Effect Modification of Interventions

Bridging the Gap Between Clinical Studies and the Individual Patient

A. F. Schmidt

Effect Modification of Interventions: Bridging the Gap Between Clinical Studies and the Individual Patient

ISBN 9789039361535

Cover: A.F.Schmidt (design), Rob de Voogd ZZAPBACK (foto).
Lay-out: A.F.Schmidt.
Printed by: Ipskamp Drukkers.

The studies in this thesis were funded by Research Focus Areas funding of the Utrecht University and was a collaboration between the faculties of medicine (Julius Center for Health Sciences and Primary Care), science (Utrecht Institute for Pharmaceutical Sciences), and veterinary medicine (Department of Farm Animal Health). Financial support by the Julius Center for Health Sciences and Primary Care and by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged. Additional financial support for the printing of this thesis was provided by Koninklijke Nederlandse Maatschappij ter bevordering der Pharmacie (KNMP), ChipSoft, Boehringer Ingelheim and GlaxoSmithKline.

Effect Modification of Interventions

Bridging the Gap Between Clinical Studies and the Individual Patient

Modificatie van effecten van interventies en het toepassen van onderzoeksresultaten
bij de behandeling van individuele patiënten
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Utrecht
op gezag van de rector magnificus
prof. dr. G.J. van der Zwaan,
ingevolge het besluit van
het college voor promoties
in het openbaar te verdedigen op

23 juni 2014 des middags
te 2.30 uur

door

Amand Floriaan Schmidt
geboren 22 oktober 1986
te Amsterdam

Promotoren:	Prof. dr. A.W. Hoes
	Prof. dr. M. Nielen
Co-promotoren	Dr. R.H.H. Groenwold
	Dr. O.H. Klungel

Voor Elke

Contents

Part I

General introduction

The safety and efficacy of medical interventions can be explored in various types of clinical studies. Randomized clinical trials (RCTs) are commonly used to determine intended treatment effects and can sometimes provide information on frequently occurring unintended effects (notably Type A adverse events). Nonrandomized studies (e.g., cohort or case-control studies) can provide information on the occurrence of unintended effects (i.e., type B adverse events), but also on intended effects(1;2). Clinical studies (randomized or nonrandomized) are typically designed to provide estimates of the average (intended or unintended) treatment effect. However, patients, health care professionals, regulators, and researchers recognize that treatment effects may not be constant across a wide range of potential users (3-7).

When treatment effects differ between subgroups of patients, this is often referred to as effect modification, interaction, or heterogeneity of treatment effects. For example, a recent RCT comparing 6 months to 12 months anti-platelet therapy on preventing cardiovascular endpoints (8), showed that in patients with diabetes the 6 months regime increased the risk of a cardiovascular event by 216%. However, in patients without diabetes the 6 months regime reduced the risk by 56%. Of the patients included in the trial 38% had diabetes, on average therefore, the 6 months regimen increased the risk by 49%. In this example, the effect of anti-platelet therapy differs between subgroups based on diabetes status, i.e., there is effect modification by diabetes.

Whether effect modification is present depends on the effect measure that is considered. For example, if the treatment effect expressed as a risk ratio is constant across subgroups, the risk differences will possibly differ between subgroups (9-12). Such “effect modification” is therefore also referred to as effect measure modification (9). Here, we consider mainly situations in which researchers choose a specific effect measure (e.g. odds ratio, risk ratio or risk difference) prior to analysing the study and thus consider relevant any effect modification of that particular effect measure.

In the presence of effect modification, treatment effects differ between subgroups of subjects, in which case average treatment effects are non-informative and cannot be generalized to populations of future users. For example, the average treatment effect observed in the aforementioned study of anti-platelet therapy neither applies to patients with diabetes nor to patients without. In this case, only estimates of treatment effects that are stratified for those

subgroups are relevant for future users.

When study results do not suggest any effect modification, the main treatment effect found in a study can more likely be generalized beyond the population included in the study (because there is no direct reason to believe the treatment may act differently in other subjects). However, generalizability of treatment effects also depends on similarities between the study population and future users with respect to e.g. availability of treatments, adherence, and possible unidentified effect modification. A detailed understanding of the biological effects by which the treatment acts can be of help when considering the latter.

Previous research focussed either on generalizability (3;13-15) or on effect modification (5;6;16-20). However, as was just discussed, these topics are very closely interlinked; treating these topics separately is at its best inefficient and at its worst prevents estimation of the true treatment effect in terms of actual patient benefit (or harm). As the example of antiplatelet therapy shows, ignoring effect modification may result in treating subjects in whom the treatment is ineffective (but may have adverse effects and puts a financial burden on healthcare) or not treating subjects in whom the treatment is effective. Thus, patients are suboptimally treated when effect modification is not recognized and appropriately taken into account.

This thesis will address the issue of effect modification of treatment effects with the ultimate goal to identify subgroups of patients who should ideally be treated (because the treatment is effective) and those patients for whom it would be better to refrain from treatment. First, methods to detect treatment effect modification will be evaluated. Second, we apply a method to optimize targeting of treatment based on multiple patient characteristics. Third, we will address the question of how to assess generalizability of study results. Finally, a framework to assess effect modification as well as generalizability is provided.

Thesis outline

The thesis outline is as follows. Methods to detect treatment effect modification are presented in part II. In chapter 1, a simulation study is used to evaluate frequently used interaction tests. In chapter 2, reported interaction effects from randomized studies are compared to interaction effects from nonrandomized studies. Subsequently, in chapter 3, a simulation study is presented evaluating a Bayesian approach to combine nonrandomized information and

randomized studies in an effort to increase the power of interaction tests. In part III, a method to identify multivariable subgroups is applied to an empirical example of canine osteosarcoma patients. In chapter 4 we describe a study to predict the probability of cancer recurrence or death in dogs with osteosarcoma. In chapter 5 this predicted individual probability is used to determine in which patients chemotherapy treatment is beneficial. In part IV of this thesis, generalizability is addressed. In chapter 6 a systematic review on justification of exclusion criteria is presented. This review explores how generalizability is affected by heterogeneity of patient populations. In chapter 7, we show that explicitly modelling treatment effect modification might increase comparability between results from randomized and nonrandomized studies. An approach to explore generalizability of treatment effect estimates within and between studies is described in chapter 8. In part V of the thesis a framework is presented to assess effect modification and generalizability.

Reference List

1. Vandenbroucke JP. When are observational studies as credible as randomised trials? Lancet 2004 May 22;363(9422):1728-31.
2. Vandenbroucke JP. What is the best evidence for determining harms of medical treatment? CMAJ 2006 Feb 28;174(5):645-6.
3. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005 Jan 1;365(9453):82-93.
4. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005 Jan 8;365(9454):176-86.
5. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005 Jan 8;365(9454):176-86.
6. Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. Lancet 2005 Jan 15;365(9455):256-65.
7. Bugeja G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. BMJ 1997 Oct 25;315(7115):1059.
8. Gwon HC, Hahn JY, Park KW, Song YB, Chae IH, Lim DS, et al. Six-month versus 12-month dual antiplatelet therapy after implantation of drug-eluting stents: the Efficacy of Xience/Promus Versus Cypher to Reduce Late Loss After Stenting (EXCELLENT) randomized, multicenter study. Circulation 2012 Jan 24;125(3):505-13.
9. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. Am J Epidemiol 1980 Oct;112(4):467-70.
10. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. Stat Med 1983 Apr;2(2):243-51.
11. White IR, Elbourne D. Assessing subgroup effects with binary data: can the use of different effect measures lead to different conclusions? BMC Med Res Methodol 2005;5:15.
12. Venekamp RP, Rovers MM, Hoes AW, Knol MJ. Subgroup analysis in randomized controlled trials appeared to be dependent on whether relative or absolute effect measures were used. J Clin Epidemiol 2014 Apr;67(4):410-5.
13. Rovers MM, Straatman H, Ingels K, van der Wilt GJ, van den Broek P., Zielhuis GA. Generalizability of trial results based on randomized versus nonrandomized allocation of OME infants to ventilation tubes or watchful waiting. J Clin Epidemiol 2001 Aug;54(8):789-94.
14. Van Spall HG, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. JAMA 2007 Mar 21;297(11):1233-40.
15. Dekkers OM, von EE, Algra A, Romijn JA, Vandenbroucke JP. How to assess the external validity of therapeutic trials: a conceptual approach. Int J Epidemiol 2010 Feb;39(1):89-94.
16. Dorresteijn JA, Visseren FL, Ridker PM, Wassink AM, Paynter NP, Steyerberg EW, et al. Estimating treatment effects for individual patients based on the results of randomised clinical trials. BMJ 2011;343:d5888.
17. Tanniou J, Tweel vd T, Teerenstra S, Kit CBR. Level of evidence for promising subgroup findings in an overall non-significant trial. Statistical Methods in Medical Research 2014.
18. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Medical Research Methodology 2006;6:18.
19. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. Journal of the American Medical Association 2007 Sep 12;298(10):1209-12.
20. Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 2010;11:85.

Part II

Detecting effect modification of interventions

CHAPTER 1

Exploring interaction effects in small samples increases rates of false-positive and false-negative findings: results from a systematic review and simulation study

A F Schmidt, R H H Groenwold, M J Knol, A W Hoes
M Nielen, K C B Roes, A de Boer, O H Klungel.

Journal of Clinical Epidemiology 2014
doi: 10.1016/j.jclinepi.2014.02.008.

Abstract

Objective To give a comprehensive comparison of the performance of commonly applied interaction tests.

Methods A literature review and simulation study was performed evaluating interaction tests on the odds ratio (OR) or the risk difference (RD) scales: Cochran's Q, Breslow-Day (BD), Tarone, unconditional Score, Likelihood Ratio, Wald, and RERI based tests.

Results Review results agreed with results from our simulation study. Which showed that on the OR scale, in small sample sizes (e.g. number of subjects ≤ 250) the type 1 error rates of the LR test was 0.10, the BD and Tarone tests showed results around 0.05. On the RD scale, the LR and RERI tests had error rates around 0.05. On both scales tests did not differ regarding power. When exposure prevented the outcome RERI based tests were relatively underpowered (e.g., N = 100, RERI power = 5% vs. Wald power =18%).With increasing sample size difference decreased.

Conclusions In small samples interaction tests differed. On the OR scale the Tarone and Breslow-Day tests are recommend. On the RD scale, the LR and RERI based tests performed best. However, RERI based tests are underpowered compared to other tests when exposure prevent the outcome and sample size is limited.

Background

When studying the effect of medical treatments physicians may wonder whether the effect differs between groups of patients. For example, the effects of aspirin in preventing myocardial infarctions may be different in men compared to women (1). To explore whether treatment effects indeed differ between subgroups of patients, one can stratify the study population according to the subgroup of interest. An interaction test can then be performed, which tests whether treatment interacts with certain patient characteristics (e.g. gender) and thus whether treatment effects indeed differs between subgroups (2;3).

Presence of interaction depends on the type of effect measure that quantifies the relation between treatment and outcome (4;5). For example, in case of a binary outcome (e.g., myocardial infarction) an interaction can be present on the odds ratio (multiplicative) scale but absent on the risk difference (additive) scale, or vice versa.

Previously, the performance of interaction tests was assessed using simulation studies (6-11). Most studies focused on interaction tests using odds ratios (ORs) and no single study compared all the commonly used interaction tests together in one scenario. We aimed to provide a comprehensive comparison of commonly applied test on the OR scale and the RD scale (specifically the Cochran's Q, Breslow-Day, Tarone, unconditional Score, Likelihood Ratio, and Wald test and tests based on the Relative Excess Risk due to Interaction or RERI). First, a systematic review was conducted providing an overview of previous simulations studies. Obviously, each simulation study used different simulation scenarios which could potentially explain any dissimilarity in performance between interaction tests. Therefore, in a second part we conducted a simulation study to compare all of the previously mentioned interaction tests under equal simulation conditions.

Methods

The review and subsequent simulation study evaluated the following asymptotic interaction tests: on the OR scale the Cochran's Q (Q), Breslow-Day (BD), Tarone, unconditional Score (Score), Likelihood Ratio (LR) and the Wald test were compared. For the RD scale we compared the Q, LR, and the Wald test and tests based on the RERI. To our knowledge no variance estimator is available for the BD, Tarone and Score tests using the RD scale, therefore these tests were not assed for the RD scale. Similarly, the RERI is specifically proposed

for estimating interaction on a RD scale using risk ratios (RR) and, therefore, was only evaluated on the RD scale. For the formulae of these interaction tests we refer to **Appendix I**. In both the review and the subsequent simulation study we focused on sparse data scenarios because asymptotic tests differ in such settings. In small sample sizes power is often limited therefore, while exploring both power and type 1 error rates we focus on the latter.

Systematic review

Using the following search terms in title or abstract, Medline was searched (date: 24-05-2013):

(homogeneity OR modification OR interaction OR synergism OR antagonism) AND (simulation OR "monte carlo") AND (effect OR test OR statistic OR power OR significance)

Papers were screened and included when they [1] presented results from a simulation study, [2] assessed the performance of the previously mentioned interaction tests for dichotomous outcomes, and [3] were published in English. This was supplemented with a Scopus (12) based cross-reference search.

Simulation study

A simulation study was performed to assess the statistical performance of the previously mentioned interaction tests. Most evaluated tests are only applicable to categorical data and therefore all simulations were based on scenarios with two dichotomous exposure (i.e., *X* and *S*) and a dichotomous outcome. In such settings subjects can be in one of four possible exposure categories, indicated by *i* = 0 or 1 if exposure to *X* is absent or present and *j* = 0 or 1 depending on the absence or presence of exposure *S*. The corresponding outcome probabilities are indicated by P_{ij}. Initially, six scenarios (A-F, see **Table 1**) were created and each scenario was studied using different sample sizes (i.e., the number of observations (*N*) was set to 50, 100, 250, 500, 1,000 or 2000 in different simulations). The number of observations was equally distributed over the 4 exposure categories by setting the prevalence (F_{ij}) of every exposure type to 0.25 (i.e., F_{ij} represent the fraction that every exposure type (ij = 00, 11, 10 or 01) contributes to N; hence F_{ij} sums to 1). The expected number of events and non-events can then be calculated: *number of event* $= P_{ij} * N * F_{ij}$ *; number of non-events* $= (1 - P_{ij}) * N * F_{ij}$. In **Table 2** a numerical example of scenarios B is given. This example

shows how the different parameters impact expected cell counts and expected interaction effects. In order to assess the impact of the fraction F_{ij} on interaction test performance the prevalence of the combined exposure group (F_{11}) was varied from 0.05 to 0.25, where the complement (i.e., 0.25 - F_{11}) was equally divided over the remaining F_{ij}.

Table 1 Six simulations scenarios used to evaluated interaction test on the Risk Difference and Odds Ratios scale*.

Scenario *(tests used)*	P_{00}	P_{01}	P_{10}	P_{11}	***RD interaction***	***OR interaction***
A *(RD.tests)*	*0.30*	*0.30*	*0.30*	*0.50*	*0.20*	*2.33*
B *(RD.tests)*	*0.30*	*0.30*	*0.30*	*0.10*	*-0.20*	*0.26*
C *(OR tests)*	*0.25*	*0.25*	*0.25*	*0.40*	*0.15*	*2.00*
D *(OR tests)*	*0.25*	*0.25*	*0.25*	*0.14*	*-0.11*	*0.49*
E *(OR and RD tests)*	*0.20*	*0.20*	*0.20*	*0.20*	*0.00*	*1.00*
F *(OR and RD tests)*	*0.80*	*0.80*	*0.80*	*0.80*	*0.00*	*1.00*

* P_{00} = Risk in unexposed, P_{10} = Risk among subjects exposed to factor X, P_{01} = Risk among subjects expose to factor S, P_{11} = Risk in subjects exposed to both factors, **RD interaction**= interaction magnitude on the risk difference scale (0 = no interaction), **OR interaction** = interaction magnitude on the odds ratio scale (1 = no interaction). Note that Risk Difference (RD) interaction tests were only applied to A, B , E and F, similarly Odds Ratio (OR) interaction tests were applied to the C, D, E and F scenarios.

The above scenarios were based on the P_{ij} probabilities and interaction effects given in **Table 1**. To further explore empirical power under different interaction effects, P_{ij} was set to 0.5 except for the event probability of the combined exposure group (P_{11}) which varied between 0.05 to 0.95. This was done for the setting in which F_{ij} = 0.25 and *N* equalled 100, 250, 500 or 1,000. In two final simulations (where F_{ij} = 0.25 and *N* equalled 100, 250, 500 or 1,000) P_{11} was set to 0.05 or 0.95, in each case P_{00} ranged from 0.05 to 0.95. These two scenarios can be viewed as more extreme versions of the previous scenarios where the interaction effects are allowed to be larger and data more sparse.

In scenarios E and F (**Table 1**) type 1 error rates *(i.e., the probability to incorrectly reject the null hypothesis)*, using an alpha of 0.05, were determined. Power *(i.e., the probability to detect an effect when it is present)* was evaluated for the RD scale in scenarios A and B and in scenarios C and D for the OR scale. To more closely compare the RERI to the other RD interact tests the root mean squared error (RMSE) was calculated for the RERI and the Wald tests. RMSE is the squared root of the sum of the squared bias and the variance, measuring both bias and variance on the original scale of the measurements. Bias is defined as the difference between the estimated interaction effect and the true interaction effect. Variance is the variance of the estimated interaction effect over all simulation replications. Customarily, these measures of the performance are combined in the mean squared error (MSE) as the squared bias plus the variance. The squared root of the MSE (the RMSE) is on same

scale as the interaction effect and therefore easier to interpret. Each scenario was based on four draws (one for every type of exposure) from a binomial distribution with P_{ij} as the event probabilities and $N^* F_{ij}$ as the number of subjects per exposure type. All simulations were replicated 10,000 resulting in a 95 percent confidence interval (95%CI) width of approximately 0.009 (i.e., lower and upper bounds: 0.046; 0.054) around a type 1 error rate of 0.05. The standard error of the empirical power was at most $(0.5 * 0.5 / 10{,}000)^{1/2} = 0.005$, resulting in a maximum 95% CI width of 0.020.

In the generated scenarios the previously mentioned interaction test were compared. The LR test was applied as a model comparison test: comparing a model with a product term to

Table 2 The expected cell counts and interaction effects for scenario B with exposure prevalence set to 0.25 for each exposure type and a sample size of 1,000.

Simulation parameters	*Total sample size*	*Event probabilities*	*Exposure prevalence*
	N = 1000	P_{ij} *= 0.30, 0.30, 0.30, 0.10*	F_{ij} *= 0.25, 0.25, 0.25, 0.25*

X exposure status	***S exposure status***	***Events***	***Non-events***
1	*1*	$25 = 0.10 * \frac{1000}{1/0.25}$	$225 = (1 - 0.10) * \frac{1000}{1/0.25}$
1	*0*	$75 = 0.30 * \frac{1000}{1/0.25}$	$175 = (1 - 0.30) * \frac{500}{1/0.25}$
0	*1*	$75 = 0.30 * \frac{1000}{1/0.25}$	$175 = (1 - 0.30) * \frac{500}{1/0.25}$
0	*0*	$75 = 0.30 * \frac{1000}{1/0.25}$	$175 = (1 - 0.30) * \frac{500}{1/0.25}$

$$OR_{00} = \frac{75/175}{75/175};\ OR_{10} = \frac{75/175}{75/175};\ OR_{01} = \frac{75/175}{75/175};\ OR_{11} = \frac{25/225}{75/175}; interaction\ OR = \frac{OR_{11}}{OR_{10} * OR_{01}}$$

$$RD_{00} = \frac{75}{75+175} - \frac{75}{75+175};\ RD_{10} = \frac{75}{75+175} - \frac{75}{75+175};\ RD_{01} = \frac{75}{75+175} - \frac{75}{75+175};\ RD_{11} = \frac{25}{75+225} - \frac{75}{75+175}; interaction\ RD = RD_{11} - RD_{01} - RD_{10}$$

a model without a product term. On the OR scale, logistic models were compared. On the RD scale, the LR test was based on Poisson models with an identity link and robust variance estimators, specifically the Heteroscedasticity-Consistent covariance estimator 0, i.e., HC0 (13). Two RERI based tests were evaluated, the first using the delta ($RERI_{delta}$) estimator (14)

and the second the bootstrap percentile (15) variance estimator ($RERI_{bs}$). For the $RERI_{delta}$ test a Poisson model with log link and robust variance estimators was used, the $RERI_{bs}$ test was estimated from the crude 2 by 2 by 2 table and used 1,000 bootstraps.

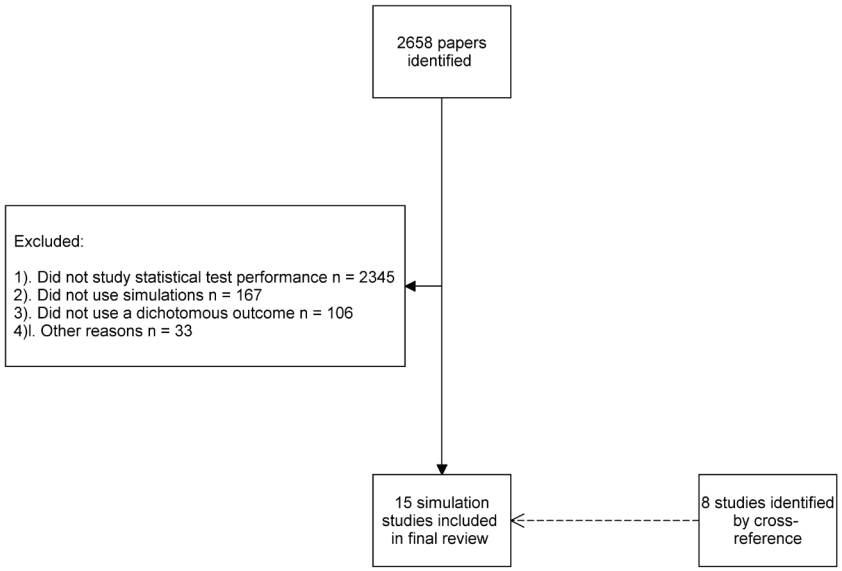
Computations were performed using the R statistical software package version 3.0.0 (16); R code is available upon request.

Results

Systematic review

We identified 15 studies that evaluated the performance of interaction tests (**Figure 1**). Of these, 6 focused on interaction tests on the RD scale (14;16-20), 7 focused on interaction tests on the OR scale (6;7;9-11;21;22) and 2 addressed both (8;23). The results of these studies are summarized in **Appendix Table 1**.

Figure 1 Flow of studies in the Medline search for simulation studies on the performance of interaction test.



The 9 studies that studied the OR scale, explored a large number of scenarios, ranging from scenarios with expected cell counts of 1 to scenarios with expected cell counts of 55 or more (9;10). Studies that explored the LR tests showed that in most scenarios its type 1 error rate was larger than 0.05. In the same scenarios the Tarone, BD and score tests showed error rates closer to 0.05 (6;7;9;10). In extreme settings with expected cell counts of 1-5 the LR

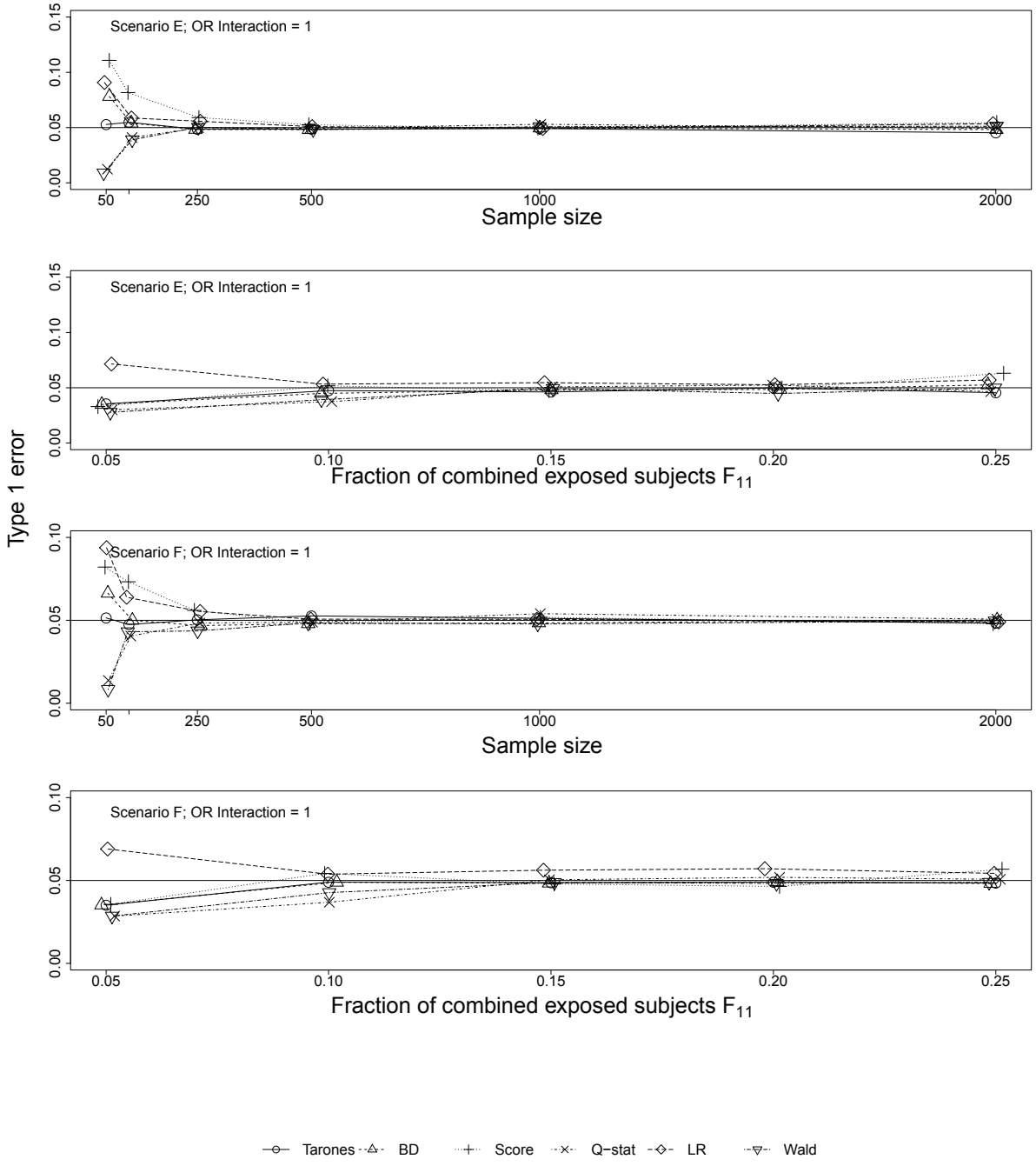
type 1 error rate could be as high as 0.97 while in the same settings the BD and Tarone tests had error rates of 0.44. In three other studies (21-23) the LR had type 1 error rates < 0.05, however these only explored large sample situations. The Wald and Q statistics also showed type 1 error rates close to 0.05 in relatively non-sparse data settings (e.g.,75 cases and 150 controls) (8;21). Generally, power did not differ much between the tests studied, differing usually not more than 5%, except in very small sample sizes.

The 8 studies that evaluated RD interaction tests performance used a great number of scenarios, similar to the OR scenarios, including scenarios with expected cell counts of 1 to simulations with 1,000 cases and controls (18;23). Two studies explored the LR tests performance for the RD scale and showed type 1 error rates close to 0.05 (16) or lower (23), the first study used a small amount of replications (maximum 600) (16) the latter used 5000 repetitions with at least 500 subjects (23). The Q tests on the RD scale was evaluated by three studies (18-20). All three studies showed that in sparse data the type 1 error rate was often higher than 0.05, however generally not larger than 0.06. In scenarios with expected cell counts of 1 and a large number of subgroups the type 1 error rate was seriously inflated: 0.60 (19). A single study explored the Wald test for the RD scale and showed type 1 error rates below 0.05 (8). RERI based interaction tests were evaluated by three studies which showed that the type 1 error rate (or the 95% confidence interval coverage) was below 0.05 (or the coverage rate was above the 95%), for example a type 1 error rate of 0.025 or a coverage rate of 97% (14;16;17). All of the simulations studies that evaluated the RERI based tests used the OR as an approximation of the RR. (14;17). Power did not differ much between RD interaction tests and was mostly driven by sparseness of the data and interaction magnitude.

Simulation study

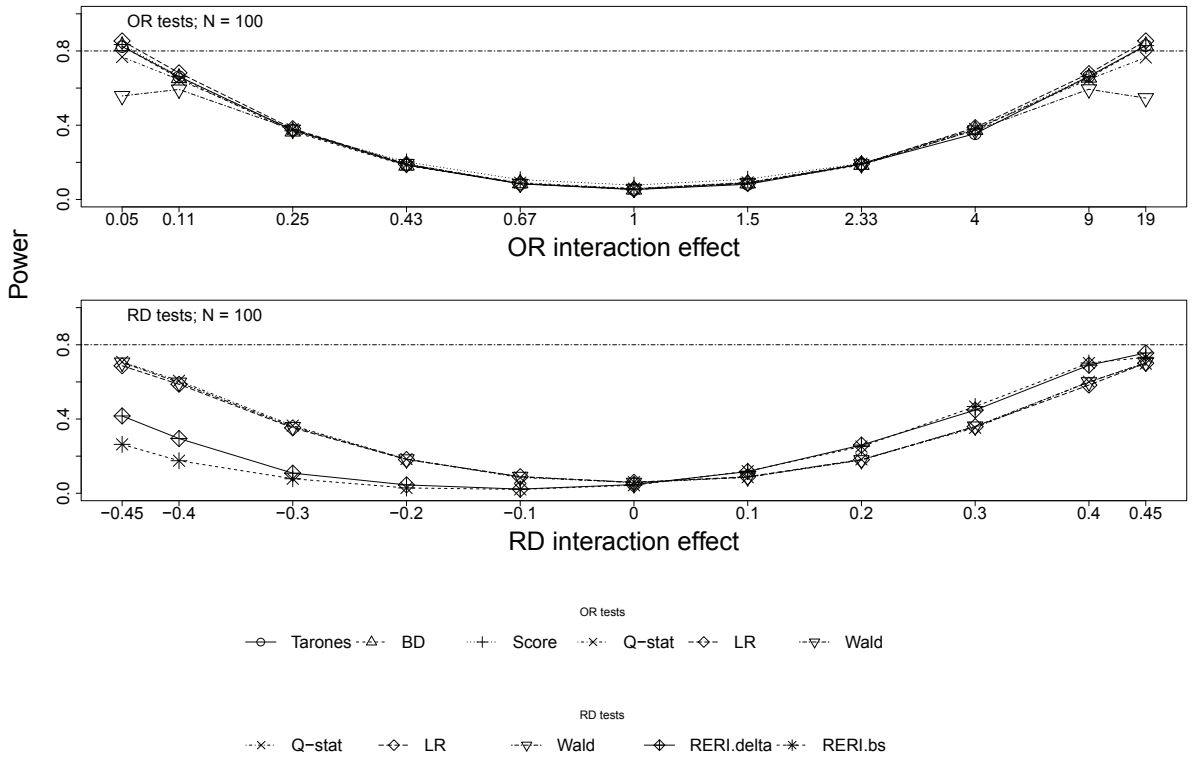
Simulations on the OR scale showed that in small sample sizes (N ≤ 250 or when one of the exposure groups contributed ≤ 0.10 to the overall N) the Score and LR test had error rates above 0.05, sometimes as high as 0.10 (**Figure 2**). The Q and Wald tests displayed type 1 error rates below 0.05, while the BD and Tarone tests had type 1 error rates closest to 0.05. Power did not markedly differ between tests (**Appendix III, Figure a**), for example the maximum difference in power for scenario C using 50 subjects was 0.07. **Figure 3** shows that in samples of 100 subjects a high power, for example 80%, was only reached when the interaction effects were very large; interaction OR of 0.05 or 19. Note that the symmetry around the

Figure 2 Type 1 error rates of odds ratio (OR) interaction tests evaluated in simulation scenarios E en F



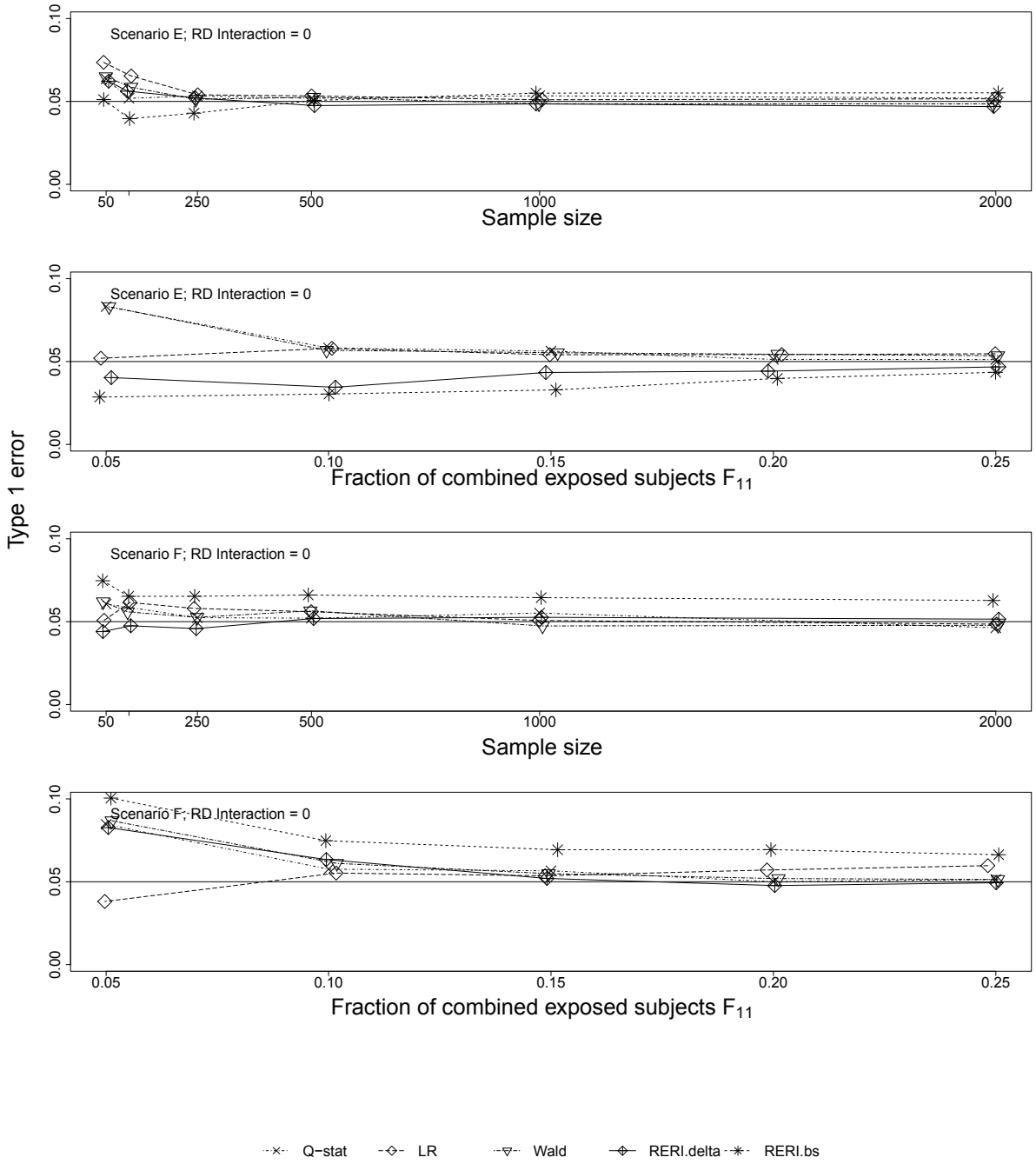
N.B. in the upper parts of scenarios E and F the sample size was increased from 50 observations to 2,000. In the bottom part sample size was fixed at 250 observations and the fraction of F11 was increased from 0.05 to 0.25. Thus the relative number of subjects that were exposed to both factors increased from 5 percent to 25 percent of the total sample size of 250. All simulations were repeated 10,000 times.

Figure 3 Power of odd ratio (OR) and risk difference (RD) based interaction tests under different interaction effect sizes



N.B. the upper panel depicts power of the interaction tests evaluated using different interaction effect sizes, the bottom panel shows RD interaction tests performance. All simulations were carried out using a sample size of 100 subjects which was equally divided over the exposure categories (i.e., F11 = 0.25) and repeated 10,000 times.

Figure 4 Type 1 error rates of risk difference (RD) interaction tests evaluated in simulation scenarios E en F



N.B. in the upper parts of scenarios E and F the sample size was increased from 50 observations to 2,000. In the bottom part sample size was fixed at 250 observations and the fraction of F11 was increased from 0.05 to 0.25. Thus the relative number of subjects that were exposed to both factors increased from 5 percent to 25 percent of the total sample size of 250. All simulations were repeated 10,000 times.

1 is expected because the interaction effects in positive and negative settings only differ regarding their sign, e.g., 1/19 = 0.0526. Due to empty cells counts in scenarios of 50 subjects the Q tests did not converge in 25% of the cases and the BD, Tarone and Score tests failed up to 10% of the replications. In these settings the Wald and LR tests did converge because these were implemented using generalized linear models.

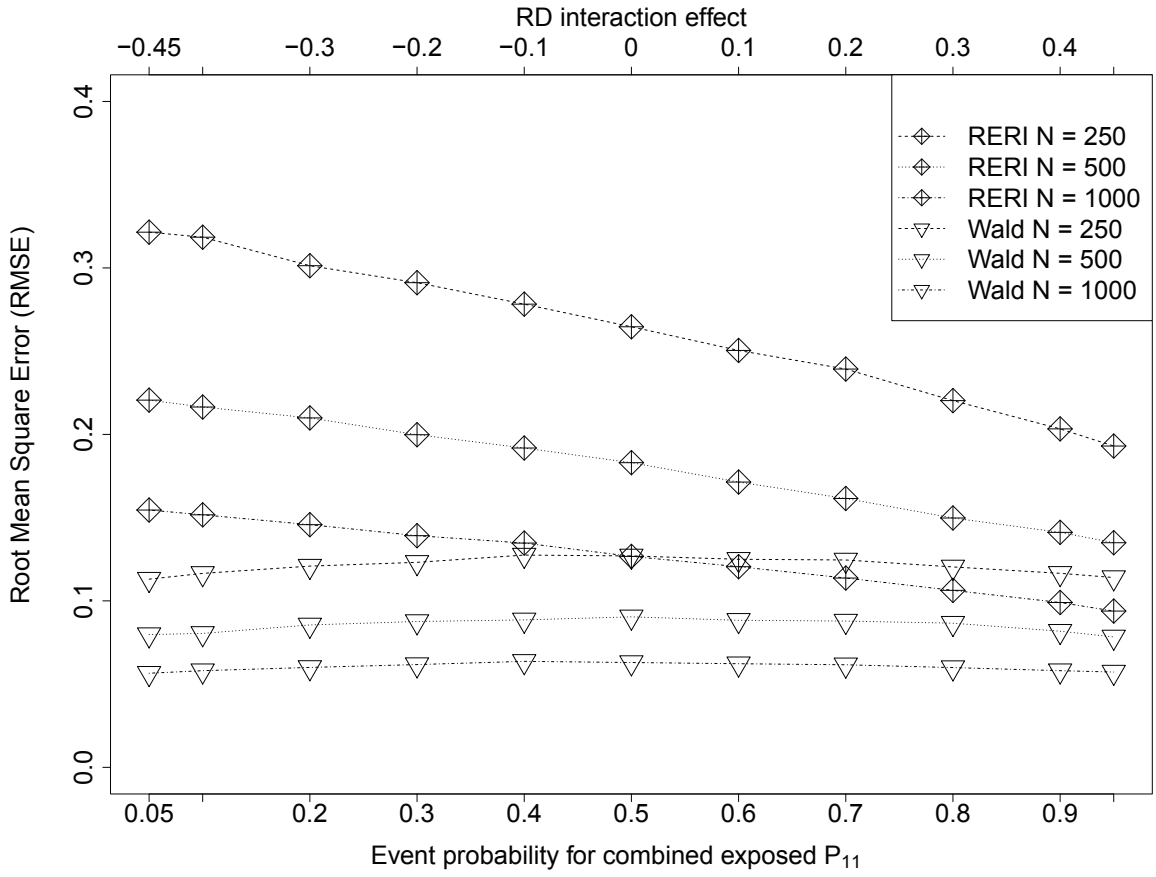
Simulations on the RD scale showed that in sample sizes of 250 subjects or less (or when the prevalence of combined exposure subjects was ≤ 0.15) the type 1 error rates were typically above 0.05 (**Figure 4**). In such settings the LR and the $RERI_{delta}$ tests had type 1 error rates closest to 0.05, but in some scenarios also showed type 1 error rates above 0.05. Results of scenarios A and B showed that, excluding the RERI based tests, there was little difference in power between the RD tests (**Appendix III, Figure b)**. **Figure 2** shows that the RERI based tests were relatively) underpowered (compared to the other interaction tests on the RD scale) when the interaction effect was negative. For example, when the interaction RD = -0.3 the power of the Wald test = 0.36 and the power of the RERI.bs test = 0.11. This difference decreased as sample size increased to 1,000 (**Appendix III, Figures b and c**). Due to empty cells counts in scenarios of 50 subjects the LR tests failed up to 15% of the times, the other RD tests failed in less than 5% of the replications.

In negative interaction settings, the RERI tests had higher root mean squared error (RMSE) values than the Wald test (**Figure 5**). For example when the interaction effect was 0.00 and N = 250 RMSE was considerably higher for the RERI.delta test (0.27) than for the Wald test (0.13). Again, asymptotically the difference in RMSE between the RERI and the Wald test minimalized (**Figure 5**). Similar results (regarding RMSE and power in positive and negative settings) were observed in simulations where P_{11} was fixed at 0.05 or 0.95 and instead P_{00} was iterated from 0.05 to 0.95 (data available upon request).

Discussion

Our simulation study showed that in small sample sizes (on the OR scale) the Breslow-Day (BD) and Tarone's test had type 1 error rates closest to 0.05 while the Likelihood Ratio (LR) test had type 1 error rates as high as 0.10. On the RD scale simulation results revealed that RD interaction tests frequently had type 1 error rates > 0.05. Of all the RD tests evaluated the LR and $RERI_{delta}$ tests had type 1 error rate closest to 0.05.

Figure 5 Root Mean Squared Error of the RERI.delta and Wald.Z tests under different interaction effect sizes.



N.B. Simulations were repeated with sample sizes of 250, 500 and 1,000 subjects and with different prevalence's of combined exposed subjects. Each simulations was replicated 10,000 times.

Additionally, our simulations showed that the RERI based tests were relatively underpowered (as compared to the other RD interaction tests) in the presence of negative interaction effects, this difference decreased as sample size increased to 1,000.

Results of our simulation study were generally supported by the review results. For example both the review and simulation study showed that on the OR scale the BD and Tarone tests had type 1 error rates closest to 0.05. However, there are also differences between the review and simulations results. The most important difference is that our simulation showed a relative lack of power for the RERI based tests, as compared to the other RD tests, in negative interaction settings with less than 1,000 subjects. Note that the RERI uses relative measures, such as the RR, to calculate the RD interaction. Studies included in the review did not show such results. This difference can be explained by the scenarios that were considered. In our simulations negative interaction settings were based on scenarios where compared to the unexposed, exposure prevented the outcome; i.e., taking the unexposed group as the

reference the exposure effects was RR < 1. The studies included in the review never used scenarios where (combined) exposure was protective. Instead these studies created negative interaction scenarios by setting $RR_{11} < RR_{10}$ and/or RR_{01}, but at the same time ensuring that all RRs > 1. Previously, it has been recognized that the RERI should only be used when all types of exposure increase the outcome risk (58). To achieve this it is recommended to recode the exposure so that the reference category is always the group with the smallest incidence (59). We showed that given sufficient sample size the RERI based tests performed similar to the other RD interaction tests even when exposure prevents the outcome. Note, that in case-control studies the RERI is often the only test available to explore RD interaction, making recoding an important consideration.

Several limitations and strengths warrant discussion. First, although we searched systematically, we concede that we may have missed studies. Our review and simulation study showed comparable results. Therefore, it seems unlikely that the overall conclusions would materially change by including additional studies. Second, we recognize that instead of using asymptotic tests, exact tests perform better with respect to type I error control. However, power of exact tests may be lower – in a situation where by design there usually already is a lack of power. Third, we concede that in this study the number of simulations scenarios was limited and that it would be interesting to explore scenarios with larger disbalance between subgroups and different event probabilities. Furthermore, we simulated scenarios with expected cell counts as low as 1.25 which might seem unrealistic. However, when adjusting for multiple (co)variables these scenarios might occur more often than initially expected, making these simulations potentially very relevant. Fourth, some might question the use of Poisson or binomial models because these are known to provide impossible estimates (e.g., probabilities outside the range 0-1) or to not converge at all (60). Obviously, as with all models, it is advisable to check the plausibility of derived estimates. In settings where estimates are (expected to be) implausible, methods based on calculating the *"marginal probabilities of success"* using logistic regression models might be preferable (61-63). Fifth, in spare data settings (e.g., N = 50) some tests failed. Often, tests based on generalized linear models (glms) did converge and tests based on 2 by 2 by 2 tables failed, due to empty cells. For comparison the Wald test was estimated based on glms as well as 2 by 2 by 2 tables, where the latter did indeed fail more often. Despite this increase in failed tests, results were equal up to two decimals (results not shown). Thus it seems unlikely that differences in failure rate

can explain our results. Obviously, performing interaction testing (or any testing at all) is problematic when empty cells exist and researchers should generally reconsider testing in such settings. On the other hand empty or sparse cell counts could also be due to large (interaction) effects which are important to report. Finally, an often heard comment on interaction testing in small sample sizes is that "when a significant result is found, despite low power, this interaction effect must therefore be present". However, this study showed that, depending on the test used and to some extend the scale chosen, the type 1 error rate can be high thus invalidating the previous comment.

Based on our results we recommend the following. First, when sufficient sample size is available (e.g., 500-1000 subjects) all interaction test perform similarly, hence the choice of interaction tests is irrelevant here. Second, in smaller sample sizes power is limited (unless a large interaction effect is present) and type 1 error rates are high, hence exploring interactions in such settings might not be appropriate. Furthermore, when deciding whether sample size is sufficient, researchers should also consider the distribution across exposure categories and across other potentially relevant (confounding) variables. Fourth, in sparse data settings the Tarone and Breslow-Day tests on the OR scale and the LR or $RERI_{delta}$ tests on the RD scale should be preferred because these tests have type 1 error rates closest to 0.05. Finally, users of the RERI based tests should be aware of its behavior when exposure is protective and should consider recoding the statistic or use one of the other RD tests.

Conclusions

In small sample sizes (e.g., N < 1000) the Tarone and Breslow-Day tests are preferred when assessing interaction on the odds ratio scale. On the risk difference scale the Likelihood Ratio and $RERI_{delta}$ are the preferred tests for interaction. However, when exposure is preventative for the outcome, RERI based tests are relatively underpowered compared to other interaction tests unless sample size is large. Recoding the exposure so that the RERI interaction effect becomes positive will resolve this problem.

Abbreviations

Cochrane'Q (Q).

Likelihood Ratios (LR).

Breslow-Day (BD).

Relative Excess Risk due to Interaction (RERI).

RERI tests using a variance estimate based on the delta method ($RERI_{delta}$).

RERI tests using a variance estimate based on the bootstrap percentile ($RERI_{bs}$).

Odd Ratio (OR).

Risk Difference (RD).

Risk Ratio (RR).

Reference List

1. Berger JS, Roncaglioni MC, Avanzini F, Pangrazzi I, Tognoni G, Brown DL. Aspirin for the primary prevention of cardiovascular events in women and men: a sex-specific meta-analysis of randomized controlled trials. JAMA 2006 Jan 18;295(3):306-13.
2. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol 2010 Aug;63(8):e1-37.
3. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. BMJ 2003 Jan 25;326(7382):219.
4. Rothman KJ. Epidemiology: An Introduction. New York: Oxford University Press; 2002.
5. White IR, Elbourne D. Assessing subgroup effects with binary data: can the use of different effect measures lead to different conclusions? BMC Med Res Methodol 2005;5:15.
6. Paul SR, Donner A. Small sample performance of tests of homogeneity of odds ratios in K 2 x 2 tables. Stat Med 1992 Jan 30;11(2):159-65.
7. Paul SR, Donner A. A comparison of tests of homogeneity of odds ratios in K 2 x 2 tables. Stat Med 1989 Dec;8(12):1455-68.
8. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. Stat Med 1983 Apr;2(2):243-51.
9. Jones MP, O'Gorman TW, Lemke JH, Woolson RF. A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. Biometrics 1989 Mar;45(1):171-81.
10. O'Gorman TW, Woolson RF, Jones MP, Lemke JH. Statistical analysis of K 2 x 2 tables: a comparative study of estimators/test statistics for association and homogeneity. Environ Health Perspect 1990 Jul;87:103-7.
11. Reis IM, Hirji KF, Afifi AA. Exact and asymptotic tests for homogeneity in several 2 x 2 tables. Stat Med 1999 Apr 30;18(8):893-906.
12. Scopus www.scopus.com. 12-9-2011.
13. Cribari-Neto F, Silva W. A new heteroskedasticity-consistent covariance matrix estimator for the linear regression model. AStA Advances in Statistical Analysis 2011;95(2):129-46.
14. Assmann SF, Hosmer DW, Lemeshow S, Mundt KA. Confidence intervals for measures of interaction. Epidemiology 1996 May;7(3):286-90.
15. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2009.
16. Hogan MD, Kupper LL, Most BM, Haseman JK. Alternatives to Rothman's approach for assessing synergism (or antagonism) in cohort studies. Am J Epidemiol 1978 Jul;108(1):60-7.
17. Zou GY. On the estimation of additive interaction by use of the four-by-two table and beyond. Am J Epidemiol 2008 Jul 15;168(2):212-24.
18. Lipsitz SR, Dear KB, Laird NM, Molenberghs G. Tests for homogeneity of the risk difference when data are sparse. Biometrics 1998 Mar;54(1):148-60.
19. Lui KJ, Kelly C. A revisit on tests for homogeneity of the risk difference. Biometrics 2000 Mar;56(1):309-15.
20. Zhang L, Yang H, Cho I. Test homogeneity of risk difference across subgroups in clinical trials. J Biopharm Stat 2009 Jan 7;(1520-5711 (Electronic)).
21. Takkouche B, Cadarso-Suarez C, Spiegelman D. Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. Am J Epidemiol 1999 Jul 15;150(2):206-15.
22. Bagheri Z, Ayatollahi SM, Jafari P. Comparison of three tests of homogeneity of odds ratios in multicenter trials with unequal sample sizes within and among centers. BMC Med Res Methodol 2011;11:58.
23. Starr JR, McKnight B. Assessing interaction in case-control studies: type I errors when using both additive and multiplicative scales. Epidemiology 2004 Jul;15(4):422-7.

24. Knol MJ, VanderWeele TJ, Groenwold RH, Klungel OH, Rovers MM, Grobbee DE. Estimating measures of interaction on an additive scale for preventive exposures. Eur J Epidemiol 2011 Jun;26(6):433-8.
25. Localio AR, Margolis DJ, Berlin JA. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. J Clin Epidemiol 2007 Sep;60(9):874-82.
26. Austin PC. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. J Clin Epidemiol 2010 Jan;63(1):2-6.
27. Austin PC, Laupacis A. A tutorial on methods to estimating clinically and policy-meaningful measures of treatment effects in prospective observational studies: a review. Int J Biostat 2011;7(1):6.

Appendix I formulae of interaction statistics.

