Differences in interaction and subgroup-specific effects were observed between randomized and nonrandomized studies in three empirical examples

Amand F. Schmidt\textsuperscript{a,b,c,*}, Maroeska M. Rovers\textsuperscript{d}, Olaf H. Klungel\textsuperscript{a,b}, Arno W. Hoes\textsuperscript{a}, Mirjam J. Knol\textsuperscript{a}, Mirjam Nielen\textsuperscript{c}, Antonius de Boer\textsuperscript{b}, Rolf H.H. Groenwold\textsuperscript{a,b}

\textsuperscript{a}Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, The Netherlands
\textsuperscript{b}Division of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute for Pharmaceutical Sciences, PO Box 80082, 3508 TB Utrecht, The Netherlands
\textsuperscript{c}Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, Yalelaan 7, 3584 CL Utrecht, The Netherlands
\textsuperscript{d}Departments of Operating Rooms and Health Evidence, Radboud University Nijmegen Medical Centre, Geert Grooteplein 21, 6525 EZ Nijmegen, The Netherlands

Accepted 20 August 2012; Published online 16 March 2013

Abstract

**Objective:** To determine the comparability of subgroup-specific and interaction effects (differences between subgroups) between different study designs.

**Study Design and Setting:** We compared effects of interventions based on observational studies, randomized clinical trials (RCTs), and individual patient data meta-analyses (IPDMAs) of RCTs (reference) on three clinical topics: (1) mammography screening and breast cancer mortality, (2) coronary artery bypass surgery (CABG) and all-cause mortality, and (3) statins and incidence of major coronary events.

**Main, subgroup-specific, and interaction effects were compared.**

**Results:** Main and subgroup-specific effects were comparable with respect to the direction of the effects. Differences in the magnitude of subgroup-specific effects in observational studies yielded different interactions compared with those in IPDMA. In the mammography example, the ratio of risk ratios (RRR) (i.e., interaction effect) among observational studies was 1.46 [95% confidence interval (CI): 1.09, 1.96] compared with an IPDMA effect of 1.10 (95% CI: 0.89, 1.37). For the CABG studies, the observational RRR was 1.03 (95% CI: 0.84, 1.26), whereas in the IPDMA, this was 1.40 (95% CI: 1.08, 1.81). Finally, in the statin example, the RRR was 1.35 (95% CI: 1.13, 1.61) and 0.90 (95% CI: 0.84, 0.97) for observational studies and IPDMA, respectively.

**Conclusion:** Main and subgroup-specific effects based on observational data were similar to main and subgroup-specific effects in IPDMAs based on RCTs, yet interactions differed. © 2013 Elsevier Inc. All rights reserved.

**Keywords:** Interaction; Effect modification; Individual patient data meta-analyses; Subgroups; Nonrandomized; Randomized

1. Background

Randomized clinical trials (RCTs) are the gold standard to evaluate the effects of medical interventions. Typically, randomized trials provide an estimate of the intervention effect that applies to the average patient included in the study. However, today’s clinical practice is shifting more and more toward individually tailored care. Personalized care requires knowledge on the effects of medical interventions at an individual rather than at a patient population level [1]. Compared with main effect estimates, subgroup analyses move toward a more personalized estimate.

A distinction can be made between subgroup-specific effects (i.e., effects within subgroups of patients) and interaction effects (i.e., difference between subgroup-specific effects). When exploring subgroups, one may stratify the
study population, which decreases the sample size. Hence, differences in effects between subgroups are more likely to occur simply because of chance, and therefore, it is recommended to perform a formal test of interaction [2,3]. Furthermore, if one is interested to test whether effects of medical interventions differ between subgroups (i.e., interaction effects), exploring subgroup-specific effects is inappropriate [2,4].

