S1* presents details of the populations, interventions, and comparators included in each study. Briefly, six topics pertained exclusively to STEMI populations, seven exclusively to NSTEMI/UA populations, and four to mixed populations (STEMI and NSTEMI/UA). The treatments investigated included pharmacological interventions [antiplatelet agents ($n = 3$), ACE inhibitors ($n = 1$), and lipid-lowering medications (primarily statins and the timing of their administration, $n = 3$)] and non-pharmacological approaches [cardiac rehabilitation involving exercise component ($n = 1$) and alternative revascularization strategies ($n = 9$)].

Eligible observational studies were generally large with a median of 2310 (range: 193–38 395 and 25–75th percentile: 1003–4892) and 5194 (range: 324–126 128, 25–75th percentile: 872–8769)

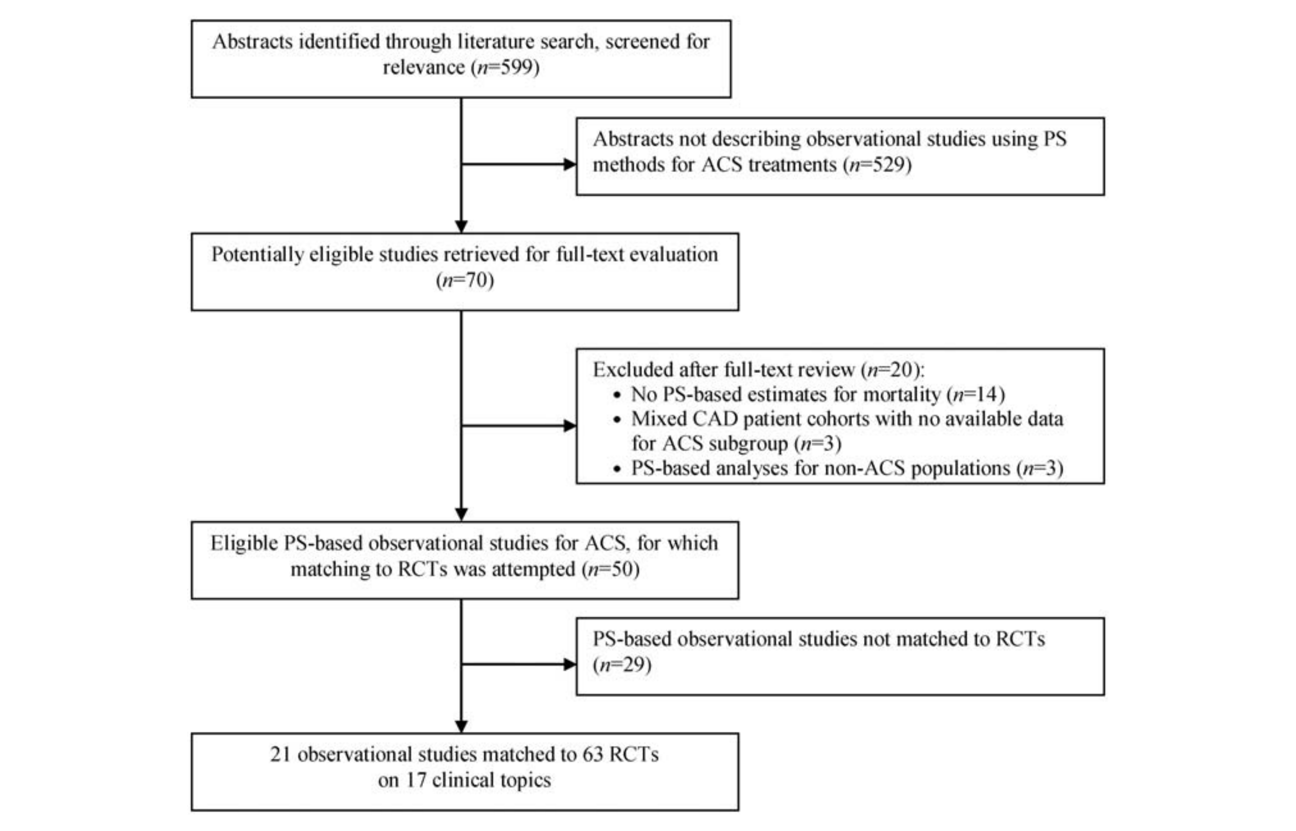
Abstracts identified through literature search, screened for relevance (n=599)
Abstracts not describing observational studies using PS methods for ACS treatments (n=529)
Potentially eligible studies retrieved for full-text evaluation (n=70)
Excluded after full-text review (n=20):
• No PS-based estimates for mortality (n=14)
• Mixed CAD patient cohorts with no available data for ACS subgroup (n=3)
• PS-based analyses for non-ACS populations (n=3)
Eligible PS-based observational studies for ACS, for which matching to RCTs was attempted (n=50)
PS-based observational studies not matched to RCTs (n=29)
21 observational studies matched to 63 RCTs on 17 clinical topics