General definitions

Let there be k = 1,2, …,n subjects. For all n subjects the outcome y is either present $y_k = 1$ or absent $y_k = 0$. These n subjects are exposed i = 1 or unexposed i = 0 to X. Similarly, these same n subjects can be exposed j = 1 or unexposed j = 0 to factor S. The number of exposure categories does not necessarily have to be the same for X and S, however in the following we assume both have two categories (indicated by 0 or 1). The highest exposure category is indicated by L_i for exposure X and L_j for exposure S. Ignoring the possibility of interaction with X the main (or common) exposure effect of S equals $\hat{\varphi}$, which can be the ln(odds ratio) (ln(OR)), ln(risk ratio) (ln(RR)) or risk difference (RD), where $\hat{\varphi}$ can be the maximum likelihood estimator (ML) or the Mantel-Haenszel estimator (MH). Stratifying $\hat{\varphi}$ for X results in subgroup-specific effects of exposure S, which are represented by $\hat{\beta}_i$; $\hat{\beta}_0 = \text{the exposure effect of S in subgroup X} = 0$; $\hat{\beta}_1 = \text{the exposure effect of S in subgroup X} = 1$. The difference between these two subgroup effects is measured by the interaction effect $\hat{\beta}_{int} = \hat{\beta}_{i=1} - \hat{\beta}_{i=0}$. Estimates of the variation of the above defined effect estimates are abbreviated with var e.g., $\text{var}(\hat{\beta}_{i=0})$. The estimated standard error of an estimated effects is indicated by se. Note that while in this appendix S is treated as the main exposure of interest, without any loss of generalizability both X and S can be of equal importance or alternatively X could be the exposure of interest with S as subgroup indicator.

Wald (1)

$$Wald = \hat{\beta}_{ML,int} / SE(\hat{\beta}_{ML,int})$$

Where $se(\hat{\beta}_{ML,int}) = \sqrt{var(\hat{\beta}_{ML,i=0}) + var(\hat{\beta}_{ML,i=1})}$. Under the null hypothesis the Wald statistic follows a Z-distribution.

Likelihood ratio test (LR) (2)

$$LR = -2 * (LL_{m=0} - LL_{m=1})$$

Here the LL indicates log-likelihood from a model (m) with (1) or without (0) a product term for the S by X interaction. LL is calculated as $LL = y_k * \ln(\hat{p}_k) + (1 - y_k) * \ln(1 - \hat{p}_k)$ with y_k representing an individuals' outcome and p_k the estimated outcome probability for individual k. Under the null hypothesis the LR statistic follows a central $\chi^2 - distribution.$

Breslow-Day test (3)

$$BD = \sum_{i=0}^{L_i} \frac{\left(c_i - E\left[c_i | e^{\hat{\varphi}_{mh}}\right]\right)^2}{var(c_i | e^{\hat{\varphi}_{mh}})}$$

Here c_i represents the number of observed subjects exposed to S that experienced an event in subgroup i, with i =0 if unexposed to X and i = 1 if exposed to X. The expected number of exposed cases under the assumption of no interaction equals $E\left[c_i | e^{\hat{\varphi}_{mh}}\right]$. Where $e^{\hat{\varphi}_{mh}}$ represent the OR. $E\left[c_i | e^{\hat{\varphi}_{mh}}\right]$ is obtained by solving the quadratic equation $E[c_i](h_i - t_i + E[c_i]) = e^{\hat{\varphi}_{mh}}(g_i - E[c_i])(t_i - E[c_i])$, where $E[c_i]$ is the expected frequency of c_i, h_i the observed number of total non-cases in subgroup i, t_i the number of observed exposed subjects in subgroup i, g_i the number of observed cases in subgroup i and $\hat{\varphi}_{mh}$ is the Mantel-Haenszel estimator of the ln OR. The variance estimator $var(c_i | e^{\hat{\varphi}_{mh}})$ is obtained by $var(c_i | e^{\hat{\varphi}_{mh}}) = \left[\frac{1}{E[c_i]} + \frac{1}{g_i - E[c_i]} + \frac{1}{h_i - E[c_i]} + \frac{1}{h_i - t_i + E[c_i]}\right]^{-1}$. Under the null hypothesis the BD statistic follows a central $\chi^2 - distribution.$

Tarone's test (3)

$$Tarone = BD - \frac{\left(\sum_{i=0}^{L_i} c_i - \sum_{i=0}^{L_i} E\left[c_i | e^{\widehat{\varphi}_{mh}}\right]\right)^2}{var(c_i | e^{\widehat{\varphi}_{mh}})}$$

All notations are the same as the notation described under the BD test. Under the null hypothesis the Tarone statistic follows a central $\chi^2 - distribution.$

Score statistic (4)

$$S = \sum_{i=0}^{L_i} \frac{\left(c_i - E\left[c_i | e^{\widehat{\varphi}_{ML}}\right]\right)^2}{var(c_i | e^{\widehat{\varphi}_{ML}})}$$

The Score statistic is similar to the BD statistic, but instead of using the Mantel-Haenszel estimator of the OR, the maximum likelihood estimator is used. Under the null hypothesis the Score statistic follows a central $\chi^2 - distribution.$

Cochran's Q (5)

$$Q = \sum_{i=0}^{L_i} w_i * \left(\hat{\beta}_{ML,i} - \hat{\varphi}_{ML}\right)^2$$

Here $w_i = 1/se\left(\hat{\beta}_{ML,i}\right)^2$ and the common effect estimate $\hat{\varphi} = \frac{\sum_{i=0}^{L_i} w_i * \hat{\beta}_{ML,i}}{\sum_{i=0}^{L_i} w_i}$. Depending on the outcome measure chosen, $\hat{\varphi}$ represents the ln(OR) or the RD of exposure S. Under the null hypothesis the Q statistic follows a central $\chi^2 - distribution.$

RERI based test (6;7)

$$RERI.test = \frac{\widehat{RERI}}{se(\widehat{RERI})}$$

The $\widehat{RERI} = e^{\hat{\beta}_{ML,1} + \hat{\beta}_{ML,2} + \hat{\beta}_{ML,int}} - e^{\hat{\beta}_{ML,1}} - e^{\hat{\beta}_{ML,2}} + 1$,with $\hat{\beta}_{ML,1}, \hat{\beta}_{ML,2}, \hat{\beta}_{ML,int}$corresponding to the maximum likelihood estimates of exposure X, exposure S, and their product term, respectively. These estimates are derived using a generalized linear model with a Poisson distribution and log link resulting in $\hat{\beta}$ equaling the ln(RR) and $e^{\hat{\beta}}$ the RR. However, assuming a disease incidence < 10% the ln(OR) approximates the ln(RR) and logistic regression models might also be considered. Regardless of the model used standard error estimates $se(\widehat{RERI})$ can be derived using the delta method (6): $\sqrt{var\left(\widehat{RERI}\right)} = h_1^2 * var\left(\hat{\beta}_1\right) + h_2^2 * var\left(\hat{\beta}_2\right) + h_{int}^2 * var\left(\hat{\beta}_{int}\right) + 2h_1h_2\hat{\sigma}_{1,2} + 2h_1h_{int}\hat{\sigma}_{1,int} + 2h_2h_{int}\hat{\sigma}_{2,int}$ where $h = -e^{\hat{\beta}}$ and $\hat{\sigma} = \text{covariance}$. The $se(\widehat{RERI})$ can also be estimated using bootstrapping: If Z equals the bootstrap sample distribution, then let $Z_{0.025}$ and $Z_{0.975}$ represent the 2.5 percentile and the 97.5 percentile. The standard error using the bootstrap percentiles (7) can be calculated by $se\left(\widehat{RERI}\right) = (Z_{0.975} - Z_{0.025})/(2 * 1.96)$. Under the null hypothesis the RERI based test statistics follow a Z-distribution.

Appendix Table 1 Results of the systematic review on simulation studies evaluating asymptotic interaction tests*.

Study	Scale	Tests	Results
Case-control or Cohort simulations			
Hogan 1978 (8)	Risk Difference	LR S-index	Scenario's Simulations were replicated 600, 400 or 200 times; N_K was set to 20 or 50, K = 2 and interaction effects were set to 0. Type 1 error rate Using these simulations the LR test was compared to the Synergy index, a statistic based on the RERI. Overall the nominal type 1 error rate of the S-index was always lower than 0.05 and the LR test generally showing error rate around 0.05. Power Not studied.
Greenland 1983 (9)	Odds Ratio Risk Difference	Wald	Scenario's Scenario's were repeated 1000 times with N_{cases}= 75, 150 or 300 and $N_{controls}$= 150, 300 or 600. OR interaction effect differed from LN -1.10 to LN 0.29, RD interaction effect differed from -2 to 10; for each outcome scale 10 scenarios were created. Type 1 error rate The multiplicative Wald statistic did not deviate from 0.05. The additive Wald tests (assuming an outcome prevalence of < 10% approximately equal to the a RD Wald test) showed type 1 error rates below 0.05 in small sample settings. Power The additive Wald test had a maximum power of 98%. The multiplicative Wald tests statistic reached 94%.In small samples, when N_{cases}=75,$N_{controls}$=150, power of both Wald tests never exceeded 60%.
Paul 1989 (10)	Odds Ratio	LR Tarone Score	Scenario's Simulations were iterated 500 times, the number of strata K were set to 3, 6 or 12. Event probability in the unexposed was set to 0.3, 0.5 and 0.7 for the first 3 K and repeated when K > 3. N_K = 120, N_K = 60, 120 and 180 or N_K = 20, 120, 220 and repeated depending on the size of K. In within strata balanced settings $N_{exposed}$ = $N_{unexposed}$. In within strata unbalanced settings $N_{exposed}$ = 20 and $N_{unexposed}$ = 100. Stratum specific effect differed from 1 (no effect) to 4.0. Depending on the scenarios some strata were set to 1 and other strata to > 1. Type 1 error rate In dependent of scenario chosen the Tarone and Score had lower type 1 error rates than the LR test which maximized at 0.064. Power Power hardly differed between tests. Only in settings with 12 strata and a large difference between the number of exposed and unexposed subjects (20 vs 100) did the difference increase. In such setting the LR tests had a power of 45%, the Score 39% and the Tarone 40%.
Jones 1989 (3) **and** **O'Gorman 1990 (11)**	Odds Ratio	LR BD Tarone	Scenario's Both studies used the same simulation study using 1000 repetitions, K was set to 2, 4, 8, 16, 32 or 64; N=256 was equally divided between cases and controls and K (so large K decreased the subjects per stratum); the exposure probability for controls was set to 0.05 or 0.30 or generated from a uniform distribution using the same values. The common odds ratio was set to 1, 4, 16 or randomly drawn from a log-normal (0,1), exponential(1/7), uniform(1,4) or (1,16) or from a two-point distribution with OR 1 and as second maximum an OR of 4 or 16. Type 1 error rate In very extreme scenarios of K = 64 and exposure probability of 0.3, the LR statistic showed a type I error rate of 0.97, while the Tarone and BD tests had error rates of 0.44. In general the LR had larger type 1 error rates (above 0.05) than the Tarone or BD which showed error rates closest to 0.05. Power Power for the LR was discarded because error rate were >0.07 in al scenarios. For the BD and the Tarone tests power almost never exceeded

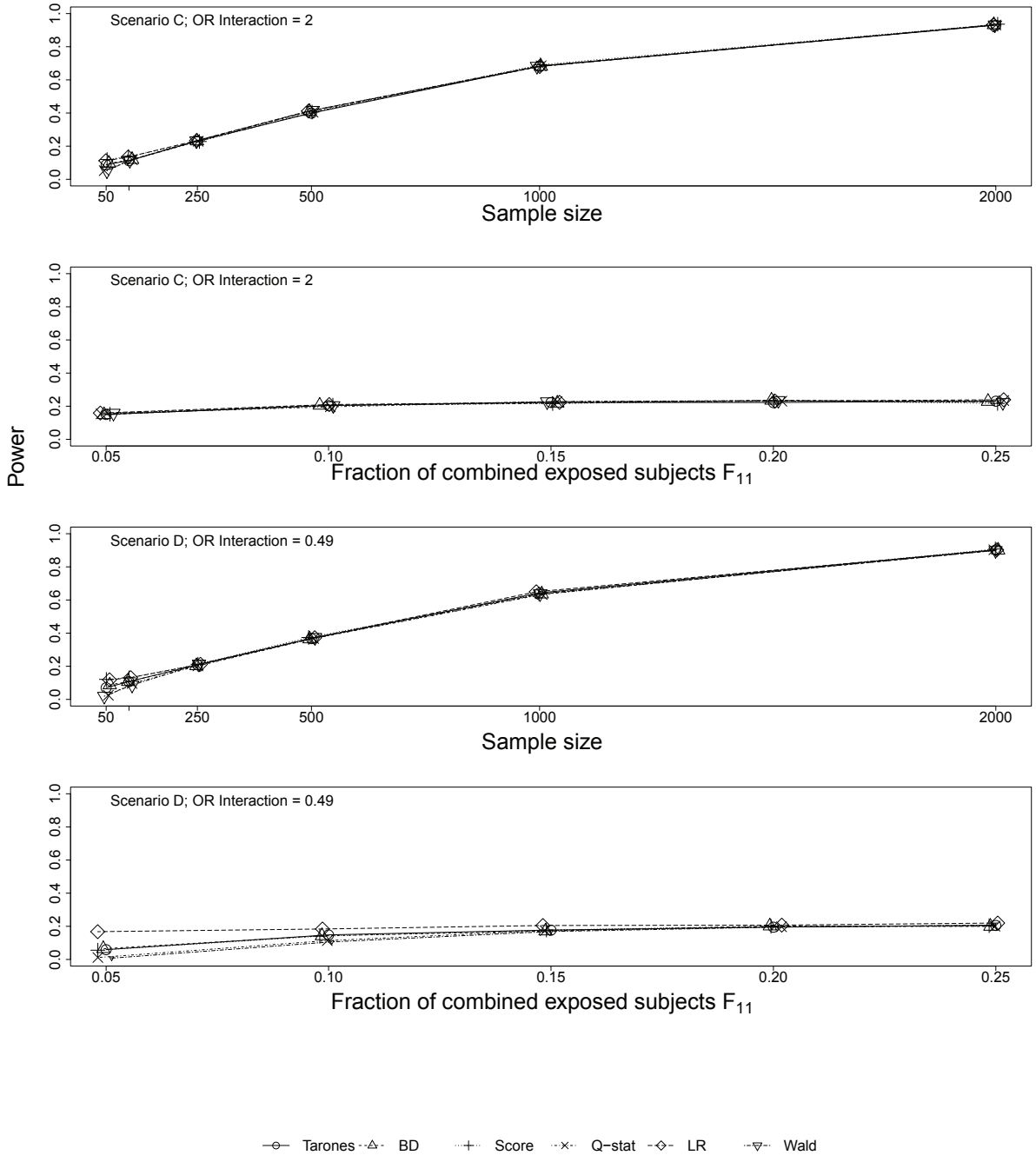
			60% and there was hardly any difference between the tests.
Paul 1992 (4)	Odds Ratio	LR Tarone BD Score	Scenario's Simulations were iterated 500 times, the number of strata K were set to 3, 6 or 12. Event probability in the unexposed was set to 0.3, 0.5 and 0.7 for the first 3 K and repeated in K > 3. N_K = 120, N_K = 60, 120 and 180 or N_K = 5, 10 or 20. In between strata unbalanced settings N_K = was set to 5, 10 and 20 and repeated according to the size of K. Type 1 error rate In dependent of scenario chosen the BD, Tarone and Score had lower type 1 error rates (< 0.05) than the LR test which almost always showed error rates above 0.05 Power Not explored.
Assman 1996 (7)	Risk Difference	RERI with delta variances estimator and bootstrapped percentile variance estimator.	Scenario's All scenario's were replicated 300 times, the number of N_{cases}= $N_{controls}$= 250. N was disblanced over the exposure categories, with 10% being exposed to exposure 1, 20% to exposure 2, 10% to both exposures and the remaining to neither factor. RERI interaction effect estimates differed from 12.0 to -4.0 and also included scenarios with no interaction effect i.e., 0. 95% Coverage rates In scenarios without interaction the coverage rates were > 95% for the delta method and ≤95% for the bootstrapped percentile method. Overall when the interaction effect was negative coverage rates <95% with the percentile method being closer to the 95%. In positive interaction settings. Independent of the variance estimator used coverage rates were skewed, when the interaction effect were positive the upper tail was generally larger, in negative settings the lower tail was biggest.
Reis 1999 (12)	Odds Ratio	LR BD Score	Scenario's The simulation setups were repeated 1000 times. A large number of scenario's were created (please see Reis 1999 for details) differing in the number of K strata set to 4,6,8,1 or 12; the number of exposed subjects per strata was set to5, 10, 20 ,30 or 50; the ratio of exposed to unexposed subjects differed from 1, 2 to 3; the outcome probability in unexposed subjects was set to 0.05, 0.10, 0.15 or 0.20 and the stratum specific odds ratio varied from 1 (no interaction effect) to 7. Finally, N_K sizes differed for example from K=8 setting 2 strata to 80 subjects and the remainder to 10. Type 1 error rate Independent of sample size, K or effect size, the Score and BD showed type 1 error rates around 0.05. Conversely to this, the LR test always had error rate above 0.05, maximizing at 0.16. Power When the ratio of exposed to unexposed subjects was 1 power never exceeded 40%. When the number of exposed subjects increased to 3 times the number of controls empirical power was around 50%, with the LR test showing a power of 55%. In scenarios with 2 large strata (K) and different number of smaller K (2 to 10), power increased, with a maximum of 75%.
Starr 2004 (13)	Odds Ratio Risk Difference	LR	Scenario's All simulations were based on 5000 repetitions, N samples size varied from 500 to 2500 equally divided between cases and controls. N was unbalanced between the exposure categories, with 10, 15 or 20 percent being exposed to exposure 1 or to exposure 2, 5, 10 or 15 percent to both exposures and 75, 60 or 45 percent remained free from exposure. In all scenario's the interaction effect was 0, but main effects were set to1, 2, 5, or 10. LR tests were based on logistic regression models with and without an interaction term. Using the similarity of the OR on the natural logarithm scale (i.e., odds difference) to the risk difference when disease incidence is low (e.g., < 10%) logistic regression models were also used to test for the presence of interaction on the risk difference scale. Type 1 error rate. Independent off scenarios both the multiplicative LR(β_3) and the additive LR(α_3) tests showed type error rates below or at 0.05, especially when the exposure probability was 0.10 and the overall N was set to 500. As N increased to 2500 or the exposure probability > 0.10 type 1 error rates increased to 0.05. Only in the presence high outcome prevalence (20%

			exposed to either factor 1 or 2 and 15% to both) and high main effects did the type 1 error rate increase to values above 0.05 with a maximum of 7.3. Power Not studied.
Zou 2008 (14)	Risk Difference	RERI with MOVER variance estimator and with the Delta variance estimator.	Scenario's All scenario's were replicated 1000 times, the number of N_{cases}= $N_{controls}$= 250 or 1000. N was disbalanced over the exposure categories, with 10% being exposed to exposure 1, 20% to exposure 2, 10% to both exposures and the remaining to neither factor. RERI interaction effect estimates differed from 12.0 to -4.0 and also included scenarios with no interaction effect i.e., 0. 95% Coverage rates In scenarios without interaction the coverage rates were always above the 95% maximizing in scenario's at 99% with 250 cases and control for the delta method and 95.9% for the MOVER method. Overall the MOVER method had coverage closer to the 95% than the delta method and both methods came closer to the 95% as the number of cases and controls increased to 1000 In the presence of a negative interaction effects and independent of the variance estimator used the coverage rate was > 95%. This was less pronounced in the scenario with 1000 cases and controls, and also when using the MOVER method. In the presence of large positive interaction effects (in the range of 12 to 5.25) the coverage rate dropped below the 95% otherwise the coverage rate was > 95%.
Meta-analysis or multi-centre trial settings			
Lipsitz 1998 (15)	Risk Difference	Q	Scenario's Type 1 error simulation were repeated 1825 times, scenario's that explored empirical power were repeated 1000 times. K was set at 8, 16,32 or 48 and N_K = 4,8,16,32, 64. The interaction effects (between study variance) differed either from 0 to 0.2 or in sparse data settings from 0 to 10, K was either set to 100 with N_K = 100 or to 32 and N_K = 16 Type 1 error rate The type 1 error rate was lowest (0.059) when K = 32 and N_N = 64. In extreme scenarios (K = 48 and N_K = 4), the error rate increased to 0.58. Power With a between study variance of 6 and K=32 and N_K=16 the Q statistic reached a power of 80%. In scenario of K = 100 and N_K= 100, a power of 80% was reached with a between study variance of 0.06
Takkouche 1999 (16)	Odds Ratio	Q LR	Scenario's K centers was set to 7, 20 or 40, N_K = 5000. interaction effect (variance between studies) was set at 0.0, 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.75, 0.8 or 0.9. Center size was based on published meta-analysis data and greatly varied but did not result in small sample sizes. Type 1 error rate Generally the type I error rate of the Q-statistic was approximately 0.05, with a maximum of 0.055 when K=40. In the same scenario the LR had a type 1 error rate of 0.32. Overall the LR tests had type 1 error rates below that of the Q, which decreased with decreasing K. Power When the between study variance of effect estimate was 0.25 power was 38%, the LR was 31%. When the between study variance increased to 0.75 a power of 100% was reached with a K of 40. In same scenario the power of the LR test was 99.9%.
Lui 2000 (17)	Risk Difference	Q	Scenario's Simulations were repeated 10000 times, K centers was set to 8, 16 , 32, 96. The number of subjects were center N_K set to 4,8, 16,32 or 64, ICC was set to 2/5, 1/6 or 1/20. Type 1 error rate Generally the Q statistics error rate was >0.05. In scenarios of K = 48, ICC = 1/6 and N_K = 4 the type 1 error rate maximized at 0.61. Overall the type 1 error rate decreased to 0.05 as N_K= increases to 64. Power

			Not studied.
Zhang 2009 (18)	Risk difference	Q	Scenario's Simulations were replicated 10000 times, the number of strata K was set to 8, 16,32 or 48 and N_K was set to 4,8, 16, 32, 64 or sample from a uniform distribution with 4 and 50 as minimum and maximum. Interaction effect were randomly sample from a uniform distribution of 0, 1. Scenarios used to determine the empirical power set K to 8 or 32 and N_K to 10, 15, 20, 25, 30, 35 or 40. Type 1 error rate The Q statistic almost never showed a type 1 error rate <0.05. As the number of subjects per strata increased to 32 the type I error rate was closest to 0.05. The largest type 1 error rates of 0.09 occurred in scenarios with K = 48 and N_K = 4, as N_K = increased to 64 type 1 error rates decreased to 0.05 Power When K was 8 the Q statistic reaches a maximum power of 40% however when K = 32, power reached 80% at a N_K of 30 and maximized at 85% with N_K = 40.
Bagheri 2011 (19)	Odds Ratio	LR BD	Scenario's All scenarios were resample 1000 times, the number of subjects per center N_K was set to 40, 100, 200 and the K centers to 4,6, or 8, center random effects was set to 0.1 or 0.5. The treatment by center interaction was set to 0, 0.2, 0.4, 0.6, and 0.8. Apart from scenarios with equal N_K and balance between $N_{exposed}$ and $N_{unexposed,}$, scenarios were created where $N_{exposed}$:$N_{unexposed}$ allocated unequally by the ratio 3:1. Finally, among center disbalance was created by creating on N_K with 5, 7 or 9 times the subjects as the other centers. Type 1 error rate Generally the type 1 error rate of the LR was below 0.05, the lowest error rate of 0.29 occurred in scenarios with K = 4 centers, and 0.1 random center effect. The type 1 error rate of the LR increased to 0.05 as K increased and or the random center effect was set to 0.5. The highest value of 0.59 was reached in the scenario where there was within center inequality. The BD tests had always had larger type 1 error rates. In scenarios with equal sample size or disbalance between treatment arms the BD type 1 error rate overshot 0.05 mark as K increased, with a maximum of 0.64 Power The BD test usually had more power than the LR test, this was largest in small sample sizes of N_K = 40, K = 4 and interaction effect = 0.2. In such scenarios power was 13% and 22% for the LR and BD tests. In scenarios of N_K =200, K = 8 and interaction effect = 0.8 power could be as high as 99% and 99,5% for the LR and BD tests. This was not markedly effect by sample size balance.

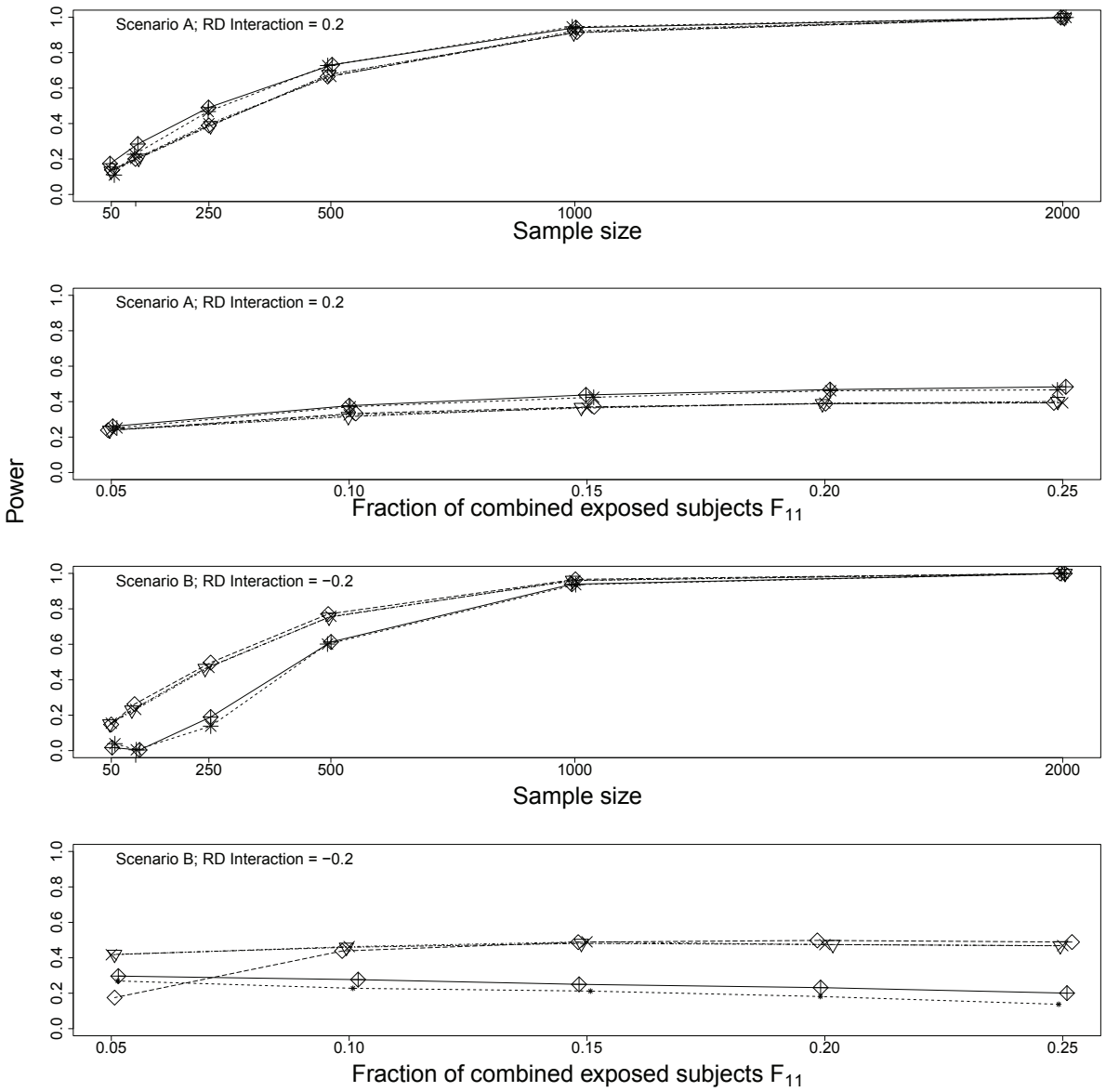
**We only review the performance of the: Breslow-Day, Tarone, unconditional Score, Cochrane's Q, log-likelihood ratio, Wald, and RERI based tests. Strata=K; subjects per stratum=N_K; N=total sample size intra class correlation=ICC; Odd Ratio=OR. * All nominal type 1 error rates were 0.05.*

Figure a Power of OR based interaction tests based on simulation scenarios C and D.



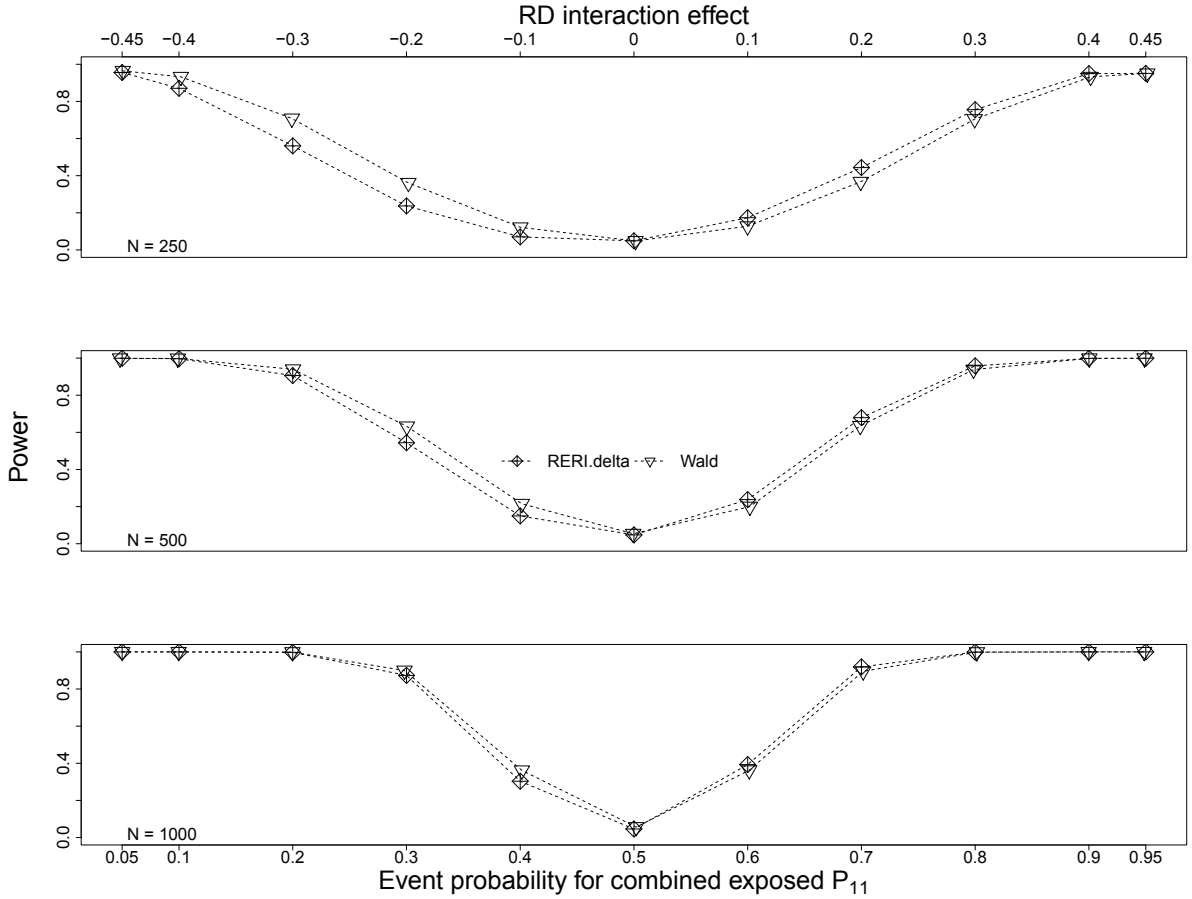
N.B. sample size was increased from 50 observations to 2,000. In the bottom part of scenario C and D sample size was fixed at 250 observations and the fraction of F11 was increased from 0.05 to 0.25. Thus the relative number of subjects that were exposed to both factors increased from 5 present to 25 percent of the total sample size of 250. All simulations were repeated 10,000 times.

Figure b Power of RD based interaction tests based on simulation scenarios A and B.



N.B. sample size was increased from 50 observations to 2,000. In the bottom part of scenario A and B sample size was fixed at 250 observations and the fraction of F11 was increased from 0.05 to 0.25. Thus the relative number of subjects that were exposed to both factors increased from 5 present to 25 percent of the total sample size of 250. All simulations were repeated 10,000 times.

Figure c Power of the RERI based test compared to the Wald test based on simulations using different interaction effect sizes.



N.B. In the upper panel sample size was 250, the middle panel shows results based on 500 subjects and the bottom panel shows results using 1,000 subjects. All simulations were repeated 10,000 times.

CHAPTER 2

Similarity of interaction and subgroup-specific effects in randomized and non-randomized studies: three empirical examples

A F Schmidt, M M Rovers, O H Klungel, A W Hoes, M J Knol
M Nielen, A de Boer, R H H Groenwold.

Journal of Clinical Epidemiology 2013 66(6)
doi: 10.1016/j.jclinepi.2012.08.008

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, RCTs, and Individual Patient Data Meta-Analyses of RCTs (IPDMAs; reference) on three clinical topics: [1] mammography screening and breast cancer mortality; [2] CABG and all-cause mortality; [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 to IPDMA. In the mammography example the Ratio of Risk Ratio's (RRR) (i.e., interaction effect) among observational studies was 1.46 (95%CI 1.09;1.96) compared to 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), in the IPDMA this was 1.40 (95%CI 1.08;1.1.81). 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.

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 towards 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 to main effect estimates, subgroup analyses move towards 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 due to 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 due to 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 feasibly. Alternatively, observational (i.e., non-randomized) 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.

Even though 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 as compared to 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 non-randomized 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).

Methods

Search strategy

IPDMAs were identified using the "IPD Cochrane Methods Group" website (17) and the 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 (CI); (4) were based on RCTs; and (5) written in English.

Subsequently we searched for (additional) RCTs and observational papers. First, we searched MEDLINE and the CENTRAL databases with an adapted search strategy, used by the original IPDMAs to also include observational studies (**Appendix I**). This search was supplemented with a Scopus (18) cross-reference search. We performed this strategy on five pre-selected 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: CABG vs. 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 treatment effects; (4) used a RCT, cohort study or case-control design; and (5) were written in English. We deemed an example viable if we found 2 or more observational studies that were comparable with the IPDMA. Since 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, pooled RCT estimates were compared to IPDMA estimates (26;27).

Statistical analysis

Extracted data were analyzed using R, version 2.10 for windows (28). When available, we used effect measures that were adjusted for baseline covariates.

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 (HR) (29-32) or odds ratios (OR; for case-control studies), with 95% confidence intervals (95%CI). In all cases where 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 (I^2) and tau-squared (τ^2) (34)) were calculated and pooled effects were estimated using fixed and random effects models. In al 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 to the reference group; RRR>1 indicates a larger effect of treatment in one subgroup compared to the reference group. To obtain a standard error (s.e.) for the RRR, the square root was taking from the sum of the stratum specific variances. For the observational and RCT data, the RRR was based on a random effect model. In the results section we state which reference groups were used.

Results

Effect of mammography screening by age.

The IPDMA of Nystrom et al.(23), determined the effects of mass mammography screening versus no-screening on breast cancer mortality. To study the effectiveness of mammography screening in younger women Nystrom stratified the results by age, <50 years and ≥50 years, which resulted in a non-significant interaction test. The data of the IPDMA could be compared with 6 trials (of which 4 were included in the IPDMA) (35-40) and 6 observational studies (one cohort study (41) and five case-control studies (42-46)).The IPDMA included 247 010 women of whom 1642 died of breast cancer, whereas in the RCTs 392 483 women participated of whom 1645 died of breast cancer. The observational studies included 233 791 women of whom 4498 died of breast cancer. Overall, the included studies were similar regarding the intervention, control group and outcome parameter (see **Appendix II**).

The pooled main effect of mammography screening on breast cancer mortality (**Figure 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 (**Figure 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 the effect 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 to 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.

Effect of CABG (versus PCI) by diabetes status

Hlatky et al. (25), studied the effect CABG versus Percutaneous Coronary Intervention (PCI) on all-cause mortality using an IPDMA (including 10 trials). Hlatky stratified the main effects for numerous baseline characteristics including presence of diabetes, which produced a significant interaction (p=0.014).

Figure 1 Effect of mammography screening on breast cancer mortality, stratified by study design. RR< 1 = protective effect of screening.

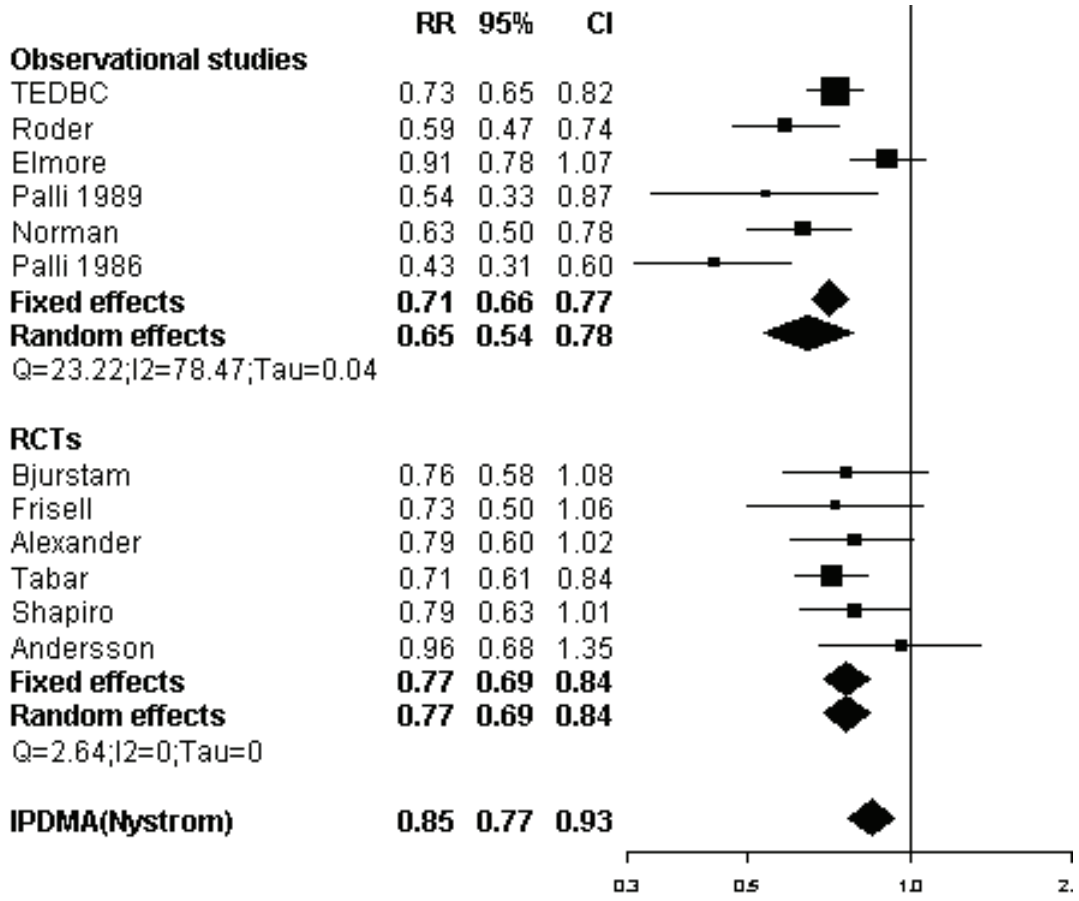
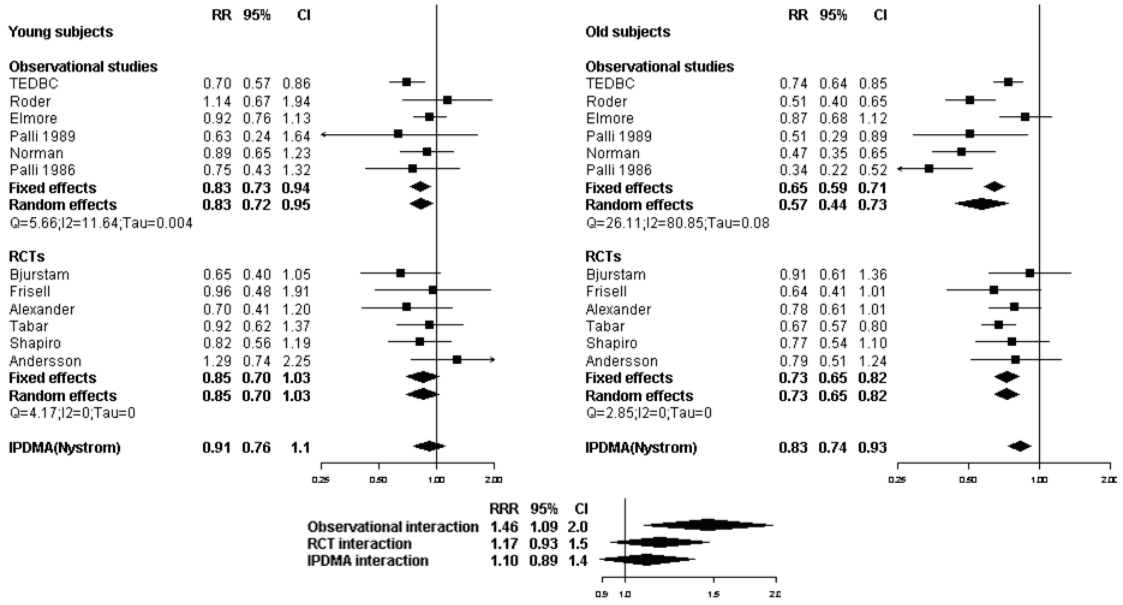


Figure 2 Effects of mammography screening on breast cancer mortality in strata of younger and older subjects.

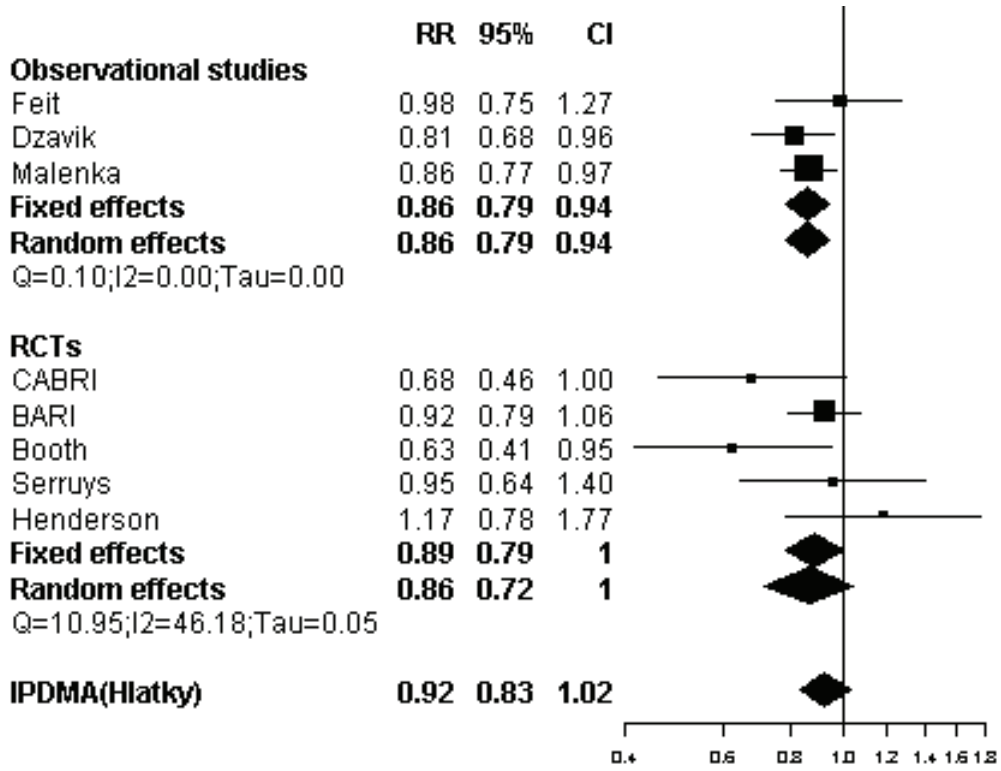


N.B. RR< 1 = protective effect of screening. Extreme values were truncated.
Interaction effects are the ratio of effect in younger (<50 years) divided by effect in older (≥50 years) subjects.

The data of the IPDMA could be compared with 5 trials (47-51) (which were all included in the IPDMA), and 3 observational cohort studies (31;32;52). The IPDMA included 7812 subjects who underwent a CABG or PCI procedure, of whom1203 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 study design type: 6561 (27.76%) in the IPDMA, 5197 (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 **Figure 3**.

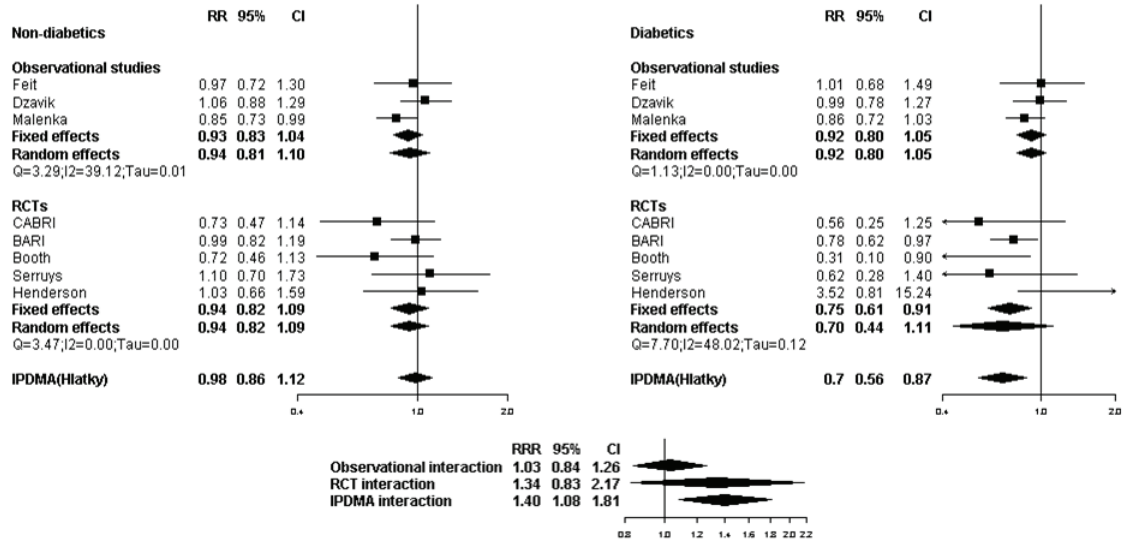
Figure 3 Effect of CABG versus PCI on all-cause mortality, stratified by study design. RR< 1 = protective effect of CABG.



The effect estimates of CABG (versus. PCI) in the group of non-diabetic patients were similar in the cohort studies, the RCTs, and the IPDMA, showing no effect (**Figure 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 non-diabetics compared to diabetics (**Figure 4**) showed that performing CABG (vs. PCI) was more effective in preventing all-cause mortality in diabetics compared to non-diabetics: 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 to the IPDMA effect, albeit non-significant, whereas the interaction effect based on observational studies (RRR=1.03, 95%CI 0.84;1.26) was smaller.

Figure 4 Effects of CABG versus PCI on all-cause mortality in strata of non-diabetics and diabetics.



N.B. RR< 1 = protective effect of CABG. Extreme values were truncated. Interaction effects are ratio of the effect in non-diabetics by effect in diabetics.

Effect of statin therapy by age

The Cholesterol Treatment Trialists' (CTT) Collaborators (24) IPDMA, studied the effect of statin therapy versus 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. Since 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 6 RCTs (30;53-57) (of which 5 were included in the IPDMA), and 4 observational studies (three cohort studies (29;58;59) and one case-control study(60)). The IPDMA consisted of 90 056 subjects of whom 7757 developed the outcome of interest. In the RCTs 70 877 subjects were included of whom 8192 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; 2485 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 II**).

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 (**Figure 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 (**Figure 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 versus 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.

Figure 5 Effect of statin therapy on cardiovascular endpoints, stratified by study design. RR < 1 = protective effect of statins.

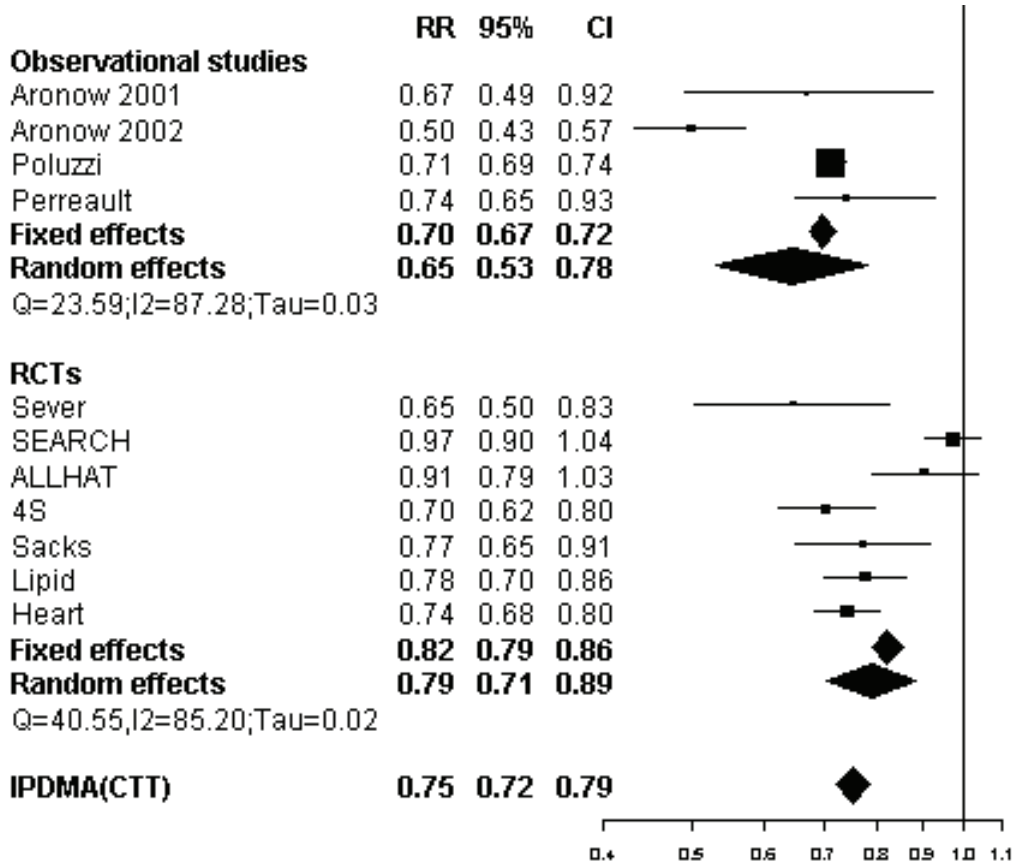
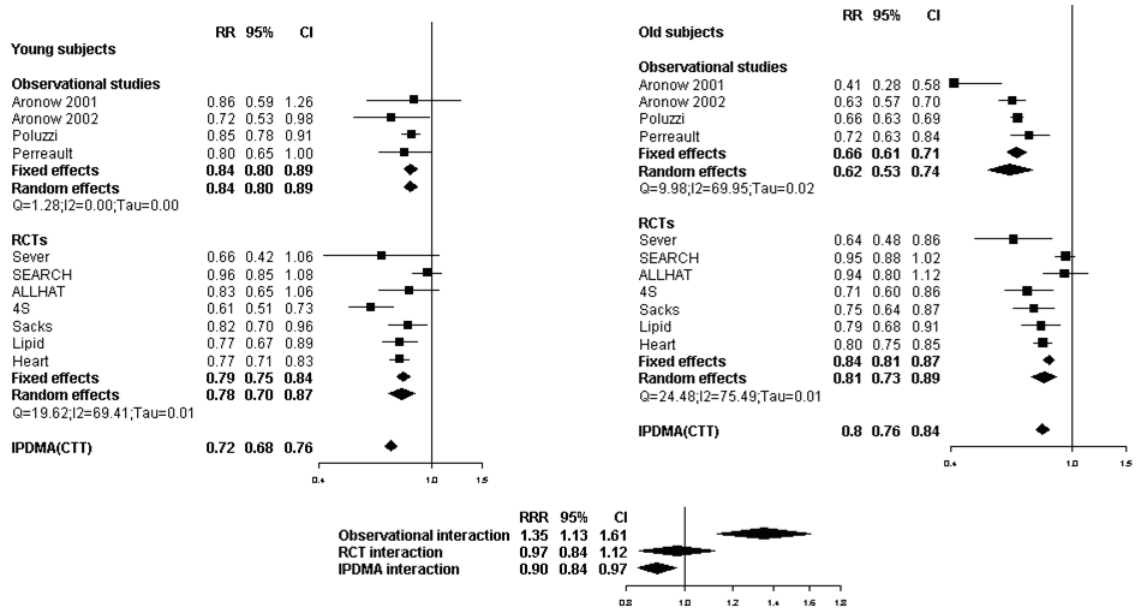


Figure 6 Effects of statin therapy on cardiovascular endpoints in strata of younger and older subjects.