An individual RCT is often underpowered to detect interaction effects because of sample size constraints [4]. Alternatively, data from multiple RCTs can be pooled in an individual patient data meta-analysis (IPDMA) [5,6], in which interaction effects can be evaluated. Nevertheless, conducting an IPDMA, based on RCTs, is not always feasible. Alternatively, observational (i.e., nonrandomized) studies, which typically comprise larger sample sizes, can be used to explore subgroup effects. Observational data, however, have limitations [7], such as the potential for confounding.

Although numerous techniques and designs have been proposed to control for confounding [8], few can account for unobserved (i.e., unmeasured) confounding. In particular, observational studies of intended effects of interventions are at risk for confounding bias [9–12]. Moreover, observational studies are also hampered by other problems such as the potential for selection and information bias. A related issue is that RCTs tend to include healthier subjects compared with the general patient population [13]. On the other hand, several authors showed that high-quality observational studies of intended effects often display main effects that are comparable with those obtained from RCTs [14–16]. Whether observational-based (i.e., based on nonrandomized data), subgroup-specific, and interaction effect estimates can also approximate results of RCTs or IPDMA of RCTs remains unknown. We therefore conducted a review of three clinical examples to evaluate the comparability of effect estimates obtained from different study designs (e.g., observational, RCT, and IPDMA).

2. Methods

2.1. Search strategy

IPDMAs were identified using the “IPD Cochrane Methods Group” Web site [17] and MEDLINE database. IPDMAs were deemed suitable if they (1) explored subgroups based on patient characteristics at baseline, (2) allowed for direct comparison of subgroup-specific effects, (3) reported sufficient data to calculate point estimates of the treatment effects with confidence intervals (CIs), (4) were based on RCTs, and (5) were written in English.

Subsequently, we searched for (additional) RCTs and observational articles. First, we searched MEDLINE and the CENTRAL databases with an adapted search strategy, used by the original IPDMAs to also include observational studies (Appendix A at www.jclinepi.com). This search was supplemented with a Scopus [18] cross-reference search. We performed this strategy on five preselected domains: mammography screening in breast cancer mortality, antibiotics in rhinosinusitis, antibiotics in acute otitis media, phenytoin in epileptics, and carboplatin in ovarium cancer survival [19–23]. This strategy only yielded enough observational studies for the mammography [23] example to facilitate a meaningful comparison.

Additionally, we searched MEDLINE for IPDMAs and systematic reviews that included RCTs and/or observational studies. The reference lists of these reviews were searched for relevant publications, which we subsequently retrieved and screened for inclusion. We used Scopus to search for additional references. This search resulted in two additional post hoc examples: coronary artery bypass surgery (CABG) vs. percutaneous coronary intervention (PCI) on all-cause mortality and statin therapy in the prevention of cardiovascular events [24,25].

RCTs and observational studies were included when they (1) investigated similar patients, interventions, and outcomes as the IPDMA; (2) investigated similar subgroup-specifics that allowed for direct comparison of treatment effects; (3) allowed calculation of point estimates and CI of the treatments effects; (4) used an RCT, cohort
study, or case—control design; and (5) were written in English. We deemed an example viable if we found two or more observational studies that were comparable with the IPDMA. Because a meta-analysis based on individual patient data from RCTs allows the researcher to uniformly apply subgroup-specific cut-off points, choose similar endpoints and adjustment for confounding when necessary, we considered IPDMAs as the reference standard [6]. To check whether pooled estimates of reported studies could approximate IPDMA results, we explored reported RCT estimates, including the IPDMA RCTs [26,27].

2.2. Statistical analysis

Extracted data were analyzed using R, version 2.10 for windows [28]. When available, we used effect measures that were adjusted for possible differences in baseline covariates between comparison groups.