Figure 1 Search strategy flow. ACS, acute coronary syndromes; CAD, coronary artery disease; PS, propensity score; RCTs, randomized controlled trials.

patients on the experimental and control arms, respectively. RCTs had smaller sample sizes: median of 176 (range: 12–4722 and 25–75th percentile: 64–372) and 177 (range: 17–4739 and 25–75th percentile: 67–452) patients on the experimental and control arms, respectively. Nineteen of the observational studies had a prospective design.

Comparison of observational studies and randomized controlled trials for long-term mortality

Long-term mortality was assessed in 10 of the 17 topics (14 observational studies/43 RCTs). The median duration of follow-up in the examined studies was 12 months (25–75th percentile: 9–12). RRs ranged between 0.44 and 0.99 (median = 0.69) for observational studies and between 0.10 and 3.10 (median = 0.77) for RCTs (*Figure 2A*). Estimates from observational studies and RCTs had opposite directions in one topic (choice of IIb/IIIa inhibitors for primary percutaneous coronary intervention). PS-based analyses suggested a greater benefit for experimental treatments compared with RCTs in 8 of the 10 long-term mortality comparisons (sign test $P = 0.109$). The difference between PS- and RCT-derived estimates was statistically significant in 1 of the 10 topics (exercise rehabilitation for AMI, *Figure 2B*).

Comparison of observational studies and randomized controlled trials for short-term mortality

Short-term mortality was assessed in 7 of the 17 topics (7 observational studies/20 RCTs). Seventeen studies assessed in-hospital or 7-day mortality, and the remaining studies examined 30-day mortality. RRs ranged between 0.45 and 0.89 (median = 0.62) for observational studies and between 0.10 and 4.12 (median = 0.75) for RCTs (*Figure 3A*). Estimates from observational studies and RCTs had opposite directions in two topics (statin timing for AMI and invasive vs. conservative strategy for NSTE-ACS). PS-based analyses suggested a greater benefit for experimental treatments compared with RCTs in five of the seven short-term mortality comparisons (sign test $P = 0.45$). The difference between PS- and RCT-derived estimates was statistically significant in one of the seven topics (invasive vs. conservative strategy for NSTE-ACS, *Figure 3B*).

Considering both long- and short-term mortality topics combined, there was some suggestion that observational studies tended to report more extreme results in favour of experimental treatments (13 of 17 topics, sign test $P = 0.049$). In 6 out of 17 topics, the actual point estimates from PS-based analyses implied much more protective effects (ratio of PS- to RCT-based RRs

Table I Summary of treatment comparisons investigated by observational studies using propensity scores and matched randomized trials