N.B. RR< 1 = protective effect of statins. Extreme values were truncated. Interaction effects are the ratio of the effect in younger subjects (<60 to 70 years) divided by effect in older subjects (>60 to 70 years).

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 to 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, as well as 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 favours 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 to the observational studies (61;62). The (TEDBC) (41) observational study analyzed their data based on screening versus non-screening center (an analysis more similar to ITT) and found no interaction effect, which is in line with the 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 CABG 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 burdened diabetics), were more likely to receive CABG intervention. Similar, in the statin example only the Poluzzi (59) cohort study adjusted for confounding in subgroups, for instance by using a categorized age variable (<50, 50-65,65-80,>80). However, this does not sufficiently exclude residual confounding. In this case, lack of adjustment revealed a healthy user bias, where 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 to 18 years in the mammography example, 5.6 to 10.4 years in the CABG example and 0.5 to 8 years in the statin example. Furthermore, while treatment and outcomes were very similar in the mammography and CABG examples (appendix II), in the statin example RCTs used

placebo or active comparison groups, whereas observational studies used no or diminished treatment adherence as 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 towards 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 1 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 effect 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 to RCTs and the IPDMA estimates. Similarity of main and subgroup-specific effects across designs therefore does not imply similarity of interaction effects.

Reference List

1. Groenwold RH, Moons KG, Peelen LM, Knol MJ, Hoes AW. Reporting of treatment effects from randomized trials: a plea for multivariable risk ratios. Contemp Clin Trials 2011 May;32(3):399-402.
2. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. BMJ 2003 Jan 25;326(7382):219.
3. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol 2010 Aug;63(8):e1-37.
4. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 2004 Mar;57(3):229-36.
5. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010;340:c221.
6. Thompson SG, Higgins JP. Treating individuals 4: can meta-analysis help target interventions at individuals most likely to benefit? Lancet 2005 Jan 22;365(9456):341-6.
7. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. Med Care 2010 Jun;48(6 Suppl):S114-S120.
8. Klungel OH, Martens EP, Psaty BM, Grobbee DE, Sullivan SD, Stricker BH, et al. Methods to assess intended effects of drug treatment in observational studies are reviewed. J Clin Epidemiol 2004 Dec;57(12):1223-31.
9. Sacks H, Chalmers TC, Smith H, Jr. Randomized versus historical controls for clinical trials. Am J Med 1982 Feb;72(2):233-40.
10. Grobbee DE, Hoes AW. Confounding and indication for treatment in evaluation of drug treatment for hypertension. BMJ 1997 Nov 1;315(7116):1151-4.
11. Groenwold RH, Hak E, Hoes AW. Quantitative assessment of unobserved confounding is mandatory in nonrandomized intervention studies. J Clin Epidemiol 2009 Jan;62(1):22-8.
12. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001 Aug 15;286(7):821-30.
13. Van Spall HG, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. JAMA 2007 Mar 21;297(11):1233-40.
14. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. Am J Ophthalmol 2000 Nov;130(5):688.
15. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000 Jun 22;342(25):1887-92.
16. Shikata S, Nakayama T, Noguchi Y, Taji Y, Yamagishi H. Comparison of effects in randomized controlled trials with observational studies in digestive surgery. Ann Surg 2006 Nov;244(5):668-76.
17. IPD Cochrane Methods Group http://www.ctu.mrc.ac.uk/cochrane/ipdmg/DBIPD.asp. 2011.
18. Scopus www.scopus.com.
19. Young J, De SA, Merenstein D, van Essen GA, Kaiser L, Varonen H, et al. Antibiotics for adults with clinically diagnosed acute rhinosinusitis: a meta-analysis of individual patient data. Lancet 2008 Mar 15;371(9616):908-14.
20. Tudur SC, Marson AG, Williamson PR. Phenytoin versus valproate monotherapy for partial onset seizures and generalized onset tonic-clonic seizures. Tudur Smith Catrin , Marson Anthony G , Williamson Paula R Phenytoin versus valproate monotherapy for partial onset seizures and generalized onset tonic clonic seizures Cochrane Database of Systematic Reviews : Reviews 2001 Issue 4 John Wiley & Sons , L 2001.
21. Aabo K, Adams M, Adnitt P, Alberts DS, Athanazziou A, Barley V, et al. Chemotherapy in advanced

ovarian cancer: four systematic meta-analyses of individual patient data from 37 randomized trials. Advanced Ovarian Cancer Trialists' Group. Br J Cancer 1998 Dec;78(11):1479-87.
22. Rovers MM, Glasziou P, Appelman CL, Burke P, McCormick DP, Damoiseaux RA, et al. Antibiotics for acute otitis media: a meta-analysis with individual patient data. Lancet 2006 Oct 21;368(9545):1429-35.
23. Nystrom L, Andersson I, Bjurstam N, Frisell J, Nordenskjold B, Rutqvist LE. Long-term effects of mammography screening: updated overview of the Swedish randomised trials. Lancet 2002 Mar 16;359(9310):909-19.
24. Cholesterol Treatment Trialists' (CTT) Collaborators. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. Lancet 2005 Oct 8;366(9493):1267-78.
25. Hlatky MA, Boothroyd DB, Bravata DM, Boersma E, Booth J, Brooks MM, et al. Coronary artery bypass surgery compared with percutaneous coronary interventions for multivessel disease: a collaborative analysis of individual patient data from ten randomised trials. Lancet 2009 Apr 4;373(9670):1190-7.
26. Stewart LA, Parmar MK. Meta-analysis of the literature or of individual patient data: is there a difference? Lancet 1993 Feb 13;341(8842):418-22.
27. Koopman L, van der Heijden GJ, Hoes AW, Grobbee DE, Rovers MM. Empirical comparison of subgroup effects in conventional and individual patient data meta-analyses. Int J Technol Assess Health Care 2008;24(3):358-61.
28. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2009.
29. Aronow HD, Topol EJ, Roe MT, Houghtaling PL, Wolski KE, Lincoff AM, et al. Effect of lipid-lowering therapy on early mortality after acute coronary syndromes: an observational study. Lancet 2001 Apr 7;357(9262):1063-8.
30. ALLHAT Collaborative Research Group. Major outcomes in moderately hypercholesterolemic, hypertensive patients randomized to pravastatin vs usual care: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT-LLT). JAMA 2002 Dec 18;288(23):2998-3007.
31. Dzavik V, Ghali WA, Norris C, Mitchell LB, Koshal A, Saunders LD, et al. Long-term survival in 11,661 patients with multivessel coronary artery disease in the era of stenting: a report from the Alberta Provincial Project for Outcome Assessment in Coronary Heart Disease (APPROACH) Investigators. Am Heart J 2001 Jul;142(1):119-26.
32. Malenka DJ, Leavitt BJ, Hearne MJ, Robb JF, Baribeau YR, Ryan TJ, et al. Comparing long-term survival of patients with multivessel coronary disease after CABG or PCI: analysis of BARI-like patients in northern New England. Circulation 2005 Aug 30;112(9 Suppl):I371-I376.
33. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Philadelphia: Lippencott Williams & Wilkins; 2008.
34. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002 Jun 15;21(11):1539-58.
35. Frisell J, Lidbrink E, Hellstrom L, Rutqvist LE. Followup after 11 years--update of mortality results in the Stockholm mammographic screening trial. Breast Cancer Res Treat 1997 Sep;45(3):263-70.
36. Bjurstam N, Bjorneld L, Warwick J, Sala E, Duffy SW, Nystrom L, et al. The Gothenburg Breast Screening Trial. Cancer 2003 May 15;97(10):2387-96.
37. Tabar L, Vitak B, Chen HH, Prevost TC, Duffy SW. Update of the Swedish Two-County Trial of breast cancer screening: histologic grade-specific and age-specific results. Swiss Surg 1999;5(5):199-204.
38. Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F, et al. Mammographic screening and mortality from breast cancer: the Malmo mammographic screening trial. BMJ 1988 Oct 15;297(6654):943-8.

39. Alexander FE, Anderson TJ, Brown HK, Forrest AP, Hepburn W, Kirkpatrick AE, et al. 14 years of follow-up from the Edinburgh randomised trial of breast-cancer screening. Lancet 1999 Jun 5;353(9168):1903-8.
40. Shapiro S, Venet W, Strax P, Venet L, Roeser R. Selection, follow-up, and analysis in the Health Insurance Plan Study: a randomized trial with breast cancer screening. Natl Cancer Inst Monogr 1985 May;67:65-74.
41. TEDBC. 16-year mortality from breast cancer in the UK Trial of Early Detection of Breast Cancer. Lancet 1999 Jun 5;353(9168):1909-14.
42. Palli D, Del Turco MR, Buiatti E, Carli S, Ciatto S, Toscani L, et al. A case-control study of the efficacy of a non-randomized breast cancer screening program in Florence (Italy). Int J Cancer 1986 Oct 15;38(4):501-4.
43. Palli D, Rosselli del TM, Buiatti E, Ciatto S, Crocetti E, Paci E. Time interval since last test in a breast cancer screening programme: a case-control study in Italy. J Epidemiol Community Health 1989 Sep;43(3):241-8.
44. Roder D, Houssami N, Farshid G, Gill G, Luke C, Downey P, et al. Population screening and intensity of screening are associated with reduced breast cancer mortality: evidence of efficacy of mammography screening in Australia. Breast Cancer Res Treat 2008 Apr;108(3):409-16.
45. Elmore JG, Reisch LM, Barton MB, Barlow WE, Rolnick S, Harris EL, et al. Efficacy of breast cancer screening in the community according to risk level. J Natl Cancer Inst 2005 Jul 20;97(14):1035-43.
46. Norman SA, Russell LA, Weber AL, Coates RJ, Zhou L, Bernstein L, et al. Protection of mammography screening against death from breast cancer in women aged 40-64 years. Cancer Causes Control 2007 Nov;18(9):909-18.
47. CABRI, Kurbaan AS, Bowker TJ, Ilsley CD, Sigwart U, Rickards AF. Difference in the mortality of the CABRI diabetic and nondiabetic populations and its relation to coronary artery disease and the revascularization mode. Am J Cardiol 2001 Apr 15;87(8):947-50.
48. BARI. The final 10-year follow-up results from the BARI randomized trial. J Am Coll Cardiol 2007 Apr 17;49(15):1600-6.
49. Booth J, Clayton T, Pepper J, Nugara F, Flather M, Sigwart U, et al. Randomized, controlled trial of coronary artery bypass surgery versus percutaneous coronary intervention in patients with multivessel coronary artery disease: six-year follow-up from the Stent or Surgery Trial (SoS). Circulation 2008 Jul 22;118(4):381-8.
50. Serruys PW, Ong AT, van Herwerden LA, Sousa JE, Jatene A, Bonnier JJ, et al. Five-year outcomes after coronary stenting versus bypass surgery for the treatment of multivessel disease: the final analysis of the Arterial Revascularization Therapies Study (ARTS) randomized trial J Am Coll Cardiol 2005 Aug 16;46(4):575-81.
51. Henderson RA, Pocock SJ, Sharp SJ, Nanchahal K, Sculpher MJ, Buxton MJ, et al. Long-term results of RITA-1 trial: clinical and cost comparisons of coronary angioplasty and coronary-artery bypass grafting. Randomised Intervention Treatment of Angina. Lancet 1998 Oct 31;352(9138):1419-25.
52. Dzavik V, Ghali WA, Norris C, Mitchell LB, Koshal A, Saunders LD, et al. Long-term survival in 11,661 patients with multivessel coronary artery disease in the era of stenting: a report from the Alberta Provincial Project for Outcome Assessment in Coronary Heart Disease (APPROACH) Investigators. Am Heart J 2001 Jul;142(1):119-26.
53. Malenka DJ, Leavitt BJ, Hearne MJ, Robb JF, Baribeau YR, Ryan TJ, et al. Comparing long-term survival of patients with multivessel coronary disease after CABG or PCI: analysis of BARI-like patients in northern New England. Circulation 2005 Aug 30;112(9 Suppl):I371-I376.
54. Feit F, Brooks MM, Sopko G, Keller NM, Rosen A, Krone R, et al. Long-term clinical outcome in the Bypass Angioplasty Revascularization Investigation Registry: comparison with the randomized trial. BARI Investigators. Circulation 2000 Jun 20;101(24):2795-802.

55. SEARCH Collaborative Group, Bowman L, Wallendszus K, Bulbulia R, Rahimi K, Haynes R, et al. Intensive lowering of LDL cholesterol with 80 mg versus 20 mg simvastatin daily in 12,064 survivors of myocardial infarction: a double-blind randomised trial
63. Lancet 2010 Nov 13;376(9753):1658-69.
56. ALLHAT Collaborative Research Group. Major outcomes in moderately hypercholesterolemic, hypertensive patients randomized to pravastatin vs usual care: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT-LLT). JAMA 2002 Dec 18;288(23):2998-3007.
57. Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). Lancet 1994 Nov 19;344(8934):1383-9.
58. LIPID Study Group. Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels. The Long-Term Intervention with Pravastatin in Ischaemic Disease (LIPID) Study Group. N Engl J Med 1998 Nov 5;339(19):1349-57.
59. Heart Protection Study Collaborative Group. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet 2002 Jul 6;360(9326):7-22.
60. Sever PS, Dahlof B, Poulter NR, Wedel H, Beevers G, Caulfield M, et al. Prevention of coronary and stroke events with atorvastatin in hypertensive patients who have average or lower-than-average cholesterol concentrations, in the Anglo-Scandinavian Cardiac Outcomes Trial--Lipid Lowering Arm (ASCOT-LLA): a multicentre randomised controlled trial. Lancet 2003 Apr 5;361(9364):1149-58.
61. Aronow HD, Topol EJ, Roe MT, Houghtaling PL, Wolski KE, Lincoff AM, et al. Effect of lipid-lowering therapy on early mortality after acute coronary syndromes: an observational study. Lancet 2001 Apr 7;357(9262):1063-8.
62. Aronow WS, Ahn C. Incidence of new coronary events in older persons with prior myocardial infarction and serum low-density lipoprotein cholesterol > or = 125 mg/dl treated with statins versus no lipid-lowering drug. Am J Cardiol 2002 Jan 1;89(1):67-9.
63. Poluzzi E, Piccinni C, Carta P, Puccini A, Lanzoni M, Motola D, et al. Cardiovascular events in statin recipients: impact of adherence to treatment in a 3-year record linkage study. Eur J Clin Pharmacol 2011 Apr;67(4):407-14.
64. Perreault S, Ellia L, Dragomir A, Cote R, Blais L, Berard A, et al. Effect of statin adherence on cerebrovascular disease in primary prevention. Am J Med 2009 Jul;122(7):647-55.
65. Berry DA. Benefits and risks of screening mammography for women in their forties: a statistical appraisal. J Natl Cancer Inst 1998 Oct 7;90(19):1431-9.
66. Hernan MA, Hernandez-Diaz S. Beyond the intention-to-treat in comparative effectiveness research. Clin Trials 2011 Oct 7.
67. Malenka DJ, Leavitt BJ, Hearne MJ, Robb JF, Baribeau YR, Ryan TJ, et al. Comparing long-term survival of patients with multivessel coronary disease after CABG or PCI: analysis of BARI-like patients in northern New England. Circulation 2005 Aug 30;112(9 Suppl):I371-I376.
68. Poluzzi E, Piccinni C, Carta P, Puccini A, Lanzoni M, Motola D, et al. Cardiovascular events in statin recipients: impact of adherence to treatment in a 3-year record linkage study. Eur J Clin Pharmacol 2011 Apr;67(4):407-14.

Appendix I Search strategy for the mammography example.

IPDMA	MEDLINE	CENTRAL	Combined
Nystrom et al., 2002[I]	(breast neoplasms [MeSH] OR "breast cancer"[tiab] OR mammography[MeSH] OR "mammograph*" [tiab]) AND (mass screening[MeSH] OR "screen*" [tiab]) AND ("randomized controlled trial" [tiab] OR "randomized clinical trial" [tiab] OR "randomised controlled trial" [tiab] OR "randomised clinical trial" [tiab] OR RCT* [tiab] OR trial* [tiab] OR observational [tiab] OR retrospective [tiab] OR cohort [tiab] OR case-control [tiab] OR nonrandomized [tiab]OR nonrandomised [tiab] OR non-randomized [tiab] OR non-randomised [tiab]) Results:1490	("breast neoplasms" OR "breast cancer" OR mammography OR "mammograph*") AND ("mass screening" OR "screen*") Results:377	Screened on title, abstract and full text: 1468 Included studies:12

[I] Search strategy adapted from Gotzsche and Nielsen 2011

Appendix II Baseline characteristics of included studies

Study	Design	N	Intervention	Comparison	Subgroups	N of variables adjusted for	N of variables adjusted for in stratified analyses	Outcome
Mammography studies								
Nystrom 2002	IPDMA	247010	Mammography	No screening	Age (<50/≥50)	1	0	Breast cancer death
Bjurstam 2003	RCT	51611	Mammography	No screening	Age (<50/≥50)	1	1	Breast cancer death
Frisell 1997	RCT	60261	Mammography	No screening	Age (<50/≥50)	0	0	Breast cancer death
Alexander 1999	RCT	44268	Mammography and clinical breast examincation	No screening	< Age (<50/≥50)	1	1	Breast cancer death
Tabar 1999	RCT	133065	Mammography	No screening	Age (<50/≥50)	0	0	Breast cancer death
Shapiro 1985	RCT	60995	Mammography and clinical breast examincation	No screening	Age (<50/≥50)	0	0	Breast cancer death
Andersson 1988	RCT	42283	Mammography	No screening	Age (<55/≥55)	0	0	Breast cancer death
TEDBC 1999	Cohort study	222446	Mammography	No screening	Age (<50/≥50)	3	3	Breast cancer death
Roder 2008	Case-control	1964	Participation in mammography screening program	Not Participation in mammography screening program	Age (<50/≥50)	3	3	Breast cancer death
Elmore 2005	Case-control	3852	Mammography +/- self breast examination	No screening	Age (<50/≥50)	7	7	Breast cancer death
Palli 1989	Case-control	618	Mammography	No screening	< Age (<50/≥50)	11	11	Breast cancer death
Norman 2007	Case-control	4549	Mammography	No screening in 2 years prior to indexing date	Age (<50/≥50)	14	14	Breast cancer death
Palli 1986	Case-control	342	Mammography	No screening	Age (<50/≥50)	5	5	Breast cancer death
CABG studies								

Hlatky 2009	IPDMA	7812	CABG	PCI	Diabetes (yes/no)	0	0	Mortality
CABRI 2001	RCT	1054	CABG	PCI	Diabetes (yes/no)	0	0	Mortality
BARI 2007	RCT	1829	CABG	PCI	Diabetes (yes/no)	0	0	Mortality
Booth 2008	RCT	988	CABG	PCI	Diabetes (yes/no)	0	0	Mortality
Serruys 2005	RCT	1205	CABG	PCI	Diabetes (yes/no)	0	0	Mortality
Henderson 1998	RCT	1011	CABG	PCI	Diabetes (yes/no)	0	0	Mortality
Feit 2000	Cohort study	1814	CABG	PCI	Diabetes (yes/no)	14	0	mortality
Dzavik 2001	Cohort study	7322	CABG	PCI	Diabetes (yes/no)	18	0	mortality
Malenka 2005	Cohort study	14493	CABG	PCI	Diabetes (yes/no)	11	1	mortality
Statin studies								
CTT 2005	IPDMA	90056	Statins	Placebo/usual care/no treatment/active comparison	Age (≤65/>65)	0	0	CHD death+MI
Sever 2003	RCT	10305	Atorvastatin 10 mg	Placebo	Age (≤60/>60)	0	0	CHD death+MI
SEARCH 2010	RCT	12064	Simvastatin 80mg	Simvastatin 20 mg	Age (<60/≥60)	0	0	CHD death+MI, stroke, CABG
ALLHAT 2002	RCT	10355	Pravastatin 40 mg	Usual care	Age (<65/≥65)	0	0	CHD death+MI
4S 1994	RCT	4444	Simvastatin 20 mg	Placebo	Age (<60/≥60)	0	0	CHD death+MI
Sacks 1996	RCT	4159	Pravastatin 40 mg	Placebo	Age (<60/≥60)	0	0	CHD death+MI+CABG PTCA
Lipid 1998	RCT	9014	Pravastatin 40 mg	Placebo	Age (<65/≥65)	0	0	CHD death+MI
Heart 2002	RCT	20536	Simvastatin 40 mg	Placebo	Age (<65/≥65)	0	0	CHD death+MI

Aronow 2001	Cohort study	20809	Statins	No statins	Age (<65/≥65)	32	0	Mortality
Aronow 2002	Cohort study	1410	Statins	No statins	Age (≤70/>70)	6	0	CHD death+MI
Poluzzi 2011	Cohort study	48386	≥80% statin adherence	≤80% statin adherence	Age (≤65/>65	3	3	Non-fatal CHD events
Perreault 2009	Nested case-control	28334	≥80% statin adherence	≤20% statin adherence	Age (<65/≥65)	8	unclear	Non-fatal cerebrovascular events

CHAPTER 3

Increasing efficiency of post-launch RCTs to detect treatment effect modification

A F Schmidt, I Klugkist, O H Klungel, M Nielen
A de Boer, A W Hoes, R H H Groenwold.

submitted

Abstract

Background Findings from nonrandomized studies on safety or efficacy of treatment in patient subgroups may trigger post-launch randomized clinical trials (RCTs). In the analysis of such RCTs, results from nonrandomized studies are typically ignored, however incorporating prior evidence could increase power of post-launch RCTs.

Objective To study the trade-off between bias and power of Bayesian RCT analysis when incorporating information from nonrandomized studies as prior information.

Methods A simulation study was conducted to compare frequentist with Bayesian analysis using non-informative and informative priors. Scenarios were based on a RCT that showed that the effect of rosiglitazone on bone fractures is modified by gender: odds ratio (OR) 1.00 in men, 2.23 in women, and interaction OR of 0.45. Simulations varied in sample size, proportion of women, agreement between nonrandomized and RCT data and the hyperparameter of the prior distributions.

Results The frequentist and non-informative Bayesian analysis both yielded unbiased effects estimates. Results from informative Bayesian analyses were biased, e.g. interaction OR 0.71 for optimistic prior. However, due to a reduction in posterior variance, power increased from 44% to 93% when using an optimistic prior. Type 1 error rates were generally around 5%. However, when the informative priors were in the opposite direction of the RCT data (e.g., interaction OR>1.0 instead of <1.0), type 1 error rates could be 100% and power 0%.

Conclusion When prior information and interaction effects in the available data are in the same direction, Bayesian methods incorporating data from nonrandomized studies can meaningfully increase power of interaction tests in post-launch RCTs.

Background

Randomized clinical trials (RCTs) are the gold standard to assess effects of interventions (e.g., a drug or surgical procedure). An important reason for this is that random allocation of treatment prevents confounding. It is well known, however, that RCTs are usually underpowered to detect adverse events or treatment effect modification (i.e., underpowered to detect differences in treatment effects between patient subgroups) (1;2). It is therefore possible that signals of effect modification or adverse events are only detected after marketing of the treatment in post-launch nonrandomized studies (e.g., case-control or cohort studies). Such signals might then lead to the initiation of new (subgroup-specific) RCTs in an attempt to provide more evidence on which patient group benefit most from the treatment.

An example of treatment effect modification is the effect of the oral antidiabetic drug rosiglitazone, a type of thiazolidinedione (TZDs), on bone fractures: rosiglitazone increases the risk of bone fractures in women but not in men (3-5) (which was already observed in the pre-launch RCT). Customarily, post-launch RCTs are analysed without incorporating information from nonrandomized studies. On the other hand, including information from nonrandomized studies in the analysis of RCTs may decrease the required number of subjects for that trial. This may be particularly important in studies of treatment effect modification or adverse advents, since these typically require considerably more patients than studies exploring average treatment effects (1;2).

The validity of an RCT analysis that incorporates information from nonrandomized studies will heavily depend on the quality (i.e., validity) of the latter. With the advent of healthcare databases (6) developed specifically for research, like CPRD (www.cprd.com) or PHARMO (www.pharmo.nl), the quality of nonrandomized studies may have increased by allowing better control for confounders through improved data quality and data analytical techniques, more appropriate selection of comparators, and decreased likelihood of loss to follow-up. Furthermore, if bias is constant across subgroups this will cancel out when assessing e.g. the ratio of treatment effects in two patient subgroups, resulting in correct estimate of interaction. For example, assume that the true risk ratio (RR) in men equals 1.00 and 2.00 in women; the interaction effect is 1.00/2.00 = 0.50. If both RRs were biased upward by 40% e.g. due to confounding, the interaction effect would still be (1.00 * 1.40)/(2.00 * 1.40) = 0.50. Unfortunately, the assumption of equal bias is untestable. However, even when this assumption

is not exactly met, estimates of interaction effects will likely be less biased than main effect estimates. Given that most RCTs lack power to detect interaction effects, and the increased potential to adjust for confounding in nonrandomized studies, it might be time to more seriously consider combing results from both sources.

One way of incorporating information from nonrandomized studies in RCT analysis is by means of Bayesian statistics (7). Although not the only way to incorporate results from previous studies, Bayesian methods are intuitively appealing because of their ease of reweighing prior knowledge, depending on for example its perceived relevance or validity (8). To explore the trade-off between bias and power when including results of nonrandomized studies in the analysis of a post-launch RCT, we conducted a simulation study based on the aforementioned example of effect modification by gender in a study of the adverse effect of rosiglitazone on the risk of bone fracture.

Methods

A simulation study was performed. In the following, we first describe the clinical example of rosiglitazone use and the risk of hip fracture, which was the starting point for the simulations. Second, we describe the classical (frequentist) and Bayesian analyses that were performed on the simulated data. Finally, we describe the different parameters that were varied across simulations.

Clinical example: rosiglitazone and hip fractures

RCT data were simulated based on the empirical example of the effect of rosiglitazone on bone fractures. This effect was modified by gender. Specifically, effect estimates were used from a recent meta-analysis (5) of five RCTs comparing TZDs to placebo or an active comparator (e.g. metformin) on the risk of bone fractures (**Table 1**). This meta-analysis used data from 11,401 subjects, of whom 346 experienced a bone fracture. Among the 11,401 subjects, the male to female ratio was 1.59. The odds ratio (OR) of fracture for TZD in men was 1.00 (95% confidence interval [95%CI] 0.73 ; 1.39), while the OR in women was 2.23 (95%CI 1.65 ; 3.01), resulting in an interaction effect of 0.45 (95%CI 0.29 ; 0.70). Information from nonrandomized studies was based on a recent individual patient data meta-analysis (IPDMA) by Bazelier et.al (3), which included data from three nonrandomized studies comparing TZDs to other oral anti-diabetic treatments. In total 1,637,084 patients were included, with a male

to female ratio of 1.05, of whom 32,244 experienced a bone fracture during follow-up. This IPDMA was used because it presented subgroup-specific effect estimates after confounding adjustment: OR 1.05 (95%CI 0.96 ; 1.14) in men, OR 1.44 (95%CI 1.35 ; 1.53) in women, and an interaction OR of 1.05/1.44 = 0.73 (95%CI 0.66 ; 0.81).

Table 1 Empirical data on the adverse effect of thiazolidinediones (TZDs) on bone fracture incidence within male and female patient subgroups; effects presented as odds ratio and 95% confidence interval*.

	RCT data		Nonrandomized data	
	TZDs effects on bone fractures	Interaction effects	TZDs effects on bone fractures	Interaction effects
Subgroup 1(men)	1.00 (0.73; 1.39)	0.45 (0.29; 0.70)	1.05 (0.96; 1.14)	0.73 (0.66; 0.81)
Subgroup 2(women)	2.23 (1.65; 3.01)	Reference	1.44 (1.35; 1.53)	Reference

** The empirical RCT data was used to simulate a new RCT and was combined with prior information based on the nonrandomized data. RCT data consisted of 11,401 subjects, of whom 346 experienced a bone fracture; the nonrandomized data consisted of 1,637,084 patients of whom 32,244 experienced a bone fracture. Presented RCT effect estimates are unadjusted for covariables, while nonrandomized estimates are adjusted for potential confounders. Interaction effects are the ratio of the subgroup specific effects and measure how much the treatment effect is modified by gender.*

Statistical analyses

RCTs are typically analysed without taking prior knowledge into account i.e., a frequentist approach. This frequentist analysis was considered as the reference. We assumed that the gender-specific treatment effects (the natural logarithm of the OR $[\overline{y}_i]$) were approximately normally distributed and Wald based 95%CI were constructed. Here $\overline{y}_i$ indicates the ln(OR) of rosiglitazone in males (i = 1) or females (i = 2) based on the simulated RCT data. Note, that the ORs can be interpreted as risk ratios, since the incidence of bone fractures is low (<10%) (9). Interaction effects were calculated by dividing the treatment effect among men by that among women. The variance in interaction effect was derived by taking the sum of the variances $[s_i^2]$ in gender-specific treatment effect estimates (10).

Bayesian methods were implemented by combining the above described gender specific likelihoods with conjugate (normally distributed) gender-specific priors: $N(\overline{x}_i, t_i^2)$ with the hyperparameter $\overline{x}_i$ representing the gender-specific mean ln(OR) of treatment and t_i^2 the corresponding variance in $\overline{x}_i$. Given that the variance (t_i^2) of an OR depends on its mean, we assumed this variance to be known and no prior distribution was defined. Since the nonrandomized IPDMA (3) only reported confounding adjusted effect estimates (ORs) and not the absolute fracture risk in different subgroups, we did not use beta-distributions in the Bayesian analysis.

Information from nonrandomized studies was incorporated in the prior distribution in different

ways. First, as a reference a non-informative prior was used, with hyperparameters $\overline{x}_i = 0$ and $t_i^2 = 10^6$. Second, three informative prior distributions were used with different precision (i.e., different variance hyperparameters) denoted sceptical, equivalent and optimistic (**Figure 1**). For the sceptical prior, a variance hyperparameter t_i^2 was used that equalled four times the variance in ln(OR) of the simulated RCT: $t_i^2 = s_i^2 * 4$. The variance hyperparameter of the equivalent prior was set equal to the variance in ln(OR) of the simulated RCT: $t_i^2 = s_i^2 * 1$. By setting the variance hyperparameter proportional to variance in the RCT treatment effect estimates, the sceptical and equivalent priors prevent the prior information from over influencing the data (11). For the optimistic prior the variance hyperparameter was set to $t_i^2 = 0.027$, which was similar to the precision of main effect of original RCT meta-analysis (5).

The hyperparameters of $\overline{x}_i$ were based on the reported nonrandomized treatment effect estimates presented in Table 1. All three informative priors mentioned above (i.e., sceptical, equivalent, and optimistic) used the same hyperparameters $\overline{x}_i$. To reflect the uncertainty in this hyperparameter, $\overline{x}_i$ was set to the point estimates [ln(1.05) in men; ln(1.44) in women], the lower bound of the 95%CI [ln(0.96); ln(1.35)] or the upper bound [ln(1.14); ln(1.53)] of the gender-specific treatment effects observed in the empirical IPDMA (**Table 1**).

Finally, using the previously defined gender-specific likelihoods and priors, the posterior distribution was estimated using equation 1.

$$\hat{\theta}_i \sim N\left(\hat{\mu}_i = \frac{s_i^2 \overline{x}_i + t_i^2 \overline{y}_i}{s_i^2 + t_i^2}, \hat{\delta}_i^2 = \frac{s_i^2 t_i^2}{s_i^2 + t_i^2} \right)$$

From this posterior distribution the mean ln(OR) and 95% credibility intervals (95%CrI) were estimated by the posterior mean $\hat{\mu}_i$ and the 2.5th and 97.5th percentiles. Posterior standard deviations for the gender-specific ORs of treatment were estimated by taking the square root of $\hat{\delta}_i^2$; following frequentist practice this will subsequently be referred to as the standard error (SE). The interaction effect was defined as the ratio of the gender-specific treatment effects, while the variance in interaction effect was estimated by summing the variances of the gender-specific treatment effect estimates.

Simulations

Based on the described empirical RCT data, a basic simulated RCT scenario was created

with gender-specific treatment effects (OR) of 1.00 in men and 2.23 in women resulting in an interaction OR of 0.45. To explore the impact of decreasing the RCT sample size, frequentist and Bayesian methods were evaluated for sample sizes of 1,000, 5,000 or 10,000 subjects for the post launch RCT. Subjects were equally divided between exposure groups and between genders. Dichotomous outcome data (bone fracture yes/no) per gender-specific subgroup were generated using random draws from a binomial distribution; with the outcome incidence in the comparator group set to 0.03 in women and 0.02 in men. All simulations were repeated 10,000 times, using the R statistical package for windows version 3.0.2 (12).

In the basic scenario, the type 1 error rate was explored among males (true OR = 1.00) and power was evaluated among females (true OR = 2.23) and for the interaction effect (i.e., 0.45). To also evaluate type 1 error rates of the interaction effect, additional scenarios were created (**Table 2**). In a second scenario the gender-specific ORs from the RCT were set to 2.23 in both subgroups, resulting in an interaction OR of 1.00. This represents no interaction in the presence of an overall main effect. In the third scenario the gender-specific ORs from the RCT were set to 1.00 in both subgroups, which represent the absence of a main, as well as an interaction effect. In a fourth scenario we explored the performance of Bayesian analysis of interaction tests when the prior information contradicted the simulated data (OR_1=1.00 and OR_2=2.23), by setting the prior hyperparameter $\overline{x}_i$ for men to ln(1.35), ln(1.44) or ln(1.53) and the prior ln(OR) for women to ln(0.96), ln(1.05) or ln(1.14). Finally, all scenarios were repeated with a different gender distribution in the simulated RCT population: 15% women, instead of 50%. **Table 2** presents an overview of the different scenarios and parameterizations.

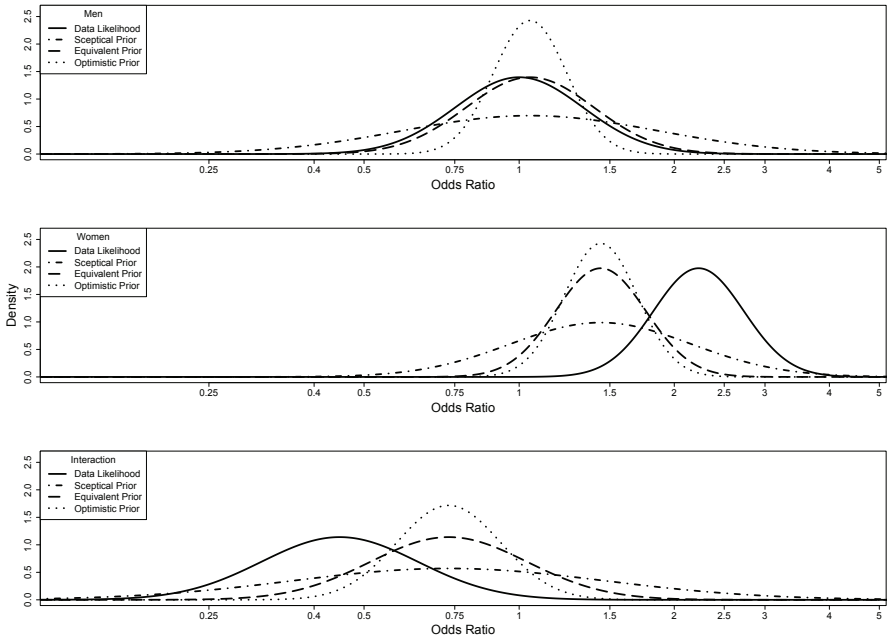
Table 2 Scenarios of a simulation study comparing frequentist to Bayesian analyses of a post-launch RCT*.

Scenario	Default (I)	II	III	IV
Sample size	1,000 5,000 10,000	1,000 5,000 10,000	1,000 5,000 10,000	1,000 5,000 10,000
Simulated OR of treatment in men	1.00	***2.23***	1.00	1.00
Simulated OR of treatment in women	2.23	2.23	***1.00***	2.23
Prior OR in men	1.05 0.96 1.14	1.05 0.96 1.14	1.05 0.96 1.14	***1.44*** ***1.35*** ***1.53***
Prior OR of treatment in women	1.44 1.35 1.53	1.44 1.35 1.53	1.44 1.35 1.53	***1.05*** ***0.96*** ***1.14***
Percentage women included*	50%	50%	50%	50%

** Note all scenarios were repeated using RCT including only 15% percent women*

Performance of the frequentist and Bayesian analyses were evaluated using power, type I error rate and bias. Power is defined as the proportion of times the 95% credibility interval (95%CrI) or 95%CI excluded an OR of 1, when there was an effect. The type 1 error rate is defined as the proportion of times the 95%CrI or 95%CI excluded an OR of 1, when in fact there was no treatment effect. Bias was defined as the mean difference between the true treatment effect (**Table 1**, RCT column) and the estimated treatment effect on the ln scale.

Figure 1. Probability density distributions of simulated RCT data (N = 5,000) and three informative priors.



The mean OR of the simulated RCT data was 1.00, 2.23 and 0.45 for the rosiglitazone effect in men, women and their interaction. For the informative priors the mean OR was set to 1.05, 1.44 and 0.73, based on results from nonrandomized studies (see table 1).

Results

In the basic scenario (**Table 3**), subgroup-specific rosiglitazone effect estimates in men were around 1.00, independent of simulated RCT size and the type of analyses. Type 1 error rates for the frequentist and non-informative priors were at most 5% and lower for the informative priors (range: 0; 3%). Subgroup-specific treatment effect estimates in women differed between the analytic methods applied and depended on the prior. For example, for a sample size of 1,000: non-informative prior (OR = 2.33), sceptical prior (OR = 2.13), equivalent prior (OR = 1.83) and optimistic prior (OR = 1.51). Stated otherwise, bias increased by adding prior knowledge, and was largest with the optimistic Bayesian analyses (bias range: -0.04; 0.12). Despite bias towards a neutral effect of 1, power improved by adding prior knowledge. For example in a RCT of 1,000 subjects, the power to detect a treatment effect in the subgroup of females was 40% for the frequentist analysis and between 87% - 100% for Bay-

esian analysis with an optimistic prior (**Table 3**). This increase in power was driven by the decrease in SE that offset the increase in bias (**Table 3**). For example, the SE of the interaction effect in a RCT of 5,000 subjects was 0.36, 0.36, 0.32, 0.25, and 0,19 for frequentist, non-informative prior, sceptical prior, equivalent prior, and optimistic prior, respectively. This also translated in an increased power for the interaction test: 62% power for the frequentist analysis compared to 83% - 93% using an optimistic prior. Finally, we note that in general the point estimates of the OR did not change with increasing RCT sample size. The one exception being the optimistic prior ORs which changed in the direction of the true treatment effect as sample size increased.

To explore the type 1 error rates of the interaction effect we simulated a RCT where the true OR in men as well as in women was 2.23 (i.e., no interaction). Type 1 error rates of the frequentist, non-informative prior and sceptical prior interaction tests were ≤ 5% (**Table 4**). Using an equivalent prior, the type 1 error rate could be as high as 10% (N = 10,000), but generally the rate was below 5% (N = 5,000 or 1,000). Type 1 error rates as high as 20% were found for the Bayesian analyses using optimistic priors. Finally, in scenarios in which the main effect was absent (i.e., OR in men and women was 1.00, see **Appendix table 1**), the type 1 error rates of the interaction tests were generally lower, with a maximum of 8% for the optimistic prior.

To assess the impact of incorrectly specifying the prior mean hyperparameter, Bayesian analyses were performed in which the priors contradicted the RCT, e.g., a prior interaction effect of 1.41 compared to the simulated OR of 0.45 (**Table 5**). As expected, the frequentist and non-informative Bayesian analyses showed ORs close to the true values. Informative Bayesian analyses, using for example a sample size of 1,000 subjects, showed ORs in men ranging from 1.07 (sceptical prior) to 1.49 (optimistic prior). Similarly, the OR in women ranged from 2.03 (sceptical prior) to 1.05 (optimistic prior). In the same scenario, interaction effects based on informative priors ranged from 0.53 (sceptical prior) to 1.26 (optimistic prior). In these cases, power and type 1 error rates were both affected, with observed powers as low as 0% (i.e., interaction effect and female subgroup) and type 1 error rates as high as 100% (male subgroup).

Table 3 Results of a simulated RCT exploring subgroup-specific and interaction effects using frequentist and Bayesian methods*.

Simulation	True OR	N = 1,000			N = 5,000			N = 10,000		
		OR(SE)	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*
Men										
Frequentist		1.00(0.71)	2%	0.00	1.00(0.29)	5%	0.00	1.00(0.21)	5%	0.00
Non-informative prior		0.99(0.72)	2%	-0.01	1.00(0.29)	4%	0.00	1.00(0.21)	5%	0.00
Sceptical prior:										
$\bar{x}_1$ = ln(1.05)		1.01(0.64)	1%	0.01	1.01(0.26)	2%	0.01	1.01(0.18)	3%	0.01
$\bar{x}_1$ = ln(0.96)		0.99(0.64)	1%	-0.01	0.99(0.26)	3%	-0.01	0.99(0.18)	3%	-0.01
$\bar{x}_1$ = ln(1.14)		1.02(0.64)	1%	0.02	1.03(0.26)	3%	0.03	1.02(0.18)	3%	0.02
Equivalent prior:	1.00									
$\bar{x}_1$ = ln(1.05)		1.02(0.51)	0%	0.02	1.03(0.21)	0%	0.03	1.02(0.15)	1%	0.02
$\bar{x}_1$ = ln(0.96)		0.98(0.51)	0%	-0.02	0.98(0.21)	1%	-0.02	0.98(0.15)	1%	-0.02
$\bar{x}_1$ = ln(1.14)		1.06(0.51)	0%	0.06	1.07(0.21)	1%	0.06	1.07(0.15)	2%	0.06
Optimistic prior:										
$\bar{x}_1$ = ln(1.05)		1.05(0.16)	0%	0.05	1.04(0.14)	0%	0.04	1.03(0.13)	0%	0.03
$\bar{x}_1$ = ln(0.96)		0.96(0.16)	0%	-0.04	0.97(0.14)	0%	-0.03	0.98(0.13)	0%	-0.03
$\bar{x}_1$ = ln(1.14)		1.13(0.16)	0%	0.12	1.11(0.14)	0%	0.10	1.08(0.13)	2%	0.08
Women										
Frequentist		2.32(0.48)	40%	0.04	2.25(0.21)	98%	0.01	2.24(0.15)	100%	0.00
Non-informative prior		2.33(0.48)	41%	0.05	2.26(0.21)	98%	0.01	2.23(0.15)	100%	0.00
Sceptical prior:										
$\bar{x}_2$ = ln(1.44)		2.13(0.43)	40%	-0.04	2.06(0.18)	99%	-0.08	2.05(0.13)	100%	-0.08
$\bar{x}_2$ = ln(1.35)		2.10(0.43)	38%	-0.06	2.03(0.18)	99%	-0.09	2.03(0.13)	100%	-0.10
$\bar{x}_2$ = ln(1.53)		2.14(0.43)	40%	-0.04	2.08(0.18)	99%	-0.07	2.07(0.13)	100%	-0.07
Equivalent prior:	2.23									
$\bar{x}_2$ = ln(1.44)		1.83(0.34)	38%	-0.20	1.80(0.15)	100%	-0.21	1.80(0.10)	100%	-0.22
$\bar{x}_2$ = ln(1.35)		1.77(0.34)	33%	-0.23	1.74(0.15)	100%	-0.25	1.74(0.10)	100%	-0.25
$\bar{x}_2$ = ln(1.53)		1.89(0.34)	45%	-0.17	1.86(0.15)	100%	-0.18	1.85(0.10)	100%	-0.18
Optimistic prior:										
$\bar{x}_2$ = ln(1.44)		1.51(0.16)	99%	-0.39	1.71(0.13)	100%	-0.27	1.84(0.11)	100%	-0.19
$\bar{x}_2$ = ln(1.35)		1.42(0.16)	87%	-0.45	1.64(0.13)	100%	-0.31	1.79(0.11)	100%	-0.22
$\bar{x}_2$ = ln(1.53)		1.59(0.16)	100%	-0.34	1.77(0.13)	100%	-0.23	1.89(0.11)	100%	-0.16
Interaction										
Frequentist		0.43(0.87)	16%	-0.04	0.44(0.36)	62%	-0.01	0.45(0.25)	89%	-0.01
Non-informative prior		0.42(0.87)	16%	-0.06	0.44(0.36)	63%	-0.01	0.45(0.25)	90%	0.00
Sceptical prior:										
= ln(0.73)		0.48(0.78)	12%	0.06	0.49(0.32)	61%	0.09	0.49(0.23)	91%	0.09
= ln(0.71)		0.47(0.78)	12%	0.05	0.49(0.32)	62%	0.08	0.49(0.23)	91%	0.09
= ln(0.75)		0.48(0.78)	12%	0.07	0.49(0.32)	62%	0.09	0.49(0.23)	90%	0.10
Equivalent prior:	0.45									
= ln(0.73)		0.56(0.62)	6%	0.22	0.57(0.25)	65%	0.24	0.57(0.18)	95%	0.24
= ln(0.71)		0.55(0.62)	6%	0.21	0.56(0.25)	67%	0.23	0.56(0.18)	96%	0.23
= ln(0.75)		0.56(0.62)	6%	0.22	0.57(0.25)	63%	0.25	0.57(0.18)	95%	0.25
Optimistic prior:										
= ln(0.73)		0.70(0.22)	8%	0.44	0.61(0.19)	88%	0.31	0.56(0.17)	99%	0.22
= ln(0.71)		0.68(0.22)	19%	0.41	0.59(0.19)	93%	0.28	0.54(0.17)	99%	0.19
= ln(0.75)		0.71(0.22)	3%	0.46	0.62(0.19)	83%	0.33	0.57(0.17)	98%	0.24

** Women contributed 50% to the overall sample size N. In all simulations Bayesian RCT analysis are compared to a frequentist analysis ignoring prior information. Subgroup-specific and interaction effect results are evaluated using the treatment OR, with the frequentist estimator reflecting the data likelihood and the Bayesian estimators the mean of the posterior distribution. Results were further evaluated using standard errors (SE) or from a Bayesian perspective standard deviations (SD), the percentage of times an OR of 1 was excluded by the 95% confidence or credibility interval (% reject). Depending on the true OR the % reject should be interpreted as power or type 1 error rate. The non-informative prior is based on a normal distribution with a mean hyperparameter of 0 and variance of 10^6; the equivalent prior is based on a normal distribution with the reported ln OR $\bar{x}_1$ and variance hyperparameter σ_1^2 equal to the simulated data; the sceptical prior uses the same mean and variance hyperparameters only now multiplying the variance by 4; the optimistic prior again uses the same point estimates from nonrandomized data but now uses a variance of 0.027 for each subgroup. All simulation results are based on 10,000 repetitions.*

Table 4 Results of a simulated RCT exploring subgroup-specific and interaction effects using frequentist and Bayesian methods in the presence of an interaction effect of 1.00*.

Simulation	True OR	N = 1,000			N = 5,000			N = 10,000		
		OR(SE)	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*
Men										
Frequentist		2.36(0.60)	28%	0.06	2.25(0.25)	91%	0.01	2.24(0.18)	100%	0.00
Non-informative prior		2.34(0.60)	27%	0.05	2.26(0.25)	92%	0.01	2.24(0.18)	100%	0.01
Sceptical prior:										
$\bar{x}_1$ = ln(1.05)		2.01(0.54)	17%	-0.10	1.95(0.22)	89%	-0.14	1.93(0.16)	99%	-0.15
$\bar{x}_1$ = ln(0.96)		1.98(0.54)	17%	-0.12	1.90(0.22)	87%	-0.16	1.89(0.16)	100%	-0.17
$\bar{x}_1$ = ln(1.14)		2.04(0.53)	19%	-0.09	1.96(0.22)	90%	-0.13	1.96(0.16)	100%	-0.13
Equivalent prior:	2.23									
$\bar{x}_1$ = ln(1.05)		1.57(0.42)	3%	-0.35	1.54(0.18)	78%	-0.37	1.54(0.12)	99%	-0.37
$\bar{x}_1$ = ln(0.96)		1.51(0.42)	2%	-0.39	1.47(0.18)	63%	-0.42	1.47(0.12)	96%	-0.42
$\bar{x}_1$ = ln(1.14)		1.63(0.42)	5%	-0.31	1.60(0.18)	86%	-0.33	1.60(0.12)	100%	-0.33
Optimistic prior:										
$\bar{x}_1$ = ln(1.05)		1.11(0.16)	0%	-0.70	1.32(0.14)	57%	-0.53	1.49(0.12)	99%	-0.40
$\bar{x}_1$ = ln(0.96)		1.02(0.16)	0%	-0.78	1.24(0.14)	18%	-0.59	1.42(0.12)	95%	-0.45
$\bar{x}_1$ = ln(1.14)		1.19(0.16)	0%	-0.62	1.39(0.14)	84%	-0.47	1.56(0.12)	100%	-0.36
Women										
Frequentist		2.32(0.48)	40%	0.04	2.25(0.21)	98%	0.01	2.24(0.15)	100%	0.01
Non-informative prior		2.32(0.48)	41%	0.04	2.25(0.21)	98%	0.01	2.24(0.15)	100%	0.00
Sceptical prior:										
$\bar{x}_2$ = ln(1.44)		2.12(0.43)	39%	-0.05	2.05(0.18)	99%	-0.08	2.05(0.13)	100%	-0.08
$\bar{x}_2$ = ln(1.35)		2.07(0.43)	37%	-0.07	2.03(0.18)	99%	-0.09	2.02(0.13)	100%	-0.10
$\bar{x}_2$ = ln(1.53)		2.13(0.43)	40%	-0.04	2.08(0.18)	99%	-0.07	2.08(0.13)	100%	-0.07
Equivalent prior:	2.23									
$\bar{x}_2$ = ln(1.44)		1.83(0.34)	39%	-0.20	1.80(0.15)	100%	-0.21	1.80(0.10)	100%	-0.22
$\bar{x}_2$ = ln(1.35)		1.77(0.34)	32%	-0.23	1.74(0.15)	100%	-0.25	1.74(0.10)	100%	-0.25
$\bar{x}_2$ = ln(1.53)		1.88(0.34)	44%	-0.17	1.85(0.15)	100%	-0.19	1.85(0.10)	100%	-0.19
Optimistic prior:										
$\bar{x}_2$ = ln(1.44)		1.51(0.16)	98%	-0.39	1.71(0.13)	100%	-0.27	1.84(0.11)	100%	-0.19
$\bar{x}_2$ = ln(1.35)		1.42(0.16)	86%	-0.45	1.64(0.13)	100%	-0.31	1.79(0.11)	100%	-0.22
$\bar{x}_2$ = ln(1.53)		1.59(0.16)	100%	-0.34	1.77(0.13)	100%	-0.23	1.89(0.11)	100%	-0.16
Interaction										
Frequentist		1.02(0.77)	4%	-0.01	1.00(0.33)	5%	0.00	1.00(0.23)	5%	0.00
Non-informative prior		1.01(0.77)	4%	-0.01	1.01(0.33)	5%	-0.01	1.00(0.23)	5%	0.00
Sceptical prior:										
= ln(0.73)		0.95(0.69)	2%	-0.07	0.95(0.29)	3%	-0.05	0.94(0.20)	4%	-0.06
= ln(0.71)		0.95(0.69)	2%	-0.06	0.94(0.29)	4%	-0.06	0.93(0.20)	4%	-0.07
= ln(0.75)		0.96(0.69)	2%	-0.07	0.94(0.29)	4%	-0.06	0.94(0.20)	4%	-0.06
Equivalent prior:	1.00									
= ln(0.73)		0.86(0.55)	1%	-0.15	0.86(0.23)	4%	-0.15	0.86(0.16)	9%	-0.16
= ln(0.71)		0.85(0.55)	1%	-0.17	0.84(0.23)	4%	-0.17	0.84(0.16)	10%	-0.17
= ln(0.75)		0.87(0.55)	1%	-0.15	0.87(0.23)	3%	-0.14	0.86(0.16)	7%	-0.15
Optimistic prior:										
= ln(0.73)		0.74(0.22)	1%	-0.30	0.77(0.19)	13%	-0.26	0.81(0.16)	15%	-0.21
= ln(0.71)		0.72(0.22)	3%	-0.32	0.76(0.19)	17%	-0.28	0.79(0.16)	20%	-0.23
= ln(0.75)		0.75(0.22)	0%	-0.28	0.79(0.19)	9%	-0.24	0.82(0.16)	12%	-0.19

** Women contributed 50% to the overall sample size N.. In all simulations Bayesian RCT analysis are compared to a frequentist analysis ignoring prior information. Subgroup-specific and interaction effect results are evaluated using the treatment OR, with the frequentist estimator reflecting the data likelihood and the Bayesian estimators the mean of the posterior distribution. Results were further evaluated using standard errors (SE) or from a Bayesian perspective standard deviations (SD), the percentage of times an OR of 1 was excluded by the 95% confidence or credibility interval (% reject). Depending on the true OR the % reject should be interpreted as power or type 1 error rate. The non-informative prior is based on a normal distribution with a mean hyperparameter of 0 and variance of 10^6; the equivalent prior is based on a normal distribution with the reported ln OR $\bar{x}_1$ and variance hyperparameter σ_1^2 equal to the simulated data; the sceptical prior uses the same mean and variance hyperparameters only now multiplying the variance by 4; the optimistic prior again uses the same point estimates from nonrandomized data but now uses a variance of 0.027 for each subgroup. All simulation results are based on 10,000 repetitions.*

Table 5 Results of a simulated RCT exploring subgroup-specific and interaction effects using frequentist and Bayesian methods showing the impact of misspecified priors*.