We used reported effect measure and if necessary calculated subgroup-specific effects based on reported data. Effects were reported in risk ratios or rate ratios (RR), hazard ratios (HRs) [29—32], or odds ratios (ORs; for case—control studies) with 95% CI. In all cases in which ORs or HRs were used, the incidence was ≤10% for both main and subgroup-specific outcomes, fulfilling the rare disease assumption [33]. Prespecified subgroups included age groups (in the examples on mammography screening and statin therapy) and diabetes presence or absence (CABG example). For the observational and RCT effects, measures of heterogeneity (Q-statistic (Q), I-squared (I²), and tau-squared (τ²) [34]) were calculated, and pooled effects were estimated using fixed- and random-effects models. In all the three design types, an interaction test was performed. This was done by taking the ratio of the stratum-specific effects [2]. This resulted in a ratio of risk ratios (RRR). When 

\[ RRR = 1, \] 

there is no interaction effect (i.e., no differences of treatment effect between subgroup); 

\[ RRR < 1 \]

indicates a smaller effect of treatment in one subgroup compared with that in the reference group; 

\[ RRR > 1 \]

indicates a larger effect of treatment in one subgroup compared with that in the reference group. To obtain a standard error (SE) for the 95% CI, we squared stratum-specific SE of which we took the square root of the sum. For the observational and RCT data, the RRR was based on the ratio of subgroup-specific random effects results. For the IPDMA, the RRR was based on the reported subgroup-specific estimates. In the Results section, we state which reference groups were used.

3. Results

3.1. Effect of mammography screening by age

The IPDMA of Nystrom et al. [23] determined the effects of mass mammography screening vs. no screening on breast cancer mortality. To study the effectiveness of
mammography screening in younger women, Nystrom stratified the results by age, <50 and ≥50 years, which resulted in a nonsignificant interaction test. The data of the IPDMA could be compared with six trials (of which four were included in the IPDMA) [35–40] and six observational studies (one cohort study [41] and five case–control studies [42–46]). The IPDMA included 247,010 women, of whom 1,642 died of breast cancer, whereas in the RCTs, 392,483 women participated, of whom 1,645 died of breast cancer. The observational studies included 233,791 women, of whom 4,498 died of breast cancer. Overall, the included studies were similar regarding the intervention, control group, and outcome parameter (see Appendix B at www.jclinepi.com).

The pooled main effect of mammography screening on breast cancer mortality (Fig. 1) in RCTs, IPDMA, and observational data were RR, 0.77 (95% CI: 0.69, 0.84); RR, 0.85 (95% CI: 0.77, 0.93); and RR, 0.65 (95% CI: 0.54, 0.78), respectively.

When data were stratified by age (women younger than 50 years and 50 years or older), a similar pattern was observed (Fig. 2). In younger women, the effects of mammography screening were similar, irrespective of type of study design. In older women, however, the effect in the individual observational studies was larger than that observed in the IPDMA, but the direction of effect was in agreement. In the IPDMA, RCTs, and observational data, the interaction effects (RRR) in young women compared with those in older women were 1.10 (95% CI: 0.89, 1.37), 1.17 (95% CI: 0.94, 1.47), and 1.46 (95% CI: 1.09, 1.96), respectively.

### 3.2. Effect of CABG (vs. PCI) by diabetes status

Hlatky et al. [25] studied the effect CABG vs. PCI on all-cause mortality using an IPDMA (including 10 trials). Hlatky stratified the main effects for numerous baseline characteristics including the presence of diabetes, which produced a significant interaction (P = 0.014). The data of the IPDMA could be compared with five trials [47–51] (which were all included in the IPDMA), and three observational cohort studies [31,32,52]. The IPDMA included 7,812 subjects who underwent a CABG or PCI procedure, of whom 1,203 died. The total sample size of the RCTs was 6,087 subjects, of whom 807 died. The cohort studies included 23,629 subjects, of whom 866 died. The number of diabetes patients varied according to the study design type: 6,561 (27.76%) in the IPDMA; 5,197 (21.99%) in the RCTs; and 11,720 (49.60%) in the cohort studies.

The pooled main effects were comparable for the different designs: RR = 0.86, 95% CI: 0.79, 0.94 (observational studies); RR = 0.86, 95% CI: 0.72, 1.00 (RCTs); and RR = 0.92, 95% CI: 0.83, 1.02 (IPDMA). See also Fig. 3.