Brief topic description[a]	PS/RCTs (*N*)	Population	Treatment comparison
Short-term mortality			
Pharmacological interventions			
Abciximab timing in pPCI	1/3	STEMI planned to be treated with pPCI	Early (pre-procedural) vs. late (peri-procedural) abciximab administration
IIb/IIIa inhibitors (NSTE-ACS)	1/2	NSTE-ACS	IIb/IIIa inhibitor administration vs. no treatment within 24 h
Statins timing (AMI)	1/2	ACS	Early statin administration vs. no treatment
Non-pharmacological interventions			
pPCI (shock)	1/1	STEMI with shock	pPCI vs. initial medical management
pPCI (elderly)	1/3	Elderly patients with STEMI	pPCI vs. thrombolysis
Invasive strategy timing (NSTE-ACS)	1/3	NSTE-ACS	Very early angiography ($<$6–12 h) vs. initial conservative management with delayed angiography
Invasive strategy (NSTE-ACS)	1/6	NSTE-ACS	Early invasive strategy with angiography (and revascularization, when indicated) vs. early conservative strategy
Long-term mortality			
Pharmacological interventions			
ACEi (AMI)	1/3	AMI	ACEi vs. no ACEi initiation during hospitalization for AMI
IIb/IIIa inhibitors type in pPCI	1/1	STEMI treated with pPCI	Eptifibatide vs. abciximab
Statins (ACS)	3/5	ACS	Statin vs. no statin initiation
Statins (NSTE-ACS)	2/2	NSTE-ACS	Statin vs. no statin initiation
Non-pharmacological interventions			
Invasive strategy (NSTE-ACS and CKD)	1/5	NSTE-ACS with CKD $\geq$ stage 3	Early invasive strategy with angiography (and revascularization, when indicated) vs. early conservative strategy
Invasive strategy timing (NSTE-ACS and elderly)	1/1	NSTE-ACS, $>$75 years	Early invasive strategy with angiography (and revascularization, when indicated) vs. early conservative strategy
DES in pPCI	2/9	STEMI treated with pPCI	DES vs. BMS
Exercise rehabilitation (AMI)	1/9	AMI	Cardiac rehabilitation programme with exercise component vs. control (no cardiac rehabilitation)
Complete PCI revascularization (STEMI)	1/2	STEMI patients with multivessel CAD	Complete vs. culprit-only PCI revascularization
Invasive strategy (NSTE-ACS)	1/6	NSTE-ACS	Early invasive strategy with angiography (and revascularization, when indicated) vs. early conservative strategy

ACEi, angiotensin-converting enzyme inhibitor; ACS, acute coronary syndromes; AMI, acute myocardial infarction; BMS, bare metal stent; CAD, coronary artery disease; CKD, chronic kidney disease; DES, drug-eluting stent; h, hours; NSTE, non-ST-elevation; PCI, percutaneous coronary intervention; pPCI, primary percutaneous coronary intervention, PS, observational studies using propensity scores; RCT, randomized controlled trial; STEMI, ST-elevation myocardial infarction.

[a] Specific ACS subtypes or population characteristics are shown in parentheses.

less than 0.70). The inverse (ratio of PS- to RCT-based RRs greater than 1.43) was not observed in any topic.

Validity of propensity score-based analyses

In general, PS-based analyses did not follow current recommendations for statistical practice (Supplementary material online, *Table S2*). In all studies that provided data for covariate selection for the construction of the PS model ($n = 10$), this selection was based on stepwise regression methods rather than pre-existing knowledge of potential confounders of the treatment–mortality association; regression methods (i.e. inclusion of the PS as a covariate in the outcome model) were used in the majority of analyses instead of the recommended matched or stratified analyses (in our main analysis 18 of 21 studies used the PS as a regression covariate). Even when matched analyses were undertaken, the balance between the matched groups was often not assessed, and the applied statistical analyses ignored the paired structure of the samples.

Regression analyses compared with propensity score-based analyses in observational studies

Eleven of the observational studies using PS methods (five for long- and six for short-term mortality) reported estimates from multivariable adjustments using regression methods (without the use of a PS). Generally, estimates of the treatment effect from regression analyses were very close to those from PS-based analyses and