Simulation	True OR	N = 1,000			N = 5,000			N = 10,000		
		OR(SE)	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*
Men										
Frequentist		0.99(0.72)	2%	-0.01	1.00(0.29)	5%	-0.01	1.00(0.21)	5%	0.00
Non-informative prior		1.00(0.72)	2%	0.00	1.00(0.29)	5%	0.00	1.00(0.21)	4%	0.00
Sceptical prior:										
μ_1 = ln(1.44)		1.08(0.64)	1%	0.07	1.08(0.26)	4%	0.07	1.07(0.18)	4%	0.07
μ_1 = ln(1.35)		1.07(0.64)	1%	0.07	1.06(0.26)	3%	0.07	1.06(0.18)	4%	0.06
μ_1 = ln(1.53)		1.08(0.64)	1%	0.08	1.09(0.26)	3%	0.08	1.09(0.18)	5%	0.08
Equivalent prior:	1.00									
μ_1 = ln(1.44)		1.19(0.51)	0%	0.18	1.20(0.21)	6%	0.18	1.20(0.15)	16%	0.18
μ_1 = ln(1.35)		1.16(0.51)	0%	0.15	1.16(0.21)	4%	0.15	1.16(0.15)	9%	0.15
μ_1 = ln(1.53)		1.24(0.51)	0%	0.22	1.23(0.21)	9%	0.22	1.24(0.15)	24%	0.21
Optimistic prior:										
μ_1 = ln(1.44)		1.41(0.16)	82%	0.34	1.32(0.14)	47%	0.34	1.25(0.13)	36%	0.22
μ_1 = ln(1.35)		1.33(0.16)	21%	0.28	1.26(0.14)	22%	0.28	1.20(0.13)	20%	0.18
μ_1 = ln(1.53)		1.49(0.16)	100%	0.40	1.38(0.14)	73%	0.40	1.30(0.13)	54%	0.26
Women										
Frequentist		2.34(0.49)	41%	0.05	2.25(0.21)	98%	0.01	2.24(0.15)	100%	0.00
Non-informative prior		2.30(0.48)	40%	0.03	2.25(0.21)	98%	0.01	2.24(0.15)	100%	0.00
Sceptical prior:										
μ_1 = ln(1.05)		1.98(0.43)	31%	-0.12	1.93(0.18)	97%	-0.14	1.92(0.13)	100%	-0.15
μ_1 = ln(0.96)		1.96(0.43)	30%	-0.13	1.89(0.18)	96%	-0.16	1.89(0.13)	100%	-0.17
μ_1 = ln(1.14)		2.03(0.43)	34%	-0.09	1.97(0.18)	98%	-0.13	1.96(0.13)	100%	-0.13
Equivalent prior:	2.23									
μ_1 = ln(1.05)		1.57(0.34)	13%	-0.35	1.54(0.15)	93%	-0.37	1.53(0.10)	100%	-0.37
μ_1 = ln(0.96)		1.50(0.34)	8%	-0.40	1.47(0.15)	86%	-0.42	1.47(0.10)	100%	-0.42
μ_1 = ln(1.14)		1.63(0.34)	18%	-0.31	1.60(0.15)	98%	-0.33	1.60(0.10)	100%	-0.33
Optimistic prior:										
μ_1 = ln(1.05)		1.14(0.16)	0%	-0.67	1.41(0.13)	91%	-0.46	1.60(0.11)	100%	-0.33
μ_1 = ln(0.96)		1.05(0.16)	0%	-0.75	1.33(0.13)	72%	-0.51	1.54(0.11)	100%	-0.37
μ_1 = ln(1.14)		1.22(0.16)	0%	-0.60	1.48(0.13)	97%	-0.41	1.66(0.11)	100%	-0.29
Interaction										
Frequentist		0.42(0.87)	16%	-0.06	0.44(0.36)	62%	-0.01	0.45(0.25)	89%	0.00
Non-informative prior		0.43(0.87)	15%	-0.03	0.45(0.36)	62%	-0.01	0.45(0.25)	89%	0.00
Sceptical prior:										
= ln(1.37)		0.55(0.78)	8%	0.20	0.56(0.32)	44%	0.22	0.56(0.23)	76%	0.22
= ln(1.41)		0.55(0.78)	8%	0.20	0.56(0.32)	43%	0.22	0.56(0.23)	75%	0.22
= ln(1.34)		0.53(0.78)	8%	0.17	0.55(0.32)	45%	0.21	0.56(0.23)	77%	0.22
Equivalent prior:										
= ln(1.37)	0.45	0.76(0.62)	1%	0.53	0.78(0.25)	8%	0.56	0.78(0.18)	21%	0.56
= ln(1.41)		0.78(0.62)	0%	0.55	0.79(0.25)	7%	0.57	0.79(0.18)	17%	0.57
= ln(1.34)		0.76(0.61)	1%	0.53	0.77(0.25)	9%	0.54	0.78(0.18)	22%	0.55
Optimistic prior:										
= ln(1.37)		1.24(0.22)	0%	1.02	0.94(0.19)	0%	0.74	0.78(0.17)	23%	0.55
= ln(1.41)		1.26(0.22)	0%	1.04	0.94(0.19)	0%	0.74	0.78(0.17)	22%	0.55
= ln(1.34)		1.22(0.22)	0%	1.00	0.93(0.19)	0%	0.73	0.78(0.17)	23%	0.55

** Women contributed 50% to the overall sample size N. In all simulations Bayesian RCT analysis are compared to a frequentist analysis ignoring prior information. Subgroup-specific and interaction effect results are evaluated using the treatment OR, with the frequentist estimator reflecting the data likelihood and the Bayesian estimators the mean of the posterior distribution. Results were further evaluated using standard errors (SE) or from a Bayesian perspective standard deviations (SD), the percentage of times an OR of 1 was excluded by the 95% confidence or credibility interval (% reject). Depending on the true OR the % reject should be interpreted as power or type 1 error rate. The non-informative prior is based on a normal distribution with a mean hyperparameter of 0 and variance of 10^6; the equivalent prior is based on a normal distribution with the reported ln OR μ_1 and variance hyperparameter σ_1^2 equal to the simulated data; the sceptical prior uses the same mean and variance hyperparameters only now multiplying the variance by 4; the optimistic prior again uses the same point estimates from nonrandomized data but now uses a variance of 0.027 for each subgroup. All simulation results are based on 10,000 repetitions.*

Table 6 Results of a simulated RCT exploring subgroup-specific and interaction effects using frequentist and Bayesian methods showing the impact of disbalance in gender subgroup sizes*.

Simulation	True OR	N = 1,000			N = 5,000			N = 10,000		
		OR(SE)	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*
Men										
Frequentist		1.00(0.53)	4%	0.00	1.00(0.22)	5%	0.00	1.00(0.16)	5%	0.00
Non-informative prior		1.00(0.53)	4%	0.00	1.00(0.22)	5%	0.00	1.00(0.16)	5%	0.00
Sceptical prior:										
θ_1 = ln(1.05)		1.01(0.47)	2%	0.01	1.01(0.20)	3%	0.01	1.01(0.14)	3%	0.01
θ_1 = ln(0.96)		1.00(0.47)	2%	0.00	0.99(0.20)	3%	-0.01	0.99(0.14)	3%	-0.01
θ_1 = ln(1.14)		1.02(0.47)	2%	0.03	1.02(0.20)	3%	0.02	1.03(0.14)	3%	0.02
Equivalent prior:	1.00									
θ_1 = ln(1.05)		1.03(0.37)	0%	0.02	1.02(0.16)	1%	0.02	1.03(0.11)	1%	0.02
θ_1 = ln(0.96)		0.98(0.37)	0%	-0.02	0.98(0.16)	1%	-0.02	0.98(0.11)	1%	-0.02
θ_1 = ln(1.14)		1.07(0.37)	0%	0.06	1.07(0.16)	1%	0.07	1.07(0.11)	3%	0.07
Optimistic prior:										
θ_1 = ln(1.05)		1.04(0.16)	0%	0.04	1.03(0.13)	0%	0.03	1.02(0.11)	1%	0.02
θ_1 = ln(0.96)		0.96(0.16)	0%	-0.04	0.97(0.13)	0%	-0.03	0.98(0.11)	1%	-0.02
θ_1 = ln(1.14)		1.13(0.16)	0%	0.12	1.09(0.13)	1%	0.09	1.07(0.11)	2%	0.06
Women										
Frequentist		2.07(0.91)	4%	-0.07	2.28(0.39)	58%	0.02	2.26(0.27)	88%	0.01
Non-informative prior		2.06(0.91)	4%	-0.08	2.30(0.39)	59%	0.03	2.26(0.27)	88%	0.01
Sceptical prior:										
θ_2 = ln(1.44)		1.92(0.81)	2%	-0.15	2.08(0.35)	60%	-0.07	2.06(0.24)	90%	-0.08
θ_2 = ln(1.35)		1.92(0.81)	2%	-0.15	2.05(0.35)	57%	-0.08	2.03(0.24)	89%	-0.09
θ_2 = ln(1.53)		1.95(0.81)	2%	-0.13	2.11(0.35)	61%	-0.06	2.09(0.24)	92%	-0.07
Equivalent prior:	2.23									
θ_2 = ln(1.44)		1.72(0.64)	0%	-0.26	1.81(0.27)	65%	-0.21	1.81(0.19)	97%	-0.21
θ_2 = ln(1.35)		1.67(0.64)	0%	-0.29	1.76(0.27)	59%	-0.24	1.75(0.19)	94%	-0.24
θ_2 = ln(1.53)		1.78(0.64)	1%	-0.23	1.87(0.27)	72%	-0.17	1.86(0.19)	98%	-0.18
Optimistic prior:										
θ_2 = ln(1.44)		1.45(0.16)	99%	-0.43	1.54(0.15)	99%	-0.37	1.62(0.14)	100%	-0.32
θ_2 = ln(1.35)		1.37(0.16)	49%	-0.49	1.46(0.15)	93%	-0.43	1.55(0.14)	99%	-0.37
θ_2 = ln(1.53)		1.54(0.16)	100%	-0.37	1.62(0.15)	100%	-0.32	1.70(0.14)	100%	-0.27
Interaction										
Frequentist		0.48(1.05)	6%	0.08	0.44(0.45)	45%	-0.03	0.44(0.31)	76%	-0.01
Non-informative prior		0.49(1.05)	6%	0.08	0.44(0.45)	46%	-0.03	0.44(0.31)	76%	-0.02
Sceptical prior:										
= ln(0.73)		0.53(0.94)	3%	0.17	0.48(0.40)	43%	0.08	0.49(0.28)	76%	0.09
= ln(0.71)		0.52(0.94)	3%	0.15	0.48(0.40)	43%	0.07	0.49(0.28)	77%	0.08
= ln(0.75)		0.53(0.94)	3%	0.16	0.49(0.40)	42%	0.08	0.49(0.28)	77%	0.09
Equivalent prior:	0.45									
= ln(0.73)		0.59(0.74)	1%	0.28	0.56(0.32)	41%	0.23	0.57(0.22)	83%	0.24
= ln(0.71)		0.59(0.74)	1%	0.27	0.56(0.32)	44%	0.21	0.56(0.22)	84%	0.22
= ln(0.75)		0.60(0.74)	1%	0.29	0.57(0.32)	38%	0.24	0.57(0.22)	80%	0.25
Optimistic prior:										
= ln(0.73)		0.72(0.22)	1%	0.47	0.67(0.20)	53%	0.40	0.63(0.18)	85%	0.34
= ln(0.71)		0.71(0.22)	3%	0.45	0.67(0.20)	54%	0.40	0.63(0.18)	85%	0.34
= ln(0.75)		0.73(0.22)	0%	0.49	0.67(0.20)	52%	0.40	0.63(0.18)	86%	0.34

** Women contributed 15% to the overall sample size N. In all simulations Bayesian RCT analysis are compared to a frequentist analysis ignoring prior information. Subgroup-specific and interaction effect results are evaluated using the treatment OR, with the frequentist estimator reflecting the data likelihood and the Bayesian estimators the mean of the posterior distribution. Results were further evaluated using standard errors (SE) or from a Bayesian perspective standard deviations (SD), the percentage of times an OR of 1 was excluded by the 95% confidence or credibility interval (% reject). Depending on the true OR the % reject should be interpreted as power or type 1 error rate. The non-informative prior is based on a normal distribution with a mean hyperparameter of 0 and variance of 10^6; the equivalent prior is based on a normal distribution with the reported ln OR θ_1 and variance hyperparameter σ^2 equal to the simulated data; the sceptical prior uses the same mean and variance hyperparameters only now multiplying the variance by 4; the optimistic prior again uses the same point estimates from nonrandomized data but now uses a variance of 0.027 for each subgroup. All simulation results are based on 10,000 repetitions.*

In scenarios in which 15% of the RCT sample were female, ,Bayesian analysis using informative priors outperformed the frequentist and non-informative prior analyses regarding power and type 1 error rate (**Table 6**). However, this difference between frequentist and informative Bayesian results was smaller than in the scenarios of equally sized subgroups. For example, in a sample size of 10,000 subjects, power of interaction tests was 85-86% using an optimistic prior and 76% using the non-informative prior. Type 1 error rates of the interaction tests in the presence of a main effect were always lower than the frequentist and non-informative prior analyses (which was 5%) (**Appendix table 2**). When there was no main effect and no interaction effect, type 1 error rates were similarly low, except for analyses using an optimistic prior; which had error rates up to 20% (**Appendix table 3**). In the scenarios in which the hyperparameters were in the opposite direction of the simulated data, power of the informative prior analysis was lower than that of a frequentist or non-informative analysis (**Appendix table 4**). For example, in a RCT of 10,000 subjects, power was 1% using an optimistic prior and 75% using a frequentist analysis.

Discussion

In this proof of principle study we showed that incorporating prior knowledge using Bayesian analysis of a post-launch RCT increases the power to detect interaction effects at the cost of increasing bias. For example, in one of our simulations (**Table 3**, N = 5,000) the power of the interaction test was 62% when using a frequentist approach compared to a power of 83% - 93% when using a Bayesian approach. This at the cost of bias between 0.28 and 0.33 (Bayesian analysis) compared to an unbiased estimate when using a frequentist approach. Alternatively phrased, to gain a similar power (62%) as the frequentist analysis, the Bayesian method required 22% - 48% less patients (i.e., 3,914 - 2,586 subjects instead of 5,000 subjects). This increase in power (or decrease in sample size) is clearly relevant for the detection of treatment effect modification and adverse events, for which RCTs are notoriously underpowered.

Previous research on Bayesian RCT analyses (7;13;14), some also focussing on effect modification (15-17), showed that non-informative priors could be used to guard against multiplicity. However, up till now none have focussed on incorporating nonrandomized study results of treatment effect modification as prior information for Bayesian analysis of RCT data.

Our study has several limitations. First, in all simulations, estimates from RCT data were considered the true treatment effect. Therefore, it should be no surprise that bias increased by adding prior information. In practice RCT estimates do not necessarily equal the true causal effect of treatment (18-20). However, due to the limitations of nonrandomized designs (notably confounding), RCTs seem to provide the most reliable approximation of the true treatment effect. Given that patients within subgroups are likely to be more similar, one might expect comparable subgroup-specific estimates from RCTs and nonrandomized studies. However, previously we showed that such similarity is very topic specific and small differences in subgroup-specific treatment estimates can cause interaction effects in opposing directions (21).

Second, Bayesian analysis of RCTs was evaluated based on power and type 1 error. These metrics are usually exclusive to frequentist analyses following the Neyman-Pearson perspective. However, given that at completion of a post-launch RCT decisions may have to be made on continued market access, e.g., by a drug regulatory agencies such as EMA of FDA, we feel that these metrics are relevant (22;23).

Third, Bayesian analyses have been criticized for basing prior information on subjective sources such as expert opinion. This study partly solved this using nonrandomized study results as prior information. However, there is still subjectivity about which nonrandomized studies to include. Similarly, weighing (i.e., the variance estimate) of the prior information is still subjective in such settings. Our study used three different informative priors (sceptical, equivalent and optimistic), which differed in variance hyperparameters. The sceptical and equivalent priors kept the variance hyperparameters proportional to the variances of the RCT effect estimates in order to prevent the prior from over influencing the data. At the same time, however, they also prevented the data from overwhelming the prior. Thus when the prior distributions do not equal the data likelihood distributions, analyses using these priors can never result in the true effect estimates (again assuming that the RCT estimate represents the truth).

Finally, it might seem inappropriate to combine information from nonrandomized studies with RCT data, because the results of RCTs are typically of higher quality. However, it also seems inappropriate to exclude information simply because treatment was not randomly allocated, especially in studies of adverse events (24-26). A more inclusive view on interventional re-

search seems sensible. Researchers should not only focus on RCTs, but include all available information from RCT and nonrandomized studies, and explore if and why results differ between designs (24-26). Following this inclusive strategy it seems obvious to include Bayesian methods for RCT analysis. This allows for a transparent way to weight the prior knowledge against the perceived chance of bias/validity.

Based on our study results we recommend the following. First, we showed that Bayesian RCT analyses, using informative priors, can increase power at the cost of an increase in bias. This trade-off between power and bias might be acceptable for interaction effects, because the power of interaction tests is notoriously limited (27). Second, when designing a RCT we suggest that if one is confident about the direction of an effect, an informative prior based on nonrandomized studies might be used. This will result in a decrease in the number of patients needed in the post-launch RCT, reduce costs and in posterior effect estimates reflecting all available (non-conflicting) evidence. In designing such a Bayesian post-launch RCT, similar to the more familiar sample size calculation, we suggest researchers to use simulations to gain insight in how the prior knowledge may influence the posterior distribution. If, after data collection, the RCT data unexpectedly contradict the prior distribution, the use of informative priors seems inappropriate. Our simulations showed a large increase in type 1 error rate and almost meaningless power. However, instead of simply ignoring the nonrandomized study results, as is current practice, we feel that in such settings it is essential to discuss and explore why RCTs and prior information differed. To be meaningful this discussion should go beyond a statement on the hierarchy of study design and the known shortcomings of nonrandomized studies (most notably potential for confounding). The possibility that the data and the prior information disagree might seem a shortcoming of Bayesian RCT analyses, but we feel that when the data contradict the prior information further research is needed. Because Bayesian methods will emphasize such contradictory findings and fuel the need for additional research, we see this as a virtue rather than a shortcoming.

Specifically in the Bayesian analysis of adverse events, it seems important to take in to account the type of adverse event (A or B) (28;29) when deciding on the hyperparameters of the informative prior. Type A adverse events result from the primary mechanism of action of the intervention, therefore confounding by indication seems more likely and this should be reflected in the prior distribution. However, for type B adverse events the underlying mecha-

nism is often unknown, thus decreasing the potential for confounding. While initiating a RCT to study adverse events (notably type B) seems unlikely because of the huge sample size required, pooling results from completed RCTs and nonrandomized studies appears highly advisable.

In conclusion, Bayesian analysis of post-launch RCTs using informative priors will likely bias estimates of treatment effects. However, when the prior information and the expected RCT results are in the same direction the decrease in variance can lead to relatively higher power of the Bayesian analysis with an acceptable degree of bias. This trade-off between power and bias might be acceptable for interaction effects because most RCTs only have limited power to detect these effects.

Reference List

1. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 2004 Mar;57(3):229-36.
2. Tsang R, Colley L, Lynd LD. Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. J Clin Epidemiol 2009 Jun;62(6):609-16.
3. Bazelier MT, de Vries F., Vestergaard P, Herings RM, Gallagher AM, Leufkens HG, et al. Risk of fracture with thiazolidinediones: an individual patient data meta-analysis. Front Endocrinol (Lausanne) 2013;4:11.
4. Home PD, Pocock SJ, Beck-Nielsen H, Curtis PS, Gomis R, Hanefeld M, et al. Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial. Lancet 2009 Jun 20;373(9681):2125-35.
5. Loke YK, Singh S, Furberg CD. Long-term use of thiazolidinediones and fractures in type 2 diabetes: a meta-analysis. CMAJ 2009 Jan 6;180(1):32-9.
6. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005 Apr;58(4):323-37.
7. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. Health Technol Assess 2000;4(38):1-130.
8. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. Stat Med 1995 Dec 30;14(24):2685-99.
9. Knol MJ, Le CS, Algra A, Vandenbroucke JP, Groenwold RH. Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression. Canadian Medical Association Journal 2012 May 15;184(8):895-9.
10. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. BMJ 2003 Jan 25;326(7382):219.
11. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. Stat Methods Med Res 2001 Aug;10(4):277-303.
12. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
13. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian Approaches to Randomized Trials. Journal of the Royal Statistical Society Series A (Statistics in Society) 1994 Jan 1;157(3):357-416.
14. Abrams K, Ashby D, Errington D. Simple Bayesian analysis in clinical trials: a tutorial. Control Clin Trials 1994 Oct;15(5):349-59.
15. Jones HE, Ohlssen DI, Neuenschwander B, Racine A, Branson M. Bayesian models for subgroup analysis in clinical trials. Clin Trials 2011 Apr;8(2):129-43.
16. Simon R. Bayesian subset analysis: application to studying treatment-by-gender interactions. Stat Med 2002 Oct 15;21(19):2909-16.
17. Chu H, Nie L, Cole SR. Estimating the relative excess risk due to interaction: a bayesian approach. Epidemiology 2011 Mar;22(2):242-8.
18. Hernan MA, Hernandez-Diaz S. Beyond the intention-to-treat in comparative effectiveness research. Clin Trials 2011 Oct 7.
19. Hernan MA. A definition of causal effect for epidemiological research. J Epidemiol Community Health 2004 Apr;58(4):265-71.
20. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000 Sep;11(5):550-60.
21. Schmidt AF, Rovers MM, Klungel OH, Hoes AW, Knol MJ, Nielen M, et al. Differences in interaction and subgroup-specific effects were observed between randomized and nonrandomized studies in three empirical examples. J Clin Epidemiol 2013 Jun;66(6):599-607.
22. Bayarri MJ, Berger JO. The Interplay of Bayesian and Frequentist Analysis. Statistical Science 2004 Feb;19(1):58-80.

23. LeBlond D. FDA Bayesian Statistics Guidance for Medical Device Clinical Trials—Application to Process Validation. Journal of Validation Technology 2010.
24. Vandenbroucke JP. Why do the results of randomised and observational studies differ? BMJ 2011;343:d7020.
25. Vandenbroucke JP. Commentary: the HRT story: vindication of old epidemiological theory. Int J Epidemiol 2004 Jun;33(3):456-7.
26. Lawlor DA, Davey SG, Ebrahim S. Commentary: the hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology? Int J Epidemiol 2004 Jun;33(3):464-7.
27. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. Stat Med 1983 Apr;2(2):243-51.
28. Grobbee DE, Hoes AW. Intervention Research: Unintended Effects. Clinical Epidemiology: Principles, Methods and Applications for Clinical Research. 2 ed. Burlington: Jones and Bartlett Learning; 2015. p. 181-214.
29. Vandenbroucke JP. What is the best evidence for determining harms of medical treatment? CMAJ 2006 Feb 28;174(5):645-6.

Appendix table 1 Results of a simulated RCT exploring subgroup-specific and interaction effects using frequentist and Bayesian methods in the presence of subgroup-specific and interaction effects of 1.00*.

Simulation	True OR	N = 1,000			N = 5,000			N = 10,000		
		OR(SE)	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*
Men										
Frequentist		0.99(0.71)	2%	-0.01	1.00(0.29)	5%	0.00	1.00(0.21)	5%	0.00
Non-informative prior		0.99(0.71)	2%	-0.01	1.01(0.29)	5%	0.01	1.00(0.21)	5%	0.00
Sceptical prior:										
$\bar{x}_2$ = ln(1.05)		1.01(0.64)	0%	0.01	1.01(0.26)	2%	0.01	1.01(0.18)	3%	0.01
$\bar{x}_2$ = ln(0.96)		0.99(0.64)	0%	-0.01	0.99(0.26)	3%	-0.01	0.99(0.18)	3%	-0.01
$\bar{x}_2$ = ln(1.14)		1.02(0.64)	0%	0.02	1.03(0.26)	3%	0.03	1.03(0.18)	3%	0.03
Equivalent prior:	1.00									
$\bar{x}_2$ = ln(1.05)		1.03(0.51)	0%	0.03	1.02(0.21)	0%	0.02	1.02(0.15)	1%	0.02
$\bar{x}_2$ = ln(0.96)		0.98(0.50)	0%	-0.02	0.98(0.21)	1%	-0.02	0.98(0.15)	1%	-0.02
$\bar{x}_2$ = ln(1.14)		1.07(0.51)	0%	0.07	1.07(0.21)	1%	0.06	1.07(0.15)	2%	0.07
Optimistic prior:										
$\bar{x}_2$ = ln(1.05)		1.05(0.16)	0%	0.05	1.04(0.14)	0%	0.04	1.03(0.13)	0%	0.03
$\bar{x}_2$ = ln(0.96)		0.96(0.16)	0%	-0.04	0.97(0.14)	0%	-0.03	0.97(0.13)	0%	-0.03
$\bar{x}_2$ = ln(1.14)		1.13(0.16)	0%	0.12	1.10(0.14)	0%	0.10	1.08(0.13)	1%	0.08
Women										
Frequentist		1.01(0.57)	3%	0.01	1.00(0.24)	5%	0.00	1.00(0.17)	5%	0.00
Non-informative prior		1.00(0.57)	3%	0.00	1.00(0.24)	4%	0.00	1.00(0.17)	5%	0.00
Sceptical prior:										
$\bar{x}_2$ = ln(1.44)		1.08(0.51)	1%	0.08	1.07(0.22)	4%	0.07	1.08(0.15)	5%	0.07
$\bar{x}_2$ = ln(1.35)		1.05(0.51)	1%	0.05	1.06(0.22)	4%	0.06	1.06(0.15)	4%	0.06
$\bar{x}_2$ = ln(1.53)		1.09(0.51)	2%	0.09	1.09(0.22)	4%	0.08	1.09(0.15)	6%	0.09
Equivalent prior:	1.00									
$\bar{x}_2$ = ln(1.44)		1.20(0.41)	1%	0.18	1.20(0.17)	10%	0.18	1.20(0.12)	27%	0.18
$\bar{x}_2$ = ln(1.35)		1.17(0.40)	0%	0.15	1.16(0.17)	6%	0.15	1.16(0.12)	16%	0.15
$\bar{x}_2$ = ln(1.53)		1.24(0.40)	1%	0.21	1.23(0.17)	15%	0.21	1.24(0.12)	40%	0.21
Optimistic prior:										
$\bar{x}_2$ = ln(1.44)		1.40(0.16)	72%	0.33	1.28(0.14)	41%	0.25	1.21(0.12)	30%	0.19
$\bar{x}_2$ = ln(1.35)		1.32(0.16)	24%	0.28	1.22(0.14)	21%	0.20	1.17(0.12)	17%	0.15
$\bar{x}_2$ = ln(1.53)		1.48(0.16)	98%	0.39	1.34(0.14)	62%	0.29	1.24(0.12)	44%	0.22
Interaction										
Frequentist		0.99(0.92)	3%	-0.01	1.00(0.38)	5%	0.00	1.00(0.27)	5%	0.00
Non-informative prior		0.99(0.92)	3%	-0.01	1.01(0.38)	5%	0.01	1.00(0.27)	5%	0.00
Sceptical prior:										
= ln(0.73)		0.93(0.83)	2%	-0.07	0.94(0.34)	3%	-0.07	0.94(0.24)	3%	-0.07
= ln(0.71)		0.94(0.82)	2%	-0.06	0.93(0.34)	3%	-0.07	0.93(0.24)	3%	-0.07
= ln(0.75)		0.93(0.83)	2%	-0.07	0.94(0.34)	3%	-0.06	0.94(0.24)	4%	-0.06
Equivalent prior:	1.00									
= ln(0.73)		0.86(0.66)	0%	-0.15	0.85(0.27)	3%	-0.16	0.85(0.19)	6%	-0.16
= ln(0.71)		0.84(0.65)	0%	-0.17	0.84(0.27)	3%	-0.17	0.84(0.19)	7%	-0.17
= ln(0.75)		0.86(0.65)	0%	-0.15	0.86(0.27)	2%	-0.15	0.86(0.19)	5%	-0.15
Optimistic prior:										
= ln(0.73)		0.75(0.22)	0%	-0.29	0.81(0.20)	4%	-0.21	0.85(0.17)	6%	-0.16
= ln(0.71)		0.73(0.22)	1%	-0.31	0.79(0.20)	7%	-0.23	0.84(0.17)	8%	-0.18
= ln(0.75)		0.77(0.22)	0%	-0.27	0.83(0.20)	3%	-0.19	0.87(0.17)	4%	-0.14

** Women contributed 50% to the overall sample size N. In all simulations Bayesian RCT analysis are compared to a frequentist analysis ignoring prior information. Subgroup-specific and interaction effect results are evaluated using the treatment OR, with the frequentist estimator reflecting the data likelihood and the Bayesian estimators the mean of the posterior distribution. Results were further evaluated using standard errors (SE) or from a Bayesian perspective standard deviations (SD), the percentage of times an OR of 1 was excluded by the 95% confidence or credibility interval (% reject). Depending on the true OR the % reject should be interpreted as power or type 1 error rate. The non-informative prior is based on a normal distribution with a mean hyperparameter of 0 and variance of 10^6; the equivalent prior is based on a normal distribution with the reported ln OR $\bar{x}_2$ and variance hyperparameter σ_2^2 equal to the simulated data; the sceptical prior uses the same mean and variance hyperparameters only now multiplying the variance by 4; the optimistic prior again uses the same point estimates from nonrandomized data but now uses a variance of 0.027 for each subgroup. All simulation results are based on 10,000 repetitions.*

Appendix table 2 Results of a simulated RCT exploring subgroup-specific and interaction effects using frequentist and Bayesian methods showing the impact of disbalance in gender subgroup sizes in the presence an interaction effect of 1.00*.

Simulation	True OR	N = 1,000			N = 5,000			N = 10,000		
		OR(SE)	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*
Men										
Frequentist		2.34(0.45)	48%	0.05	2.24(0.19)	99%	0.01	2.24(0.13)	100%	0.00
Non-informative prior		2.31(0.45)	47%	0.04	2.24(0.19)	99%	0.01	2.24(0.13)	100%	0.00
Sceptical prior:										
$\bar{x}_1$ = ln(1.05)		1.97(0.40)	38%	-0.12	1.93(0.17)	99%	-0.15	1.92(0.12)	100%	-0.15
$\bar{x}_1$ = ln(0.96)		1.94(0.40)	35%	-0.14	1.89(0.17)	98%	-0.17	1.89(0.12)	100%	-0.17
$\bar{x}_1$ = ln(1.14)		2.03(0.40)	41%	-0.10	1.96(0.17)	99%	-0.13	1.96(0.12)	100%	-0.13
Equivalent prior:	2.23									
$\bar{x}_1$ = ln(1.05)		1.56(0.31)	17%	-0.36	1.54(0.14)	97%	-0.37	1.53(0.10)	100%	-0.38
$\bar{x}_1$ = ln(0.96)		1.49(0.31)	11%	-0.40	1.47(0.14)	91%	-0.42	1.47(0.10)	100%	-0.42
$\bar{x}_1$ = ln(1.14)		1.63(0.31)	24%	-0.31	1.60(0.14)	99%	-0.33	1.60(0.10)	100%	-0.33
Optimistic prior:										
$\bar{x}_1$ = ln(1.05)		1.15(0.15)	0%	-0.66	1.45(0.12)	96%	-0.43	1.65(0.10)	99%	-0.30
$\bar{x}_1$ = ln(0.96)		1.06(0.15)	0%	-0.74	1.37(0.12)	86%	-0.49	1.59(0.10)	95%	-0.34
$\bar{x}_1$ = ln(1.14)		1.24(0.15)	1%	-0.59	1.52(0.12)	99%	-0.39	1.70(0.10	100%	-0.27
Women										
Frequentist		2.05(0.90)	4%	-0.09	2.31(0.39)	60%	0.01	2.27(0.27)	88%	0.02
Non-informative prior		2.07(0.90)	4%	-0.08	2.30(0.39)	58%	0.01	2.26(0.27)	88%	0.01
Sceptical prior:										
$\bar{x}_2$ = ln(1.44)		1.91(0.81)	2%	-0.15	2.09(0.35)	60%	-0.08	2.06(0.24)	91%	-0.08
$\bar{x}_2$ = ln(1.35)		1.88(0.81)	2%	-0.17	2.07(0.35)	58%	-0.09	2.04(0.24)	89%	-0.09
$\bar{x}_2$ = ln(1.53)		1.93(0.81)	3%	-0.14	2.13(0.35)	62%	-0.07	2.10(0.24)	92%	-0.06
Equivalent prior:	2.23									
$\bar{x}_2$ = ln(1.44)		1.72(0.64)	0%	-0.26	1.81(0.27)	66%	-0.21	1.80(0.19)	96%	-0.22
$\bar{x}_2$ = ln(1.35)		1.67(0.64)	0%	-0.29	1.76(0.27)	58%	-0.25	1.75(0.19)	94%	-0.24
$\bar{x}_2$ = ln(1.53)		1.79(0.64)	1%	-0.22	1.87(0.27)	72%	-0.19	1.86(0.19)	98%	-0.18
Optimistic prior:										
$\bar{x}_2$ = ln(1.44)		1.45(0.16)	99%	-0.43	1.54(0.15)	99%	-0.27	1.62(0.14)	100%	-0.32
$\bar{x}_2$ = ln(1.35)		1.37(0.16)	49%	-0.49	1.46(0.15)	94%	-0.31	1.55(0.14)	99%	-0.37
$\bar{x}_2$ = ln(1.53)		1.54(0.16)	100%	-0.37	1.62(0.15)	100%	-0.23	1.69(0.14)	100%	-0.28
Interaction										
Frequentist		1.15(1.01)	3%	0.14	0.97(0.43)	5%	-0.03	0.99(0.30)	5%	-0.01
Non-informative prior		1.12(1.01)	2%	0.11	0.98(0.43)	5%	-0.03	0.99(0.30)	5%	-0.01
Sceptical prior:										
$\bar{x}_3$= ln(0.73)		1.03(0.91)	1%	0.03	0.92(0.39)	3%	-0.08	0.93(0.27)	3%	-0.07
$\bar{x}_3$= ln(0.71)		1.03(0.91)	1%	0.03	0.91(0.39)	3%	-0.09	0.93(0.27)	3%	-0.07
$\bar{x}_3$= ln(0.75)		1.05(0.91)	1%	0.05	0.92(0.39)	3%	-0.08	0.93(0.27)	3%	-0.07
Equivalent prior:	2.23									
$\bar{x}_3$= ln(0.73)		0.91(0.72)	0%	-0.10	0.85(0.31)	1%	-0.17	0.85(0.21)	3%	-0.16
$\bar{x}_3$= ln(0.71)		0.89(0.72)	0%	-0.12	0.84(0.31)	1%	-0.18	0.84(0.21)	4%	-0.18
$\bar{x}_3$= ln(0.75)		0.91(0.72)	0%	-0.09	0.85(0.31)	1%	-0.16	0.86(0.21)	3%	-0.15
Optimistic prior:										
$\bar{x}_3$= ln(0.73)		0.79(0.22)	0%	-0.23	0.94(0.20)	0%	-0.06	1.02(0.17)	0%	0.02
$\bar{x}_3$= ln(0.71)		0.78(0.22)	0%	-0.25	0.94(0.20)	0%	-0.06	1.03(0.17)	0%	0.03
$\bar{x}_3$= ln(0.75)		0.80(0.22)	0%	-0.22	0.94(0.20)	0%	-0.07	1.01(0.17)	0%	0.01

** Women contributed 15% to the overall sample size N. In all simulations Bayesian RCT analysis are compared to a frequentist analysis ignoring prior information. Subgroup-specific and interaction effect results are evaluated using the treatment OR, with the frequentist estimator reflecting the data likelihood and the Bayesian estimators the mean of the posterior distribution. Results were further evaluated using standard errors (SE) or from a Bayesian perspective standard deviations (SD), the percentage of times an OR of 1 was excluded by the 95% confidence or credibility interval (% reject). Depending on the true OR the % reject should be interpreted as power or type 1 error rate. The non-informative prior is based on a normal distribution with a mean hyperparameter of 0 and variance of 10^6; the equivalent prior is based on a normal distribution with the reported ln OR $\bar{x}_i$ and variance hyperparameter s_i^2 equal to the simulated data; the sceptical prior uses the same mean and variance hyperparameters only now multiplying the variance by 4; the optimistic prior again uses the same point estimates from nonrandomized data but now uses a variance of 0.027 for each subgroup. All simulation results are based on 10,000 repetitions.*

Appendix table 3 Results of a simulated RCT exploring subgroup-specific and interaction effects using frequentist and Bayesian methods showing the impact of disbalance in gender subgroup sizes in the presence of subgroup-specific and interaction effects of 1.00*.

Simulation	True OR	N = 1,000			N = 5,000			N = 10,000		
		OR(SE)	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*
Men										
Frequentist		1.00(0.53)	3%	0.00	1.00(0.22)	5%	0.00	1.00(0.16)	5%	0.00
Non-informative prior		1.00(0.53)	4%	0.00	1.00(0.22)	5%	0.00	1.00(0.16)	5%	0.00
Sceptical prior:										
$\bar{x}_1$ = ln(1.05)		1.01(0.47)	1%	0.01	1.01(0.20)	2%	0.01	1.01(0.14)	3%	0.01
$\bar{x}_1$ = ln(0.96)		0.98(0.47)	2%	-0.02	0.99(0.20)	2%	-0.01	0.99(0.14)	3%	-0.01
$\bar{x}_1$ = ln(1.14)		1.03(0.47)	2%	0.03	1.03(0.20)	3%	0.02	1.03(0.14)	3%	0.03
Equivalent prior:	1.00									
$\bar{x}_1$ = ln(1.05)		1.03(0.37)	0%	0.02	1.03(0.16)	1%	0.02	1.02(0.11)	1%	0.02
$\bar{x}_1$ = ln(0.96)		0.98(0.37)	0%	-0.02	0.98(0.16)	1%	-0.02	0.98(0.11)	1%	-0.02
$\bar{x}_1$ = ln(1.14)		1.07(0.37)	0%	0.07	1.07(0.16)	1%	0.07	1.07(0.11)	3%	0.07
Optimistic prior:										
$\bar{x}_1$ = ln(1.05)		1.05(0.16)	0%	0.04	1.03(0.13)	0%	0.03	1.02(0.11)	1%	0.02
$\bar{x}_1$ = ln(0.96)		0.96(0.16)	0%	-0.04	0.97(0.13)	0%	-0.03	0.98(0.11)	1%	-0.02
$\bar{x}_1$ = ln(1.14)		1.13(0.16)	0%	0.12	1.09(0.13)	1%	0.08	1.06(0.11)	2%	0.06
Women										
Frequentist		1.02(1.04)	0%	0.02	1.00(0.45)	4%	0.00	1.00(0.31)	4%	0.00
Non-informative prior		1.00(1.04)	0%	0.00	1.01(0.45)	4%	0.01	1.00(0.31)	5%	0.00
Sceptical prior:										
$\bar{x}_2$ = ln(1.44)		1.07(0.93)	0%	0.07	1.07(0.41)	2%	0.07	1.08(0.28)	3%	0.07
$\bar{x}_2$ = ln(1.35)		1.06(0.93)	0%	0.06	1.06(0.41)	2%	0.06	1.06(0.28)	3%	0.06
$\bar{x}_2$ = ln(1.53)		1.10(0.93)	0%	0.10	1.09(0.41)	3%	0.09	1.09(0.28)	4%	0.08
Equivalent prior:	1.00									
$\bar{x}_2$ = ln(1.44)		1.21(0.73)	0%	0.19	1.20(0.32)	2%	0.18	1.20(0.22)	5%	0.18
$\bar{x}_2$ = ln(1.35)		1.17(0.74)	0%	0.15	1.16(0.32)	1%	0.15	1.16(0.22)	3%	0.15
$\bar{x}_2$ = ln(1.53)		1.24(0.74)	0%	0.22	1.24(0.32)	3%	0.21	1.24(0.22)	8%	0.21
Optimistic prior:										
$\bar{x}_2$ = ln(1.44)		1.43(0.16)	97%	0.35	1.38(0.15)	63%	0.32	1.33(0.15)	50%	0.28
$\bar{x}_2$ = ln(1.35)		1.34(0.16)	10%	0.29	1.30(0.15)	24%	0.26	1.26(0.15)	23%	0.23
$\bar{x}_2$ = ln(1.53)		1.51(0.16)	100%	0.41	1.46(0.15)	93%	0.38	1.39(0.15)	76%	0.33
Interaction										
Frequentist		0.99(1.18)	1%	-0.01	1.00(0.51)	5%	0.00	1.00(0.35)	5%	0.00
Non-informative prior		1.01(1.17)	1%	0.01	0.99(0.51)	4%	-0.01	1.01(0.35)	5%	0.01
Sceptical prior:										
= ln(0.73)		0.94(1.05)	0%	-0.06	0.94(0.45)	3%	-0.06	0.94(0.31)	3%	-0.06
= ln(0.71)		0.93(1.05)	0%	-0.08	0.93(0.45)	3%	-0.07	0.94(0.31)	3%	-0.07
= ln(0.75)		0.94(1.05)	0%	-0.07	0.94(0.45)	3%	-0.06	0.94(0.31)	3%	-0.06
Equivalent prior:	1.00									
= ln(0.73)		0.85(0.83)	0%	-0.16	0.86(0.36)	1%	-0.16	0.85(0.25)	3%	-0.16
= ln(0.71)		0.84(0.83)	0%	-0.18	0.84(0.36)	1%	-0.17	0.84(0.25)	3%	-0.17
= ln(0.75)		0.86(0.83)	0%	-0.15	0.86(0.36)	1%	-0.15	0.86(0.25)	3%	-0.15
Optimistic prior:										
= ln(0.73)		0.73(0.23)	0%	-0.31	0.75(0.20)	12%	-0.29	0.77(0.18)	17%	-0.26
= ln(0.71)		0.72(0.23)	1%	-0.33	0.75(0.20)	12%	-0.29	0.78(0.18)	15%	-0.25
= ln(0.75)		0.74(0.23)	0%	-0.30	0.75(0.20)	13%	-0.29	0.76(0.18)	20%	-0.27

** Women contributed 15% to the overall sample size N. In all simulations Bayesian RCT analysis are compared to a frequentist analysis ignoring prior information. Subgroup-specific and interaction effect results are evaluated using the treatment OR, with the frequentist estimator reflecting the data likelihood and the Bayesian estimators the mean of the posterior distribution. Results were further evaluated using standard errors (SE) or from a Bayesian perspective standard deviations (SD), the percentage of times an OR of 1 was excluded by the 95% confidence or credibility interval (% reject). Depending on the true OR the % reject should be interpreted as power or type 1 error rate. The non-informative prior is based on a normal distribution with a mean hyperparameter of 0 and variance of 10^6; the equivalent prior is based on a normal distribution with the reported ln OR $\bar{x}_1$ and variance hyperparameter t_1^2 equal to the simulated data; the sceptical prior uses the same mean and variance hyperparameters only now multiplying the variance by 4; the optimistic prior again uses the same point estimates from nonrandomized data but now uses a variance of 0.027 for each subgroup. All simulation results are based on 10,000 repetitions.*

Appendix table 4 Results of a simulated RCT exploring subgroup-specific and interaction effects using frequentist and Bayesian methods showing the impact of misspecified priors and disbalance in gender subgroup sizes*.

Simulation	True OR	N = 1,000			N = 5,000			N = 10,000		
		OR(SE)	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*	*OR(SE)*	*% reject*	*Bias*
Men										
Frequentist		1.00(0.53)	4%	0.00	1.00(0.22)	5%	0.00	1.00(0.16)	5%	0.00
Non-informative prior		1.00(0.53)	4%	0.00	1.00(0.22)	5%	0.00	1.00(0.16)	5%	0.00
Sceptical prior:										
$\bar{x}_c$ = ln(1.44)		1.07(0.47)	2%	0.07	1.07(0.20)	4%	0.07	1.08(0.14)	5%	0.07
$\bar{x}_c$ = ln(1.35)		1.06(0.47)	2%	0.06	1.06(0.20)	4%	0.06	1.06(0.14)	5%	0.06
$\bar{x}_c$ = ln(1.53)		1.09(0.47)	2%	0.08	1.09(0.20)	5%	0.08	1.09(0.14)	6%	0.09
Equivalent prior:	1.00									
$\bar{x}_c$ = ln(1.44)		1.20(0.37)	1%	0.18	1.20(0.16)	13%	0.18	1.20(0.11)	33%	0.18
$\bar{x}_c$ = ln(1.35)		1.16(0.37)	1%	0.15	1.16(0.16)	7%	0.15	1.16(0.11)	19%	0.15
$\bar{x}_c$ = ln(1.53)		1.24(0.37)	1%	0.22	1.24(0.16)	19%	0.21	1.24(0.11)	47%	0.21
Optimistic prior:										
$\bar{x}_c$ = ln(1.44)		1.39(0.16)	68%	0.33	1.27(0.13)	39%	0.24	1.19(0.11)	28%	0.18
$\bar{x}_c$ = ln(1.35)		1.31(0.16)	24%	0.27	1.21(0.13)	20%	0.19	1.15(0.11)	17%	0.14
$\bar{x}_c$ = ln(1.53)		1.47(0.16)	97%	0.39	1.32(0.13)	59%	0.28	1.22(0.11)	40%	0.20
Women										
Frequentist		2.06(0.90)	4%	-0.08	2.30(0.39)	59%	0.03	2.26(0.27)	88%	0.01
Non-informative prior		2.08(0.90)	4%	-0.07	2.31(0.39)	59%	0.04	2.27(0.27)	88%	0.02
Sceptical prior:										
$\bar{x}_t$ = ln(1.05)		1.79(0.81)	2%	-0.22	1.97(0.35)	50%	-0.13	1.94(0.24)	84%	-0.14
$\bar{x}_t$ = ln(0.96)		1.74(0.81)	1%	-0.25	1.92(0.35)	47%	-0.15	1.90(0.24)	82%	-0.16
$\bar{x}_t$ = ln(1.14)		1.83(0.81)	2%	-0.20	2.00(0.35)	53%	-0.11	1.97(0.24)	85%	-0.12
Equivalent prior:	2.23									
$\bar{x}_t$ = ln(1.05)		1.47(0.64)	0%	-0.42	1.55(0.27)	28%	-0.36	1.54(0.19)	70%	-0.37
$\bar{x}_t$ = ln(0.96)		1.40(0.64)	0%	-0.46	1.49(0.27)	20%	-0.40	1.47(0.19)	55%	-0.41
$\bar{x}_t$ = ln(1.14)		1.54(0.64)	0%	-0.37	1.62(0.27)	38%	-0.32	1.60(0.19)	80%	-0.33
Optimistic prior:										
$\bar{x}_t$ = ln(1.05)		1.07(0.16)	0%	-0.73	1.18(0.15)	0%	-0.64	1.29(0.14)	39%	-0.55
$\bar{x}_t$ = ln(0.96)		0.98(0.16)	0%	-0.82	1.09(0.15)	0%	-0.71	1.21(0.14)	6%	-0.61
$\bar{x}_t$ = ln(1.14)		1.16(0.16)	0%	-0.65	1.26(0.15)	8%	-0.57	1.37(0.14)	76%	-0.49
Interaction										
Frequentist		0.49(1.05)	6%	0.08	0.43(0.45)	46%	-0.03	0.44(0.31)	75%	-0.01
Non-informative prior		0.48(1.05)	6%	0.06	0.43(0.45)	46%	-0.04	0.44(0.31)	76%	-0.02
Sceptical prior:										
= ln(1.37)		0.60(0.94)	2%	0.29	0.55(0.40)	29%	0.20	0.55(0.28)	57%	0.21
= ln(1.41)		0.61(0.94)	2%	0.31	0.55(0.40)	29%	0.21	0.56(0.28)	57%	0.22
= ln(1.34)		0.60(0.94)	2%	0.28	0.54(0.40)	29%	0.19	0.55(0.28)	58%	0.21
Equivalent prior:										
= ln(1.37)	0.45	0.82(0.74)	0%	0.60	0.77(0.32)	4%	0.54	0.78(0.22)	10%	0.55
= ln(1.41)		0.83(0.75)	0%	0.61	0.78(0.32)	3%	0.55	0.79(0.22)	9%	0.56
= ln(1.34)		0.80(0.74)	0%	0.58	0.76(0.32)	4%	0.53	0.77(0.22)	12%	0.54
Optimistic prior:										
= ln(1.37)		1.30(0.22)	0%	1.06	1.07(0.20)	0%	0.87	0.92(0.18)	0%	0.72
= ln(1.41)		1.34(0.22)	1%	1.09	1.11(0.20)	0%	0.91	0.95(0.18)	0%	0.76
= ln(1.34)		1.27(0.22)	0%	1.04	1.04(0.20)	0%	0.85	0.89(0.18)	1%	0.69

** Women contributed 15% to the overall sample size N. In all simulations Bayesian RCT analysis are compared to a frequentist analysis ignoring prior information. Subgroup-specific and interaction effect results are evaluated using the treatment OR, with the frequentist estimator reflecting the data likelihood and the Bayesian estimators the mean of the posterior distribution. Results were further evaluated using standard errors (SE) or from a Bayesian perspective standard deviations (SD), the percentage of times an OR of 1 was excluded by the 95% confidence or credibility interval (% reject). Depending on the true OR the % reject should be interpreted as power or type 1 error rate. The non-informative prior is based on a normal distribution with a mean hyperparameter of 0 and variance of 10^6; the equivalent prior is based on a normal distribution with the reported ln OR $\bar{x}_t$ and variance hyperparameter s_t^2 equal to the simulated data; the sceptical prior uses the same mean and variance hyperparameters only now multiplying the variance by 4; the optimistic prior again uses the same point estimates from nonrandomized data but now uses a variance of 0.027 for each subgroup. All simulation results are based on 10,000 repetitions.*

Part III

Bridging the gap between clinical studies and individual patient care

CHAPTER 4

Prognostic factors of early metastasis and mortality in dogs with appendicular osteosarcoma after receiving surgery an individual patient data meta-analysis

A F Schmidt, M Nielen, O H Klungel, A W Hoes, A de Boer R H H Groenwold
J Kirpensteijn for the VSSO Investigators

Preventive Veterinary Medicine 2013. Nov 112(3-4);414-422
doi: 10.1016/j.prevetmed.2013.08.011

Abstract

Recently an aggregated data meta-analysis showed that serum alkaline phosphatase (SALP) and proximal humerus location are predictors for shorter survival in canine osteosarcoma. To identify additional prognostic factors of mortality and metastasis an individual patient data meta-analysis (IPDMA) was conducted. Individual patient data from 20 studies, identified via the VSSO society, were pooled. Univariable and multivariable hazard ratios (HR) for metastasis and mortality were assessed, using stratified Cox models. The study included 1405 dogs who received surgical treatment, of which the metastasis status was measured in 1155 dogs and mortality status in 1336 dogs, median survival was 256 days. High versus normal SALP and weight (kg) were associated with an increase in hazard of metastasis [HR 1.34 (95%CI 1.07; 1.68) and HR 1.02 (per kg increase) (95%CI 1.01; 1.03)] and for mortality [HR 1.43 (95%CI 1.16; 1.77) and HR 1.02 (95%CI 1.01; 1.02)]. Distal radius tumor was associated with a lower hazard of metastasis compared to other locations: HR 0.75 (95%CI 0.58; 0.96). Proximal humerus and distal femur or proximal tibia location were related with an increased mortality: HR 1.53 (95%CI 1.26; 1.84) and HR 1.23 (95%CI 1.01; 1.49) compared to other locations. Older age (years) was associated with a higher hazard for mortality [HR 1.06 per year (95%CI 1.03; 1.09)] but not for metastasis: HR 1.03 (95%CI 0.99; 1.06). These results confirm findings from a recent aggregated data meta-analysis and (in addition) showed that tumor location, SALP, weight were prognostic factors for both mortality and metastasis. Age was a prognostic factor for mortality but not for metastasis.