The effect estimates of CABG (vs. PCI) in the group of nondiabetic patients were similar in the cohort studies, RCTs, and IPDMA, showing no effect (Fig. 4). However,
in the diabetic patient subgroup, the RCTs showed a protective effect, whereas the observational studies showed no effects. This remained after pooling the subgroup-specific effects. The ratio of the effects in nondiabetics compared with diabetics (Fig. 4) showed that performing CABG (vs. PCI) was more effective in preventing all-cause mortality in diabetics compared with nondiabetics: RRR = 1.40, 95% CI: 1.08, 1.81 (in IPDMA). The interaction effect based on RCTs (RRR = 1.34, 95% CI: 0.83, 2.17) was comparable with the IPDMA effect, albeit nonsignificant, whereas the interaction effect based on observational studies (RRR = 1.03, 95% CI: 0.84, 1.26) was smaller.

### 3.3. Effect of statin therapy by age

The Cholesterol Treatment Trialists’ Collaborators [24] IPDMA studied the effect of statin therapy vs. placebo or an active comparison group on a composite of cardiovascular endpoints (n = 14 trials). The IPDMA explored numerous subgroups including a significant (P = 0.01) interaction by age (≤65/>65) on major coronary events. Because the screened RCTs and observational studies mostly reported subgroup-specific effects by age, here, we focus on this subgroup. The data of the IPDMA could be compared with six RCTs [30,53–57] (of which five were included in the IPDMA), and four observational studies (three cohort studies [29,58,59] and one case–control study [60]). The IPDMA consisted of 90,056 subjects, of whom 7,757 developed the outcome of interest. In the RCTs, 70,877 subjects were included, of whom 8,192 developed a major coronary event. The observational studies comprised 50,553 subjects, of whom 22,219 participated in cohort studies and 28,334 in the case–control study; 2,485 cases were included by these studies. The cohort study described by Poluzzi et al. did not report the number of cases for the primary prevention group in which they stratified for age. Heterogeneity in interventions, comparisons, and outcomes was large in the IPDMA, RCTs, and observational studies. For example, interventions differed in type and dosage of statins, age dichotomization...
ranged from 60 to 70 years, control groups ranged from placebo controlled to active comparison in RCTs, and from active comparison to adherence to therapy in the observational studies (see Appendix B at www.jclinepi.com).

The main effects (RRs) observed in the IPDMA, RCTs, and observational studies were 0.75 (95% CI: 0.72, 0.79), 0.79 (95% CI: 0.71, 0.89), and 0.65 (95% CI: 0.53, 0.78), respectively (Fig. 5).

When stratifying the results by age groups and pooling the individual studies, the estimates of the RCTs and observational studies were in concordance with the IPDMA (Fig. 6). The exception to this was the older subgroup in the observational study, in which the effect was smaller (but in the same direction) than the effect found in the IPDMA. The interaction effects (RRR) in young vs. older subjects were 0.90 (95% CI: 0.84, 0.97), 0.97 (95% CI: 0.84, 1.12), and 1.35 (95% CI: 1.13, 1.61) in IPDMA, RCTs, and observational studies, respectively.

4. Discussion

In the three clinical examples that we presented, main and subgroup-specific effects for observational studies were in agreement with those found in RCTs and IPDMA. This was not the case for interaction effects. In the mammography example, observational studies showed a significant interaction, whereas RCTs and the IPDMA did not. However, the interaction effect was in the same direction. In the other two examples, observational-based interaction effects showed either no effect or an effect in the opposite direction compared with RCTs and IPDMAs.