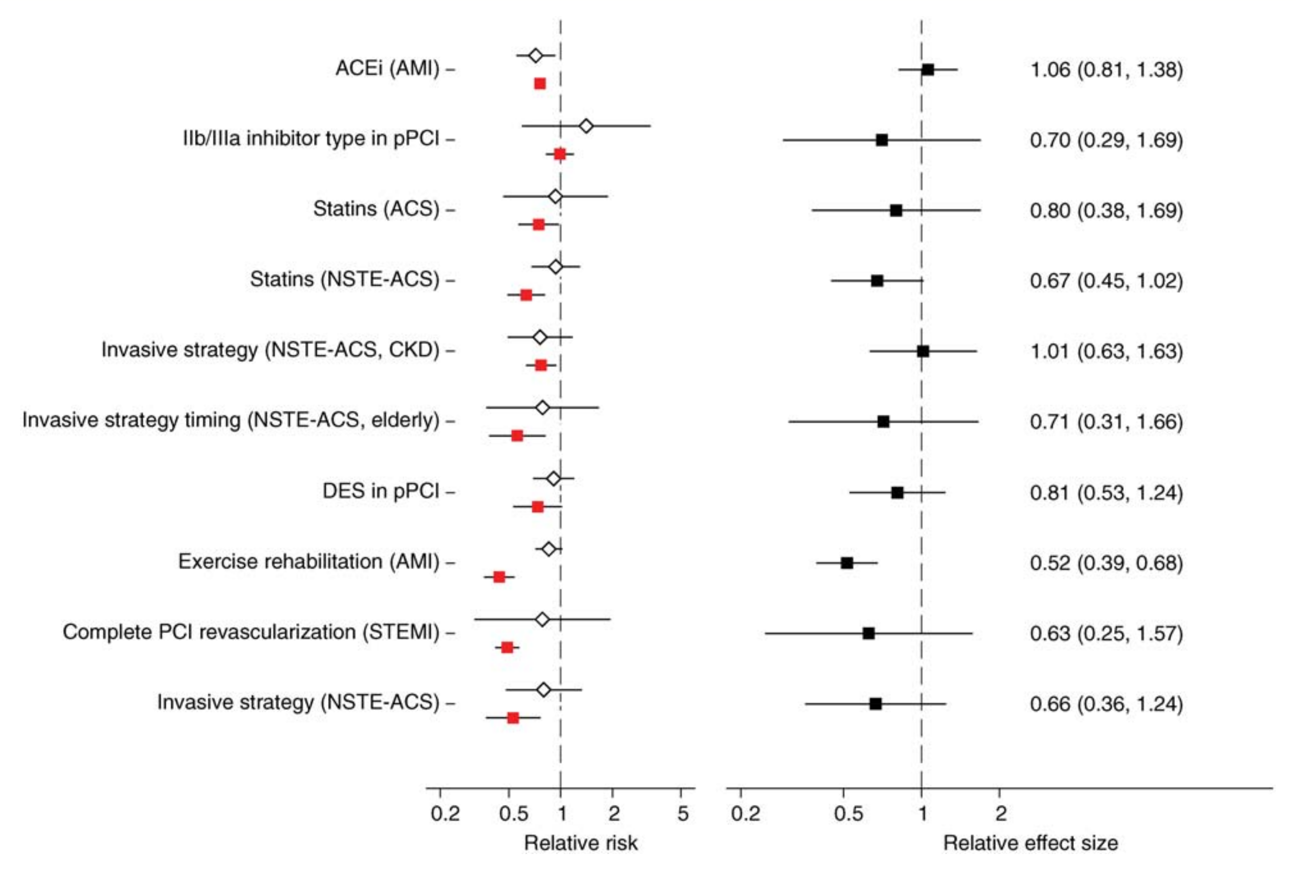

Figure 2 (*A*) Relative risks from randomized controlled trials (white diamonds) and observational studies using propensity score methods (red squares), reporting on long-term mortality. When more than two randomized or observational studies were available for a topic, estimates were obtained from random-effects meta-analysis. (*B*) Corresponding relative effect sizes (black squares) comparing the effect sizes from randomized and propensity-based observational studies on long-term mortality. In both panels, extending lines denote 95% confidence intervals of estimates. See Supplementary material online, *Table S1* for detailed study-level information. Please see *Table 1* for abbreviation definitions.

had overlapping confidence intervals (Supplementary material online, *Figure S2*).

Sensitivity analyses

Analyses based on the single largest study (observational or randomized) for each topic yielded results similar to our main analysis (1 significant discrepancy in the 17 topics, i.e. one of the discrepancies in the main analysis was eliminated by considering only the largest available RCT for invasive strategies in NSTE-ACS reporting on long-term outcomes). One or more RCTs enrolling at least 1000 participants were available in 7 of the 17 topics (2 for short-term and 5 for long-term mortality). In these cases, the results of the mega-trials were significantly different from those of PS-based analyses in two cases (the same discrepancies as in the main analysis). Repeating the meta-analyses using a fixed-effect model (instead of random-effects) also did not materially affect our results: significant discrepancies between the RRs of PS-based analyses and RCTs were observed in 3 of the 17 topics (the 2 that were discrepant in the main analysis in addition to the use of invasive vs. conservative management strategies for NSTE-ACS reporting on short-term mortality). Finally, using estimates obtained from matching or stratification for the three studies that reported results from such analyses (in addition to regression-based estimates using the PS) did not affect our results (no additional discrepancies were observed).