Introduction

Osteosarcoma (OS) is a malignant tumor of mesenchymal origin that produces osteoid. Similarities between human and canine OS are striking and include the bimodal age distribution, the high incidence of morbidity and mortality, the site of the tumor, histologic features and the response to the various treatment modalities (1;2). The biggest difference is that OS is much more common in dogs than in people. The majority of canine primary bone tumors can be classified as OS, which predominantly occurs in large and giant breeds (3-7). OS dogs, treated only by amputation, have a median survival time of five months or less, with the majority succumbing to metastatic disease (4;8;9). Due to advances in disease management overall survival can be extended to 1 year (9). Given the increased treatment options, such as adjuvant chemotherapy, 'limb-sparing' surgery and radio-ablative methods, it has become even more important to differentiate between dogs with a worse and relatively improved prognosis. Numerous studies have explored the prognostic value of, for example gender, neuter status, age or serum alkaline phosphatase (SALP), but these studies have important limitations. Most notably, the relatively small number of patients included in these studies precludes precise estimation of the prognostic consequences of these factors. A possible solution for this is collecting and pooling reported prognostic associations from individual studies. Recently, Boerman et al. (10) conducted an aggregated data meta-analysis. This meta-analysis showed that elevated SALP and location of OS in the (proximal) humerus are associated with a shorter disease free survival time. However, as Boerman acknowledges, the included studies did not analyze SALP and tumor location consistently; some explored the univariable association, while other used multivariable methods. Furthermore, other characteristics, for example age, weight and neuter status, could not be analyzed because these were not reported for all studies.

An alternative to pooling the aggregated data is to acquire the individual patient data files. An individual patient data meta-analysis (IPDMA) permits for more uniform analyses with regard to follow-up time, categorization of variables, missing values and analysis methods used (11;12). Furthermore, individual patient data allows exploring associations not reported in the original publications. Consequently, such prognostic IPDMAs are powerful tools to identify prognostic factors and subgroups of patients with different prognoses. We conducted an IPDMA in order to estimate the independent prognostic value of gender, neuter status, age, weight, breed, tumor location and SALP in predicting mortality or metastasis in canine OS.

Methods

Inclusion of individual patient data and assessment of data quality.

To explore the relations between patient characteristic and (DF) survival we identified studies via the Veterinary Society of Surgical Oncology (VSSO). In January 2012 a call for collaboration was send out to VSSO members and other veterinary researchers. We attempted to include data from as many different researchers and institutes as possible. No a priori sample size calculations were performed. Data was deemed eligible if baseline patient characteristics of OS dogs and time to event (death or metastasis) were recorded. To reduce the possibility of publication bias (13), published and unpublished studies were both eligible. Impossible or unlikely data entries were explored and remaining irregularities were discussed with the original investigators. Data were collected on gender, neuter status, age (years), weight (kg), breed (Rottweiler, Golden Retriever, Labrador Retriever, Greyhound, Doberman, Irish Setter, mixed breeds, and other breeds), tumor location (proximal humerus, distal femur or proximal tibia, distal radius, and other locations), dichotomous SALP (using study specific cut-off values for high and normal SALP levels), surgery (limb-sparing, amputations), chemotherapy (no chemotherapy, cisplatin, lobaplatin or carboplatin, doxorubicin, doxorubicin combinations), and other treatments. To prevent low cell counts we refrained from using finer categories for breed, tumor location and chemotherapy. SALP status (at baseline) was dichotomized to follow clinical practice and because continues SALP showed a positive linear sloped relationship with the outcomes that stabilized to a flat slope at high SALP values. Patients who did not receive surgery, mostly due to euthanasia (n = 197), who received an infrequently used chemotherapeutic protocol (n = 13), or who received radiation therapy (n =11) were excluded from all analyses.

Data analysis

To illustrate how patient characteristics were related to mortality or metastases at the clinically relevant time points of 5 and 12 months (4;8;9), we stratified baseline characteristics according to the outcome status (mortality and metastasis) at these time points. Univariable associations were estimated using a stratified Cox proportional hazards model (14). All the models were stratified by study to account for possible differences in baseline hazard. If a variable was missing for a certain patient, that patient was excluded from the univariable ana-

lysis (i.e., listwise deletion).

We then performed a multivariable Cox proportional hazards analysis (stratified by study) to assess the independent associations between prognostic factors and outcome. Subjects were censored if they were lost to follow-up or died (censoring for mortality was only applied in models using the metastasis outcome). Associations are given as hazard ratios (HRs) with 95% confidence intervals (95%CI) and p-values using an alpha of 0.05. For categorical variables a p-value for trend was computed and the individual associations were only explored if this overall test was significant (i.e., $p < 0.05$). All models were corrected for chemotherapy status. Variable selection for the multivariable model was based on prior knowledge, no data driven selection method was used (i.e., no stepwise selection). The proportional hazard (PH) assumption of the Cox models was checked based on Schoenfield residuals (14). For the continues variables weight and age, a linear relation with the outcome was assessed using restricted cubic splines plots (15); relations appeared to be linear. To determine how well the multivariable models discriminate between subjects with a short time to event and subjects with a longer time to event, the c-statistic (i.e., area under the receiver operator characteristic curve) was calculated (Steyerberg et al., 2010). The c-statistic represents the proportion of pairs of subjects where the subject with the longest observed time to event also received the longest predicted time to event; the c-statistic varies from 0.5 (no discrimination) and 1 (perfect discrimination).

In the multivariable analysis missing values were imputed, across studies, based on the aregImputation algorithm with ten imputations (16;17). In each of the ten imputed datasets, a multivariable Cox proportional hazards analyses was conducted and results were pooled using Rubin´s rule (18). The study by Sottnik et al., (19) (n = 69) did not provide information on time until death. Similarly, the Phillips et al., (20) (n=156) and Berg et al., (21) (n = 94) studies did not record information on time to metastasis. These studies were only used for the analyses they provided data for.

Sensitivity analysis.

Effect estimates of the association between prognostic factors and non-mortality outcomes, such as metastases, are potentially biased by competing risks. (22). In the case of time till metastasis a subject can be censored due to competing risks such as death (informative

censoring). In such a case it is obviously wrong to assume that the subjects will get the event somewhere in the future (which is assumed when censoring). If this informative censoring is systematically related to a specific group (e.g., high SALP) censoring the deceased subjects inflates the cumulative incidence and competing risk occurs. Instead of censoring subjects who die before developing a metastasis a competing risk analysis keeps these subjects in the denominator, decreasing the cumulative incidence. In canine OS, most subjects first experience a metastasis before dying; nevertheless we conducted competing risk analyses to assess how much this impacted our results (23). We also assessed the impact of missing observations through a sensitivity analysis in which a multivariate analysis was conducted using only those subjects with completely observed data. Additionally, to determine whether including subjects from small studies or unpublished studies biased our results we performed all analyses separately for large (50 or more subjects) and small studies (less than 50 subjects) and also stratified for publication status (i.e., if the study was published or not). To determine how influential the inclusion was of subjects who were not treated with chemotherapy, all analyses were also performed after excluding these patients. Finally, we assessed the impact of grouping lobaplatin and carboplatin in one group by repeated the analyses using separate categories for these chemotherapies.

All analyses were carried out with the R statistical package version 3.0.0 (24), the survival (25), the rms (26) and the Hmisc (16) packages.

Results

Data from 20 studies were included in this IPDMA, of which 11 studies were previously published (19-21;27-33). 19 studies reported solely on large breed dogs, while the unpublished study of Dr. Amsellem included 36 small breed canines. Characteristics of these studies are presented in **Table 1**. Eighteen studies (1155 patients) provided data on metastasis status and nineteen studies (1336 patients) provided information on mortality.

Univariable analysis

At 5 months, in the 550 dogs without missing data, 153 dogs developed a metastasis (**Table 2**). High weight (per kg) was related to an increase in metastasis hazard: hazard ratio (HR) 1.02 (95%CI 1.00; 1.03). Compared to the category other tumor locations, the distal radius category was associated with a decrease in hazard: HR 0.40 (95%CI 0.23; 0.68). Elevated

baseline SALP was associated with an increased hazard of metastasis: HR 2.12 (95%CI 1.52; 2.95); see **Appendix I figure A1** for the Kaplan Meier curves of SALP. Using other breeds as a reference Doberman subjects were related to a higher hazard, while mixed breed subjects were associated with a lower hazard: [HR 2.16 (95%CI 1.06; 4.42) and HR 0.49 (95%CI 0.29; 0.84)]. By 1 year of follow-up the associations for metastasis of OS were similar to the results at 5 months (**Table 2**); median DF survival was 234 days.

The median survival was 256 days, based on the 598 dogs that had no missing data. At 5 months of follow up, the prognostic factors tumor location, breed and SALP at baseline were both univariable related to mortality and the magnitude of the observed relations was similar to those for metastasis. At 1 year, weight, location, breed and SALP showed similar and significant associations as found for metastasis at 1 year (**Appendix I table A1**).

Multivariable analysis

After imputing missing values, 1155 subjects were available for the metastasis outcome (**Table 3**). By the end of follow-up 765 experienced a metastasis. Weight was associated with an increased hazard [HR 1.02 (per kg increase) (95% CI 1.01; 1.03)], as well as high SALP [HR 1.34 (95% CI 1.07;1.68)]. Compared to other tumor locations, patients with a distal radius OS were associated with a decreased hazard of metastases: HR 0.75 (95%CI 0.58; 0.96). Furthermore, the proximal humerus location was associated with an increased hazard of metastases, however this association was not significant: HR 1.21 (95% CI 0.96; 1.53). Similarly, breed was no longer significantly associated with metastasis after adjusting for other baseline characteristics.

For the outcome mortality, 1336 dogs were available for analysis, of which 1076 died. The associations between weight and mortality, and SALP and mortality were similar to those found for the outcome metastasis (**Table 3**). Compared to the category other OS locations, proximal humerus tumours were associated with a higher hazard of mortality: HR 1.53 (95%CI 1.26; 1.84). Similarly, having an OS at the distal femur or proximal tibia was related to an increased hazard: HR 1.23 (95%CI 1.01; 1.49). Finally, older aged subjects were also related to a higher hazard of mortality: HR 1.06 per year (95%CI 1.03; 1.09).

Table 1 Characteristics of studies included in the IPDMA for prognostic factors of canine osteosarcoma mortality or metastases*.

Study	*Published?*	*Design*	*N subjects*	*N mortality events*	*N metastases events*	*Maximum follow-up in days*	*Characteristics recorded*							
							Age	Weight	Gender	Neutered	SALP	Breed	Location	Chemotherapy
Amsellem	No	NR	36	24	16	2539	✓	✓	✓	✓	✓	✓	✓	✓
Bacon 1	Yes	NR	50	44	42	2192	✓	✓	✓	✓	✓	✓	✓	✓
Bacon 2	No	NR	145	113	113	2570	✓	✓	✓	✓	✓	✓	✓	✓
Berg	Yes	RCT	94	81	NA	1628	✓	✓	✓	✓	✓	✓	✓	✓
Kirpensteijn	Yes	NR	134	111	85	2428	✓	✓	✓	✓	✓	✓	✓	✓
Kow	Yes	NR	66	46	36	869	✓		✓	✓		✓	✓	✓
Kurzman 1	No	RCT	60	44	38	825	✓	✓	✓	✓		✓	✓	✓
Kurzman 2	No	RCT	36	27	22	1584	✓	✓	✓	✓		✓	✓	✓
Kurzman 3	Yes	RCT	64	55	54	986	✓	✓	✓	✓		✓	✓	✓
Kurzman 4	Yes	RCT	25	19	21	1640	✓		✓	✓		✓	✓	✓
Maritato	No	NR	63	23	5	1923	✓	✓	✓	✓		✓	✓	✓
Moore	Yes	RCT	303	273	221	2109	✓	✓	✓	✓	✓	✓	✓	✓
Morello 1	No	NR	35	29	23	2209	✓		✓			✓	✓	✓
Morello 2	No	NR	25	25	18	2023	✓		✓			✓	✓	✓
Morello 3	No	NR	6	6	4	874	✓	✓	✓			✓	✓	✓
Morello 4	No	NR	5	5	3	1737	✓	✓	✓			✓	✓	✓
Morello 5	Yes	NR	13	13	7	2210	✓	✓	✓			✓	✓	✓
Phillips	Yes	NR	156	126	NA	3463	✓	✓	✓	✓	✓	✓	✓	✓
Sottnik	Yes	NR	69	NA	49	1749	✓	✓	✓	✓	✓	✓	✓	✓
Vail	Yes	RCT	20	12	8	730	✓	✓	✓	✓	✓	✓	✓	✓

**N subjects = number of observations after excluding patients not receiving surgery, other infrequently occurring chemotherapies or radiation therapy, NA = not recorded, NR = Non Randomized study, RCT = Randomized Controlled Trial.*

Table 2 Baseline characteristics stratified for event status at 5 months and 1 year follow-up for metastases due to osteosarcoma in canines, with univariable hazard ratios (HR) and 95% confidence intervals (95%CI).

Variables	*Missing*	*5 months*			*1 year*		
		Event-free N = 872	*Event N = 283*	*HR (95%CI) p-value*	*Event-free N = 568*	*Event N = 587*	*HR (95%CI) p-value*
Mean subjects per study	-	64		-	64		-
Number of subjects without missings N(%)	-	397(46%)	153(54%)	-	241(42%)	309(53%)	-
Number of subjects from published studies N(%)	-	556(64%)	188(66%)	-	337(59%)	407(69%)	-
Follow-up days median (Q1-Q3)	55	148(148-148)	83(62-112)	-	356(172-356)	151(85.5-221)	-
Age (years) mean (sd)	20	8.29(2.66)	8.08(2.64)	1.00 (0.93;1.06) p = 0.91	8.33(2.8)	8.15(2.51)	1.00 (0.95;1.05) p = 0.99
Weight (kg) mean (sd)	200	35.64(12.47)	38.01(15.13)	1.02 (1.00;1.03) p = 0.02	35.19(13.09)	37.22(13.32)	1.02 (1.01;1.02) p < 0.01
Male gender N(%)	6	463(53%)	136(48%)	0.84 (0.61;1.15) p = 0.28	311(55%)	288(49%)	0.82 (0.65;1.03) p = 0.08
Neutered N(%)	90	674(85%)	217(81%)	0.90 (0.53;1.54) p = 0.71	439(85%)	452(82%)	0.85 (0.58;1.25) p = 0.40
High SALP N(%)	525	146(32%)	86(51%)	2.12 (1.52;2.95) p < 0.01	89(31%)	143(42%)	1.70 (1.34;2.15) p < 0.01
Breed	9			Overall p-value = 0.01			Overall p-value = 0.01
Other N(%)		293(34%)	150(41%)	Reference	208(37%)	200(34%)	Reference
Rottweiler N(%)		109(13%)	42(15%)	1.20 (0.75;1.90) p = 0.45	65(12%)	86(15%)	1.48 (1.06;2.09) p = 0.02
Golden Retriever N(%)		83(10%)	28(10%)	0.69 (0.38;1.27) p = 0.23	48(9%)	63(11%)	0.89 (0.59;1.33) p = 0.57
Labrador Retriever N(%)		75(9%)	22(8%)	0.68 (0.36;1.30) p = 0.24	41(7%)	56(10%)	0.73 (0.47;1.13) p = 0.16
Greyhound N(%)		44(5%)	17(6%)	1.11 (0.57;2.17) p = 0.76	32(6%)	29(5%)	1.09 (0.65;1.81) p = 0.75
Doberman N(%)		39(5%)	18(6%)	2.16 (1.06;4.42) p = 0.03	24(4%)	33(6%)	1.70 (0.93;3.14) p = 0.09
Irish Setter N(%)		24(3%)	5(2%)	0.56 (0.14;2.32) p = 0.42	17(3%)	12(2%)	0.73 (0.29;1.83) p = 0.51
Mixed N(%)		199(23%)	33(12%)	0.49 (0.29;0.84) p = 0.01	129(23%)	103(18%)	0.74 (0.53;1.04) p = 0.09
Tumor location	91			Overall p-value < 0.01			Overall p-value < 0.01
Other N(%)		219(27%)	91(34%)	Reference	151(29%)	159(29%)	Reference
Prox. Humerus N(%)		175(22%)	67(25%)	1.06 (0.69;1.63) p = 0.78	96(19%)	146(26%)	1.43 (1.05;1.94) p = 0.03
Dist. Femur or Prox. Tibia N(%)		175(22%)	65(24%)	1.01 (0.65;1.55) p = 0.97	106(21%)	134(24%)	1.26 (0.92;1.74) p = 0.15
Dist. Radius N(%)		228(29%)	44(16%)	0.40 (0.23;0.68) p < 0.01	159(31%)	113(20%)	0.63 (0.44;0.89) p = 0.01
Chemotherapy	30			Overall p-value = 0.02			Overall p-value = 0.15
No chemo N(%)		104(12%)	47(17%)	Reference	85(16%)	66(11%)	Reference
Cisplatin N(%)		124(15%)	40(14%)	1.14 (0.57;2.30) p = 0.71	72(13%)	92(16%)	0.92 (0.50;1.66) p = 0.77
Lobaplatin, carboplatin N(%)		61(7%)	32(11%)	0.67 (0.31;1.46) p = 0.32	39(7%)	54(9%)	1.02 (0.53;1.94) p = 0.96
Doxorubicin N(%)		348(41%)	97(35%)	0.47 (0.23;0.95) p = 0.04	218(40%)	227(39%)	0.61 (0.36;1.04) p = 0.07
Doxorubicin combinations N(%)		209(25%)	63(23%)	0.30 (0.12;0.73) p = 0.01	129(24%)	143(25%)	0.59 (0.31;1.14) p = 0.12

Table 3 Multivariable hazard ratios (HR) with 95% confidence intervals (95%CI) and p-values for the hazard of metastases or mortality, using the entire follow-up period*.

Variables	*Metastasis*	*Mortality*
	HR (95%CI), p-value	*HR (95%CI) p-value*
Number of observations (total dog years at risk)	1155 (621 years)	1336 (1042 years)
Number of events	765	1076
Age (years)	1.03 (0.99;1.06) p = 0.15	1.06 (1.03;1.09) p < 0.01
Weight (kg)	1.02 (1.01;1.03) p < 0.01	1.02 (1.01;1.02) p < 0.01
Male gender	0.91 (0.77;1.08) p = 0.29	0.95 (0.83;1.09) p = 0.50
Neutered	0.90 (0.70;1.15) p = 0.38	0.85 (0.70;1.03) p = 0.09
High SALP	1.34 (1.07;1.68) p = 0.01	1.43 (1.16;1.77) p < 0.01
Breed	Overall p-value = 0.67	Overall p-value = 0.65
Other	Reference	Reference
Rottweiler	1.00 (0.78;1.30) p = 0.98	0.98 (0.80;1.21) p = 0.87
Golden Retriever	1.09 (0.82;1.45) p = 0.56	1.04 (0.83;1.31) p = 0.73
Labrador Retriever	1.04 (0.79;1.38) p = 0.78	0.89 (0.69;1.15) p = 0.36
Greyhound	1.31 (0.91;1.89) p = 0.15	1.15 (0.86;1.56) p = 0.35
Doberman	1.02 (0.71;1.47) p = 0.93	1.10 (0.80;1.51) p = 0.56
Irish Setter	0.77 (0.44;1.38) p = 0.38	0.91 (0.57;1.45) p = 0.69
Mixed	0.92 (0.73;1.15) p = 0.44	0.89 (0.73;1.07) p = 0.22
Tumor location	Overall p-value < 0.01	Overall p-value < 0.01
Other	Reference	Reference
Prox. Humerus	1.21 (0.96;1.53) p = 0.10	1.53 (1.26;1.84) p < 0.01
Dist. Femur or Prox. Tibia	1.10 (0.88;1.39) p = 0.40	1.23 (1.01;1.49) p = 0.04
Dist. Radius	0.75 (0.58;0.96) p = 0.02	0.90 (0.74;1.10) p = 0.30
Chemotherapy	Overall p-value = 0.28	Overall p-value = 0.43
No chemo	Reference	Reference
Cisplatin	1.17 (0.75;1.82) p = 0.50	1.04 (0.67;1.60) p = 0.87
Lobaplatin, carboplatin	1.27 (0.86;1.87) p = 0.23	1.04 (0.72;1.51) p = 0.85
Doxorubicin	0.88 (0. 60;1.28) p = 0.50	0.98 (0.67;1.46) p = 0.94
Doxorubicin combinations	0.91 (0.64;1.31) p = 0.61	0.75 (0.53;1.06) p = 0.10

*results are based on a model including all variables presented, no stepwise selection was applied.

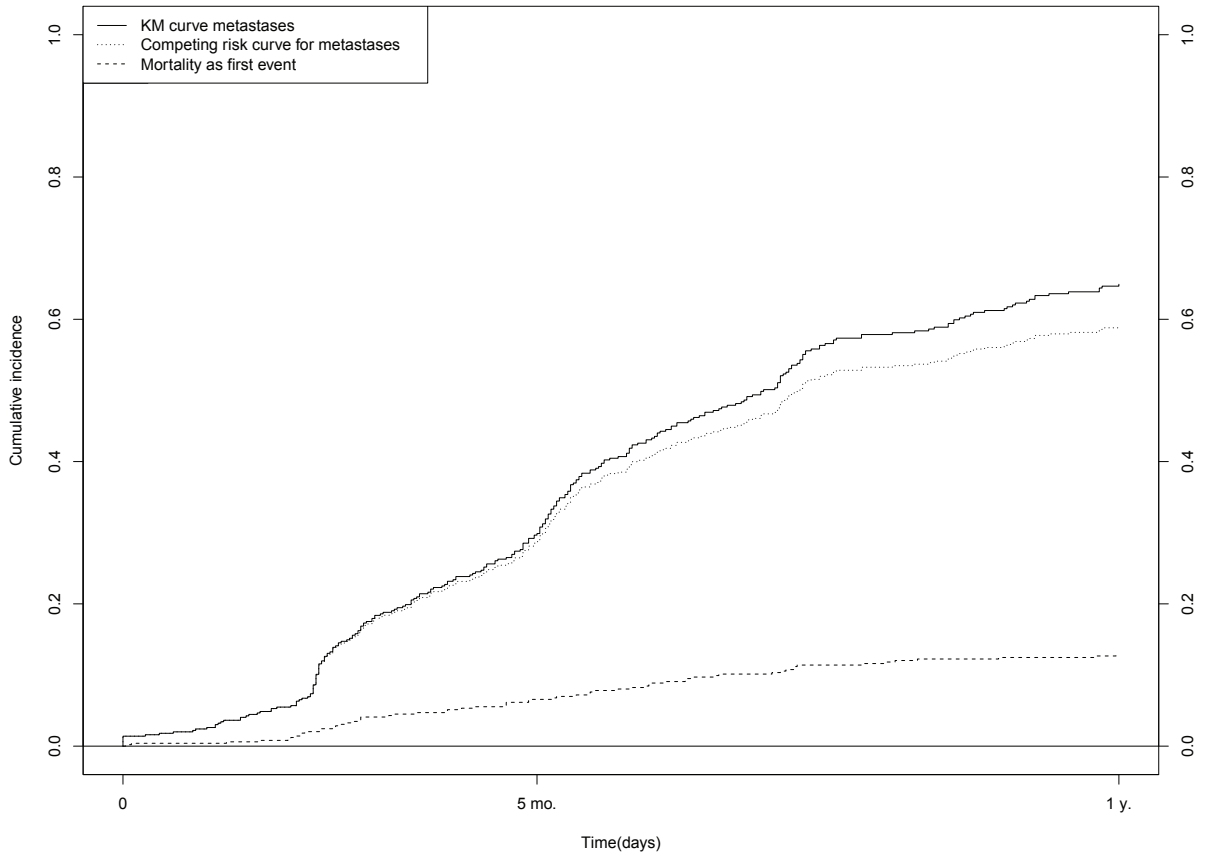
The discriminative performance of the multivariable models was modest: the model for the outcome metastasis had a c-statistic of 0.63, whereas the model for the outcome mortality had a c-statistic of 0.61.

Sensitivity analysis

Figure 1 shows the Kaplan Meier survival curves for the outcome metastasis (with and without accounting for competing risks). 29 (4%) subjects died without experiencing a metastasis event before dying. Consequently, ignoring competing risks had little impact on results: the standard Kaplan-Meier estimates only marginally overestimate the cumulative incidence of metastases when accounting for competing risks.

Additional sensitivity analyses (on the impact of missing observations, size of the included studies, publication status of the included studies, and chemotherapy status of the dogs) did not show material difference compared to the results presented in **Table 3**. Results of these sensitivity analyses are available upon request.

Figure 1 Kaplan Meier survival curves for canines with osteosarcoma



Competing risk curve for metastases, with biased Kaplan Meier curve for metastases and mortality without a metastasis. Results are based on 511 subjects without missings that had data on both mortality and metastases outcome.

Discussion

In our IPD meta-analysis on prognostic factors for metastasis and mortality among dogs with OS, weight, SALP and tumor location were independently prognostic predictors of mortality as well as metastasis. Age was significantly related with mortality only.

In accordance with the “aggregated data” meta-analysis by Boerman et al. (10), we found that elevated SALP was associated with a higher hazard of early mortality or metastasis. However, the Boerman study showed somewhat larger and less precise estimates [HR 1.62 (95%CI 1.21; 2.17) for mortality and 1.96 (95%CI 1.50; 2.56) for metastasis] compared to our results: HR 1.43 (95%CI 1.16; 1.77) and HR 1.34 (95%CI 1.07; 1.68). Compared to other tumor sites, proximal humerus and distal femur or proximal tibia OS locations were related to an increased mortality hazard, but not metastasis. This is different from the Boerman study, which concluded that proximal humerus was significantly associated with both mortality and metastasis [HR 1.86 (95%CI 1.34; 2.57) and 2.53 (95%CI 1.34; 4.77)]. Furthermore, we found that having an OS tumor at the distal radius was associated with a decreased metastasis hazard. Our IPDMA also showed that independent of breed, high weight increased the

hazard of both metastases and mortality. Possibly, this is due to the crude categorisation of the breed variable, with a large “other” category resulting in unexplained variance. Also different from the findings of the Boerman study was that we found age to be significantly related with mortality (increasing the hazard).

In this section we will discuss limitation and strengths of our study. First the number of patients with at least one missing observation was high (52% for the metastasis outcome and 57% for mortality). This was predominantly driven by SALP, which was only measured in 9 out of 20 studies. In aggregated meta-analyses, like the one conducted by Boerman et.al. (10), it is difficult to deal with missing data. In the current study we used an individual patient data meta-analyses (IPDMA) design, which allows for imputation of missing values. Like all studies with missing observations, it is possible that missingness was not only dependent on measured factors but also on unmeasured factors, thus results may still be biased even though missing data was multiple imputed. However, assuming that at least some of the missing values are dependent on measured factors, imputing missing values would likely decrease bias compared to a complete cases analysis. Secondly, several sensitivity analyses were performed, all showing similar associations as our main analysis, confirming the robustness of our findings. Third, most studies used 1 or 2 specific chemotherapy regimens, making it difficult to distinguish between chemotherapy effects and other study-specific influences. Thus, while it seems essential to include chemotherapy in modelling the independent prognostic associations between patient characteristics and outcomes, observed associations between chemotherapy and outcomes should not be interpreted causally. Fourth, none of the baseline variables, except chemotherapy (and only in some studies), were randomly allocated. Therefore, it is possible that unmeasured or residual confounding influence our results. Given that it is impossible to randomly allocate baseline characteristics such as gender or age, every study exploring these associations is potentially hampered by the possibility of confounding. Causal interpretation of observed associations might therefore lead to erroneous conclusions. For example, when, contrary to the association reported here, there is no causal relationship between weight and mortality (possible due to some unmeasured protective genetic factor that is closely related to lower weight) intervening on weight will have no effect on the outcome. However, in such a situation (no causal relationship of weight) weight will still provide useful information on the baseline risk for the outcome. Thus, the importance of causality of the here reported associations depends on the goal; either to intervene on risk

factors or to use those factors for prognostication. Fifth, (aggregated) meta-analyses can be subject to publication bias (i.e., bias due to including published studies only) (13). By recruiting data via the VSSO network, about 40% of the included subjects were from unpublished sources, making the potential for publication bias smaller. On the other hand, some researchers did not respond to our requests for collaboration therefore results presented here do not include all possible data and we cannot rule out the possibility that inclusion of more data could change our estimations. Finally, the discriminative ability of all models was modest. Including clinical predictors like grade or type of tumor could potentially increase this discriminative ability. However, in the current study this was impossible due to the large number of studies that did not record data on these variables.

Our present study used relatively new study techniques to combine individual patient data from different sources termed individual patient data meta-analysis (IPDMA). While an IPDMA requires big investments regarding time and resources we believe that the opportunities of using individual patient data compared to the alternative of relying on aggregated data outweigh this burden. An advantage of conducting an IPDMA is that one can explore relations not reported by the original authors. In our case this allowed us to estimate 7 associations while the Boerman study (10) could only explore 3 associations. Similarly, IPMDA techniques ensure that when one wants to conduct multivariable analysis all estimates are corrected for the same set of variables. Without the same corrections, one has to rely on the reported estimates and as Boerman showed it is likely that every study uses distinct sets of covariables. A third advantage is that one can uniformly recode the data which can be particularly important if different cut off values or reference categories are used by the original authors (e.g., categorizing age using a cut point of 5 or 7 years). Fourth, IPDMAs also allows one to check model assumption such as linearity or proportional hazard. Lastly by having access to individual patient data one can more easily perform subgroup analysis, sensitivity analysis and apply more refined methods to deal with problems such as missing values or competing risks. However, while we strongly recommend researcher to use IPDMA techniques in meta-analyses one should remember that the success of any meta-analysis ultimately depends on quality of the original data.

Conclusions

In conclusion, the IPDMA design used in this study allowed us to assess prognostic factors

in canines with osteosarcoma. We identified weight, SALP, and tumor location as independent prognostic factors of metastasis and mortality, while age was only associated with early mortality. This study design allowed for the application of advanced missing data techniques and multiple sensitivity analyses and showed the necessity to use individual participant data in order to comprehensively assess prognostic factors in this field of research.

Acknowledgements

We gratefully wish to acknowledge the following researchers for their willingness to collaborate:

Dr. Pierre Amsellem; Atlantic Veterinary College, University of Prince Edward Island, Charlottetown, Canada.

Dr. Nicholas Bacon; College of Veterinary Medicine, University of Florida, Gainesville, Florida, US.

Professor Dr. John Berg; Tufts Cummings School of Veterinary Medicine, North Grafton, MA, US.

Dr Kelvin Kow; College of Veterinary Medicine, University of Florida, Gainesville, Florida, US.

Dr. Ilene Kurzman; School of Veterinary Medicine University of Wisconsin, Madison, Wisconsin, US.

Dr. Karl Maritato, MedVet Medical and Cancer Center for Pets, Cincinnati, Ohio, US.

Dr. Antony Moore; Veterinary Oncology Consultants, Australia.

Dr. Emanuela Morello; School of Veterinary Medicine, Turin, Italy.

Dr Joe Sottnik; Animal Cancer Center, Colorado State University, Fort Collins, CO, US.

Professor Dr. David Vail; School of Veterinary Medicine, University of Wisconsin, Madison, Wisconsin, US.

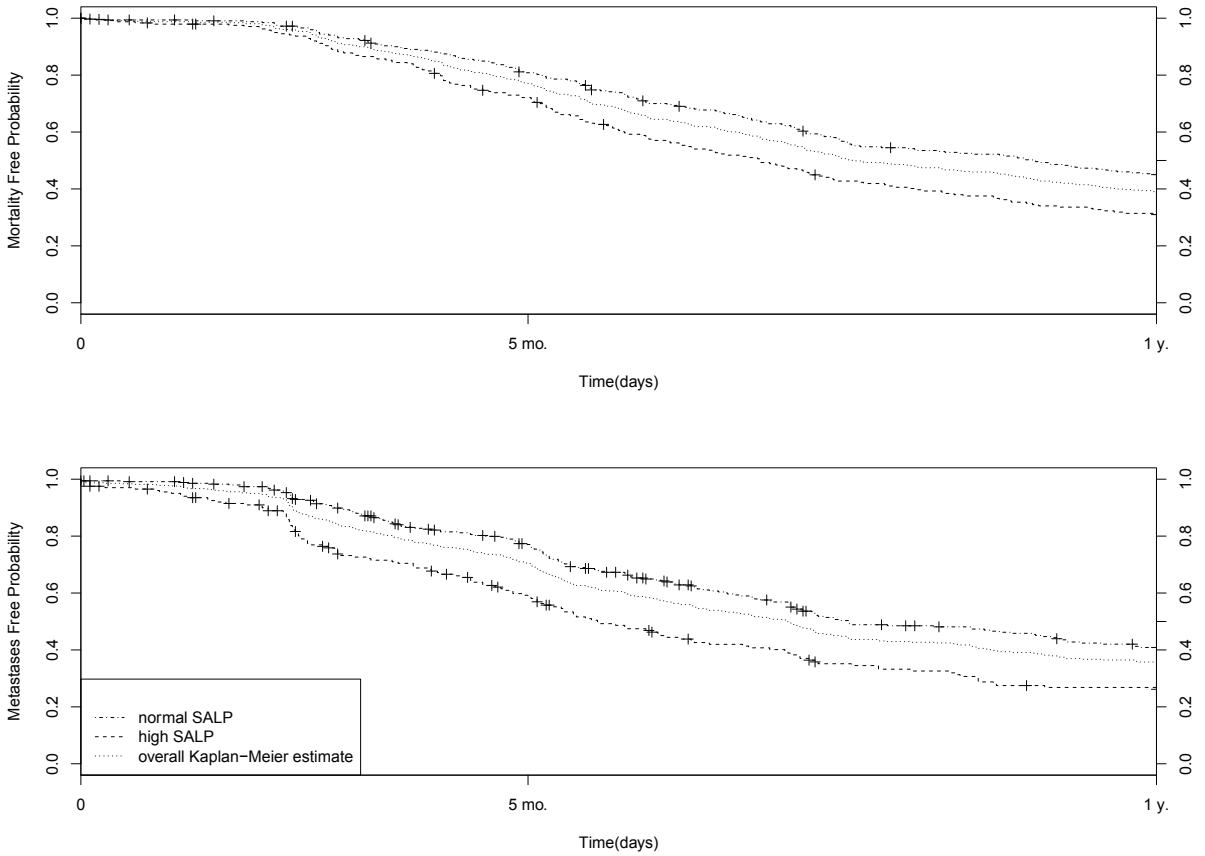
Finally we wish to thank the Veterinary Society of Surgical Oncology (VSSO) society for allowing us to approach their members for collaboration.

Reference List

1. Rowell JL, McCarthy DO, Alvarez CE. Dog models of naturally occurring cancer. Trends in Molecular Medicine 2011 Jul;17(7):380-8.
2. Withrow SJ, Wilkins RM. Cross talk from pets to people: translational osteosarcoma treatments. ILAR Journal 2010;51(3):208-13.
3. Norrdin RW, Powers BE, Torgersen JL, Smith RE, Withrow SJ. Characterization of osteosarcoma cells from two sibling large-breed dogs. American Journal of Veterinary Research 1989 Nov;50(11):1971-5.
4. Spodnick GJ, Berg J, Rand WM, Schelling SH, Couto G, Harvey HJ, et al. Prognosis for dogs with appendicular osteosarcoma treated by amputation alone: 162 cases (1978-1988). Journal of the American Veterinary Medical Association 1992 Apr 1;200(7):995-9.
5. Cooley DM, Waters DJ. Skeletal neoplasms of small dogs: a retrospective study and literature review. Journal of the American Animal Hospital Association 1997 Jan;33(1):11-23.
6. Ru G, Terracini B, Glickman LT. Host related risk factors for canine osteosarcoma. The Veterinary Journal 1998 Jul;156(1):31-9.
7. McNeill CJ, Overley B, Shofer FS, Kent MS, Clifford CA, Samluk M, et al. Characterization of the biological behaviour of appendicular osteosarcoma in Rottweilers and a comparison with other breeds: a review of 258 dogs. Veterinary and Comparative Oncology 2007 Jun;5(2):90-8.
8. Brodey RS, Abt DA. Results of surgical treatment in 65 dogs with osteosarcoma. Journal of the American Veterinary Medical Association 1976 Jun 1;168(11):1032-5.
9. Straw RC, Withrow SJ, Richter SL, Powers BE, Klein MK, Postorino NC, et al. Amputation and cisplatin for treatment of canine osteosarcoma. Journal of Veterinary Internal Medicine 1991 Jul;5(4):205-10.
10. Boerman I, Selvarajah GT, Nielen M, Kirpensteijn J. Prognostic factors in canine appendicular osteosarcoma - a meta-analysis. BMC Veterinary Research 2012;8:56.
11. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010;340:c221.
12. Stewart LA, Parmar MK. Meta-analysis of the literature or of individual patient data: is there a difference? Lancet 1993 Feb 13;341(8842):418-22.
13. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet 1991 Apr 13;337(8746):867-72.
14. Harrell FE, Jr. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. 1st ed. New York: Springer; 2001.
15. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996 Feb 28;15(4):361-87.
16. Hmisc: Harrell Miscellaneous. R package [computer program]. Version R package version 3.10-1 2012.
17. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.
18. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Medical Research Methodology 2009;9:57.
19. Sottnik JL, Rao S, Lafferty MH, Thamm DH, Morley PS, Withrow SJ, et al. Association of blood monocyte and lymphocyte count and disease-free interval in dogs with osteosarcoma. Journal of Veterinary Internal Medicine 2010 Nov;24(6):1439-44.
20. Phillips B, Powers BE, Dernell WS, Straw RC, Khanna C, Hogge GS, et al. Use of single-agent carboplatin as adjuvant or neoadjuvant therapy in conjunction with amputation for appendicular osteosarcoma in dogs. Journal of the American Animal Hospital Association 2009 Jan;45(1):33-8.
21. Berg J, Gebhardt MC, Rand WM. Effect of timing of postoperative chemotherapy on survival of

dogs with osteosarcoma. Cancer 1997 Apr 1;79(7):1343-50.
22. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. American Journal of Epidemiology 2009 Jul 15;170(2):244-56.
23. Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. British Journal of Cancer 2004 Oct 4;91(7):1229-35.
24. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
25. A Package for Survival Analysis in S. [computer program]. Version R package version 2.36-14. 2012.
26. rms: Regression Modeling Strategies. R package [computer program]. Version version 3.6-0 2012.
27. Bacon NJ, Ehrhart NP, Dernell WS, Lafferty M, Withrow SJ. Use of alternating administration of carboplatin and doxorubicin in dogs with microscopic metastases after amputation for appendicular osteosarcoma: 50 cases (1999-2006). Journal of the American Veterinary Medical Association 2008 May 15;232(10):1504-10.
28. Kirpensteijn J, Kik M, Rutteman GR, Teske E. Prognostic significance of a new histologic grading system for canine osteosarcoma. Veterinary Pathology 2002 Mar;39(2):240-6.
29. Kow K, Thamm DH, Terry J, Grunerud K, Bailey SM, Withrow SJ, et al. Impact of telomerase status on canine osteosarcoma patients. Journal of Veterinary Internal Medicine 2008 Nov;22(6):1366-72.
30. Kurzman ID, MacEwen EG, Rosenthal RC, Fox LE, Keller ET, Helfand SC, et al. Adjuvant therapy for osteosarcoma in dogs: results of randomized clinical trials using combined liposome-encapsulated muramyl tripeptide and cisplatin. Clinical Cancer Research 1995 Dec;1(12):1595-601.
31. Moore AS, Dernell WS, Ogilvie GK, Kristal O, Elmslie R, Kitchell B, et al. Doxorubicin and BAY 12-9566 for the treatment of osteosarcoma in dogs: a randomized, double-blind, placebo-controlled study. Journal of Veterinary Internal Medicine 2007 Jul;21(4):783-90.
32. Morello E, Vasconi E, Martano M, Peirone B, Buracco P. Pasteurized tumoral autograft and adjuvant chemotherapy for the treatment of canine distal radial osteosarcoma: 13 cases. Veterinary Surgery 2003 Nov;32(6):539-44.
33. Vail DM, Kurzman ID, Glawe PC, O'Brien MG, Chun R, Garrett LD, et al. STEALTH liposome-encapsulated cisplatin (SPI-77) versus carboplatin as adjuvant therapy for spontaneously arising osteosarcoma (OSA) in the dog: a randomized multicenter clinical trial. Cancer Chemotherapy and Pharmacology 2002 Aug;50(2):131-6.

Figure A1: Kaplan-Meier survival curves for mortality and metastases free survival, stratified for high or normal serum alkaline phosphatase (SALP), including the unstratified estimates



Legend: + indicates censoring.

Table A1: Baseline characteristics stratified for event status at 5 months and 1 year follow-up for mortality due to osteosarcoma in canines.

Variables	*Missing*	*5 months*			*1 year*		
		Event-free N = 1033	*Event N = 303*	*HR (95%CI) p-value*	*Event-free N = 606*	*Event N = 730*	*HR (95%CI) p-value*
Mean subjects per study	-	70		-	70		-
Number of subjects without missings N(%)	-	442(43%)	127(42%)	-	236(39%)	333(46%)	-
Number of subjects from published studies N(%)	-	702(68%)	223(74%)	-	378(62%)	547(75%)	-
Follow-up days median (Q1-Q3)	54	148(148-148)	94(63-120)	-	356(356-356)	169(107-241)	-
Age (years) mean (sd)	20	8.16(2.64)	8.19(2.77)	0.99 (0.93;1.07) p = 0.88	8.17(2.71)	8.16(2.64)	1.03 (0.99;1.08) p = 0.19
Weight (kg) mean (sd)	208	35.97(12.64)	37.03(13.14)	1.01 (1.00;1.03) p = 0.10	35.10(12.62)	37.11(12.81)	1.01 (1.00;1.02) p = 0.04
Male gender N(%)	6	545(53%)	142(47%)	0.71 (0.50;1.00) p = 0.05	311(52%)	376(52%)	0.91 (0.73;1.12) p = 0.37
Neutered N(%)	90	796(83%)	226(77%)	0.92 (0.53;1.62) p = 0.78	470(84%)	552(81%)	0.92 (0.65;1.32) p = 0.66
High SALP N(%)	566	222(38%)	95(53%)	1.81 (1.25;2.61) p < 0.01	106(33%)	211(48%)	1.71 (1.36;2.15) p < 0.01
Breed	9			Overall p-value = 0.01			Overall p-value < 0.01
Other N(%)		345(34%)	117(39%)	Reference	211(35%)	251(35%)	Reference
Rottweiler N(%)		133(13%)	47(16%)	1.02 (0.61;1.70) p = 0.95	68(11%)	112(15%)	1.39 (1.01;1.91) p = 0.04
Golden Retriever N(%)		110(11%)	30(1%)	0.81 (0.45;1.49) p = 0.51	59(10%)	81(11%)	1.03 (0.71;1.49) p = 0.90
Labrador Retriever N(%)		85(8%)	20(7%)	0.41 (0.17;0.99) p = 0.05	47(8%)	58(8%)	0.82 (0.54;1.26) p = 0.37
Greyhound N(%)		54(5%)	17(6%)	0.99 (0.46;2.15) p = 0.98	34(6%)	37(5%)	0.97 (0.57;1.64) p = 0.91
Doberman N(%)		45(4%)	21(7%)	2.46 (1.22;4.96) p = 0.01	23(4%)	43(6%)	2.47 (1.46;4.17) p < 0.01
Irish Setter N(%)		30(3%)	3(1%)	0.72 (0.17;3.00) p = 0.65	18(3%)	15(2%)	0.90 (0.39;2.07) p = 0.81
Mixed N(%)		227(22%)	43(14%)	0.55 (0.31;0.97) p = 0.04	142(24%)	128(18%)	0.71 (0.51;0.99) p = 0.04
Tumor location	82			Overall p-value < 0.01			Overall p-value < 0.01
Other N(%)		280(29%)	86(31%)	Reference	183(32%)	183(27%)	Reference
Prox. Humerus N(%)		204(21%)	85(31%)	1.82 (1.14;2.91) p = 0.01	90(16%)	199(29%)	1.80 (1.34;2.43) p < 0.01
Dist. Femur or Prox. Tibia N(%)		220(23%)	58(21%)	1.00 (0.59;1.71) p = 0.99	122(21%)	156(23%)	1.28 (0.93;1.77) p = 0.13
Dist. Radius N(%)		273(28%)	48(17%)	0.69 (0.40;1.20) p = 0.19	175(31%)	146(21%)	0.68 (0.49;0.96) p = 0.03
Chemotherapy	30			Overall p-value = 0.07			Overall p-value = 0.38
No chemo N(%)		100(10%)	51(17%)	Reference	73(12%)	78(11%)	Reference
Cisplatin N(%)		131(13%)	33(11%)	1.09 (0.48;2.45) p = 0.84	76(13%)	88(12%)	0.72 (0.38;1.34) p = 0.30
Lobaplatin, carboplatin N(%)		175(17%)	61(21%)	0.38 (0.14;1.03) p = 0.06	102(17%)	134(19%)	0.73 (0.35;1.49) p = 0.37
Doxorubicin N(%)		323(32%)	97(33%)	0.36 (0.13;0.94) p = 0.04	170(29%)	250(35%)	0.57 (0.32;1.02) p = 0.06
Doxorubicin combinations N(%)		282(28%)	53(18%)	0.59 (0.22-1.61) p = 0.31	169(29%)	166(23%)	0.58 (0.27;1.25) p = 0.16

CHAPTER 5

Which dogs with appendicular osteosarcoma benefit most from chemotherapy after surgery? results from an individual patient data meta-analysis

A F Schmidt, R H H Groenwold, P Amsellem, N Bacon, O H Klungel
A W Hoes, A de Boer, K Kow, K. Maritato, J Kirpensteijn, M Nielen

In revision

Abstract

Osteosarcoma (OS) is a malignant tumor of mesenchymal origin that produces osteoid. Given that the prognosis varies considerably between dogs, we explored whether treatment could be tailored towards prognostic subgroups of patients. For the current study, individual patient data from five nonrandomized studies were combined. Based on a multivariable prognostic model, the 5-month mortality risk was estimated. Subsequently, in surgically treated dogs, we explored whether the effects of the chemotherapeutics carboplatin, cisplatin, doxorubicin or doxorubicin combination therapy compared to no chemotherapy differed between dogs according to their baseline prognosis. For all of the four comparisons, effect estimates differed consistently according to baseline mortality risk. For example, in the comparison of carboplatin treatment vs. no chemotherapy, the overall treatment risk ratio (RR) on mortality was 0.49 (95%CI 0.25; 0.94). This effect differed according to baseline risk: with the RR ranging from 0.10 (95%CI 0.02; 0.64) to 2.38 (95%CI 0.58; 9.81); representing the treatment effect in dogs with the lowest and highest 5-month mortality risk (19% vs. 52%). Similar results were found for the other three treatment comparisons. These results indicate that the main treatment effects of chemotherapy do not necessarily apply to all patients. Specifically, dogs with a relatively low mortality risk at baseline appeared to benefit most from chemotherapy. In general, researchers should more often explore whether treatment can be tailored toward subgroups of patients.

Introduction

Osteosarcoma (OS) is a malignant tumor of mesenchymal origin that produces osteoid. In dogs, OS most frequently occurs in large and giant breeds (1-5). Dogs that are treated with amputation have a median survival time of five months, with the majority succumbing to metastatic disease (6;7). Clinical studies have shown that on average survival in OS dogs can be extended by administrating chemotherapy (8-12).

In a recent Individual Patient Data Meta-Analysis (IPDMA), we identified baseline variables that were associated with survival in dogs with osteosarcoma (13). Such a prognostic model can be used to predict a dog's risk of early mortality (14). This offers the possibility to identify subgroups of dogs according to their baseline prognosis and target treatment at those patients most likely to benefit. This can potentially prevent dogs from unnecessarily receiving treatment, which is relevant in terms of both costs and quality of life. Cleary there is a need to obtain estimates of individualized treatment effects (15-17).

In the current paper, treatment effects were individualized by determining whether dogs with a different 5-month mortality risk, reacted differently to chemotherapy treatment. Specifically, we compared the effects of carboplatin, cisplatin, doxorubicin and doxorubicin combination therapy to no chemotherapy on 5-month mortality.

Materials and Methods

The effects of the different chemotherapeutics compared to no chemotherapy were determined using individual patient data (IPD). These IPD were previously used in an IPD meta-analysis (IPDMA) combing data of 20 studies to determine prognostic factors for early mortality in dogs with osteosarcoma (13). All dogs in these studies were diagnosed with osteosarcoma and received surgical intervention (amputation or limb-spare). For the present analysis, data were used of studies that included at least five dogs on no chemotherapy and at least five dogs treated with one of the interventions of interest (i.e., carboplatin, cisplatin, doxorubicin or doxorubicin combination therapy). Of the 20 studies included in the IPDMA, five studies fulfilled this criterion; of these five studies two were previously published (18;19). Of the 5 included studies, only two include dogs on all four chemotherapies of interest (**Appendix Table 1)**. Therefore, per comparison, a different selection of patients was used (**Appendix Table 1**); this to prevent extrapolation over studies. To assess the effects of treatment

this study focused on 5-month mortality, which is regarded as a clinical relevant endpoint (2;6;8). As in our previous IPDMA study (13), missing values were imputed (20) using the aregImpute algorithm from the Hmisc package (21).