These results are in agreement with earlier studies which also found comparable main effect estimates between RCTs and observational studies [14–16]. The novelty of our study is that we compared subgroup-specific and interaction effects in IPDMAs, RCTs, and observational studies. We urge readers to be aware that similarity of effects between observational and RCT studies is topic specific and depends on the likelihood of measuring all important confounders. Because RCTs are not hampered by the potential of unmeasured confounding, they are typically preferred over observational studies to assess the effects of medical interventions. This research showed that, despite the potential of confounding by indication and inclusion of potentially different patient populations, main and subgroup-specific effects derived from observational data can, at least in our three examples, resemble IPDMA-based estimates.

Our study has several limitations that need to be addressed. An important limitation of our study is that we included only three clinical examples. Furthermore, although we tried to search systematically in the literature, we may have missed studies. Additionally, we concede that requiring our examples to comprise at least two observational studies and two RCT studies is arbitrarily chosen and increasing, this threshold would obviously decrease the number of example presented here. It seems highly unlikely, however, that these issues would lead to a bias that favors comparability of reported results.

Second, differences in confounding adjustment and other analytical discrepancies might have influenced our results. For example, all RCTs conducted an intention-to-treat (ITT) analysis, whereas most observational studies did not but conducted an as-treated analysis instead. In the mammography example, this may have led to a dilution of effects in RCTs compared with the observational studies [61,62]. The TEDBC [41] observational study analyzed their data based on screening vs. nonscreening center (an analysis more similar to ITT) and found no interaction effect, which is in line with IPDMA. Furthermore, for other examples, we were unable to extract adjusted subgroup-specific effects (either they were not presented adequately or not performed). For example, only the CAGB study by Malenka et al. [32] reported adjusted subgroup-specific effects. However, they only adjusted for “number of diseased coronary arteries.” This may have resulted in a somewhat biased subgroup-specific estimate, in which it seemed that diabetics, with a higher mortality risk (e.g., morbidity burden diabetics), were more likely to receive CAGB intervention. Similarly, in the statin example, only the cohort study by Poluzzi et al. [59] adjusted for confounding in subgroups, for instance by using a categorized age variable (<50, 50–65, 65–80, and >80 years). However, this does not sufficiently exclude residual confounding. In this case, lack of adjustment revealed a healthy user bias, in which healthy older subjects received, or complied the most with, the strictest drug therapy.

Third, apart from differences in analyses, factors such as duration of follow-up, comparison group treatment, and
outcome assessment are known reasons for discrepancies between studies. Our examples were also harmed by this; follow-up duration ranges were 8.8–18 years in the mammography example, 5.6–10.4 years in the CABG example, and 0.5–8 years in the statin example. Furthermore, although treatment and outcomes were very similar in the mammography and CABG examples (Appendix B at www.jclinepi.com), in the statin example, RCTs used placebo or active comparison groups, whereas observational studies used no or diminished treatment adherence as a comparator group.

Fourth, we concede that using IPDMA based on RCTs as a gold standard is not unattested. For example, RCTs are known to include relatively healthier patients and increase compliance toward unrealistic levels, which might be unattainably in clinical practice. Hence, estimates of treatment effects based on RCTs could overestimate the treatment effects observed in daily practice which consequently also results in differences in effect estimates.

Finally, a different issue is that exploring multiple subgroup-specific and interaction effects increases the type-I error rate. This results in confidence intervals that are smaller than 95% and therefore increase the likelihood of finding a false-positive result. The impact of multiple testing, however, is unlikely to differ between observational, RCT, and IPDMAs. Despite above described shortcomings we still found agreement for main and subgroup-specific effects across differently designed studies. However, it is possible that using more appropriate observational (IPD) data, some of these issues could be solved which in turn might increase the similarity between interaction effects based on observational and RCTs studies.

In conclusion, main and subgroup-specific effects based on reported observational data were similar in direction to those from IPDMAs. Interaction effects found in RCTs and IPDMAs were also similar. In two examples, observational-based interaction effects showed different direction of effects compared with RCTs and the IPDMA estimates. Similarity of main and subgroup-specific effects across designs therefore does not imply similarity of interaction effects.

Acknowledgments


Appendix

Supplementary material

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.jclinepi.2012.08.008.