Discussion

Summary of key findings

PS methods have been described as the observational study analogue of randomization in clinical trials, when all important confounders are accounted for in the PS model.[21] We compared the results of studies using PS methods with those of RCTs in 17 topics covering various therapeutic interventions for ACS. The comparisons of treatment effect estimates for short- and long-term mortality suggest that PS methods and RCTs are often consistent: only in 2 of the 17 comparisons, estimates from observational studies and RCTs were statistically significantly different. However, there was a substantial difference in the magnitude of the effect sizes (ratio of the PS- to RCT-based RRs lower than 0.70) in 6 of the 17 topics. We also found evidence

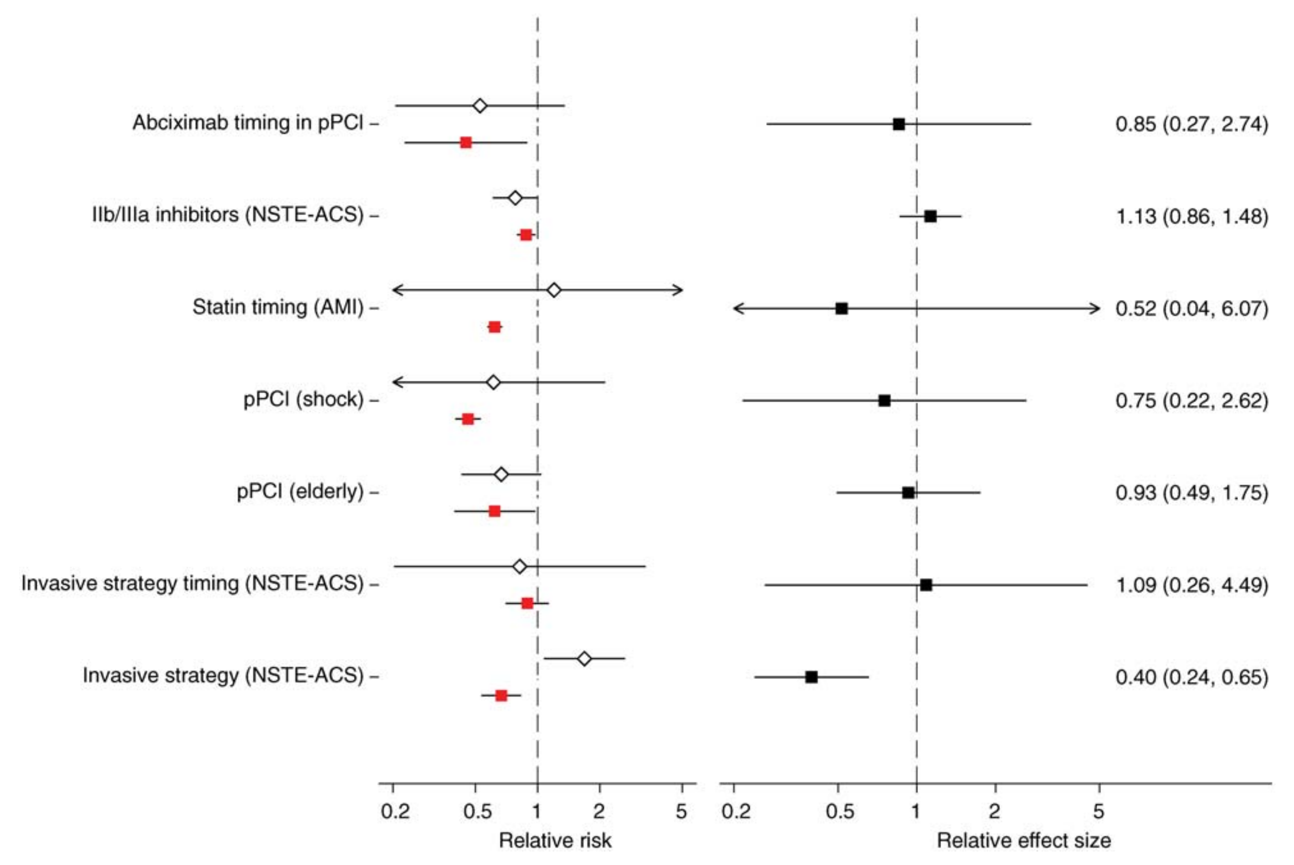

Figure 3 (*A*) Relative risks from randomized controlled trials (white diamonds) and observational studies using propensity score methods (red squares), reporting on short-term mortality. (*B*) Corresponding relative effect sizes (black squares) comparing the relative risks from randomized trials and observational studies using propensity score methods. Layout is similar to *Figure 2*. Arrows indicate lower or upper bounds of confidence intervals that were outside the plotted range of values. See Supplementary material online, *Table S1* for detailed study-level information. Please see *Table 1* for abbreviation definitions.

that PS estimates systematically overestimate treatment efficacy ($P = 0.049$). These results were robust to extensive sensitivity analyses. Given the small-to-moderate effect sizes in clinical medicine[22] and the close agreement between PS-based and regression-based estimates, differences between PS- and RCT-derived estimates cannot be ascribed to estimation properties of different PS methods.[23]

The conduct of PS-based analyses had relatively limited validity based on our assessment of a set of predefined criteria. In most cases, commonly suggested recommendations for the optimal application of PS methods were not followed. This finding is in agreement with previous assessments of published PS-based analyses[17–19] and may indicate that PS methodologies are not always optimally applied. Given that non-PS-based analyses tend to show slightly stronger effects compared with PS-based analyses,[24] it is possible that greater methodological rigour would have increased agreement with RCT results. At any rate, such are the data that exist in the literature and are available to clinicians and policy makers. For example, the American Heart Association/American College of Cardiology guidelines consider observational studies using PS methods as a source of evidence for many of their recommendations.[14]

Within observational studies, we found that standard multivariable regression analyses were generally in agreement with PS-based analyses. This is consistent with empirical investigations in diverse clinical topics.[18,24] Simulation studies suggest that PS-based methods and multivariate regression can yield very similar results when there is an adequate number of events per confounder.[25]