Data analysis

Treatment effects were estimated using Poisson regression models. Coefficients of such models can be interpreted as (the natural logarithm of) risk ratios (RR) (22;23). Poisson models produce overly conservative estimates of the standard errors therefore these were replaced by robust (Heteroscedasticity-Consistent covariance estimator 4m [i.e., HC4m]) estimates (24). Effect estimates were adjusted for the following potential confounders: gender, neuter status, tumor locations (proximal humerus, distal femur or proximal tibia, distal radius, versus other locations), age (years, continuous), weight (kg, continuous), breed (Rottweiler, Golden Retriever, Labrador Retriever, Greyhound, Doberman, Irish Setter, mixed breeds, versus other breeds) and serum alkaline phosphatase (SALP, using study specific cut-off values for high and normal SALP levels).

To determine whether chemotherapy effects differed between patients according to their baseline prognosis, the following three-step approach was applied. (15-17;25;26) First, the main or overall treatment effect was determined, without taking the possibility of differential treatment effects into account. We refer to this effect as the main treatment effect. Second, using an adapted version of the previously published prediction model (13), the baseline risk of mortality was determined. Details of the model are presented in **Table 1**. For all dogs this risk was determined under the assumption that the dogs would not be treated with chemotherapy. Third, we explored whether the effect of chemotherapy on mortality differed according to patients' baseline risk. Note that out of convenience the logit(baseline risk) will be modelled and transformed to the baseline risk where appropriate (see next section).

The logit of 5-month mortality risk was calculated conditional on no chemotherapy: $logit(baseline.risk) = logit(\hat{p}_i) = \gamma_0 + \gamma_1 * treatment(0) + ... + \gamma_j x_{ij}$. With x_{ij} representing the j th baseline characteristics of Table 1 for the i th individual and γ_j the coefficient of the relevant baseline characteristic. Theoretically the logit(baseline risk) can vary from minus to plus infinity, with zero referring to a risk of 50%. This logit(baseline risk) can be transformed to the baseline risk, bound between 0 and 1, by the following equation:

$\hat{p}_i = \frac{1}{1+e^{-logit(\hat{p}_i)}}$ [equation 1]; see **Table 1** for an example.

Whether treatment effects differed according to baseline risk was tested using a treatment by logit(baseline risk) interaction. A Poisson model including a treatment by logit(baseline risk) interaction contains the following terms: $ln(event=1) = \hat{\alpha}_0 + \hat{\alpha}_1 * treatment + \hat{\alpha}_2 * logit(\hat{p}_i) + \hat{\alpha}_3 * treatment * logit(\hat{p}_i)$, with $\hat{\alpha}_1$ representing the ln(RR) of treatment for a dog with 50% baseline risk of 5-month mortality and $\hat{\alpha}_3$

Table 1. Multivariable adjusted risk factors for 5-month mortality*.

Variables	Odds ratio (95%CI)	Regression coefficients
Intercept		γ_0 = -0.7412
Chemotherapy		
No chemotherapy	Reference	γ_1 = 0.0000
Cisplatin	0.48 (0.28;0.82)	γ_2 =-0.7427
Lobaplatin, carboplatin	0.70 (0.43;1.14)	γ_3 =-0.3574
Doxorubicin	0.63 (0.40;1.01)	γ_4 =-0.4549
Doxorubicin combinations	0.39 (0.24;0.64)	γ_5 =-0.9364
Age (years)	1.00 (0.94;1.06)	γ_6 =-0.0023
Weight (kg)	1.01 (1.00;1.03)	γ_7 =0.0129
Male gender	0.74 (0.56;0.98)	γ_8 =-0.2961
Neutered	0.68 (0.48;0.97)	γ_9 =-0.3823
High SALP	1.78 (1.18;2.68)	γ_{10} =0.5751
Breed		
Other	Reference	γ_{11} = 0.0000
Rottweiler	0.92 (0.61;1.40)	γ_{12} =-0.0789
Golden Retriever	0.91 (0.56;1.47)	γ_{13} =-0.0928
Labrador Retriever	0.74 (0.43;1.27)	γ_{14} =-0.3066
Greyhound	1.30 (0.70;2.39)	γ_{15} =0.2603
Doberman	1.39 (0.77;2.50)	γ_{16} =0.3278
Irish Setter	0.38 (0.11;1.33)	γ_{17} =-0.9577
Mixed	0.64 (0.42;0.97)	γ_{18} =-0.4453
Tumor location		
Other	Reference	γ_{19} =0.0000
Prox. Humerus	1.48 (1.01;2.68)	γ_{20} =0.3930
Dist. Femur or Prox. Tibia	0.90 (0.61;1.33)	γ_{21} =-0.1044
Dist. Radius	0.63 (0.41;0.96)	γ_{22} =-0.4633
Example patient logit(baseline risk) = -0.7412 + 0.0000*no chemotherapy(0) + -0.0023*5 years + 0.0129*25 kg + -0.2961*female(0) + -0.3823* not neutered (0) + 0.5751*high salp(1) + -0.0928*Golden Retriever(1) + 0.3902*proximal humerus(1) =0.44		
Example patient baseline risk = $1/(1+e^{-0.44})$ = 0.61		

*Numbers represent odds ratios with 95% confidence intervals (95%CI). All odds ratios were adjusted f or all other presented variables This multivariable logistic regression model is a variation of the cox proportional hazard model described in Schmidt et.al., 2013.

by how much the treatment effect changes with increasing or decreasing logit(baseline risk).

In the presence of interaction, the treatment effect of chemotherapy per unit increase (or decrease) of the baseline on the logit scale becomes:

$$RR_i = e^{\hat{\alpha}_1 * treatment(1) + \hat{\alpha}_3 * treatment(1) * logit(\hat{p}_i)}$$ [equation 2]

In the absence of interaction, $\hat{\alpha}_3$ becomes zero and can be omitted. Instead of assuming a linear effect of the interaction, as equation 2 assumes, we also determined the treatment effect per quintiles of the baseline risk. For comparisons sake these non-linear quintile specific treatment effects were compared to the linear effects from equation 2.

All tests were applied using a significance level of 0.05 and 95% confidence intervals (95%CI). Analyses were carried out using the R statistical package for windows version 3.0.2 (27) and the sandwich package (28).

Results

Results were similar across all four treatment comparisons to no chemotherapy. As an example, we focus on the effect of carboplatin compared to no chemotherapy, results from the other comparisons are presented in the **Appendix**.

Of the 199 dogs included in the carboplatin vs. no chemotherapy analysis, 47 were treated with carboplatin and 152 with no chemotherapy; within 5 months 69 dogs died. Baseline characteristics are presented in **Table 2** and for the other treatment options in the **Appendix** (**Tables A2-A4**). The crude main treatment effect of carboplatin on 5 month mortality compared to no chemotherapy was RR 0.49 (95%CI 0.26; 0.92). After adjustment for potential confounders the treatment RR was 0.49 (95%CI 0.25; 0.94), see Table 3. As previously, stated results from the other comparisons were similar; with the possible exception of the cisplatin effect, which was non-significant (**Table 3**).

Testing for treatment by baseline risk interaction revealed that the effects of carboplatin (compared to no chemotherapy) decreased with increasing baseline risk (**Table 4**); interaction P-value = 0.01. For example, in dogs with a baseline risk of 5-month mortality between 0.24 and 0.30 carboplatin therapy reduced this risk by 87% (RR 0.13, 95%CI 0.02; 1.00), whereas in dogs with a baseline risk of 5-month mortality between 0.36 and 0.43 carboplatin therapy

Table 2. Baseline characteristics of canines with osteosarcoma stratified by treatment status*.

Variables	No chemotherapy; N = 152	Carboplatin; N = 47
5-month mortality N (%)	60(39%)	9(19%)
Age (years) mean(sd)	8.88(2.84)	8.09(2.69)
Weight (kg) mean(sd)	33.17(13.69)	32.79(19.87)
Male gender N (%)	79(52%)	32(68%)
Neutered N (%)	112(74%)	36(77%)
High SALP N (%)	67(44%)	11(23%)
Breed		
Other N (%)	68(45%)	24(51%)
Rottweiler N (%)	14(9%)	6(13%)
Golden Retriever N (%)	7(5%)	0(0%)
Labrador Retriever N (%)	14(9%)	2(4%)
Greyhound N (%)	5(3%)	2(4%)
Doberman N (%)	8(5%)	4(9%)
Irish Setter N (%)	1(1%)	1(2%)
Mixed N (%)	35(23%)	8(17%))
Tumor location		
Other N (%)	88(58%)	17(36%)
Prox. Humerus N (%)	19(12%)	9(19%)
Dist. Femur or Prox. Tibia N (%)	25(16%)	5(11%)
Dist. Radius N (%)	20(13%)	16(34%)
Logit(baseline risk) mean (sd)	-0.67(0.48)	-0.87(0.49)
Baseline risk mean (sd)	0.35(0.11)	0.30(0.09)

*Serum alkaline phosphatase (SALP); N equals the number of subjects, sd equals the standard deviation. These dogs were originally included in studies by Amsellum, Bacon, Kirpenstijn, Kow and Maritato.

seemed to increase mortality risk (RR 1.86, 95%CI 0.70; 4.95). **Figure 1** shows the treatment effect of carboplatin against the baseline risk, based on a model assuming a linear increase (or decrease) in the treatment effect; model details are presented in **Table A5**. For the other treatment comparison, cisplatin, doxorubicin and doxorubicin combination, similar interactions were found (**Tables A5-A8** in the **Appendix**). Depending on the comparison the similarity between the linear and non-linear quintile specific treatment estimates, given in **Table 4** and **A6-A8**, differed.

Table 3. Treatment effect estimates of different chemotherapeutics compared to no chemotherapy on 5-month mortality*.

	Carboplatin	Cisplatin	Doxorubicin	Doxorubicin combination
Crude model				
Treatment effect	0.49 (0.26; 0.92)	0.86 (0.55; 1.36)	0.48 (0.29; 0.78)	0.61 (0.39; 0.95)
Model adjusted for confounders(except breed)				
Treatment effect	0.55 (0.28; 1.08)	0.98 (0.59; 1.60)	0.47 (0.28; 0.77)	0.62 (0.39; 0.99)
Model additionally adjusting for breed				
Treatment effect	0.49 (0.25; 0.94)	0.86 (0.54; 1.38)	0.48 (0.29; 0.78)	0.62 (0.39; 0.97)

Results presented as risk ratios (RRs) and 95 % confidence intervals (95%).

Discussion

This study showed that dogs with osteosarcoma and a relatively low 5-month mortality risk at baseline benefited more from additional chemotherapy (carboplatin, cisplatin, doxorubicin or doxorubicin combination therapy) than those with a worse baseline prognosis. In our sample dogs with a baseline risk of 36% (based on the quintile specific estimates) or lower seemed

to have benefited from additional treatment with chemotherapy.

Previous clinical studies showed that the effect of chemotherapy might be modified by another factor. One of the clearest examples of this in dogs is the synergistic effect between immunotherapy and chemotherapy (29;30). To the best of our knowledge, our study is the first to explore whether treatment effects vary according to baseline mortality risk (using multiple variables).

The current study has some limitations. First, only data from nonrandomized studies were available. Therefore, treatment effect estimates could be biased due to unobserved and residual confounding. This is most likely for the cisplatin comparison were the main treatment effect estimate did not significantly differ from a neutral risk ratio of 1, which is contrary to the expected benefit of treatment. Second, despite combining data from different individual studies, sample size was limited. This complicated confounding adjustment, interaction testing and exploration of non-linear relations. However and importantly, adjustment for clustering by study did not markedly influence results (data no shown). The limitations of sample size is clearly shown in the baseline risk quintile specific treatment effect estimates, which lacked precision (i.e., the 95% confidence intervals were wide). While there were indications to doubt that treatment effect linearly changed with baseline risk, transformations of the baseline risk did not improve model fit. Given the limited sample size we did not correct for multiple testing and subsequent studies should confirm our findings. Third, the prediction model used in this paper was not validated in external data. Another limitation is the fact that 22%, 16%, 18% and 21% of the data was missing for the carboplatin, cisplatin, doxorubicin and doxorubicin combination comparisons, respectively. Information on exposure was missing 0.5%, 1%, 3% and 5% of the times for the carboplatin, cisplatin, doxorubicin and doxorubicin combination comparisons, respectively. The outcome was not available in 6%, 7%, 10% and 10% of the data for carboplatin, cisplatin, doxorubicin and doxorubicin combination therapy, respectively. Assuming that at least some of the missing values were dependent on measured factors, missing values were imputed to reduce bias compared to a complete cases analysis (31;32). Despite these limitations our findings suggest that for all four different comparisons dogs with a relatively low 5-month mortality risk at baseline benefited most from chemotherapy.

Table 4. Treatment effect estimates of carboplatin chemotherapy compared to no chemotherapy on 5-month mortality in dogs with osteosarcoma stratified by baseline risk*.

Treatment effects	Mortality events/group size		RR (95%CI) [non-linear]*	RR (95%CI) [linear]*
	Carboplatin	*No chemotherapy*		
Effect stratified for baseline risk quintiles				
Quintile 1; Baseline risk (0.08 ; 0.24)	0/9	13/31	**NA**	**0.12 (0.03; 0.48)**
Quintile 2; Baseline risk (0.24 ; 0.30)	1/16	12/24	**0.13 (0.02; 1.00)**	**0.26 (0.11; 0.64)**
Quintile 3: Baseline risk (0.30 ; 0.36)	1/5	16/34	**0.43 (0.04; 4.44)**	**0.47 (0.24; 0.92)**
Quintile 4; Baseline risk (0.36 ; 0.43)	6/14	6/26	**1.86 (0.70; 4.95)**	**0.79 (0.40; 1.57)**
Quintile 5; Baseline risk (0.43 ; 0.63)	1/3	13/37	**0.95 (0.06; 14.02)**	**1.84 (0.61; 5.56)**

*Non-linear treatment estimates are based on the quintile specific estimates. The linear treatment estimates are based on equation 2 and treatment and interaction estimates given in appendix Table A5.

Figure 1. Estimated relative treatment effect of carboplatin therapy compared to no chemotherapy on 5-month mortality in surgically treated dogs with osteosarcoma.

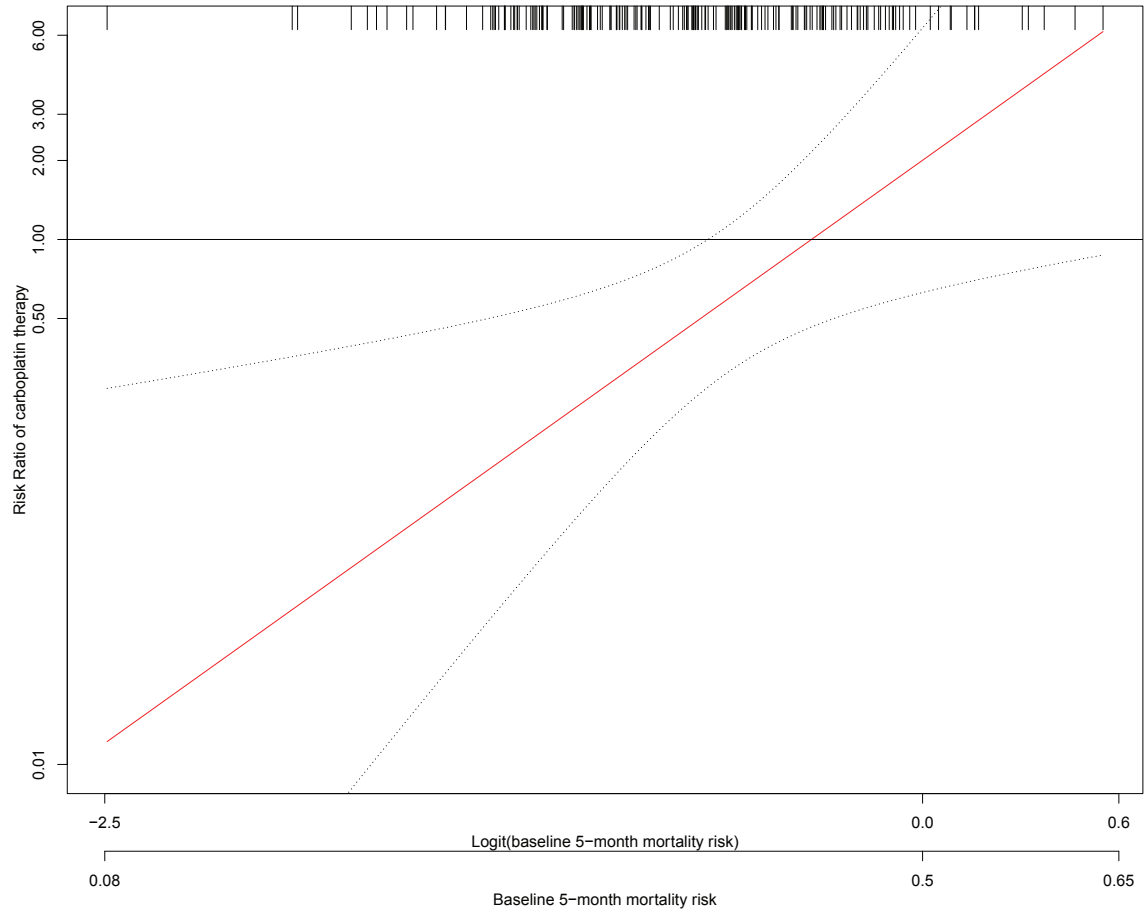


Figure shows the risk ratio (RR) of carboplatin treatment (solid line) with 95% confidence intervals (dotted lines) for different baseline risk. The horizontal solid line indicates a neutral RR of 1.00. At the top a rug plot is given, corresponding to the patient frequencies of the x-axis measurements.

As noted previously, for each comparison a different subset of the five studies was used, thus to some extend these comparisons were independent, decreasing the likelihood that results were due to a common data irregularity. Given the described shortcomings, we feel that before these findings can be used in clinical practice, additional research should try to replicate our results and specifically focus on the following: first, the prediction model used should be updated to include more clinical parameters such as tumor grading and be subsequently validated in external data. Second, these and possibly other treatment comparisons should be repeated, preferably using randomized controlled trial data to prevent confounding. Finally, if the results are replicated in external data, an algorithm such as a nomogram or a (spreadsheet) program might be developed to aid detection of patients with an increased or decreased benefit of chemotherapy in clinical practice.

Conclusions

In conclusion, surgically treated dogs with osteosarcoma who have a relatively low risk of 5-month mortality might benefit most from treatment with carboplatin, cisplatin, doxorubicin or doxorubicin combination chemotherapy.

Reference List

1. Norrdin RW, Powers BE, Torgersen JL, Smith RE, Withrow SJ. Characterization of osteosarcoma cells from two sibling large-breed dogs. American Journal of Veterinary Research 1989 Nov;50(11):1971-5.
2. Spodnick GJ, Berg J, Rand WM, Schelling SH, Couto G, Harvey HJ, et al. Prognosis for dogs with appendicular osteosarcoma treated by amputation alone: 162 cases (1978-1988). Journal of the American Veterinary Medical Association 1992 Apr 1;200(7):995-9.
3. Cooley DM, Waters DJ. Skeletal neoplasms of small dogs: a retrospective study and literature review. Journal of the American Animal Hospital Association 1997 Jan;33(1):11-23.
4. Ru G, Terracini B, Glickman LT. Host related risk factors for canine osteosarcoma. The Veterinary Journal 1998 Jul;156(1):31-9.
5. McNeill CJ, Overley B, Shofer FS, Kent MS, Clifford CA, Samluk M, et al. Characterization of the biological behaviour of appendicular osteosarcoma in Rottweilers and a comparison with other breeds: a review of 258 dogs. Veterinary and Comparative Oncology 2007 Jun;5(2):90-8.
6. Brodey RS, Abt DA. Results of surgical treatment in 65 dogs with osteosarcoma. Journal of the American Veterinary Medical Association 1976 Jun 1;168(11):1032-5.
7. Straw RC, Withrow SJ. Limb-sparing surgery versus amputation for dogs with bone tumors. Veterinary Clinics of North America: Small Animal Practice 1996 Jan;26(1):135-43.
8. Straw RC, Withrow SJ, Richter SL, Powers BE, Klein MK, Postorino NC, et al. Amputation and cisplatin for treatment of canine osteosarcoma. Journal of Veterinary Internal Medicine 1991 Jul;5(4):205-10.
9. Bailey D, Erb H, Williams L, Ruslander D, Hauck M. Carboplatin and doxorubicin combination chemotherapy for the treatment of appendicular osteosarcoma in the dog. Journal of Veterinary Internal Medicine 2003 Mar;17(2):199-205.
10. Chun R, Kurzman ID, Couto CG, Klausner J, Henry C, MacEwen EG. Cisplatin and doxorubicin combination chemotherapy for the treatment of canine osteosarcoma: a pilot study. Journal of Veterinary Internal Medicine 2000 Sep;14(5):495-8.
11. Chun R, Garrett LD, Henry C, Wall M, Smith A, Azene NM. Toxicity and efficacy of cisplatin and doxorubicin combination chemotherapy for the treatment of canine osteosarcoma. Journal of the American Animal Hospital Association 2005 Nov;41(6):382-7.
12. Vail DM, Kurzman ID, Glawe PC, O'Brien MG, Chun R, Garrett LD, et al. STEALTH liposome-encapsulated cisplatin (SPI-77) versus carboplatin as adjuvant therapy for spontaneously arising osteosarcoma (OSA) in the dog: a randomized multicenter clinical trial. Cancer Chemotherapy and Pharmacology 2002 Aug;50(2):131-6.
13. Schmidt AF, Nielen M, Klungel OH, Hoes AW, de Boer A, Groenwold RH, et al. Prognostic factors of early metastasis and mortality in dogs with appendicular osteosarcoma after receiving surgery: An individual patient data meta-analysis. Preventive Veterinary Medicine 2013 Nov 1;112(3-4):414-22.
14. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. Heart 2012 May;98(9):683-90.
15. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Medical Research Methodology 2006;6:18.
16. Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 2010;11:85.
17. Rothwell PM, Warlow CP. Prediction of benefit from carotid endarterectomy in individual patients: a risk-modelling study. European Carotid Surgery Trialists' Collaborative Group. Lancet 1999 Jun 19;353(9170):2105-10.
18. Kow K, Thamm DH, Terry J, Grunerud K, Bailey SM, Withrow SJ, et al. Impact of telome-

rase status on canine osteosarcoma patients. Journal of Veterinary Internal Medicine 2008 Nov;22(6):1366-72.
19. Kirpensteijn J, Kik M, Rutteman GR, Teske E. Prognostic significance of a new histologic grading system for canine osteosarcoma. Veterinary Pathology 2002 Mar;39(2):240-6.
20. Rubin DB. Inference and missing data. Biometrika 1976 Dec 1;63(3):581-92.
21. Hmisc: Harrell Miscellaneous. R package [computer program]. Version R package version 3.12-2 2013.
22. Knol MJ, Le CS, Algra A, Vandenbroucke JP, Groenwold RH. Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression. Canadian Medical Association Journal 2012 May 15;184(8):895-9.
23. Zou G. A modified poisson regression approach to prospective studies with binary data. American Journal of Epidemiology 2004 Apr 1;159(7):702-6.
24. Cribari-Neto F, Silva W. A new heteroskedasticity-consistent covariance matrix estimator for the linear regression model. AStA Advances in Statistical Analysis 2011;95(2):129-46.
25. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. Journal of the American Medical Association 2007 Sep 12;298(10):1209-12.
26. Dorresteijn JA, Visseren FL, Ridker PM, Wassink AM, Paynter NP, Steyerberg EW, et al. Estimating treatment effects for individual patients based on the results of randomised clinical trials. BMJ 2011;343:d5888.
27. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
28. Zeileis A. Object-Oriented Computation of Sandwich Estimators. Journal of Statistical Software 2006;16(9):1-16.
29. MacEwen EG, Kurzman ID. Canine osteosarcoma: amputation and chemoimmunotherapy. Veterinary Clinics of North America: Small Animal Practice 1996 Jan;26(1):123-33.
30. Vail DM, MacEwen EG, Kurzman ID, Dubielzig RR, Helfand SC, Kisseberth WC, et al. Liposome-encapsulated muramyl tripeptide phosphatidylethanolamine adjuvant immunotherapy for splenic hemangiosarcoma in the dog: a randomized multi-institutional clinical trial. Clinical Cancer Research 1995 Oct;1(10):1165-70.
31. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statist Med 2010 Dec 10;29(28):2920-31.
32. Groenwold RH, Donders AR, Roes KC, Harrell FE, Jr., Moons KG. Dealing with missing outcome data in randomized trials and observational studies. American Journal of Epidemiology 2012 Feb 1;175(3):210-7.

Table A1. Characteristics of studies included in the IPDMA on the effect of chemotherapy compared to no chemotherapy in canine's with osteosarcoma treated with amputation.

Study	*Published?*	*Design*	*Chemotherapeutic intervention*				
			No chemotherapy *Event/Non-event*	Carboplatin *Event/Non-event*	Cisplatin *Event/Non-event*	Doxorubicin *Event/Non-event*	Doxorubicin combinations *Event/Non-event*
Amsellem	No	NR	1/4	2/13	-	-	3/7
Bacon	No	NR	19/24	2/5	12/23	7/40	3/10
Kirpensteijn (Kirpensteijn et al., 2002)	Yes	NR	26/33	4/2	4/3	2/4	10/15
Kow (Kow et al., 2008)	Yes	NR	7/6	1/8	-	5/14	2/19
Maritato	No	NR	7/25	0/10	-	2/9	1/9

Event indicates the number of dogs who were dead at 5 months, Non-event represent the number of dogs that were alive at 5 months. NR = Non Randomized study.

Table A2. Baseline characteristics of canines with osteosarcoma stratified by treatment status.

Variables	No chemotherapy; N = 102	Cisplatin; N = 42
5-month mortality N (%)	45(44%)	16(38%)
Age (years) mean (sd)	8.83(2.74)	8.55(2.77)
Weight (kg) mean (sd)	34.61(13.45)	35.30(12.42)
Male gender N (%)	48(47%)	21(50%)
Neutered N (%)	68(67%)	35(83%)
High SALP N (%)	50(49%)	23(55%)
Breed		
Other N (%)	48(47%)	16(38%)
Rottweiler N (%)	13(13%)	1(2%)
Golden Retriever N (%)	3(3%)	7(17%)
Labrador Retriever N (%)	9(9%)	3(7%)
Greyhound N (%)	1(1%)	2(5%)
Doberman N (%)	7(7%)	1(2%)
Irish Setter N (%)	1(1%)	1(2%)
Mixed N (%)	20(20%)	11(26%))
Tumor location		
Other N(%)	63(62%)	22(52%)
Prox. Humerus N (%)	14(14%)	2(5%)
Dist. Femur or Prox. Tibia N (%)	11(11%)	9(21%)
Dist. Radius N (%)	14(14%)	9(21%)
Logit(baseline risk) mean (sd)	-0.56(0.47)	-0.72(0.57)
Baseline risk mean (sd)	0.37(0.10)	0.34(0.12)

Serum alkaline phosphatase (SALP); N equals the number of subjects, sd equals the standard deviation. These dogs were originally included in studies by Bacon and Kirpensteijn.

Table A3. Baseline characteristics of canines with osteosarcoma stratified by treatment status.

Variables	No chemotherapy; N = 147	Doxorubicin; N = 83
5-month mortality N (%)	59(40%)	16(19%)
Age (years) mean (sd)	8.88(2.88)	9.34(2.62)
Weight (kg) mean (sd)	33.91(13.28)	36.70(11.42)
Male gender N (%)	77(52%)	45(54%)
Neutered N (%)	109(74%)	73(88%)
High SALP N (%)	65(44%)	41(49%)
Breed		
Other N (%)	65(44%)	24(29%)
Rottweiler N (%)	14(10%)	12(14%)
Golden Retriever N (%)	7(5%)	11(13%)
Labrador Retriever N (%)	14(10%)	3(4%)
Greyhound N (%)	5(3%)	10(12%)
Doberman N (%)	8(5%)	1(1%)
Irish Setter N (%)	1(1%)	2(2%)
Mixed N (%)	33(22%)	20(24%))
Tumor location		
Other N (%)	85(58%)	27(33%)
Prox. Humerus N (%)	19(13%)	22(27%)
Dist. Femur or Prox. Tibia N (%)	23(16%)	22(27%)
Dist. Radius N (%)	20(14%)	12(14%)
Logit(baseline risk) mean (sd)	-0.66(0.48)	-0.62(0.52)
Baseline risk mean (sd)	0.35(0.11)	0.36(0.11)

Serum alkaline phosphatase (SALP); N equals the number of subjects, sd equals the standard deviation. These dogs were originally included in studies by Bacon, Kirpensteijn, Kow and Maritato.

Table A4. Baseline characteristics of canines with osteosarcoma stratified by treatment status.

Variables	No chemotherapy; N = 152	Doxorubicin combination chemotherapy; N = 79
5-month mortality N (%)	60(39%)	19(24%)
Age (years) mean (sd)	8.88(2.84)	8.12(2.93)
Weight (kg) mean (sd)	33.17(13.69)	35.82(15.55)
Male gender N (%)	79(52%)	45(57%)
Neutered N (%)	112(74%)	63(80%)
High SALP N (%)	67(44%)	28(35%)
Breed		
Other N (%)	68(45%)	35(44%)
Rottweiler N (%)	14(9%)	10(13%)
Golden Retriever N (%)	7(5%)	5(4%)
Labrador Retriever N (%)	14(9%)	3(4%)
Greyhound N (%)	5(3%)	2(3%)
Doberman N (%)	8(5%)	1(1%)
Irish Setter N (%)	1(1%)	5(6%)
Mixed N (%)	35(23%)	18(23%))
Tumor location		
Other N (%)	88(58%)	26(33%)
Prox. Humerus N (%)	19(12%)	15(19%)
Dist. Femur or Prox. Tibia N (%)	25(16%)	18(23%)
Dist. Radius N (%)	20(13%)	20(25%)
Logit(baseline risk) mean (sd)	-0.67(0.48)	-0.81(0.59)
Baseline risk mean (sd)	0.35(0.11)	0.32(0.12)

Serum alkaline phosphatase (SALP); N equals the number of subjects, sd equals the standard deviation. These dogs were originally included in studies by Amsellem, Bacon, Kirpensteijn, Kow and Maritato.

Table A5. Treatment effect and interaction effect estimates of cisplatin chemotherapy compared to no chemotherapy on 5-month mortality in dogs with osteosarcoma modified by baseline risk.

	Carboplatin	Cisplatin	Doxorubicin	Doxorubicin combination
Model with baseline risk by treatment interaction				
Treatment effect	2.00 (0.63; 6.39)	1.49 (0.76; 2.92)	0.93 (0.49; 1.77)	1.11 (0.60; 2.06)
Interaction effects	7.73 (1.56; 38.45)	2.42 (0.88; 6.65)	3.62 (1.15; 11.34)	2.38 (1.12; 5.05)

Results presented as risk ratios (RRs) and 95 % confidence intervals (95%). The treatment effect estimates can be interpreted as the treatment effect when the logit(baseline risk) equals 0 or stated otherwise when the baseline risk for 5-month mortality is 0.50. Assuming linearity the interaction effects indicates how much the treatment effect changes with changing logit(baseline risk); a 95%CI excluding the 1 indicates a significant interaction.

Table A6. Treatment effect estimates of cisplatin chemotherapy compared to no chemotherapy on 5-month mortality in dogs with osteosarcoma stratified by baseline risk.

Treatment effects	**Mortality events/group size**		**RR (95%CI) [non-linear]***	**RR (95%CI) [linear]***
	Cisplatin	*No chemotherapy*		
Effect stratified for baseline risk quintiles				
Quintile 1; Baseline risk (0.11; 0.25)	3/12	10/17	**0.43 (0.13; 1.35)**	**0.46 (0.18; 1.20)**
Quintile 2; Baseline risk (0.26; 0.33)	4/11	8/18	**0.82 (0.29; 2.28)**	**0.68 (0.37; 1.25)**
Quintile 3: Baseline risk (0.33; 0.38)	1/4	12/24	**0.50 (0.04; 5.89)**	**0.88 (0.55; 1.40)**
Quintile 4; Baseline risk (0.38; 0.46)	3/7	7/22	**1.35 (0.40; 4.53)**	**1.13 (0.69; 1.83)**
Quintile 5; Baseline risk (0.47; 0.63)	5/8	8/21	**1.64 (0.70; 3.83)**	**1.60 (0.77; 3.35)**

**Non-linear treatment estimates are based on the quintile specific estimates. The linear treatment estimates are based on equation 2 and treatment and interaction estimates given in appendix Table A5*

Table A7. Treatment effect estimates of doxorubicin chemotherapy compared to no chemotherapy on 5-month mortality in dogs with osteosarcoma stratified by baseline risk.

Treatment effects	Mortality events/group size		RR (95%CI) [non-linear]*	RR (95%CI) [linear]*
	Doxorubicin	*No chemotherapy*		
Effect stratified for baseline risk quintiles				
Quintile 1; Baseline risk (0.11; 0.26)	1/14	14/32	**0.16 (0.02; 1.35)**	**0.16 (0.05; 0.58)**
Quintile 2; Baseline risk (0.26; 0.32)	3/18	15/28	**0.31 (0.10; 0.99)**	**0.30 (0.13; 0.66)**
Quintile 3: Baseline risk (0.32; 0.37)	2/16	11/30	**0.34 (0.08; 1.50)**	**0.40 (0.22; 0.75)**
Quintile 4; Baseline risk (0.37; 0.45)	3/17	10/29	**0.51 (0.15; 1.73)**	**0.57 (0.34; 0.95)**
Quintile 5; Baseline risk (0.45; 0.63)	7/18	9/28	**1.21 (0.53; 2.78)**	**1.00 (0.51; 1.97)**

**Non-linear treatment estimates are based on the quintile specific estimates. The linear treatment estimates are based on equation 2 and treatment and interaction estimates given in appendix Table A5*

Table A8. Treatment effect estimates of carboplatin chemotherapy compared to no chemotherapy on 5-month mortality in dogs with osteosarcoma stratified by baseline risk.

Treatment effects	Mortality events/group size		RR (95%CI) [non-linear]*	RR (95%CI) [linear]*
	Doxorubicin combinations	*No chemotherapy*		
Effect stratified for baseline risk quintiles				
Quintile 1; Baseline risk (0.08; 0.24)	3/18	11/29	**0.44 (0.13; 1.46)**	**0.32 (0.15; 0.71)**
Quintile 2; Baseline risk (0.24; 0.30)	5/17	16/29	**0.53 (0.23; 1.26)**	**0.47 (0.27; 0.82)**
Quintile 3: Baseline risk (0.30; 0.36)	1/16	14/30	**0.13 (0.02; 1.08)**	**0.60 (0.38; 0.94)**
Quintile 4; Baseline risk (0.36; 0.43)	5/17	6/29	**1.42 (0.48; 4.18)**	**0.75 (0.48; 1.19)**
Quintile 5; Baseline risk (0.43; 0.67)	5/11	13/35	**1.22 (0.52; 2.87)**	**1.12 (0.60; 2.08)**

**Non-linear treatment estimates are based on the quintile specific estimates. The linear treatment estimates are based on equation 2 and treatment and interaction estimates given in appendix Table A5*

Part IV

Generalizability of the effects of interventions

CHAPTER 6

The generalizability of randomized controlled trial results of the effects of beta-blockers compared to diuretics on the risk of non-fatal myocardial infarction

A F Schmidt, R H H Groenwold, F Gueyffier, A W Hoes, A de Boer
M Nielen, O H Klungel

In revision

Abstract

Purpose To explore the generalizability of the RCT effect estimates of atenolol and propranolol compared to diuretics on the risk of non-fatal myocardial infarction (MI) and whether generalizability differed between age groups.

Methods The effect of beta-blocker versus diuretic use on the risk of MI was estimated using data from two RCTs (antihypertensive MRC trials), a case-control study, and a cohort study. Treatment effect modification by age was assessed.

Results The estimated effects of propranolol compared to diuretics did not differ between designs: HR 0.86 (95%CI 0.59; 1.26) and HR 0.61 (95%CI 0.26; 1.42), for the RCT and cohort study, respectively. While the atenolol effect estimates (compared to diuretics) differed from the propranolol estimates, they were similar across the differently designed studies: HR 2.17 (95%CI 1.11; 4.23) in the RCT, HR 1.61 (95%CI 1.06; 2.45) in the cohort, and OR 0.96 (95%CI 0.43; 2.15) in the case-control study. Results from the cohort study indicate that the effect of atenolol might change with age: treatment by age (per 10 years) interaction HR 0.63 (95%CI 0.40; 0.99). Similar (though non-significant) interaction effects were observed in the RCT and case-control data (for atenolol and propranolol.

Conclusions The RCT effect estimates of propranolol and atenolol compared to diuretics were similar to those in nonrandomized data suggesting generalizability of effect estimates to less controlled settings. Compared to diuretics, atenolol increased the hazard of MI; propranolol on the other hand seemed to be equally effective. An age interaction could not be excluded.

Background

In comparative effectiveness research, randomized clinical trials (RCTs) are considered the gold standard to assess treatment effects. One reason for this is that random allocation of treatment prevents confounding (i.e., treatment groups will have the same baseline risk for the outcome). It is well known however, that patients included in RCTs may differ from those included in nonrandomized studies. The latter include “real life” patients, whereas RCTs may include highly selected patient populations, as a result of strict in- and exclusion criteria (1-3). If treatment effects differs between patients included and excluded, effect estimates based on RCTs are not generalizable to all patients(4;5).

To explore generalizability, researchers have compared results from RCTs to results from nonrandomized studies and in some cases found comparable results (6-10).Empirical evidence showed that generalizability might be subgroup specific(10). Previously, the generalizability of a number of interventions (e.g., statin, beta-blocker and CABG therapy) has been explored. However, generalizability of subgroup effects has not been previously addressed.

We assessed to what extent the RCT effect estimates of propranolol and atenolol compared to diuretics on the risk of non-fatal myocardial infarction (MI) are generalizable to estimates based on nonrandomized data. Furthermore, we also assessed whether the relative effectiveness and generalizability of treatment effects was constant across patients subgroup defined by age (11). For this, individual patient data was used from two RCTs and two nonrandomized studies based on electronic health care record databases (12;13).

Methods

To assess the generalizability of the treatment effect estimates of propranolol and atenolol compared to diuretics on the risk of non-fatal myocardial infarction (MI) individual patient data was obtained from the INDANA (14) and the PHARMO(15) databases. In these data, we first estimated the effect of beta-blockers (propranolol or atenolol) compared to diuretics. Second, we explored whether effect estimates were constant across age categories. Third, recognizing that propranolol and atenolol might differ in relative effectiveness, both compounds were individually compared to diuretics and effect modification by age was explored.

Data sources

From the INDANA database we extracted data from the two MRC antihypertensive trials (16;17), which were used as the reference standard in the current study. The MRC trials recruited untreated hypertensive patients. The first RCT (MRC younger) included patients aged 35 through 64 years with diastolic blood pressure (DBP) between 90 and 109 and a systolic blood pressure (SBP) < 200 mmHg. The second trial (MRC older) recruited patients aged 65 through 74 years with a DBP < 115 and with a SBP between 160 and 209 mmHg. In the MRC younger, patients were randomized to propanolol, bendrofluazide or placebo. In the MRC older, atenolol, hydrochlorothiazide (or hydrochlorothiazide with amiloride) or placebo were given. For the current comparison, patients randomized to placebo (n = 10,867) as well as those aged below 35 (n = 20) and above 75 years (n = 10) were excluded, resulting in 10,853 patients within the age inclusion criteria of the original trial. In the MRC trials, outcomes were assessed by the researchers using general practitioners' documents, hospital files, and ECGs.

As nonrandomized counterparts, we used a case-control and a cohort study, both sampled from the PHARMO database. The PHARMO database includes drug-dispensing histories from a sample of Dutch community pharmacies (including about 2,000,000 subjects (15)) that are linked to the national hospital discharge registry. Drug-dispensing is registered using the Anatomical Therapeutic Chemical (ATC) classification system. Subjects were eligible for the nonrandomized studies when they were incident users (at least 1 year without antihypertensive drug) starting on mono-therapy and were registered in the PHARMO database for at least 1 year. Atenolol use was defined as ATC code C07AB03, propranolol users were selected based on ATC C07AA05. Use of diuretic was defined based on the ATC codes C03AA03, C03AA04, C03BA04, C03BA05, C03BA11, C03DB01, C03EA01, C03EA03. See **Appendix Table 1** for number of users per compound. MI events were retrieved from hospital discharge records (ICD-9 code 410).

From the PHARMO database, 11,471 patients aged between 35 and 75 were sampled. In the cohort study information was available on age, sex, diabetes status and drug dispensing and hospital discharge records.

To explore the relevance of additional confounding adjustment the Utrecht Cardiovascular

Pharmacogenetic (UCP) study was used. This study was designed as a case-control study nested among antihypertensive drug users in the PHARMO database (15). This study included additional information on the baseline lifestyle factors: BMI, smoking habits, hypercholesterolemia, physical activity, and alcohol use. Cases were hospitalized for a first MI and prior to the event were on antihypertensive medication. Control patients were selected using incidence sampling and met the same criteria as cases but were not hospitalized for MI. Cases and controls were matched on age, sex and pharmacy location. After excluding subjects aged younger than 35 years (n = 7) and older than 75 years (n =112), the total number of patients consisted of 150 cases and 916 controls.

Data Analysis

All analyses were done using the R statistical package, version 3.0.2 (18) and the survival (19) package. RCT data were analysed according to the intention-to-treat principle. In the cohort study, exposure status was based on the first prescription of beta-blocker or diuretics during follow-up (and considered constant during follow-up). In the case-control study, exposure status was based on the prescription filled in the 90 days prior to the index date (i.e., becoming a case or control).

Both the RCTs and the cohort data were analysed using multivariable Cox proportional hazard models (20), with calendar time as time axis. Subjects were censored when lost to follow-up, or at study end (22-01-1985, 31-07-1990 and 31-03-2005, for the younger, older MRC trials and the PHARMO cohort). The proportional hazard assumption was assessed by graphing time against the scaled Schoenfeld residuals (a time varying beta coefficient) and by testing whether there was an interaction with time using a global chi-square test; no deviations were observed. The case-control study was analysed using conditional multivariable logistic regression. Effect estimates are presented as hazard ratios (HRs) or odds ratios (ORs) with 95% confidence intervals (95%CI). Observations with missing values were excluded from the analysis (complete case analysis), which resulted in exclusion of 3%, 17%, and 0% of the subjects from the RCT, case-control, and cohort data, respectively.

Treatment effect estimates were adjusted for baseline variables using multivariable regression models. Three different models were constructed with increasing adjustments. Model 1 estimated the association of exposure with the outcome including variables for age and

gender. Model 2, which was only fitted for the RCTs and case-control study, additionally included BMI, smoking status, cholesterol (or hypercholesterolemia), and in the case-control study also physical activity and alcohol use. Note that, because these variables were not available in the cohort data, model 2 was not fitted for the cohort study. Model 3 corrected for a dichotomous morbidity or co-medication variable (defined subsequently), which indicated the presence or absence of co-morbidity or co-medication at first drug use.

The dichotomous co-morbidity or co-medication variable mimicked the exclusion criteria of the MRC trials (see **Appendix**) (16;17). Co-morbidity was based on the following ICD-9 codes: essential hypertension (401), secondary hypertension (402 through 405), liver cirrhosis (571), nephrotic syndrome (581), (nephrogenic) diabetes insipidus (253.5 and 588,1), nephrolithiasis (592.0 and 274.11), angina pectoris (413), tachycardia (785.0), migraine (346.0 through 346.9; excluding 346.4), gout (274), intermittent claudication (440.21), COPD, chronic bronchitis, asthma (490 through 496), malignancies (140 through 239; excluding 210 through 229), stroke (434.91, 434.11, 430, 431, 432), transient ischemic attack (TIA) (435), late effects of cerebrovascular disease (438), heart failure (428), cardiac dysrhythmias (427), or because of admission for a previous MI (410 and 412). Similarly Co-medication use was defined using the ACT codes: antidepressants (N06A), statins (C10), nitrates (C01D), drugs for obstructive airway diseases (R03), or anti-diabetic drugs (A10).The presence of co-medication was defined as any filled prescription in the 6 months prior to first antihypertensive prescription. Similarly, co-morbidity was based on any hospital prior to the first antihypertensive prescription registered. Since the case-control study was nested in the PHARMO database, the co-morbidity and/or co-medication variable was also based on data prior to first antihypertensive drug use.

A next step in the analysis was to assess whether treatment effects changed by patient subgroups defined by age and whether this impacted generalizability. To assess this, an age (continuous) by treatment interaction term was added to the previously defined models. To simplify interpretation of the interaction coefficient results are presented per 10 years and centred at 65 years; i.e., 65 was subtracted from every age measurement centring zero at 65 and divided by 10. Deviations of linearity of the interaction effect was tested using a likelihood ratio test comparing a linear model with a model including a restricted cubic spline for the interaction effect with five knots; no significant deviations were found.

Due to the age inclusion criteria of the RCTs, propranolol was only administered to subjects aged < 65 and atenolol only to subjects ≥ 65 years. Therefore, the treatment by age interaction effects could not differentiate between age, center or treatment effects. Furthermore, there is evidence that the relative effectiveness of atenolol and propranolol might differ (21-23). To explore this, all analyses (for the RCTs, cohort and case-control study) were repeated comparing atenolol to diuretics or propranolol to diuretics (as two separate comparisons).

Results

Baseline characteristics of included patients are presented in **Table 1**. The MRC trials included 10,863 subjects of whom 159 (1.5%) experienced a non-fatal myocardial infarction (MI). In the cohort study 0.9% (N = 103) experienced an MI event, whereas in the case-control study 150 subjects were included as cases and 916 as controls. In the RCTs 1095 (28 MI events; 2.6%) subjects were randomized to atenolol and 4389 (56 MI events; 1.3%) to propranolol (**Table 1**). In the cohort study 4154 subjects started on atenolol (58 MI events 1.4%) and 2341 on propranolol (7 MI events 0.3%). Of the 150 cases included in the case control study, 71 used atenolol and 17 propranolol.

Table 1 Baseline characteristics of the RCTs, the cohort study and the case-control study, stratified for type of beta-blocker used.

	RCT Atenolol			*RCT Propranolol*		
	Diuretics N=1077	Atenolol N=1095	Missing	Diuretics N=4288	*Propranolol* N=4389	Missing
Non-fatal MI, %	15 (1%)	28 (3%)	0	60 (1%)	56 (1%)	0
Follow-up time (years) median (IQR)	5.5 (4.1;6.5)	5.4 (3.4;6.5)	0	5.0 (4.0;5.5)	5.0 (4.0;5.5)	0
Men (%)	452 (42%)	453 (41%)	0	2231 (52%)	2276 (52%)	0
Age (years) median (IQR)	70.6 (68.2;72.7)	70.4 (68.1;72.7)	0	52.7 (46.7;58.2)	52.7 (46.8;58.2)	0
BMI (kg/m$^{\wedge 2}$) mean (sd)	26.5 (3.9)	26.6 (4.0)	3	27.2 (4.2)	27.1 (4.2)	5
Smokers (%)	230 (21%)	244 (22%)	1	1258 (29%)	1224 (28%)	47
Serum cholesterol (mmol/L) mean (sd)	6.5 (1.3)	6.5 (1.2)	9	6.5 (1.1)	6.5 (1.2)	233
Co-medication or morbidity (%)	0 (0%)	0(0%)	0	0 (0%)	0 (0%)	0
Complete observations (%)	1072 (100%)	1087 (99%)	13	4144 (97%)	4253 (97%)	280
	Cohort study Atenolol			*Cohort study Propranolol*		
	Diuretics N=4976	Atenolol N=4154	Missing	Diuretics N=4976	*Propranolol* N=2341	Missing
Non-fatal MI, %	38 (1%)	58 (1%)	0	38 (1%)	7 (0%)	0
Follow-up time (years) median (IQR)	2.7 (1.2;5.0)	3.0 (1.4;4.9)	0	2.7 (1.2;5.0)	3.1 (1.5;5.0)	0
Men (%)	1674 (34%)	1828 (44%)	0	1674 (34%)	686 (29%)	1
Age (years) median (IQR)	57.3 (49.2;62.4)	53.8 (53.8;62.4)	0	57.3 (49.2;54.5)	46.4 (40.3;54.5)	0
Co-medication or morbidity (%)	1109 (22%)	677 (16%)	0	1109 (22%)	396 (17%)	0
Complete observations (%)	4976 (100%)	4154 (100%)	0	4976 (100%)	2340 (100%)	0
	Case-control study Atenolol			*Case-control study Propranolol*		
	Controls N=851	Cases N=133	Missing	Controls N=476	Cases N=79	Missing
Beta-blocker use (%)	440 (52%)	71 (53%)	0	65 (14%)	17 (22%)	0
Diuretic use (%)	411 (48%)	62 (47%)	0	411 (86%)	62 (78%)	0
Men (%)	602 (71%)	99 (74%)	0	346 (73%)	57 (72%)	0
Age (years) median (IQR)	59.2 (52.8;66.7)	61.2 (54.0;67.0)	0	59.2 (52.2;67.6)	62.3 (55.4;67.6)	0
BMI (kg/m$^{\wedge 2}$) mean (sd)	27.1 (3.9)	28.1 (6.0)	39	27.2 (4.1)	26.7 (3.6)	21
Smokers (%)	507 (63%)	90 (70%)	46	291 (65%)	54 (73%)	30
Alcohol consumer (%)	325 (39%)	40 (31%)	18	198 (42%)	21 (29%)	14
Hypercholesterolemia (%)	267 (32%)	52 (39%)	7	162 (34%)	29 (37%)	3
Co-medication or morbidity (%)	173 (20%)	34 (26%)	0	92 (19%)	22 (28%)	0
Complete observations (%)	726 (85%)	99 (74%)	159	400 (84%)	52 (66%)	103

Unless otherwise stated figures are absolute numbers. % = column percentage; IQR = Inter Quartile Range; sd = standard deviation.

Comparing propranolol and atenolol to diuretics showed no significant effect on MI: HR 1.10 (95%CI 0.80;1.53) for the RCTs, HR 1.33 (95% 0.88; 2.00) for the cohort study and OR 1.06 (95%CI 0.48; 2.30) for the case-control study (**Appendix Table 2**). Additional adjusting for baseline characteristics resulted in a small shift away from the neutral effect in the nonrandomized studies: HR 1.33 (95%CI 0.88; 2.01) and OR 1.16 (95%CI 0.50; 2.73) for the cohort and case-control studies. In the analysis of the RCT data, including a treatment by age (per 10 years) interaction term showed that as age increases the treatment HR increases as well: interaction HR 1.32 (95%CI 0.94; 1.85). The interaction effects in the nonrandomized studies were in opposite direction: HR 0.69 (95%CI 0.44; 1.06) for the cohort study and OR 0.81 (95%CI 0.36; 1.82) for the case-control study. After dichotomizing age (<65 years, and >= 65 years, i.e., stratifying for trial in the RCT data) the p-values of the interaction term were 0.04, <0.01 and 0.82 for the RCT, cohort and the case-control studies.

Given that propranolol and atenolol potentially differ in relative effectiveness, the analyses were repeated per compound (note that the propranolol RCT only included subjects aged <65 and the atenolol RCT subjects were aged between 65 and 75). Compared to diuretics propranolol seemed to be equally effective in preventing MI: HR 0.86 (95%CI 0.59; 1.26) for the RCT and HR 0.61 (95%CI 0.26; 1.43) for the cohort study (**Table 2**). Due to data sparseness, the conditional logistic regression models failed to converge for the case-control study. Interaction effects of the RCT and the cohort study pointed towards an (non-significant) increase in treatment effect with increasing age: HR 0.89 (95%CI 0.52; 1.53) per 10 years for the RCT and HR 0.63 (95%CI 0.29; 1.38) per 10 years for the cohort study (Models 1, Table 2). Compared to diuretics atenolol increased the hazard of MI in the RCT and the cohort study: HR 2.17 (95%CI 1.11; 4.23) and HR 1.61 (95%CI 1.06; 2.45). In the case-control the OR was 0.96 (95%CI 0.43; 2.15). Additional adjustment for baseline characteristics did not result in a marked change in estimate (**Table 2**). Interaction effects per 10 years were HR 0.86 (95%CI 0.06; 11.56) for the RCT, HR 0.63 (95%CI 0.40; 0.99) for the cohort study and OR 0.78 (95CI 0.35; 1.76) for the case-control study.

Discussion

This study showed that the RCTs effect estimates of the beta-blockers atenolol and propranolol compared to diuretics on non-fatal myocardial infraction (MI) were comparable to those estimated in a cohort study using electronic health care record databases. The case-control

study lacked precision therefore, generalizability to these patients could not be definitively shown. Exploring generalizability within patient subgroups defined by age, showed that there was some indication that the propranolol effects changed with age. However, generalizability between age groups could not be rejected. Finally, we showed that, in the RCT and nonrandomized data propranolol and atenolol were differently associated with the incidence of MI.

Table 2 Main and continuous age by treatment effects of propranolol versus diuretic treatment or atenolol versus diuretic treatment on non-fatal myocardial infarction in different study designs.

Model	RCT		Cohort study		Case-control study	
	Main Effects	*Interaction effects*	*Main Effects*	*Interaction effects*	*Main Effects*	*Interaction effects*
Propranolol						
Propranolol Model 1	0.86 (0.59; 1.26)	-	0.61 (0.26; 1.42)	-	N.A.	-
Propranolol Model 2	0.83 (0.57; 1.22)	-	-	-	N.A.	-
Propranolol Model 3	-	-	0.61 (0.26; 1.42)	-	N.A.	-
Propranolol Model 1 with age interaction	0.76 (0.37; 1.56)	0.89 (0.52; 1.53)	0.41 (0.12; 1.38)	0.63 (0.29; 1.38)	N.A.	N.A.
Propranolol Model 2 with age interaction	0.65 (0.31; 1.56)	0.80 (0.47; 1.38)	-	-	N.A.	N.A.
Propranolol Model 3 with age interaction	-	-	0.41 (0.12; 1.38)	0.63 (0.29; 1.38)	N.A.	N.A.
Atenolol						
Atenolol Model 1	2.17 (1.11; 4.23)	-	1.61 (1.06; 2.45)	-	0.96 (0.43; 2.15)	-
Atenolol Model 2	2.30 (1.16; 4.54)	-	-	-	0.92 (0.38; 2.23)	-
Atenolol Model 3	-	-	1.62 (1.06; 2.47)	-	0.95 (0.39; 2.33)	-
Atenolol Model 1 with age interaction	2.36 (0.46; 12.16)	0.86 (0.06; 11.56)	1.35 (0.85; 2.15)	0.63 (0.40; 0.99)	0.82 (0.32; 2.13)	0.78 (0.35; 1.76)
Atenolol Model 2 with age interaction	2.41 (0.47; 12.36)	0.92 (0.07; 12.11)	-	-	0.81 (0.29; 2.31)	0.81 (0.32; 2.05)
Atenolol Model 3 with age interaction	-	-	1.36 (0.86; 2.17)	0.63 (0.40; 0.99)	0.85 (0.30; 2.44)	0.83 (0.33; 2.10)

***Reported associations are hazard ratios for the RCT and cohort study and odds ratios for the case-control study (including 95% confidence intervals)**. RCT models were stratified for center. Case-control models were conditioned on the matching variables: age (categories of 1 year), sex and region. Model 1 consist of treatment, outcome, age and sex. Model 2, only fit for the RCT and case-control studies, additionally adjusted for BMI, smoking and cholesterol and in the case control study physical activity, and alcohol use. Model 3 additionally included the co-medications and comorbidity indicator The age interaction effects are per 10 years and centred at 65. N.A. not available due to small sample size.*

Previously, others have also found comparability between RCTs and nonrandomized studies (6;7). However, these studies used aggregated data, complicating exploration of interaction effects and consistency across subgroups (5;10). Our findings are in line with previous studies that found no significant difference of atenolol compared to placebo on myocardial infraction: risk ratio 0.99 (95%CI 0.83; 1.19) (24), and showed that atenolol was inferior to diuretics in preventing MI (21-23). Most of these RCTs included subjects <65 or >65 and therefore consistency across age groups could not be fully explored in these studies. In the present study, the cohort and case-control studies included subjects aged between 35 and 75 years allowing exploration of the treatment effects across different age groups. The propranolol interaction effects seemed to imply that if there was a treatment by age interaction the treatment effects would increase with increasing age: HR 0.80 (95%CI 0.47; 1.38) for the RCT and HR 0.63 (95%CI 0.29; 1.38) for the cohort study. The atenolol by age interaction terms indicated that as age increased the (relative to diuretics) harmful effect [HR 2.41 (95%CI 0.47; 12. 36) for the RCT] decreased to a HR 1 as aged increased: HR 0.92

(95%CI 0.07; 12.11) for the RCT, HR 0.63 (95%CI 0.40; 0.99) for the cohort study and OR 0.83 (95%CI 0.33; 2.10) for the case-control study. Despite a significant interaction effect in the cohort study, the other interaction effects were non-significant and homogeneity of treatment effects could not be rejected. Therefore, focusing on the main effect estimates seems appropriate. These showed that, compared to diuretics, atenolol increased the hazard of MI in the RCTs and in the cohort study. Propranolol seemed to be equally effective as diuretics in preventing MI in both the RCT and the cohort study. However, due to a lack of precision, a harmful effect could not be ruled out (the upper bound was 1.26 in the RCT and 1.42 in the cohort). Results of these main effects were similar across study designs indicating generalizability of the RCT effect estimates to less controlled settings.

Despite careful considerations, the current study suffers from a few limitations. In the RCTs, all patients received therapy due to hypertension. Contrary to this, the prescription indication was not recorded in the nonrandomized studies. Given that cardiovascular risk increases which age it seems likely that the proportion of subjects receiving treatment for cardiovascular diseases increased with age, potentially explaining the observed interaction effects in the cohort study. Alternatively, the observed interaction effects in the nonrandomized data, might be due to a relation between unobserved confounding and age. Including a treatment by age interaction could than result in an erroneous interaction effect reflecting this bias. Finally, we should note that no correction was made for multiple testing which might provide an explanation for the significant atenolol by age interaction effect. Other differences between RCTs and nonrandomized studies include differences in outcome and exposure definitions. In the RCTs MI was assessed using data from multiple sources. In the nonrandomized studies the outcome was based on hospital discharge diagnoses only. Previously, the sensitivity of the hospital discharge data for acute MI was found to be 84% (25). Second, in the nonrandomized studies, exposure was defined as a filled pharmacy prescription. Clearly, filling a prescription is different from taking the medication however this is obviously very similar to the ITT analysis used in the RCT studies. Regardless, a validation study found that the positive predictive value for both prescription types of a filled prescription was 100%, compared to a home inventory (26). Finally, exposure in the case-control study was based on current use at the index-date which is not the same as in the RCTs and cohort study where exposure was based on starting with a beta-blockers or diuretic. This might explain why results of the cohort study was more in line with the RCT results than the case-control study results.

All effect estimates presented in this study were adjusted for baseline characteristics. Covariate adjustment in RCT differs from covariate adjustment in nonrandomized studies. In RCTs, adjusting for baseline variables is typically done to increase power (27), whereas in nonrandomized studies results are adjusted to control for potential confounding (28;29). Adjusted effect estimates were derived using regression analysis (conditional logistic regression model and Cox's proportional hazards model). Due to the limitations of the available data, confounder adjustment differed between case-control and cohort study, the former allowing to adjust for more variables. In the case-control data, additional adjustment for confounding did not result in markedly different estimates. Given that both the cohort study and the case-control study were sampled from the same database, it is likely that further covariate adjustment in the cohort study would follow the direction of the case-control study. Besides confounding and effect modification, obviously other factors such as information or misclassification bias may also affect the comparability of results based on RCT and nonrandomized data. Despite, not considering these biases, comparable estimates were observed.

We conclude that the RCT effect estimates of propranolol and atenolol compared to diuretics were similar to those in nonrandomized data indicating generalizability of effect estimates. Compared to diuretics, atenolol increased the hazard of non-fatal myocardial infarction; propranolol on the other hand seemed to be equally effective as diuretics. While there was, some indication that the relative effectiveness of both beta-blockers changed with age, homogeneity of treatment effect across age groups could not be rejected.

Reference List

1. Van Spall HG, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. JAMA 2007 Mar 21;297(11):1233-40.
2. Bugeja G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. BMJ 1997 Oct 25;315(7115):1059.
3. Schmidt AF, Groenwold RH, van Delden JJ, van der DY, Klungel OH, Roes KC, et al. Justification of exclusion criteria was underreported in a review of cardiovascular trials. J Clin Epidemiol 2014 Mar 5.
4. Pressler TR, Kaizar EE. The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. Stat Med 2013 Apr 1.
5. Schmidt AF, Hoes AW, Groenwold RH. Comments on 'The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias' by Taylor R. Pressler and Eloise E. Kaizar, Statistics in Medicine 2013. Stat Med 2014 Feb 10;33(3):536-7.
6. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000 Jun 22;342(25):1887-92.
7. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. Am J Ophthalmol 2000 Nov;130(5):688.
8. Weiner MG, Xie D, Tannen RL. Replication of the Scandinavian Simvastatin Survival Study using a primary care medical record database prompted exploration of a new method to address unmeasured confounding. Pharmacoepidemiol Drug Saf 2008 Jul;17(7):661-70.
9. Schneeweiss S, Patrick AR, Sturmer T, Brookhart MA, Avorn J, Maclure M, et al. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. Med Care 2007 Oct;45(10 Supl 2):S131-S142.
10. Schmidt AF, Rovers MM, Klungel OH, Hoes AW, Knol MJ, Nielen M, et al. Differences in interaction and subgroup-specific effects were observed between randomized and nonrandomized studies in three empirical examples. J Clin Epidemiol 2013 Jun;66(6):599-607.
11. Gueyffier F, Bulpitt C, Boissel JP, Schron E, Ekbom T, Fagard R, et al. Antihypertensive drugs in very old people: a subgroup meta-analysis of randomised controlled trials. INDANA Group. Lancet 1999 Mar 6;353(9155):793-6.
12. Raftery J, Roderick P, Stevens A. Potential use of routine databases in health technology assessment. Health Technol Assess 2005 May;9(20):1-iv.
13. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005 Apr;58(4):323-37.
14. Gueyffier F, Boutitie F, Boissel JP, Pocock S, Coope J, Cutler J, et al. Effect of antihypertensive drug treatment on cardiovascular outcomes in women and men. A meta-analysis of individual patient data from randomized, controlled trials. The INDANA Investigators. Ann Intern Med 1997 May 15;126(10):761-7.
15. van Wieren-de Wijer DBMA, Maitland-van der Zee AH, de Boer A, Stricker BH, Kroon AA, de Leeuw PW, et al. Recruitment of participants through community pharmacies for a pharmacogenetic study of antihypertensive drug treatment. Pharm World Sci 2009 Apr;31(2):158-64.
16. MRC trial of treatment of mild hypertension: principal results. Medical Research Council Working Party. Br Med J (Clin Res Ed) 1985 Jul 13;291(6488):97-104.
17. Medical Research Council trial of treatment of hypertension in older adults: principal results. MRC Working Party. BMJ 1992 Feb 15;304(6824):405-12.
18. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.
19. A Package for Survival Analysis in S. [computer program]. Version R package version 2.36-14. 2012.
20. Harrel JrFE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regres-

sion, and Survival Analysis. 1st ed. New York: Springer; 2001.
21. Wilhelmsen L, Berglund G, Elmfeldt D, Fitzsimons T, Holzgreve H, Hosie J, et al. Beta-blockers versus diuretics in hypertensive men: main results from the HAPPHY trial. J Hypertens 1987 Oct;5(5):561-72.
22. Wikstrand J, Warnold I, Tuomilehto J, Olsson G, Barber HJ, Eliasson K, et al. Metoprolol versus thiazide diuretics in hypertension. Morbidity results from the MAPHY Study. Hypertension 1991 Apr;17(4):579-88.
23. Holme I. MAPHY and the two arms of HAPPHY. JAMA 1989 Dec 15;262(23):3272-4.
24. Carlberg B, Samuelsson O, Lindholm LH. Atenolol in hypertension: is it a wise choice? Lancet 2004 Nov 6;364(9446):1684-9.
25. Merry AH, Boer JM, Schouten LJ, Feskens EJ, Verschuren WM, Gorgels AP, et al. Validity of coronary heart diseases and heart failure based on hospital discharge and mortality data in the Netherlands using the cardiovascular registry Maastricht cohort study. Eur J Epidemiol 2009;24(5):237-47.
26. Lau HS, de BA, Beuning KS, Porsius A. Validation of pharmacy records in drug exposure assessment. J Clin Epidemiol 1997 May;50(5):619-25.
27. Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. J Clin Epidemiol 2004 May;57(5):454-60.
28. Klungel OH, Martens EP, Psaty BM, Grobbee DE, Sullivan SD, Stricker BH, et al. Methods to assess intended effects of drug treatment in observational studies are reviewed. J Clin Epidemiol 2004 Dec;57(12):1223-31.
29. Grobbee DE, Hoes AW. Confounding and indication for treatment in evaluation of drug treatment for hypertension. BMJ 1997 Nov 1;315(7116):1151-4.

Appendix 1 Exclusion criteria of the MRC younger and older trial.

Both trials excluded patients based on (suspected) secondary hypertension, antihypertensive drug use during the run-in phase, heart failure, treatment for angina, stroke in the previous 3 months, prior MI, impaired renal function, diabetes, asthma, malignancies, and potassium ≤ 3.4 or > 5 mmol/l. Furthermore, the MRC younger trial also excluded patients if they were pregnant, had gout, or suffered from intermittent claudication.

Appendix table 1 Type of medication used

Type of medication	*MRC younger*	*MRC older*	*Cohort*	*Case-control*
Atenolol	0	1095	4154	511
Propranolol	4391	0	2341	82
Bendrofluazide	4289	0	0	0
Amiloride/hydrochlorthiazide*	0	1078	3413	310
Chlorothiazide	0	0	10	0
Chlortalidone	0	0	1197	92
Indapamide	0	0	87	25
Epitizide and potassium-sparing agents	0	0	269	46

Following the RCT amiloride or hydrochlorthiazide was used as a single category consisting of the ATC codes: C03AA03, C03DB01 and C03EA01.

Appendix table 2 Main and continuous age by treatment effects of propranolol and atenolol versus diuretic treatment on non-fatal myocardial infarction in different study designs.

Model	RCT		Cohort study		Case-control study	
	Main Effects	*Interaction effects*	*Main Effects*	*Interaction effects*	*Main Effects*	*Interaction effects*
Model 1	1.10 (0.80; 1.53)	-	1.33 (0.88;2.00)	-	1.06 (0.48; 2.30)	-
Model 2	1.09 (0.79; 1.52)	-	-	-	1.11 (0.48; 2.58)	-
Model 3	-	-	1.33 (0.88;2.01)	-	1.16 (0.50; 2.73)	-
Model 1 with age interaction	1.34 (0.90; 2.01)	1.32 (0.94; 1.85)	1.16 (0.73; 1.82)	0.69 (0.44; 1.06)	0.93 (0.37; 2.31)	0.81 (0.36; 1.82)
Model 2 with age interaction	1.34 (0.89; 2.03)	1.33 (0.94; 1.88)	-	-	0.91 (0.34; 2.46)	0.72 (0.29; 1.77)
Model 3 with age interaction	-	-	1.16 (0.74; 1.83)	0.69 (0.44; 1.06)	0.99 (0.36; 2.71)	0.76 (0.31; 1.88)

Reported associations are hazard ratios for the RCT and cohort study and odds ratios for the case-control study (including 95% confidence intervals). *RCT models were stratified for center. Case-control models were conditioned on the matching variables: age (categories of 1 year), sex and region. Model 1 consist of treatment, outcome, age and sex. Model 2, only fit for the RCT and case-control studies, additionally adjusted for BMI, smoking and cholesterol and in the case control study physical activity, and alcohol use. Model 3 additionally included a co-medications and comorbidity indicator. The age interaction effects are per 10 years and centred at 65.*

CHAPTER 7

Justification of exclusion criteria was underreported in a review of cardiovascular trials

A F Schmidt, R H H Groenwold, J J M van Delden, Y van der Does, O H Klungel
K C B Roes, A W Hoes, R van der Graaf

Journal of Clinical Epidemiology 2014 Jun;67(6):635-644
dio:10.1016/j.jclinepi.2013.12.005.

Abstract

Objective Ethical guidelines for human subjects research require that the burdens and benefits of participation are equally distributed. This study aimed to provide empirical data on exclusion of trial participants and reasons for this exclusion. As a secondary objective we assessed to what extent exclusion affects generalizability of study results.

Study Design and Setting Review of trials on secondary prevention of cardiovascular events.

Results 113 trials were identified, of which 112 reported exclusion criteria. One study justified the exclusion criteria applied. Ambiguous exclusion criteria due to the opinion of the physician (28/112 = 25%) or physical disability (12/112 =11%) were reported. Within groups of trials that studied similar treatments (i.e., beta-blocker, clopidogrel or statins therapy) baseline characteristics differed between trials. For example, the proportion of women ranged between 23.1%-47.4%, 2.1%-38.9%, and 10.6%-50.6% for the clopidogrel, beta-blocker, and statin trials, respectively. Nevertheless, no evidence was found for heterogeneity of treatment effects.

Conclusion Almost none of the papers justified the applied exclusion criteria. No evidence was found that inclusion of dissimilar participants affected generalizability. To allow for a normative discussion on equitable selection of study populations, researchers should not only report exclusion criteria but also the reasons for using these criteria.

Background

International ethical guidelines for medical research involving humans widely acknowledge that inclusion of human beings for research purposes has to be justified (1;2). Inclusion of human participants in medical research, such as randomized clinical trials (RCTs), is ethically difficult since we ´use´ humans primarily for the purposes of science and society (3;4). Moreover, there have been serious wrongdoings and highly controversial cases in the past (5). Because of the ethical and historical complexity many have felt, and still feel, that specific groups should not be included in clinical trials, such as (pregnant) women, children, and people from low- and middle income countries (2;5;6). However, such exclusion practices have resulted in underrepresentation in research of certain groups (7). Therefore, current versions of ethical guidelines require not only justification of inclusion but also of exclusion (2;8;9). For instance, the Council for International Organizations of Medical Sciences (CIOMS) guideline for biomedical research involving human beings requires that "Groups or communities to be invited to be subjects of research should be selected in such a way that the burdens and benefits of the research will be equitably distributed. The exclusion of groups or communities that might benefit from study participation must be justified." (2) Likewise, the Canadian Tri-Council Policy Statement (TCPS) 2 (9) stresses that "taking into account the scope and objectives of their research, researchers should be inclusive in selecting participants. Researchers shall not exclude individuals […] unless there is a valid reason for the exclusion.

Although ethical concerns of inappropriate exclusion of trial populations are expressed in guidelines (9), it is currently unknown to what extent benefits and burdens of research are equally distributed. It is not straightforward to evaluate the current selection of study location and population, because trial databases do not require reporting which potential study populations have been excluded and why. In addition, considerations on equitable distribution of burdens and benefits may be part of the evaluation of study protocols in research ethics committees, but the notes of these meetings are usually not publicly available. Therefore, a logical first step to assess in what proportion of studies unbalanced selection of patient groups was applied are literature reviews on reporting exclusion criteria and the grounds for using these criteria. Information on exclusion criteria is likely to be available in papers since both the CONsolidated Standards of Reporting Trials (CONSORT) and the Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) statements require the reporting of exclusion criteria (10;11). Although these data not necessarily reveal whether study popu-

lations have been deliberately excluded, they may nevertheless show whether and to what extent the reasons for exclusion of trial participants have been transparent on a more general level and hence whether there are concerns of unjustifiable exclusion of study populations. This paper is the first to study the use of exclusion criteria in this way. There have been previous studies on the use of exclusion criteria, but they have focused on their unnecessary use (12-14). Thus far no study explored whether researchers themselves justified the exclusion criteria that were applied. In this paper, we will report both the current status of the application of exclusion criteria and the justification of exclusion criteria using reported data from RCTs on secondary prevention of cardiovascular events. We have chosen studies on this topic since we expected a large number of trials from a large number of research groups, thus increasing representativeness of the sample.

Apart from ethical reasons for justifying exclusion there are also methodological reasons: if certain patient populations are not represented in RCTs, this may reduce generalizability of trial results. The previous studies that explored the application of exclusion criteria assumed that any exclusion of potential subjects hampers generalizability (12-14). However, studies comparing RCTs and nonrandomized studies (typically with less stringent in- and exclusion criteria) found little differences in the treatment effects (15-18). Therefore, as a secondary objective of our study, we assessed to what extent inclusion of different patient groups affects generalizability and thus results in different treatment effects.

Methods

Review of trials on secondary prevention of cardiovascular events

We conducted a review to assess the current practice in reporting and justification of exclusion criteria in RCTs. We focused on the rationale for excluding groups of subjects by extracting information on included subjects and reported exclusion criteria. Using the query described in **Appendix I** we searched Medline (using PubMed) for papers indexed from 01-10-2010 till 31-05-2012. Based on title and abstract, we identified RCTs on secondary prevention of cardiovascular events, which was defined as trials including patients with a stroke, myocardial infarction (MI), heart failure, cardiac arrhythmia, peripheral vascular disease, or patients undergoing coronary artery bypass grafting (CABG) or percutaneous coronary intervention (PCI). In addition, participants had to be randomized to one of the following

treatments: statins, platelet aggregation inhibitors, beta blockers, angiotensin converting enzyme inhibitors, or angiotensin II receptor blockers and a placebo or active comparator. Furthermore, the papers needed to be written in English, and describe a single trial. To allow for a fair comparison between different paper types (e.g., main analyses of trial results vs. post hoc analyses) we also searched for design papers and primary publications using cross-references and trial registries. Information from different sources related to a single trial was combined into a single entry.

Data Extraction

Of the included papers, data were extracted on applied exclusion criteria, the justification of exclusion criteria, baseline characteristics, and treatment effect estimates. The number of exclusion criteria reported by a single paper is often large; to limit this number, data was extracted on 18 a priori defined criteria with the option to include more if relevant. In the case of inclusion criteria, we defined the opposite as an exclusion criterion. For example, if for a particular trial age of 65 years and older was reported as an inclusion criterion, age below 65 years was considered as an exclusion criterion. Finally, we determined whether a rationale for exclusion criteria was provided. This was defined as any point by point explanation about why exclusion criteria were applied. This allowed us to differentiate between papers that mentioned very explicit criteria like any contraindication or high risk for loss to follow-up, but otherwise offered no explanation, from studies that did offer justification. Exclusion criteria were divided in those needing justification and those that were self-explanatory. Criteria were judged to be self-explanatory when there was one obvious explanation for excluding these patients. For example, a self-explanatory exclusion criterion would be a contraindication to the medication under study (e.g. allergy). The rationale behind excluding such patients is obviously safety concerns and patients with contraindications would not be considered future users.

Obviously, RCTs exploring different treatments are also likely to differ regarding exclusion criteria and groups of participants included. Therefore, to allow for a fair comparison we selected (from the larger overall review) three groups of trials that explored the effect of clopidogrel (n = 20), beta-blocker (n = 6), or statin therapy (n = 13). These (phase III) trials compared treatment (clopidogrel, beta-blocker, or statin therapy) to an active control or placebo, on (composite) outcomes including death, myocardial infarction, or stroke. Most studies rando-

mized participants to add-on treatment, for example, adding clopidogrel to aspirin treatment and comparing this with aspirin plus placebo or usual care.

Data Analysis

All analyses were performed using R for Windows, version 3.0.2 (19). The flowchart of the Medline search was created using the Diagram Designer program (20). Baseline characteristics and effect estimates were pooled and weighted by the number of subjects. To assess generalizability of treatment effect estimates, we compared baseline characteristics of study participant across trials and determined the heterogeneity of treatment effects. We chose not to (statistically) test for the presence of treatment heterogeneity. Instead, treatment heterogeneity was quantified using the I^2 statistic (21) and its precision by a 95% confidence interval (95%CI). The I^2 statistic represents the percentage of variation in effect estimates across studies explained by actual differences (i.e., not due to chance). An I^2 value of 0%-25%, 25%-50%, 50%-75% and >75% can be interpreted as no, low, moderate or high heterogeneity (22). Additionally, we explored if there were any signs of treatment effect modification by age and proportion women (i.e., if there was a trend of increasing or decreasing treatment effect dependent on age or gender) (23). These baseline characteristics were chosen because we expected them to be uniformly reported. While we did not expect a large number of trials to exclude subjects based on age or gender, we do expect that exclusion due to other reasons will impact the gender and age distribution. For example, if the number of co-morbidities increases with age excluding subjects based on any co-morbidity will decrease the average age in the study sample. Therefore, mean age and gender are used as proxies for differences in the application of exclusion criteria. Finally, in order to evaluate the possibility that treatment heterogeneity was dependent on exclusion criteria and not (or not only) on baseline characteristics we evaluated whether treatment effects changed when stratifying for three exclusion criteria. These criteria (exclusion due to: any medication usage at baseline, non-naïve for Intervention, opinion of physician) were selected because they were applied around 50% of the times thus ensuring approximately equally sized strata.

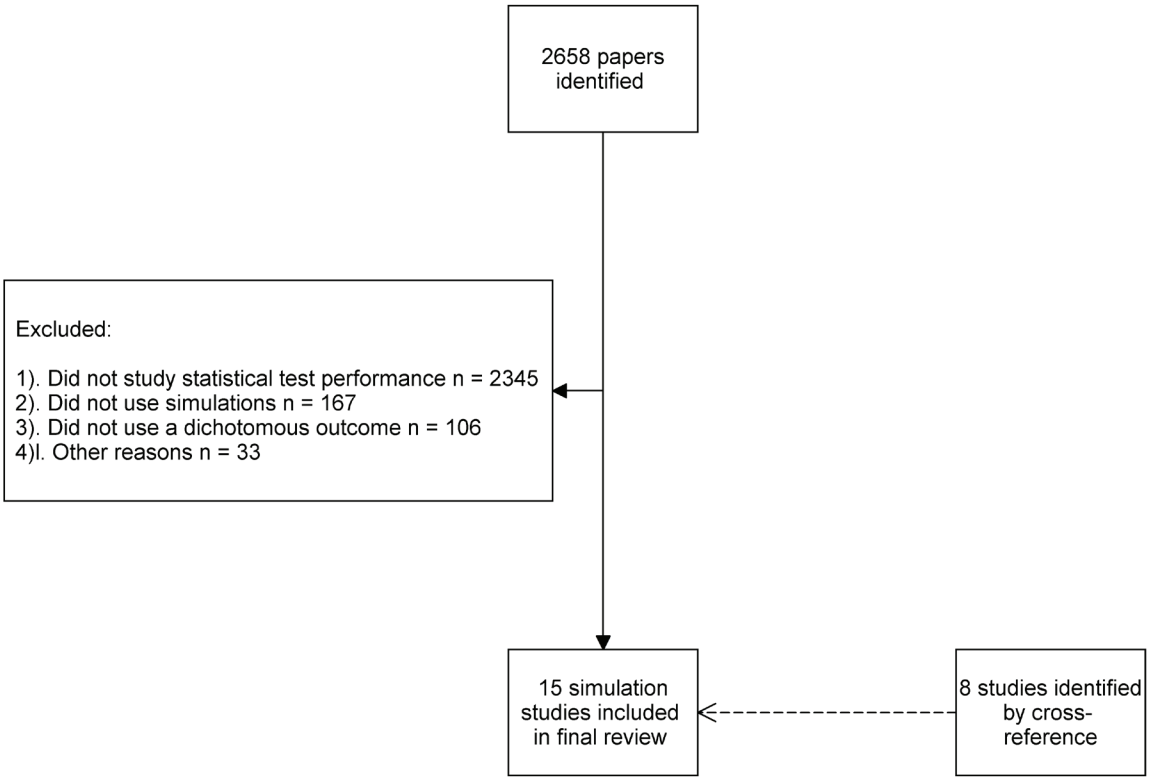
Results

Description of included trials

The Medline search resulted in 3001 potentially relevant papers, of which 113 were inclu-

ded (see **Figure 1** for the flow and **Appendix II** for references of the 113 included papers). Among the 113 included RCTs, 17 (15%) papers reported only on the design of the study. Characteristics of trial participants were reported in 96 (85%) of the papers, which included a median of 447 subjects (interquartile range (IQR): 192-2141).

Figure 1. Flow of identified publications in the systematic review of RCTs in secondary prevention of cardiovascular events.



Reporting and justification of exclusion criteria

Exclusion criteria were reported in 112 (99%) papers, a median of 6 exclusion criteria were reported per paper (IQR: 4-6; range: 0-12). The prevalence of different exclusion criteria is presented in **Table 1** and stratified for criteria needing justification and those criteria that are self-explanatory. Self-explanatory exclusion criteria included exclusion because of high bleeding risk (56/112 = 50%), contraindications for the studied intervention (73/112 = 65%) and impaired renal (63/112 = 56%) or liver (59/112 = 53%) function. Exclusion criteria that potentially require justification are exclusion of pregnant or fertile women (reported by 53 (47%) of the trials) and exclusion of lactating women (30 (27%), all of which also excluded the category pregnant or fertile women). Studies not excluding lactating or fertile/pregnant women reported a relatively high median age (64 years; IQR 62-67, for the first group and 64

Table 1. Reported exclusion criteria (with percentage and 95% confidence interval) in RCTs in secondary prevention of cardiovascular events (published Oct 2010 – May 2012)*.

	All RCTs (N=112)**	RCT of beta-blocker therapy (N=6)	RCTs of clopidogrel therapy (N=20)	RCTs of statin therapy (N=13)
Self-explanatory criteria				
Contraindication to intervention	73 (65% 95CI 56; 74)	3 (50% 95CI 10; 90)	15 (75% 95CI 56; 94)	6 (46% 95CI 27; 81)
Any impaired renal condition	63 (56% 95CI 47; 65)	4 (67% 95CI 29; 100)	6 (30% 95CI 10; 50)	11 (85% 95CI 65; 100)
Any impaired liver condition	59 (53% 95CI 43; 62)	5 (83% 95CI 54; 100)	4 (20% 95CI 2; 38)	12 (92% 95CI 78; 100)
High risk of bleeding	56 (50% 95CI 41; 59)	0	13 (65% 95CI 44; 86)	1 (8% 95CI 0; 22)
Criteria requiring justification				
Age below 18	78 (70% 95CI 61; 78)	4 (67% 95CI 29; 100)	13 (65% 95CI 44; 86)	9 (69% 95CI 44; 94)
Other age restrictions	41 (37% 95CI 28; 46)	3 (50% 95CI 10; 90)	6 (30% 95CI 10; 50)	6 (46% 95CI 19; 73)
Pregnant and/or fertile	53 (47% 95CI 38; 57)	0	7 (35% 95CI 14; 56)	6 (46% 95CI 19; 73)
Lactating women	30 (27% 95CI 19; 35)	0	4 (20% 95CI 2; 38)	4 (31% 95CI 6; 56)
Female gender	1 (1% 95CI 0; 3)	0	0	0
Male gender	0	0	0	0
Any medication usage at baseline	80 (71% 95CI 63; 80)	4 (67% 95CI 29; 100)	17 (85% 95CI 69; 100)	7 (54% 95CI 27; 81)
Non-naïve for Intervention	40 (36% 95CI 27; 45)	3 (50% 95CI 10; 90)	9 (45% 95CI 23; 67)	9 (69% 95CI 44; 94)
Opinion of physician	28 (25% 95CI 17; 33)	3 (50% 95CI 10; 90)	5 (25% 95CI 6; 44)	3 (23% 95CI 0; 46)
Indication for either treatment arm	13 (12% 95CI 6; 18)	0	0	1 (8% 95CI 0; 22)
Likely to be lost to follow-up	9 (8% 95CI 3; 13)	0	3 (15% 95CI 0; 31)	0
Short life expectancy	45 (40% 95CI 31; 49)	2 (33% 95CI 0; 71)	7 (35% 95CI 14; 56)	3 (23% 95CI 0; 46)
Lack of cognition or mental impairment	16 (14% 95CI 8; 21)	2 (33% 95CI 0; 71)	0	1 (8% 95CI 0; 22)
Physical disability	12 (11% 95CI 5; 16)	0	0	1 (8% 95CI 0; 22)

* note 95% confidence intervals are based on the asymptotic Wald method and values below zero and above 100 were truncated.
**One study did not report any exclusion criteria.

years; IQR 62-67 for the latter group), indicating that only a small number of women would be affected by excluding lactating or fertile/pregnant women. Other criteria needing justification are exclusion due to impaired cognition (16/112 = 14%), physical disability (12/112 =11%), medication use at baseline (80/112 = 71%), non-naivety to the studied intervention (40/112 = 36%), specific indication for either treatment arm (13/112 = 12%), short life expectancy (45/112 = 40%), based on the opinion of the physician (28/112 = 25%) or exclusion due to an increased risk of being lost to follow-up (9/112 = 8%). Furthermore, children (i.e., participants aged < 18 years) were excluded in 78 (70%) trials and exclusion based age, other than age <18, was reported in 41 trials (36%), often resulting in the inclusion of older subjects. A

single trial (1%) mentioned excluding women. In a sensitivity analysis we explored whether exclusion criteria differed between publications from journals with a high (>5) and low (5≤) impact factor (**Appendix III**). This revealed that trials published in journals with a higher impact factor tended to apply more exclusion criteria.

Because differences in exclusion criteria in such a large group of trials might occur because of differences in e.g. treatment or outcome under study, we also explored the exclusion criteria mentioned in a subset of trials, i.e., trials on clopidogrel (n = 20), beta-blocker (n = 6), or statin therapy (n = 13). These trials were similar in the studied treatments as well as the outcomes (see **Appendix IV** for details). Given this similarity one might also except that within each group of trials similar exclusion criteria were applied. This was indeed the case for some criteria, for example 12 of the 13 statin trials excluded subjects with liver impairment at baseline. Contrary to this, some exclusion criteria were more variably applied, within groups of trials. For example, within the group of clopidogrel trials, 9 (45%) RCTs excluded non-naive subjects, whereas 11 (55%) included such participants. In the group of statin trials, 6 (46%) trials excluded pregnant or fertile women. Similarly within the beta-blocker RCTs 3 out of 6 (50%) focused on a specific age group of adults.

Only one paper reported a rationale for the applied exclusion criteria. This particular study assessed the effect of clopidogrel in patients undergoing CABG and used a non-fatal endpoint (24;25). The authors explain that subjects with a current malignancy were excluded by stating: “Higher risk of early postoperative mortality”(25).

Generalizability

The baseline characteristics of included trials showed a large range in patient characteristics between trials (**Table 2**), also when focusing on the subsets of trials on clopidogrel, beta-blocker, or statin therapy. For example, the proportion of women included ranged from 23.1%-47.4% for the clopidogrel, 2.1%-38.9% for the beta-blocker, and 10.6%-50.6% for the statin RCTs. Other examples could be differences in the proportion of subjects with e.g. diabetes, hypertension, and hypercholesterolemia (**Table 2**). Note that the minimum included proportion of women was 2.1% (despite the fact that this trial reported to exclude women).

Table 2. Baseline characteristic of study participants in RCTs on beta-blocker, clopidogrel or statin therapy in secondary prevention of cardiovascular events (published Oct 2010 – May 2012)*.

	Beta-Blockers RCTS		Clopidogrel RCTs		Statin RCTs		All RCTs	
Baseline characteristics	**Range (min-max)**	**N of studies (N=6)**	**Range (min-max)**	**N of studies (N=17)**	**Range (min-max)**	**N of studies (N=12)**	**Range (min-max)**	**N of studies (N=96)**
Number of subjects	70-2708	6	60-13608	17	44-9251	12	36-26449	96
Women	2.1%-38.9%	6	23.1%-47.4%	17	10.6%-50.6%	12	2.1%-78.4%	96
Mean age (years)	46.6-75.7	6	59.0-68.6	16	58.4-71.0	12	46.6-81.0	88
Mean weight (kg)	84.1	1	87.8	1	78.4-85.7	2	75.2-92.8	15
Mean height (cm)	172.7	1	-	0	170.7-172.5	2	167.8-172.7	7
Mean BMI (kg/m^2)	25.7-28.0	4	24.2-30.0	10	23.0-28.8	4	23.0-31.9	46
Currently smoking	4.2%-17.5%	3	12.8%-49.8%	15	12.9%-58.7%	9	4.2%-59.8%	73
Previously smoking	73%	1	-	0	37.1%-63.5%	2	23.7%-72.5%	13
Never smoked	-	0	-	0	24%	1	23.5%-48.5%	7
White race	70%	1	88.8-93.7	3	90.9-94.3	2	69.9%-98.8%	21
Black race	24%	1	1%	1	-	0	0.9%-100%	10
Asian race	-	0	7%	1	-	0	5.0%-100%	8
Hispanic race	6%	1	-	0	-	0	3.0%-10.3%	3
Diabetes	20.6%-35.5%	4	19.4%-45.1%	16	14.7%-44.0%	10	1.5%-100%	85
Hypercholesterolemia	43.0%-63.0%	4	15.7%-82.6%	16	28.5%-40.7%	2	10.8%-87.9%	56
Hypertension	59.0%-82.9%	4	40.1%-88.8%	15	43.0%-82.7%	10	29.9%-90.4%	75
Mean serum cholesterol (mmol/L)	4.8	1	4.0	1	4.5-6.2	4	3.8-6.2	11
Mean systolic BP (mmHg)	113.6-140.9	6	129.4-130.0	2	127.0-138.9	4	96.9-145.3	38
Mean diastolic BP (mmHg)	71.0-90.8	4	77.9-80.0	2	71.1-81.8	4	60.5-90.8	30
Mean heart rate	73.0-81.5	6	75.0	1	67.0-68.0	2	64.2-91.1	26
History of MI	40.0%-50.3%	3	3.8%-53.6%	12	18.5%-81.0%	6	1.5%-81.0%	57
History of ST	10%	1	2.0%-7.7%	6	4.9%-9.4%	2	2.0%-99.8%	30

*Displayed information is based on papers that allowed for extraction of patient characteristics. The range gives minimum and maximum mean or mean percentage if there were 2 or more RCTs included. If only 1 RCTs reported on the respective characteristics the mean is presented. BMI=body mass index, BP= blood pressure, MI=Myocardial infarction, ST=stroke.

Among the clopidogrel trials, 13 (80%) studies allowed for extraction of the treatment effect. Of the 7 trials not reporting outcome data, 3 were design papers and in 3 trials no outcomes were observed (e.g., due to the outcome being of secondary importance). The reported risk ratio (RR) for clopidogrel versus active or placebo add-on therapies ranged between 0.13 and 0.99 [pooled RR = 0.77, 95%CI 0.67; 0.88], for the composite endpoint of mortality, MI, stroke, revascularization and stent-thrombosis (see **Appendix IV**). Plotting the RR against the proportion of women or mean age of the included subjects did not show any dependency (**Figure 2**). This is in line with the I^2 statistic, which indicated little heterogeneity (I^2 16%, 95CI 0; 35).

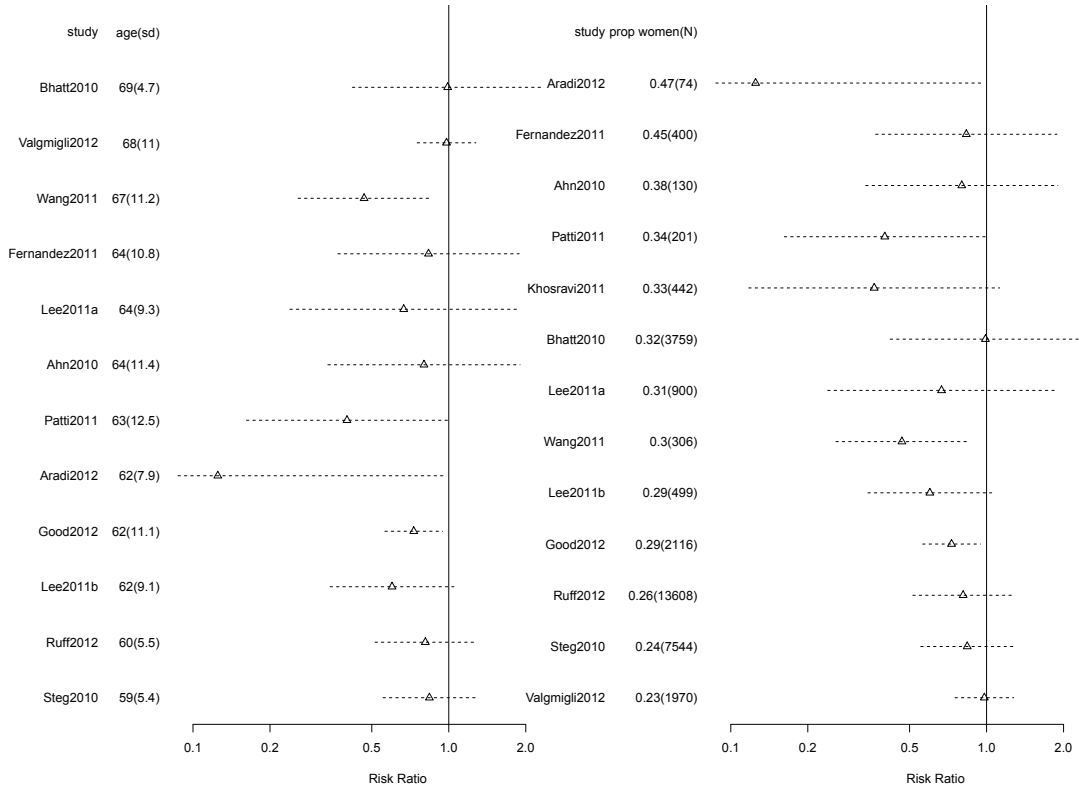
For the group of statin trials, treatment effects could be extracted from 12 (92%) papers. The RR for the composite endpoint of mortality, MI, stroke, and revascularization, ranged

between 0.25 and 1.50 (pooled RR = 0.82, 95%CI 0.75; 0.91). Graphics did not suggest any dependency between gender or age and the treatment effect (**Figure 3**) and neither did the I^2 statistic (I^2 12%, 0; 31).

Among the beta-blocker trials, using data from 5 (83%) studies, the RR for the mortality and/or MI endpoint ranged between 0.68 and 2.25 (pooled RR = 0.91, 95%CI 0.68; 1.21). After excluding the most extreme observation of 2.25, the range was RR 0.68 to 0.94. As with the two previous examples, the graphical display (**Figure 4**) as well as the I^2 statistic did not suggest any heterogeneity (I^2 0%, 95CI 0; 100). However, due to small sample size the 95%CI was large indicating a lack of precision.

Finally, we explored whether treatment effects were dependent on the following exclusion criteria: 'any medication at baseline', non-naïve for intervention' or 'opinion of physician'; see **Appendix V**. No dependency between the treatment effect estimates and exclusion criteria was observed.

Figure 2 Forrest plot of the effect of clopidogrel on the composite endpoint of mortality, myocardial infarction, stroke, revascularization and stent-thrombosis, ordered by the proportion of women or mean age of the individual trials.



Legend figure 2: Triangles indicate treatment effects (risk ratio). Horizontal bars indicate 95% confidence intervals of the risk ratio's. N reflects the sample size including both genders, SD indicate the standard deviation. One of the clopidogrel trials did not report mean age and was excluded.

Figure 3. Forrest plot of the effect of statins on the composite endpoint of mortality, myocardial infarction and revascularization, ordered by the proportion of women or mean age of the individual trials.

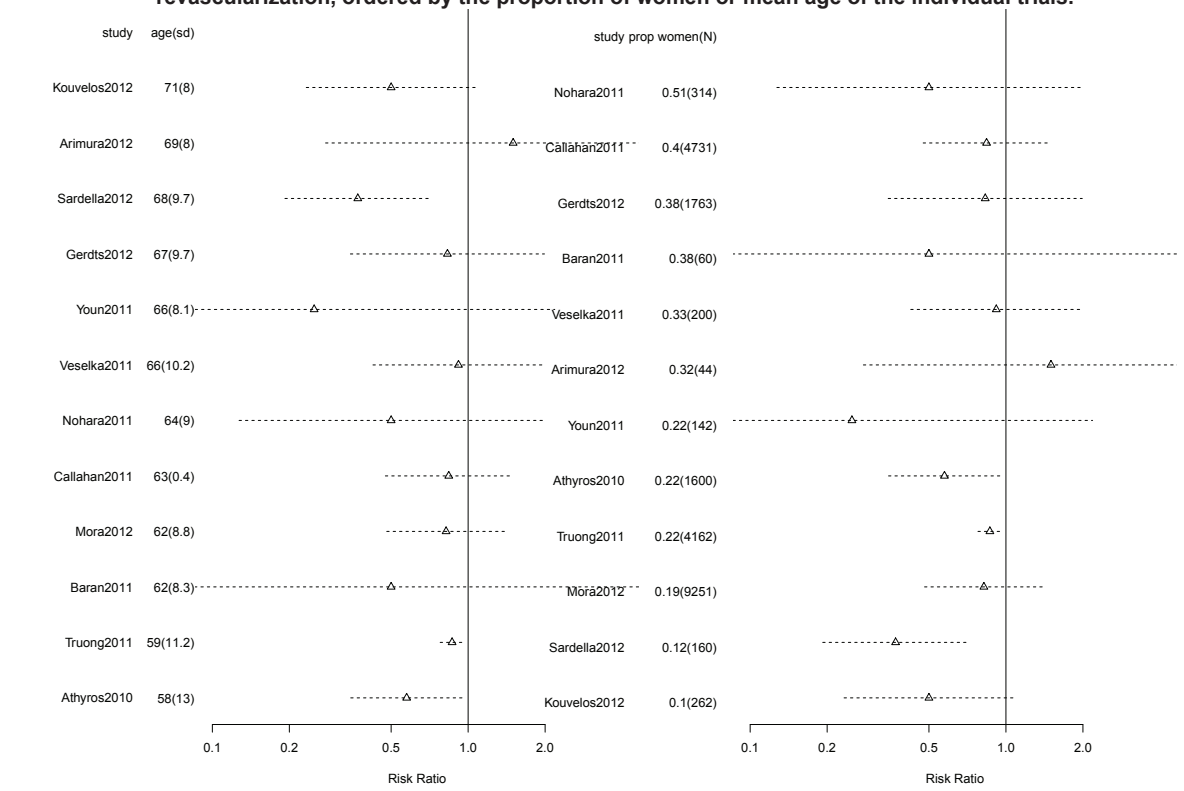
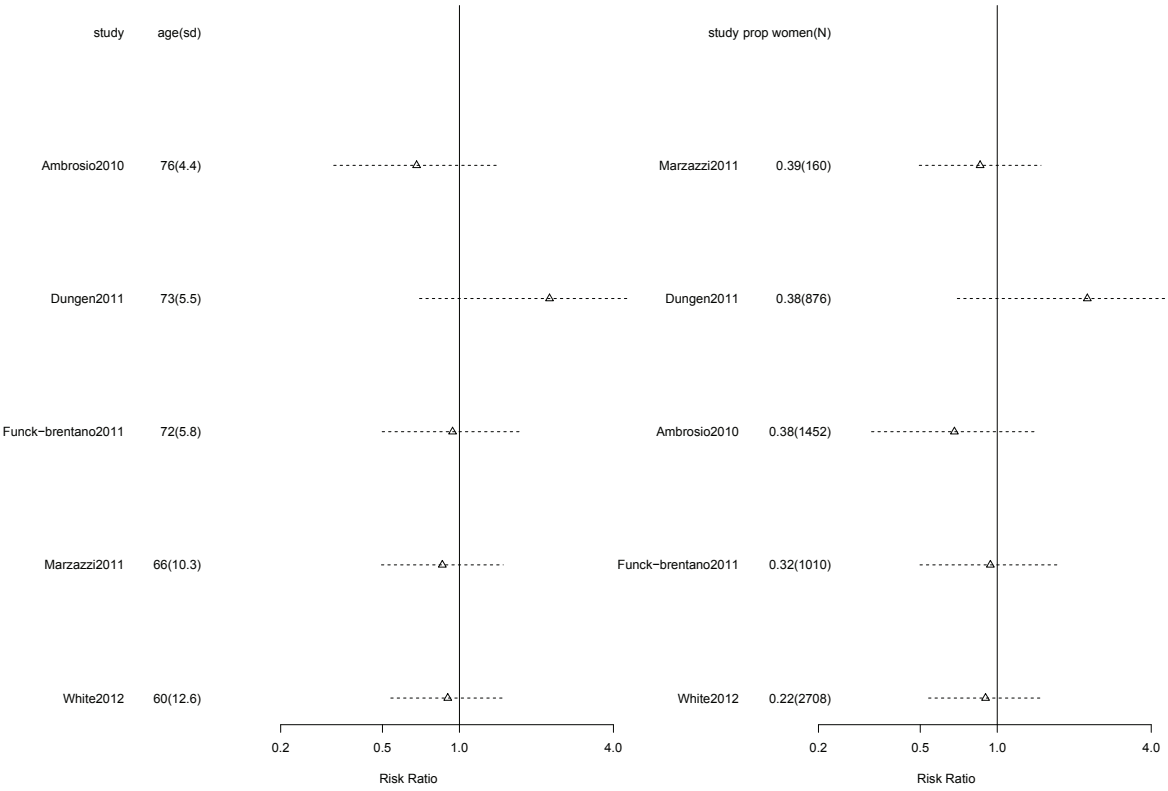


Figure 4. Forrest plot of the effect of beta-blocker on the composite endpoint of mortality and myocardial infarction, ordered by the proportion of women or mean age of the individual trials.



Legend Figure 3 and 4: Triangles indicate treatment effects (risk ratio). Horizontal bars indicate 95% confidence intervals of the risk ratio's. N reflects the sample size including both genders, SD indicate the standard deviation.

Discussion

Key findings of this study are that 1.) a rationale for exclusion criteria is hardly ever reported; and 2.) the applied exclusion criteria differed considerably between studies exploring the same treatment, yet despite differences in baseline characteristics between these studies there was no evidence for impaired generalizability. In the following, we will discuss these findings.

Although almost all RCTs in our review of secondary prevention of cardiovascular events reported exclusion criteria (112/113 = 99%), only 1 paper provided a reason for a specific criterion that had been applied. Therefore, it is difficult to assess whether the inclusion and exclusion of trial participants in these trials was justified. We categorized exclusion criteria in those needing justification and those that are self-explanatory. Obviously, this is to some extend an arbitrary decision and other categorizations are also possible. However, we expect that most would agree that justification is not needed for excluding patients due to safety reasons such as contra-indications for the intervention under study. We deemed other criteria less self-explanatory and these would require justification. For example, in some occasions participants were excluded because of a 'short life expectancy' (45/112 = 40%). Exclusion for this reason possibly has to do with statistical power when studying non-mortality outcomes. This can, however, also be interpreted as gatekeeping (26), meaning that some groups of subjects may have been eligible to participate but have nonetheless been excluded. Another example is 'opinion of the physician' (28/112 = 25%). This criterion may imply that researchers in their roles as physicians have made individualized judgments for patients, which should typically be avoided in a research context. Other exclusion criteria, for example those relating to age or pregnancy, might be seen as self-explanatory by some. However, we still viewed these as needing justification because there can be multiple non-exclusive reason for applying these criteria. For example children could be excluded because treatment effectiveness was expected to differ but also because of the (administrative) burden of including children. Similarly, excluding pregnant/fertile, or lactating women could be due to expected adverse event or other reasons such as the need for closer monitoring which might be infeasible during the trial. In our sample of trials almost half of the papers excluded pregnant/fertile women and one fourth excluded lactating women. However, due to the relatively older target population of studies of secondary prevention of cardiovascular events it remains uncertain whether indeed women have been excluded inappropriately and if so how many have

been unfairly excluded. In the introduction we have already mentioned that these reasons for exclusion of potential trial participants are probably not published elsewhere. Hence, there is a potential risk of unjustifiable exclusion.