Observational studies may overestimate treatment effects

Although PS-based analyses and RCTs produced results that were statistically consistent, the discrepancies in specific cases (all pertaining to non-pharmacological interventions) cannot be ignored. Moreover, the point estimates of the relative effect sizes comparing treatment effects in observational studies vs. RCTs were often far from the null, despite being statistically non-significant. We found evidence that observational studies reported more extreme treatment effects. Several potential explanations for such discrepancies exist. First, studies may have enrolled populations with different underlying characteristics, not captured by our matching algorithm. In such case, observational and randomized studies are estimating different underlying parameters, and any discrepancies cannot be used to infer whether one study design is more valid than the other.

Secondly, RCTs analyse patients in the group they were allocated to (per the intention-to-treat principle), whereas observational studies by definition perform as-treated analyses. These analytic approaches are fundamentally distinct; their numerical results deviate with increasing number of people who crossover between treatments. For example, in RCTs comparing invasive revascularization with medical therapy, patients in the medical therapy arm may crossover to the revascularization arm for several reasons. Such crossovers can attenuate the treatment effect in RCTs analysed by intention-to-treat (compared with an as-treated analysis)[26,27] and may partly explain why RCTs tended to have more conservative estimates, since most of the included RCTs reported intention-to-treat analyses. However, the exact definition of intention-to-treat analyses was often unclear or inconsistent between studies.[28]

Thirdly, publication bias and selective outcome reporting may be another explanation for the observed discrepancies, particularly since they may affect observational studies more than RCTs. Conducting studies based on protocols predefining the analyses to be performed and making all results available would eliminate this potential source of discrepancies; however, it is not possible to assess to what extent these biases could explain our results.[29]

Finally, discrepancies may be due to the presence of biases in observational studies that cannot be accounted by PS methods, such as selection bias, outcome ascertainment bias, immortal-time bias (particularly in studies examining the effects of the timing of interventions), or because of residual confounding; the latter is more of a concern in PS analyses of large administrative databases, in which many important variables may not be captured at all or may be measured with noise. In the presence of such biases, estimates from observational studies cannot be considered valid. However, because the 'true' magnitude of treatment effects is unknowable, critical appraisal of the actual implementation of design and analysis, both for observational studies and for RCTs, is necessary on a case-by-case basis.[30]

Strengths and limitations

The strengths of the present study include the focus on a homogeneous clinical condition (ACS) and selecting studies in which mortality was the outcome of interest. By focusing on diseases for which the existing evidence basis is extensive, our work overcame the limitations in previous studies comparing observational and randomized studies among very heterogeneous clinical conditions.[31] Furthermore, by focusing on mortality as the outcome of interest, we were able minimize the possibility of outcome misclassification and differential ascertainment as potential explanations for any observed discrepancies.

Our work has several limitations. The process of matching observational studies to RCTs is inherently subjective. We minimized subjectivity by performing the matching in duplicate and then having a third reviewer with context expertise verify the matching results. In many cases, our criteria for similarity are narrower than the eligibility criteria of several published meta-analyses (for example, we matched studies on ACS subtypes, whereas meta-analyses often considered studies of STE- and NSTE-ACS together), and our results were robust to extensive sensitivity analyses. Our matching algorithm may not be exhaustive; a full systematic review for PS-based studies and RCTs for all ACS interventions is not feasible. Further, the subset of matched observational studies may not be representative of all observational studies using PS methods. Nonetheless, our approach represents a replicable, systematic way for comparing the study designs of interest and covered many commonly used interventions for ACS. We chose mortality as our outcome of interest, a relatively rare outcome for which the treatment effects estimates from RCTs were relatively imprecise. However, mortality is arguably the clinical outcome of primary interest to patients and decision makers, and ascertainment of deaths is also less susceptible to reporting or misclassification biases; as such, estimates of treatment effects on mortality may be the most suitable for systematic comparisons across study designs.

Implications for clinical practice and future research

Our work has important implications for clinical practice and future research, particularly in regard to comparative effectiveness. Observational data that are representative of current clinical practice are becoming increasingly available through several sources, such as prospectively maintained registries or electronic medical record systems. Observational studies using modern design methods such as PS matching may provide an efficient way for evaluating the effects of interventions in typical clinical settings,[32,33] providing timely decision-relevant information and helping to prioritize which research needs to address with more resource-intensive RCTs.

The suggestion that the results of PS-based analyses and matched RCTs are generally consistent does not mean that decision-makers should uncritically rely on observational evidence, given the indications that the latter overestimate treatment efficacy. Randomization can guarantee that the compared groups are (on average) balanced on both observed and unobserved covariates and that the treatment effect estimates are unbiased. In this regard, RCTs remain superior to observational studies, despite advances in the design and analyses of the latter. Results from well-conducted and analysed observational studies can, however, be judiciously used to supplement insufficient or pending randomized evidence, to inform decisions that are unlikely to be examined in RCTs, and to help set the future research needs agenda.

Supplementary material

Supplementary material is available at *European Heart Journal* online.

Funding

This study was supported in part by grant UL1 RR025752 from the National Center for Research Resources to Tufts-Clinical Translational

Science Institute. The content is solely the responsibility of the authors and does not represent the official views of the National Center for Research Resources or the National Institutes of Health. The funder did not participate in the design, conduct, analysis, interpretation of the study, or the decision to submit the manuscript for publication.

Conflict of interest: none declared.

References

1. Roger VL, Go AS, Lloyd-Jones DM, Adams RJ, Berry JD, Brown TM, Carnethon MR, Dai S, De Simone G, Ford ES, Fox CS, Fullerton HJ, Gillespie C, Greenlund KJ, Hailpern SM, Heit JA, Ho PM, Howard VJ, Kissela BM, Kittner SJ, Lackland DT, Lichtman JH, Lisabeth LD, Makuc DM, Marcus GM, Marelli A, Matchar DB, McDermott MM, Meigs JB, Moy CS, Mozaffarian D, Mussolino ME, Nichol G, Paynter NP, Rosamond WD, Sorlie PD, Stafford RS, Turan TN, Turner MB, Wong ND, Wylie-Rosett J. Heart disease and stroke statistics—2011 update: a report from the American Heart Association. *Circulation* 2011;**123**:e18–e209.
2. Kent DM, Trikalinos TA. Therapeutic innovations, diminishing returns, and control rate preservation. *JAMA* 2009;**302**:2254–2256.
3. Rosenbaum PR. *Observational Studies*. 2nd ed. New York, USA: Springer; 2002.
4. Byar DP. Why databases should not replace randomized clinical trials. *Biometrics* 1980;**36**:337–342.
5. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;**342**:1887–1892.
6. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;**342**:1878–1886.
7. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, Contopoulos-Ioannidis DG, Lau J. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;**286**:821–830.
8. MacLehose RR, Reeves BC, Harvey IM,, Sheldon TA, Russell IT, Black AM. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technol Assess* 2000;**4**:1–154.
9. Ioannidis JP, Haidich AB, Lau J. Any casualties in the clash of randomised and observational evidence? *BMJ* 2001;**322**:879–880.
10. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70**:41–55.
11. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med* 2007;**26**:20–36.
12. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health* 2000; **21**:121–145.
13. Heinze G, Juni P. An overview of the objectives of and the approaches to propensity score analyses. *Eur Heart J* 2011;**32**:1704–1708.
14. Wright RS, Anderson JL, Adams CD,, Bridges CR, Casey DE Jr, Ettinger SM, Fesmire FM, Ganiats TG, Jneid H, Lincoff AM, Peterson ED, Philippides GJ, Theroux P, Wenger NK, Zidar JP, Jacobs AK. 2011 ACCF/AHA Focused Update of the Guidelines for the Management of Patients With Unstable Angina/Non-ST-Elevation Myocardial Infarction (Updating the 2007 Guideline): a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *Circulation* 2011;**123**:2022–2060.
15. Stukel TA, Fisher ES, Wennberg DE,, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA* 2007;**297**:278–285.
16. Foster C, Mistry NF, Peddi PF, Shivak S. *The Washington Manual of Medical Therapeutics*. 33rd ed. St. Louis, MO, USA: Lippincott Williams & Wilkins; 2010.
17. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med* 2008;**27**:2037–2049.
18. Sturmer T, Joshi M, Glynn RJ,, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol* 2006;**59**:437–447.
19. Austin PC. Primer on statistical interpretation or methods report card on propensity-score matching in the cardiology literature from 2004 to 2006: a systematic review. *Circ Cardiovasc Qual Outcomes* 2008;**1**:62–67.
20. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;**7**: 177–188.
21. Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *J Am Stat Assoc* 2000;**95**:573–585.
22. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* 1984;**3**:409–422.
23. Austin PC, Grootendorst P, Normand SL, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Stat Med* 2007;**26**:754–768.
24. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol* 2005;**58**:550–559.
25. 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. *Am J Epidemiol* 2003;**158**:280–287.
26. Brophy JM, Belisle P, Joseph L. Evidence for use of coronary stents. *A hierarchical Bayesian meta-analysis. Ann Intern Med* 2003;**138**:777–786.
27. Ottervanger JP, Armstrong P, Barnathan ES,, Boersma E, Cooper JS, Ohman EM, James S, Wallentin L, Simoons ML. Association of revascularisation with low mortality in non-ST elevation acute coronary syndrome, a report from GUSTO IV-ACS. *Eur Heart J* 2004;**25**:1494–1501.
28. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ* 1999;**319**:670–674.
29. Sterne JA, Sutton AJ, Ioannidis JP,, Terrin N, Jones DR, Lau J, Carpenter J, Rucker G, Harbord RM, Schmid CH, Tetzlaff J, Deeks JJ, Peters J, Macaskill P, Schwarzer G, Duval S, Altman DG, Moher D, Higgins JP. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;**343**:d4002.
30. Concato J, Lawler EV, Lew RA,, Gaziano JM, Aslan M, Huang GD. Observational methods in comparative effectiveness research. *Am J Med* 2010;**123**(Suppl. 1): e16–e23.
31. Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? *N Engl J Med* 2000;**342**:1907–1909.
32. Schneeweiss S, Rassen JA, Glynn RJ,, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 2009;**20**:512–522.
33. Rassen JA, Avorn J, Schneeweiss S. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. *Pharmacoepidemiol Drug Saf* 2010;**19**:848–857.