In three groups of RCTs of the same treatment, there were no uniform application of exclusion criteria and baseline characteristics differed considerably between studies. Despite this the observed treatment effects were similar across trials. This suggests that findings from these trials can be generalized across groups of participants; i.e., inclusion of different groups of participants did not seem to impair generalizability of treatment effects. Generalizability was assessed using the I^2 statistics and by graphically determining whether there was a trend between treatment effect estimates and the baseline characteristics age and gender; no trend was found. However, it could be possible that researchers did not a priori expect difference in treatment effects between age or gender subgroups, thus allowing for differences in exclusion rates for the different subgroups. On the other hand, it seems likely that age and gender are related to other patient characteristics such as frailty and polypharmacy. Thus age and gender might still be used as proxies for treatment effect modification between treatment and other baseline characteristics. We focused, however, on age and gender, because these patient characteristics were reported by almost all trials.

Despite careful considerations, this review potentially suffers from a few weaknesses. Our review focused on trials on secondary prevention of cardiovascular events. Our findings may therefore be only applicable to this particular clinical domain and not to other domains. In addition, we conducted our search using the Medline database only. Hence, RCTs not indexed by Medline were not included in our review. Most if not all journals with a high impact factor are indexed by Medline. However, this is not the case for lower impact journal. Inclusion of more lower impact publications, from other databases, could possibly change our findings. As a sensitivity analysis we therefore stratified exclusion criteria by the impact factor (>5 vs. ≤5), which suggested that lower impact publication reported less exclusion criteria. Thus the percentage of reported exclusion criteria is possibly somewhat inflated by only using Medline(PubMed). However, it seems unlikely that searching additional databases would markedly increase the percentage of papers justifying exclusion criteria. Similarly, we recognize that our Medline search, of which 4% of the hits were included, might have been overly sensitive. While inefficient, this might nevertheless reduce the likelihood of excluding relevant

publications. In the current review we observed differences between baseline characteristics of trial populations. It seems likely that these differences are not only explained by different application of exclusion criteria but also (partially) reflect differences in available patient population. Regardless, of the causes of these differences in baseline characteristics we did not find any indication of treatment heterogeneity depending on these differences. This heterogeneity was assessed using I^2 and by determining whether there was a trend in the treatment effect estimates per study and the mean age and the proportion of women. Due to the relative small number of studies precision of these methods was sometimes lacking. This was most pronounced in the beta-blocker example were the 95%CI of the I^2 included 0% and 100%. In the clopidogrel and statin examples the precision was higher indicating an upper level of 35% heterogeneity. Given these limitations, we cannot conclude that there is in fact no treatment heterogeneity but merely that we could not detect any. Another issue is that the pooling of baseline characteristics, treatment effect estimates, and the exploration of heterogeneity are based on meta-analysis methods. Given that our interest is not on estimating any clinically relevant treatment effect estimates no assessment of risk of bias of the individual studies was performed. On the other hand, little heterogeneity was found between the three trial subgroups, indicating that if a bias assessment would have been applied results would not differ markedly.

Although ethical guidelines require justification of exclusion, this study shows that authors do not feel obliged to provide this rationale. However, if potential subjects are excluded because of expected differences in treatment effectiveness (or harm) it seems very relevant to report this information. It might be beneficial when reporting guidelines like the CONSORT or the SPIRIT statement would recommend such reporting. In fact the recent SPIRIT statement on RCT protocols (11) advices researchers to do just this and states: “Certain eligibility criteria warrant explicit justification in the protocol, particularly when they limit the trial sample to a narrow subset of the population” (11). In line with the SPIRIT statement, we feel that the quality of RCT reporting would improve when researchers report justification of exclusion criteria. Whether this is done in the RCT protocol, the primary publication, trial registries or in any other publicly available documentation is of secondary importance.

Conclusion

Although ethical guidelines require justification of exclusion of study populations (8;10), this

study shows that authors do not feel obliged to provide this rationale in their papers. In line with these guidelines we emphasize that researchers should not only report exclusion criteria but also discuss why these exclusion criteria were used and to what extent exclusion of those subjects could affect generalizability of treatment effects. Explicitly reporting both exclusion criteria and rationales for these criteria may decrease the use of ambiguous exclusion criteria, or at the very least readers can more easily judge whether exclusion of groups of patients was justified.

Reference List

1. WMA. Declaraction of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. Seoul, Korea 2010.
2. Council for International Organizations of Medical Sciences (CIOMS). International ethical guidelines for biomedical research involving human subjects. Geneva; 2002.
3. Jonas H. Reflections on experimenting with human subjects. Daedalus, the Journal of the American Academy of Arts and Sciences 1969;98:219-47.
4. Graaf vdR, van Delden JJ. On using people merely as a means in clinical research. Bioethics 2012 Feb;26(2):76-83.
5. Emanuel EJ, Grady C. Four paradigms of clinical research and research oversight. Camb Q Healthc Ethics 2007;16(1):82-96.
6. Baylis F, Halperin SA. Research involving pregnant women: trials and tribulations. Clinical Investigation 2012 Feb 1;2(2):139-46.
7. The Oxford Textbook of Clinical Research Ethics. Reprint ed. Oxford New York: Oxford University Press; 2011.
8. Policy and Guidelines on the Inclusion of Women and Minorities as subjects in clinical research. National Institute of Health.; 1993.
9. Tri-Council Policy Statement (TCPS) 2: Ethical Conduct for Research Involving Humans. 2010.
10. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol 2010 Aug;63(8):e1-37.
11. Chan AW, Tetzlaff JM, Gotzsche PC, Altman DG, Mann H, Berlin JA, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ 2013;346:e7586.
12. Lee PY, Alexander KP, Hammill BG, Pasquali SK, Peterson ED. Representation of elderly persons and women in published randomized trials of acute coronary syndromes. JAMA 2001 Aug 8;286(6):708-13.
13. Van Spall HG, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. JAMA 2007 Mar 21;297(11):1233-40.
14. Bugeja G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. BMJ 1997 Oct 25;315(7115):1059.
15. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000 Jun 22;342(25):1887-92.
16. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. Am J Ophthalmol 2000 Nov;130(5):688.
17. Rovers MM, Straatman H, Ingels K, van der Wilt GJ, van den Broek P., Zielhuis GA. Generalizability of trial results based on randomized versus nonrandomized allocation of OME infants to ventilation tubes or watchful waiting. J Clin Epidemiol 2001 Aug;54(8):789-94.
18. Schmidt AF, Rovers MM, Klungel OH, Hoes AW, Knol MJ, Nielen M, et al. Differences in interaction and subgroup-specific effects were observed between randomized and nonrandomized studies in three empirical examples. J Clin Epidemiol 2013 Jun;66(6):599-607.
19. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.
20. Diagram Designer [computer program]. Version 1.25 MeeSoft; 2012.
21. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002 Jun 15;21(11):1539-58.
22. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-60.
23. Groenwold RH, Rovers MM, Lubsen J, van der Heijden GJ. Subgroup effects despite homogeneous heterogeneity test results. BMC Med Res Methodol 2010;10:43.

24. Kulik A, Le May MR, Voisine P, Tardif JC, Delarochelliere R, Naidoo S, et al. Aspirin plus clopidogrel versus aspirin alone after coronary artery bypass grafting: the clopidogrel after surgery for coronary artery disease (CASCADE) Trial. Circulation 2010 Dec 21;122(25):2680-7.
25. Kulik A, Le MM, Wells GA, Mesana TG, Ruel M. The clopidogrel after surgery for coronary artery disease (CASCADE) randomized controlled trial: clopidogrel and aspirin versus aspirin alone after coronary bypass surgery [NCT00228423]. Curr Control Trials Cardiovasc Med 2005 Oct 11;6:15.
26. Sharkey K, Savulescu J, Aranda S, Schofield P. Clinician gate-keeping in clinical research is not ethically defensible: an analysis. J Med Ethics 2010 Jun;36(6):363-6.

Appendix I

Medline search strategy for randomized controlled trials on secondary prevention of cardiovascular events.

Search date:31-05-2012

(statin* OR hydromethylglytaryl OR ”CoA reductase inhibitors” OR “hydromethylglytaryl CoA” OR “hydromethylglytaryl Coenzyme A” OR hydromethylglytaryl-CoA OR HMG-COA OR “HMG COA” OR fluvastatin OR simvastatin OR Zocor OR lipex OR pravastatin OR lipostat OR atorvastatin OR Lipitor OR lovastatin OR cerivastatin OR rosuvastatin OR crestor OR “beta blocker” OR “beta blockers” OR beta-blockers OR “beta blockade” OR “adrenergic beta antagonist” OR “adrenergic beta antagonists” OR beta-antagonist OR beta-antagonists OR “beta adrenergic” OR “adrenergic receptor blockader” OR “adrenergic receptor blockaders” OR blocker OR blockers OR beta-adrenergic OR “blocking agent” OR “blocking agents” OR “beta adrenergic” OR acebutolol OR spectral OR atenolol OR Tenormin OR betaxolol OR kerlon OR bisoprolol OR bisobloc OR emcor OR carvedilol OR eucardic OR celiprolol OR dilanorm OR labetalol OR trandate OR metoprolol OR lopresor OR selokeen OR nebivolol OR nebilet OR oxprenolol OR trasicor OR pindolol OR viskeen OR propranolol OR Inderal OR sotalol OR sotacor OR esmolol OR brevibloc OR “platelet aggregation inhibitor” OR “platelet aggregation inhibitors” OR “platelet antiaggregant” OR “platelet antiaggregants” OR “platelet inhibitor” OR “platelet inhibitors” OR antiplatelet OR “platelet antagonist” OR “platelet antagonists” OR “acetylsalicylic acid” OR asprin OR “aspro cardio” OR acylpyrin OR aloxiprimum OR colfarit OR dispril OR easprin OR ecotrin OR endosprin OR magnecyl OR micristin OR polopirin OR polopiryna OR solprin OR solusan OR zorprin OR acetysal OR cardegic OR “calcium carbasalate” OR ascal OR “calcium acetylsalicylic carbamidate” OR clopidogrel OR Plavix OR iscover OR dipyridamol OR persantin OR asasantin OR dipyridamole OR miosen OR cleridium OR cerebrovase OR antistenocardin OR curantil OR curantyl OR kurantil OR persantine OR dipiradol OR “platelet IIb/IIIa receptor” OR “platelet IIb/IIIa receptors” OR abciximab OR reopro OR eptifibatine OR integrilin OR tirofiban OR aggrastat OR “angiotensin converting enzyme inhibitor” OR “angiotensin converting enzyme inhibitors” OR ACE OR kininase II OR “angiotensin converting” OR angiotensin OR benazepril OR cibacen OR captopril OR capoten OR cilazapril OR vascace OR enalapril OR renitec OR fosinopril OR newace OR lisonopril OR novatec OR Zestril OR perindopril OR coversyl OR quinapril OR acupril OR ramipril OR tritace OR trandolapril OR gopten OR zofendopril OR zofil OR Sartan* OR “angiotensin receptor antagonist” OR “angiotensin receptor antagonists” OR angiotensin OR “receptor blocker” OR “angiotensin receptor blockers” OR “angiotensin II receptor antagonist” OR “angiotensin II receptor antagonists” OR “angiotensin II receptor blocker” OR “angiotensin II receptor blockers” OR AT1 OR candesartan OR atacand OR eprosartan OR teveten OR irbesartan OR aprovel OR losartan OR cozaar OR telmisartan OR micardis OR valsartan OR diovan OR olmesartan)

AND (CVA* OR stroke* OR cerebrovascular OR “brain vascular” OR apoplexy OR infarct* OR “heart attack” OR “heart attacks” OR cardiac OR coronary OR cardiovascular OR “cardio vascular”)

AND (Randomized [tiab] OR randomised [tiab] OR RCT [tiab] OR trial [tiab] OR placebo [tiab] OR controlled [tiab] OR characteristics [tiab] OR baseline [tiab])

AND (2010/09/22:2012/05/31[edat])

NOT (Review OR meta-analysis)

Appendix II Trial subgroups of secondary prevention of cardiovascular events RCTs, with intervention, comparison and outcome.

Study	Intervention	Comparison	Outcomes						Baseline data	Outcome data
			Mortality	Myocardial Infarction	Stroke	Revascular-ization	Stent thrombosis	Other		
Beta-Blocker RCTs										
Ambrosio 2010 (#4)	Nebivolol	Placebo	✓	✓				✓	✓	✓
Dungen 2011 (#24)	Bisoprolol	Carvedilol	✓						✓	✓
Funck-Brentano 2011(#28)	Bisoprolol	Enalapril	✓					✓	✓	✓
Marazzi 2011 (#63)	Nebivolol	Carvedilol	✓	✓				✓	✓	✓
White 2012 (#107)	Bucindolol	Placebo	✓					✓	✓	✓
Wojnicz 2010 (#109)	Carvedilol	Verapamil	✓					✓	✓	
Clopidogrel RCTs										
Ahn 2010 (#2)	Aspirin, Clopidogrel, Cilastazol	Aspirin, Clopidogrel	✓	✓	✓	✓			✓	✓
Aradi 2012 (#6)	Clopidogrel 150 mg	Clopidogrel 75 mg	✓	✓		✓			✓	✓
Bhatt 2010 (#14)	Clopidogrel, Omeprazole	Clopidogrel, Placebo	✓	✓	✓	✓			✓	✓
Collet 2011 (#22)	Tailored Aspirin, Clopidogrel	Usual dose Aspirin, Clopidogrel	✓	✓	✓	✓	✓			
Fernandez 2011 (#27)	Clopidogrel 600 mg	Clopidogrel 300 mg	✓	✓	✓	✓	✓	✓	✓	✓
Good 2012 (#34)	Clopidogrel	Placebo	✓	✓	✓				✓	✓
Hazarbasanov 2012 (#39)	Tailored Clopidogrel	Clopidogrel 75 mg	✓	✓	✓		✓		✓	✓
Jin 2012 (#44)	Aspirin, Clopidogrel, Cilastazol	Aspirin, Clopidogrel	✓	✓		✓	✓	✓	✓	✓
Khosravi 2011 (#46)	Clopidogrel (Osvix)	Clopidogrel (Plavix)	✓	✓	✓	✓	✓	✓	✓	✓
Lee 2011(#55)	Aspirin, Clopidogrel, Cilastazol	Aspirin, Clopidogrel, Placebo	✓	✓		✓			✓	✓
Lee 2011 (#54)	Aspirin, Clopidogrel, Cilastazol	Aspirin, Clopidogrel	✓	✓	✓				✓	✓
Leonardi 2012 (#58)	Cangrelor	Clopidogrel	✓			✓	✓			
Mauri 2010 (#64)	30 months Aspirin, Clopidogrel	12 months Aspirin, Clopidogrel	✓	✓	✓					
Meng 2010 (#66)	Clopidogrel (Talcom)	Clopidogrel (Plavix)	✓	✓		✓	✓		✓	
Patti 2011 (#79)	Clopidogrel 600 mg	Clopidogrel 300 mg	✓	✓	✓	✓			✓	✓
Roghani 2011 (#83)	Clopidogrel 150 mg	Clopidogrel 75 mg	✓						✓	✓
Ruff 2012 (#84)	Prasugrel	Clopidogrel	✓	✓	✓			✓	✓	✓
Steg 2010 (#87)	Ticagrelor	Clopidogrel	✓	✓	✓				✓	✓
Valgimigli 2012 (#99)	12 months Aspirin, Clopidogrel	6 months Aspirin, Clopidogrel	✓	✓	✓				✓	✓
Wang 2011 (#103)	Tailored Clopidogrel	Usual dose Clopidogrel	✓	✓	✓	✓			✓	✓
Statin RCTs										
Arimura 2012 (#7)	Atorvastatin, Ezetimibe	Atorvastatin	✓	✓		✓			✓	✓

Athyros 2010 (#8)	Tailored Atorvastatin	Usual dose Atorvastatin	✓	✓	✓	✓		✓	✓	✓
Baran 2011 (#10)	Atorvastatin	Placebo	✓	✓	✓		✓		✓	✓
Callahan 2011 (#18)	Atorvastatin	Placebo			✓				✓	✓
Gerdts 2012 (#29)	Simvastatin, Ezetimibe	Placebo	✓	✓	✓			✓	✓	✓
Kouvelos 2012 (#49)	Rovastatin, Ezetimibe	Rosuvastatin	✓	✓	✓			✓	✓	✓
Liu 2012 (#61)	Atorvastatin	No statins	✓	✓		✓				
Mora 2012 (#72)	Atorvastatin 80 mg	Atorvastatin 10 mg	✓	✓	✓			✓	✓	✓
Nohara 2012 (#75)	Rosuvastatin	Pravastatin	✓	✓		✓		✓	✓	✓
Sardella 2012 (#86)	Rosuvastatin	No statins	✓	✓	✓	✓			✓	✓
Truong 2011 (#95)	Atorvastatin	Pravastatin	✓	✓	✓	✓		✓	✓	✓
Veselka 2011 (#101)	Atorvastatin	No statins	✓	✓					✓	✓
Youn 2011 (#111)	Rovastatin	No statins	✓	✓		✓			✓	✓

Appendix III

#1. Aasa M, Dellborg M, Herlitz J, Svensson L, Grip L. Risk reduction for cardiac events after primary coronary intervention compared with thrombolysis for acute ST-elevation myocardial infarction (five-year results of the Swedish early decision reperfusion strategy [SWEDES] trial). Am J Cardiol 2010 Dec 15;106(12):1685-91.

#2 Ahn CM, Hong SJ, Park JH, Kim JS, Lim DS. Cilostazol reduces the progression of carotid intima-media thickness without increasing the risk of bleeding in patients with acute coronary syndrome during a 2-year follow-up. Heart Vessels 2011 Sep;26(5):502-10.

#3 Alexander JH, Lopes RD, James S, Kilaru R, He Y, Mohan P, et al. Apixaban with antiplatelet therapy after acute coronary syndrome. N Engl J Med 2011 Aug 25;365(8):699-708.

#4 Ambrosio G, Flather MD, Bohm M, Cohen-Solal A, Murrone A, Mascagni F, et al. beta-blockade with nebivolol for prevention of acute ischaemic events in elderly patients with heart failure. Heart 2011 Feb;97(3):209-14.

#5 Antolovic D, Rakow A, Contin P, Ulrich A, Rahbari NN, Buchler MW, et al. A randomised controlled pilot trial to evaluate and optimize the use of anti-platelet agents in the perioperative management in patients undergoing general and abdominal surgery--the APAP trial (ISRCTN45810007). Langenbecks Arch Surg 2012 Feb;397(2):297-306.

#6 Aradi D, Rideg O, Vorobcsuk A, Magyarlaki T, Magyari B, Konyi A, et al. Justification of 150 mg clopidogrel in patients with high on-clopidogrel platelet reactivity. Eur J Clin Invest 2012 Apr;42(4):384-92.

#7 Arimura T, Miura SI, Ike A, Sugihara M, Iwata A, Nishikawa H, et al. Comparison of the efficacy and safety of statin and statin/ezetimibe therapy after coronary stent implantation in patients with stable angina. J Cardiol 2012 Apr 28.

#8 Athyros VG, Tziomalos K, Gossios TD, Griva T, Anagnostis P, Kargiotis K, et al. Safety and efficacy of long-term statin treatment for cardiovascular events in patients with coronary heart disease and abnormal liver tests in the Greek Atorvastatin and Coronary Heart Disease Evaluation (GREACE) Study: a post-hoc analysis. Lancet 2010 Dec 4;376(9756):1916-22.

#9 Bakker EJ, Ravensbergen NJ, Voute MT, Hoeks SE, Chonchol M, Klimek M, et al. A randomised study of perioperative esmolol infusion for haemodynamic stability during major vascular surgery; rationale and design of DECREASE-XIII. Eur J Vasc Endovasc Surg 2011 Sep;42(3):317-23.

#10 Baran C, Durdu S, Dalva K, Zaim C, Dogan A, Ocakoglu G, et al. Effects of Preoperative Short Term Use of Atorvastatin on Endothelial Progenitor Cells after Coronary Surgery: A Randomized, Controlled Trial. Stem Cell Rev 2011 Nov 11.

#11 Becattini C, Agnelli G, Schenone A, Eichinger S, Bucherini E, Silingardi M, et al. Aspirin for preventing the recurrence of venous thromboembolism. N Engl J Med 2012 May 24;366(21):1959-67.

#12 Benavente OR, White CL, Pearce L, Pergola P, Roldan A, Benavente MF, et al. The Secondary Prevention of Small Subcortical Strokes (SPS3) study. Int J Stroke 2011 Apr;6(2):164-75.

#13 Beygui F, Vicaut E, Ecollan P, Machecourt J, Van BE, Zannad F, et al. Rationale for an early aldosterone blockade in acute myocardial infarction and design of the ALBATROSS trial. Am Heart J 2010 Oct;160(4):642-8.

#14 Bhatt DL, Cryer BL, Contant CF, Cohen M, Lanas A, Schnitzer TJ, et al. Clopidogrel with or without omeprazole in coronary artery disease. N Engl J Med 2010 Nov 11;363(20):1909-17.

#15 Biondi-Zoccai G, Valgimigli M, Margheri M, Marzocchi A, Lettieri C, Stabile A, et al. Assessing the role of eptifibatide in patients with diffuse coronary disease undergoing drug-eluting stenting: The INtegrilin plus STenting to Avoid myocardial Necrosis Trial. Am Heart J 2012 May;163(5):835-7.

#16 Bousser MG, Amarenco P, Chamorro A, Fisher M, Ford I, Fox KM, et al. Terutroban versus aspirin in patients with cerebral ischaemic events (PERFORM): a randomised, double-blind, parallel-group trial. Lancet 2011 Jun 11;377(9782):2013-22.

#17 Bowling CB, Sanders PW, Allman RM, Rogers WJ, Patel K, Aban IB, et al. Effects of enalapril in systolic heart failure patients with and without chronic kidney disease: Insights from the SOLVD

Treatment trial. Int J Cardiol 2012 Jan 16.
#18 Callahan A, Amarenco P, Goldstein LB, Sillesen H, Messig M, Samsa GP, et al. Risk of stroke and cardiovascular events after ischemic stroke or transient ischemic attack in patients with type 2 diabetes or metabolic syndrome: secondary analysis of the Stroke Prevention by Aggressive Reduction in Cholesterol Levels (SPARCL) trial. Arch Neurol 2011 Oct;68(10):1245-51.
#19 Chimowitz MI, Lynn MJ, Derdeyn CP, Turan TN, Fiorella D, Lane BF, et al. Stenting versus aggressive medical therapy for intracranial arterial stenosis. N Engl J Med 2011 Sep 15;365(11):993-1003.
#20 Cho YR, Kim YD, Park TH, Park K, Park JS, Baek H, et al. The impact of dose of the angiotensin-receptor blocker valsartan on the post-myocardial infarction ventricular remodeling: study protocol for a randomized controlled trial. Trials 2011;12:247.
#21 Cice G, Di BA, D'Isa S, D'Andrea A, Marcelli D, Gatti E, et al. Effects of telmisartan added to Angiotensin-converting enzyme inhibitors on mortality and morbidity in hemodialysis patients with chronic heart failure a double-blind, placebo-controlled trial. J Am Coll Cardiol 2010 Nov 16;56(21):1701-8.
#22 Collet JP, Cayla G, Cuisset T, Elhadad S, Range G, Vicaut E, et al. Randomized comparison of platelet function monitoring to adjust antiplatelet therapy versus standard of care: rationale and design of the assessment with a double randomization of (1) a fixed dose versus a monitoring-guided dose of aspirin and clopidogrel after DES implantation, and (2) treatment interruption versus continuation, 1 year after stenting (ARCTIC) study. Am Heart J 2011 Jan;161(1):5-12.
#23 Deja MA, Kargul T, Domaradzki W, Stacel T, Mazur W, Wojakowski W, et al. Effects of preoperative aspirin in coronary artery bypass grafting: A double-blind, placebo-controlled, randomized trial. J Thorac Cardiovasc Surg 2012 May 1.
#24 Dungen HD, Apostolovic S, Inkrot S, Tahirovic E, Topper A, Mehrhof F, et al. Titration to target dose of bisoprolol vs. carvedilol in elderly patients with heart failure: the CIBIS-ELD trial. Eur J Heart Fail 2011 Jun;13(6):670-80.
#25 El-Atat F, Sarkar K, Kodali V, Karajgikar R, Jakkulla M, Mares A, et al. A randomized pilot trial for aggressive therapeutic approaches in aspirin-resistant patients undergoing percutaneous coronary intervention. J Invasive Cardiol 2011 Jan;23(1):9-13.
#26 Erdmann E, Spanheimer R, Charbonnel B. Pioglitazone and the risk of cardiovascular events in patients with Type 2 diabetes receiving concomitant treatment with nitrates, renin-angiotensin system blockers, or insulin: results from the PROactive study (PROactive 20). J Diabetes 2010 Sep;2(3):212-20.
#27 Fernandez A, Aboodi MS, Milewski K, Delgado JA, Rodriguez A, Granada JF. Comparison of adverse cardiovascular events and bleeding complications of loading dose of clopidogrel 300 mg versus 600 mg in stable patients undergoing elective percutaneous intervention (from the CADICE study). Am J Cardiol 2011 Jan;107(1):6-9.
#28 Funck-Brentano C, van Veldhuisen DJ, van d, V, Follath F, Goulder M, Willenheimer R. Influence of order and type of drug (bisoprolol vs. enalapril) on outcome and adverse events in patients with chronic heart failure: a post hoc analysis of the CIBIS-III trial. Eur J Heart Fail 2011 Jul;13(7):765-72.
#29 Gerdts E, Rossebo AB, Pedersen TR, Boman K, Brudi P, Chambers JB, et al. Impact of baseline severity of aortic valve stenosis on effect of intensive lipid lowering therapy (from the SEAS study). Am J Cardiol 2010 Dec 1;106(11):1634-9.
#30 Gheorghiade M, Albaghdadi M, Zannad F, Fonarow GC, Bohm M, Gimpelewicz C, et al. Rationale and design of the multicentre, randomized, double-blind, placebo-controlled Aliskiren Trial on Acute Heart Failure Outcomes (ASTRONAUT). Eur J Heart Fail 2011 Jan;13(1):100-6.
#31 Gibson CM, Mega JL, Burton P, Goto S, Verheugt F, Bode C, et al. Rationale and design of the Anti-Xa therapy to lower cardiovascular events in addition to standard therapy in subjects with acute coronary syndrome-thrombolysis in myocardial infarction 51 (ATLAS-ACS 2 TIMI 51) trial: a

randomized, double-blind, placebo-controlled study to evaluate the efficacy and safety of rivaroxaban in subjects with acute coronary syndrome. Am Heart J 2011 May;161(5):815-21.

#32 Gibson CM, Maehara A, Lansky AJ, Wohrle J, Stuckey T, Dave R, et al. Rationale and design of the INFUSE-AMI study: A 2 x 2 factorial, randomized, multicenter, single-blind evaluation of intracoronary abciximab infusion and aspiration thrombectomy in patients undergoing percutaneous coronary intervention for anterior ST-segment elevation myocardial infarction. Am Heart J 2011 Mar;161(3):478-86.

#33 Goette A, Schon N, Kirchhof P, Breithardt G, Fetsch T, Hausler KG, et al. Angiotensin II-antagonist in paroxysmal atrial fibrillation (ANTIPAF) trial. Circ Arrhythm Electrophysiol 2012 Feb;5(1):43-51.

#34 Good CW, Steinhubl SR, Brennan DM, Lincoff AM, Topol EJ, Berger PB. Is there a clinically significant interaction between calcium channel antagonists and clopidogrel?: results from the Clopidogrel for the Reduction of Events During Observation (CREDO) trial. Circ Cardiovasc Interv 2012 Feb 1;5(1):77-81.

#35 Goto K, Nikolsky E, Lansky AJ, Dangas G, Witzenbichler B, Parise H, et al. Impact of smoking on outcomes of patients with ST-segment elevation myocardial infarction (from the HORIZONS-AMI Trial). Am J Cardiol 2011 Nov 15;108(10):1387-94.

#36 Granger CB, Alexander JH, McMurray JJ, Lopes RD, Hylek EM, Hanna M, et al. Apixaban versus warfarin in patients with atrial fibrillation. N Engl J Med 2011 Sep 15;365(11):981-92.

#37 Gwon HC, Hahn JY, Park KW, Song YB, Chae IH, Lim DS, et al. Six-month versus 12-month dual antiplatelet therapy after implantation of drug-eluting stents: the Efficacy of Xience/Promus Versus Cypher to Reduce Late Loss After Stenting (EXCELLENT) randomized, multicenter study. Circulation 2012 Jan 24;125(3):505-13.

#38 Hart RG, Diener HC, Yang S, Connolly SJ, Wallentin L, Reilly PA, et al. Intracranial Hemorrhage in Atrial Fibrillation Patients During Anticoagulation With Warfarin or Dabigatran: The RE-LY Trial. Stroke 2012 Jun;43(6):1511-7.

#39 Hazarbasanov D, Velchev V, Finkov B, Postadjian A, Kostov E, Rifai N, et al. Tailoring clopidogrel dose according to multiple electrode aggregometry decreases the rate of ischemic complications after percutaneous coronary intervention. J Thromb Thrombolysis 2012 Jan 15.

#40 Hermanides RS, van HG, Ottervanger JP, de Boer MJ, Dill T, Hamm C, et al. The impact of age on effects of pre-hospital initiation of high bolus dose of tirofiban before primary angioplasty for ST-elevation myocardial infarction. Cardiovasc Drugs Ther 2011 Aug;25(4):323-30.

#41 Hermanides RS, van Werkum JW, Ottervanger JP, Breet NJ, Gosselink AT, van Houwelingen KG, et al. The effect of pre-hospital glycoprotein IIb-IIIa inhibitors on angiographic outcome in STEMI patients who are candidates for primary PCI. Catheter Cardiovasc Interv 2012 May 1;79(6):956-64.

#42 Hirohata A, Yamamoto K, Miyoshi T, Hatanaka K, Hirohata S, Yamawaki H, et al. Four-year clinical outcomes of the OLIVUS-Ex (impact of Olmesartan on progression of coronary atherosclerosis: evaluation by intravascular ultrasound) extension trial. Atherosclerosis 2012 Jan;220(1):134-8.

#43 Homma S, Thompson JL, Pullicino PM, Levin B, Freudenberger RS, Teerlink JR, et al. Warfarin and aspirin in patients with heart failure and sinus rhythm. N Engl J Med 2012 May 17;366(20):1859-69.

#44 Jin EZ, Yu LH, Li XQ. Loading effect of 200 mg cilostazol on platelet inhibition in patients undergoing percutaneous coronary intervention. Int Heart J 2012;53(1):1-4.

#45 Khattab AA, Ndrepepa G, Schulz S, Neumann FJ, Mehilli J, Buttner HJ, et al. Statin effect on thrombin inhibitor effectiveness during percutaneous coronary intervention: a post-hoc analysis from the ISAR-REACT 3 trial. Clin Res Cardiol 2011 Jul;100(7):579-85.

#46 Khosravi AR, Pourmoghadas M, Ostovan M, Mehr GK, Gharipour M, Zakeri H, et al. The impact of generic form of Clopidogrel on cardiovascular events in patients with coronary artery stent: results of the OPCES study. J Res Med Sci 2011 May;16(5):640-50.

#47 Kim HK, Hong YJ, Jeong MH, Kim W, Kim SS, Ko JS, et al. Two-year clinical outcome after

carvedilol-loaded stent implantation in patients with coronary artery disease. Korean J Intern Med 2011 Mar;26(1):41-6.
#48 Kim JS, Park SM, Kim BK, Ko YG, Choi D, Hong MK, et al. Efficacy of clotinab in acute myocardial infarction trial-ST elevation myocardial infarction (ECLAT-STEMI). Circ J 2012;76(2):405-13.
#49 Kouvelos GN, Arnaoutoglou EM, Matsagkas MI, Kostara C, Gartzonika C, Bairaktari ET, et al. Effects of Rosuvastatin With or Without Ezetimibe on Clinical Outcomes in Patients Undergoing Elective Vascular Surgery: Results of a Pilot Study. J Cardiovasc Pharmacol Ther 2012 May 9.
#50 Krum H, Massie B, Abraham WT, Dickstein K, Kober L, McMurray JJ, et al. Direct renin inhibition in addition to or as an alternative to angiotensin converting enzyme inhibition in patients with chronic systolic heart failure: rationale and design of the Aliskiren Trial to Minimize OutcomeS in Patients with HEart failuRE (ATMOSPHERE) study. Eur J Heart Fail 2011 Jan;13(1):107-14.
#51 Kulik A, Le May MR, Voisine P, Tardif JC, Delarochelliere R, Naidoo S, et al. Aspirin plus clopidogrel versus aspirin alone after coronary artery bypass grafting: the clopidogrel after surgery for coronary artery disease (CASCADE) Trial. Circulation 2010 Dec 21;122(25):2680-7.
#52 Lavitola PL, Sampaio RO, Oliveira WA, Boer BN, Tarasoutchi F, Spina GS, et al. Warfarin or aspirin in embolism prevention in patients with mitral valvulopathy and atrial fibrillation. Arq Bras Cardiol 2010 Dec;95(6):749-55.
#53 Lawrence J, Pogue J, Synhorst D, Adalet K, Atar D, Avezum A, et al. Apixaban versus aspirin in patients with atrial fibrillation and previous stroke or transient ischaemic attack: a predefined subgroup analysis from AVERROES, a randomised trial. Lancet Neurol 2012 Mar;11(3):225-31.
#54 Lee SP, Bae JW, Park KW, Rha SW, Bae JH, Suh JW, et al. Inhibitory interaction between calcium channel blocker and clopidogrel. -Efficacy of cilostazol to overcome it-. Circ J 2011;75(11):2581-9.
#55 Lee SW, Park SW, Kim YH, Yun SC, Park DW, Lee CW, et al. A randomized, double-blind, multicenter comparison study of triple antiplatelet therapy with dual antiplatelet therapy to reduce restenosis after drug-eluting stent implantation in long coronary lesions: results from the DECLARE-LONG II (Drug-Eluting Stenting Followed by Cilostazol Treatment Reduces Late Restenosis in Patients with Long Coronary Lesions) trial. J Am Coll Cardiol 2011 Mar 15;57(11):1264-70.
#56 Lee YS, Bae HJ, Kang DW, Lee SH, Yu K, Park JM, et al. Cilostazol in Acute Ischemic Stroke Treatment (CAIST Trial): a randomized double-blind non-inferiority trial. Cerebrovasc Dis 2011;32(1):65-71.
#57 Lengenfelder B, Stoerk S, Boes L, Strotmann J, Ertl G, Voelker W, et al. Long-term reduction of mortality in the 4-year follow up of tirofiban therapy in elective percutaneous coronary interventions (TOPSTAR) trial. J Invasive Cardiol 2011 Apr;23(4):128-32.
#58 Leonardi S, Mahaffey KW, White HD, Gibson CM, Stone GW, Steg GW, et al. Rationale and design of the Cangrelor versus standard therapy to acHieve optimal Management of Platelet InhibitiON PHOENIX trial. Am Heart J 2012 May;163(5):768-76.
#59 Li WM, Yang XC, Wang LF, Ge YG, Wang HS, Xu L, et al. Comparison of tirofiban combined with dalteparin or unfractionated heparin in primary percutaneous coronary intervention of acute ST-segment elevation myocardial infarction patients. Chin Med J (Engl) 2011 Oct;124(20):3275-80.
#60 Liu CP, Lin MS, Chiu YW, Lee JK, Hsu CN, Hung CS, et al. Additive benefit of glycoprotein IIb/IIIa inhibition and adjunctive thrombus aspiration during primary coronary intervention: results of the Initial Thrombosuction and Tirofiban Infusion (ITTI) trial. Int J Cardiol 2012 Apr 19;156(2):174-9.
#61 Liu P, Jiang J, Li J, Hong T, Zhang Y, Yu R, et al. Intensive statin therapy for Chinese patients with coronary artery disease undergoing percutaneous coronary intervention (ISCAP study): rationale and design. Catheter Cardiovasc Interv 2012 May 1;79(6):967-71.
#62 Mantz J, Samama CM, Tubach F, Devereaux PJ, Collet JP, Albaladejo P, et al. Impact of preoperative maintenance or interruption of aspirin on thrombotic and bleeding events after elective non-cardiac surgery: the multicentre, randomized, blinded, placebo-controlled, STRATAGEM trial. Br J Anaesth 2011 Dec;107(6):899-910.
#63 Marazzi G, Volterrani M, Caminiti G, Iaia L, Massaro R, Vitale C, et al. Comparative long

term effects of nebivolol and carvedilol in hypertensive heart failure patients. J Card Fail 2011 Sep;17(9):703-9.

#64 Mauri L, Kereiakes DJ, Normand SL, Wiviott SD, Cohen DJ, Holmes DR, et al. Rationale and design of the dual antiplatelet therapy study, a prospective, multicenter, randomized, double-blind trial to assess the effectiveness and safety of 12 versus 30 months of dual antiplatelet therapy in subjects undergoing percutaneous coronary intervention with either drug-eluting stent or bare metal stent placement for the treatment of coronary artery lesions. Am Heart J 2010 Dec;160(6):1035-41, 1041.

#65 McMurray JJ, Dunselman P, Wedel H, Cleland JG, Lindberg M, Hjalmarson A, et al. Coenzyme Q10, rosuvastatin, and clinical outcomes in heart failure: a pre-specified substudy of CORONA (controlled rosuvastatin multinational study in heart failure). J Am Coll Cardiol 2010 Oct 5;56(15):1196-204.

#66 Meng K, Lu SZ, Zhu HG, Chen X, Ge CJ, Song XT. Use of tailored loading-dose clopidogrel in patients undergoing selected percutaneous coronary intervention based on adenosine diphosphate-mediated platelet aggregation. Chin Med J (Engl) 2010 Dec;123(24):3578-82.

#67 Mitchell JE, Tam SW, Trivedi K, Taylor AL, O'Neal W, Cohn JN, et al. Atrial fibrillation and mortality in African American patients with heart failure: results from the African American Heart Failure Trial (A-HeFT). Am Heart J 2011 Jul;162(1):154-9.

#68 Mizuma H, Inoue T, Takano H, Shindo S, Oka T, Fujimatsu D, et al. Rationale and design of a study to evaluate effects of pitavastatin on Japanese patients with chronic heart failure: the pitavastatin heart failure study (PEARL study). Int J Cardiol 2012 Apr 19;156(2):144-7.

#69 Moliterno DJ. A randomized two-by-two comparison of high-dose bolus tirofiban versus abciximab and unfractionated heparin versus bivalirudin during percutaneous coronary revascularization and stent placement: the tirofiban evaluation of novel dosing versus abciximab with clopidogrel and inhibition of thrombin (TENACITY) study trial. Catheter Cardiovasc Interv 2011 Jun 1;77(7):1001-9.

#70 Monaco M, Di TL, Pinna GB, Lillo S, Schiavone V, Stassano P. Combination therapy with warfarin plus clopidogrel improves outcomes in femoropopliteal bypass surgery patients. J Vasc Surg 2012 May 1.

#71 Montalescot G, Zeymer U, Silvain J, Boulanger B, Cohen M, Goldstein P, et al. Intravenous enoxaparin or unfractionated heparin in primary percutaneous coronary intervention for ST-elevation myocardial infarction: the international randomised open-label ATOLL trial. Lancet 2011 Aug 20;378(9792):693-703.

#72 Mora S, Wenger NK, Demicco DA, Breazna A, Boekholdt SM, Arsenault BJ, et al. Determinants of Residual Risk in Secondary Prevention Patients Treated With High- Versus Low-Dose Statin Therapy: The Treating to New Targets (TNT) Study. Circulation 2012 Apr 24;125(16):1979-87.

#73 Morrow DA, Braunwald E, Bonaca MP, Ameriso SF, Dalby AJ, Fish MP, et al. Vorapaxar in the secondary prevention of atherothrombotic events. N Engl J Med 2012 Apr 12;366(15):1404-13.

#74 Nakagomi A, Kodani E, Takano H, Uchida T, Sato N, Ibuki C, et al. Secondary preventive effects of a calcium antagonist for ischemic heart attack: randomized parallel comparison with beta-blockers. Circ J 2011;75(7):1696-705.

#75 Nohara R, Daida H, Hata M, Kaku K, Kawamori R, Kishimoto J, et al. Effect of intensive lipid-lowering therapy with rosuvastatin on progression of carotid intima-media thickness in Japanese patients: Justification for Atherosclerosis Regression Treatment (JART) study. Circ J 2012;76(1):221-9.

#76 Ohlmann P, Reydel P, Jacquemin L, Adnet F, Wolf O, Bartier JC, et al. Prehospital abciximab in ST-segment elevation myocardial infarction: results of the randomized, double-blind MISTRAL study. Circ Cardiovasc Interv 2012 Feb 1;5(1):69-76, S1.

#77 Okada T, Yamamoto H, Okimoto T, Otsuka M, Ishibashi K, Dohi Y, et al. Beneficial effects of valsartan on target lesion revascularization after percutaneous coronary interventions with bare-metal stents. Circ J 2011;75(7):1641-8.

#78 Oldgren J, Budaj A, Granger CB, Khder Y, Roberts J, Siegbahn A, et al. Dabigatran vs. placebo in patients with acute coronary syndromes on dual antiplatelet therapy: a randomized, double-blind, phase II trial. Eur Heart J 2011 Nov;32(22):2781-9.
#79 Patti G, Barczi G, Orlic D, Mangiacapra F, Colonna G, Pasceri V, et al. Outcome comparison of 600- and 300-mg loading doses of clopidogrel in patients undergoing primary percutaneous coronary intervention for ST-segment elevation myocardial infarction: results from the ARMYDA-6 MI (Antiplatelet therapy for Reduction of MYocardial Damage during Angioplasty-Myocardial Infarction) randomized study. J Am Coll Cardiol 2011 Oct 4;58(15):1592-9.
#80 Patti G, Pasceri V, D'Antonio L, D'Ambrosio A, Macri M, Dicuonzo G, et al. Comparison of Safety and Efficacy of Bivalirudin Versus Unfractionated Heparin in High-Risk Patients Undergoing Percutaneous Coronary Intervention (from the Anti-Thrombotic Strategy for Reduction of Myocardial Damage During Angioplasty-Bivalirudin vs Heparin Study). Am J Cardiol 2012 May 12.
#81 Price MJ, Berger PB, Teirstein PS, Tanguay JF, Angiolillo DJ, Spriggs D, et al. Standard- vs high-dose clopidogrel based on platelet function testing after percutaneous coronary intervention: the GRAVITAS randomized trial. JAMA 2011 Mar 16;305(11):1097-105.
#82 Rafiq S, Johansson PI, Zacho M, Stissing T, Kofoed K, Lilleoer NB, et al. Thrombelastographic haemostatic status and antiplatelet therapy after coronary artery bypass surgery (TEG-CABG trial): assessing and monitoring the antithrombotic effect of clopidogrel and aspirin versus aspirin alone in hypercoagulable patients: study protocol for a randomized controlled trial. Trials 2012 Apr 27;13(1):48.
#83 Roghani F, Hemmat A, Golabchi A. Can doubling the maintenance dose of clopidogrel prevent from early stent thrombosis after the primary percutaneous coronary intervention? ARYA Atheroscler 2011;7(1):18-23.
#84 Ruff CT, Giugliano RP, Antman EM, Murphy SA, Lotan C, Heuer H, et al. Safety and efficacy of prasugrel compared with clopidogrel in different regions of the world. Int J Cardiol 2012 Mar 22;155(3):424-9.
#85 Sandset EC, Murray G, Boysen G, Jatuzis D, Korv J, Luders S, et al. Angiotensin receptor blockade in acute stroke. The Scandinavian Candesartan Acute Stroke Trial: rationale, methods and design of a multicentre, randomised- and placebo-controlled clinical trial (NCT00120003). Int J Stroke 2010 Oct;5(5):423-7.
#86 Sardella G, Conti G, Donahue M, Mancone M, Canali E, De CC, et al. Rosuvastatin pre-treatment in patients undergoing elective PCI to reduce the incidence of myocardial periprocedural necrosis. The ROMA trial. Catheter Cardiovasc Interv 2012 Apr 19.
#87 Steg PG, James S, Harrington RA, Ardissino D, Becker RC, Cannon CP, et al. Ticagrelor versus clopidogrel in patients with ST-elevation acute coronary syndromes intended for reperfusion with primary percutaneous coronary intervention: A Platelet Inhibition and Patient Outcomes (PLATO) trial subgroup analysis. Circulation 2010 Nov 23;122(21):2131-41.
#88 Steg PG, Mehta SR, Jukema JW, Lip GY, Gibson CM, Kovar F, et al. RUBY-1: a randomized, double-blind, placebo-controlled trial of the safety and tolerability of the novel oral factor Xa inhibitor darexaban (YM150) following acute coronary syndrome. Eur Heart J 2011 Oct;32(20):2541-54.
#89 Sun JC, Teoh KH, Lamy A, Sheth T, Ellins ML, Jung H, et al. Randomized trial of aspirin and clopidogrel versus aspirin alone for the prevention of coronary artery bypass graft occlusion: the Preoperative Aspirin and Postoperative Antiplatelets in Coronary Artery Bypass Grafting study. Am Heart J 2010 Dec;160(6):1178-84.
#90 Suzuki S, Sayama T, Nakamura T, Nishimura H, Ohta M, Inoue T, et al. Cilostazol improves outcome after subarachnoid hemorrhage: a preliminary report. Cerebrovasc Dis 2011;32(1):89-93.
#91 Swedberg K, Komajda M, Bohm M, Borer J, Robertson M, Tavazzi L, et al. Effects on Outcomes of Heart Rate Reduction by Ivabradine in Patients With Congestive Heart Failure: Is There an Influence of Beta-Blocker Dose?: Findings From the SHIFT (Systolic Heart failure treatment with the I(f) inhibitor ivabradine Trial) Study. J Am Coll Cardiol 2012 Feb 22.

#92 Tousek P, Osmancik P, Paulu P, Kocka V, Widimsky P. Clopidogrel up-titration versus standard dose in patients with high residual platelet reactivity after percutaneous coronary intervention: a single-center pilot randomised study. Int J Cardiol 2011 Jul 15;150(2):231-2.

#93 Tousek P, Rokyta R, Tesarova J, Pudil R, Belohlavek J, Stasek J, et al. Routine upfront abciximab versus standard periprocedural therapy in patients undergoing primary percutaneous coronary intervention for cardiogenic shock: The PRAGUE-7 Study. An open randomized multicentre study. Acute Card Care 2011 Sep;13(3):116-22.

#94 Trenk D, Stone GW, Gawaz M, Kastrati A, Angiolillo DJ, Muller U, et al. A Randomized Trial of Prasugrel Versus Clopidogrel in Patients With High Platelet Reactivity on Clopidogrel After Elective Percutaneous Coronary Intervention With Implantation of Drug-Eluting Stents: Results of the TRIGGER-PCI (Testing Platelet Reactivity In Patients Undergoing Elective Stent Placement on Clopidogrel to Guide Alternative Therapy With Prasugrel) Study. J Am Coll Cardiol 2012 Apr 9.

#95 Truong QA, Murphy SA, McCabe CH, Armani A, Cannon CP. Benefit of intensive statin therapy in women: results from PROVE IT-TIMI 22. Circ Cardiovasc Qual Outcomes 2011 May;4(3):328-36.

#96 Uchiyama S, Ikeda Y, Urano Y, Horie Y, Yamaguchi T. The Japanese aggrenox (extended-release dipyridamole plus aspirin) stroke prevention versus aspirin programme (JASAP) study: a randomized, double-blind, controlled trial. Cerebrovasc Dis 2011;31(6):601-13.

#97 Ussia GP, Scarabelli M, Mule M, Barbanti M, Sarkar K, Cammalleri V, et al. Dual antiplatelet therapy versus aspirin alone in patients undergoing transcatheter aortic valve implantation. Am J Cardiol 2011 Dec 15;108(12):1772-6.

#98 Valgimigli M, Campo G, Gambetti S, Bolognese L, Ribichini F, Colangelo S, et al. Three-year follow-up of the MULTIcentre evaluation of Single high-dose Bolus TiRofiban versus Abciximab with Sirolimus-eluting STEnt or Bare-Metal Stent in Acute Myocardial Infarction StudY (MULTISTRATEGY). Int J Cardiol 2011 Aug 22.

#99 Valgimigli M, Campo G, Monti M, Vranckx P, Percoco G, Tumscitz C, et al. Short- Versus Long-Term Duration of Dual-Antiplatelet Therapy After Coronary Stenting: A Randomized Multicenter Trial. Circulation 2012 Apr 24;125(16):2015-26.

#100 van Kuijk JP, Voute MT, Flu WJ, Schouten O, Chonchol M, Hoeks SE, et al. The efficacy and safety of clopidogrel in vascular surgery patients with immediate postoperative asymptomatic troponin T release for the prevention of late cardiac events: Rationale and design of the Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echo-VII (DECREASE-VII) trial. Am Heart J 2010 Sep;160(3):387-93.

#101 Veselka J, Zemanek D, Hajek P, Maly M, Adlova R, Martinkovicova L, et al. Effect of two-day atorvastatin pretreatment on long-term outcome of patients with stable angina pectoris undergoing elective percutaneous coronary intervention. Am J Cardiol 2011 May 1;107(9):1295-9.

#102 Wang TY, White JA, Tricoci P, Giugliano RP, Zeymer U, Harrington RA, et al. Upstream clopidogrel use and the efficacy and safety of early eptifibatide treatment in patients with acute coronary syndrome: an analysis from the Early Glycoprotein IIb/IIIa Inhibition in Patients with Non-ST-Segment Elevation Acute Coronary Syndrome (EARLY ACS) trial. Circulation 2011 Feb 22;123(7):722-30.

#103 Wang XD, Zhang DF, Zhuang SW, Lai Y. Modifying clopidogrel maintenance doses according to vasodilator-stimulated phosphoprotein phosphorylation index improves clinical outcome in patients with clopidogrel resistance. Clin Cardiol 2011 May;34(5):332-8.

#104 Wang Y, Johnston SC. Rationale and design of a randomized, double-blind trial comparing the effects of a 3-month clopidogrel-aspirin regimen versus aspirin alone for the treatment of high-risk patients with acute nondisabling cerebrovascular event. Am Heart J 2010 Sep;160(3):380-6.

#105 Weber M, Bhatt DL, Brennan DM, Hankey GJ, Steinhubl SR, Johnston SC, et al. High-sensitivity C-reactive protein and clopidogrel treatment in patients at high risk of cardiovascular events: a substudy from the CHARISMA trial. Heart 2011 Apr;97(8):626-31.

#106 White H, Held C, Stewart R, Watson D, Harrington R, Budaj A, et al. Study design and ratio-

nale for the clinical outcomes of the STABILITY Trial (STabilization of Atherosclerotic plaque By Initiation of darapLadIb TherapY) comparing darapladib versus placebo in patients with coronary heart disease. Am Heart J 2010 Oct;160(4):655-61.

#107 White M, Desai RV, Guichard JL, Mujib M, Aban IB, Ahmed MI, et al. Bucindolol, Systolic Blood Pressure, and Outcomes in Systolic Heart Failure: A Prespecified Post Hoc Analysis of BEST. Can J Cardiol 2012 May;28(3):354-9.

#108 Wiviott SD, Flather MD, O'Donoghue ML, Goto S, Fitzgerald DJ, Cura F, et al. Randomized trial of atopaxar in the treatment of patients with coronary artery disease: the lessons from antagonizing the cellular effect of Thrombin-Coronary Artery Disease Trial. Circulation 2011 May 3;123(17):1854-63.

#109 Wojnicz R, Nowak J, Lekston A, Wilczewski P, Nowalany-Kozielska E, Streb W, et al. Therapeutic window for calcium-channel blockers in the management of dilated cardiomyopathy: a prospective, two-centre study on non-advanced disease. Cardiology 2010;117(2):148-54.

#110 Yamashita T, Inoue H, Okumura K, Kodama I, Aizawa Y, Atarashi H, et al. Randomized trial of angiotensin II-receptor blocker vs. dihydropiridine calcium channel blocker in the treatment of paroxysmal atrial fibrillation with hypertension (J-RHYTHM II study). Europace 2011 Apr;13(4):473-9.

#111 Youn YN, Park SY, Hwang Y, Joo HC, Yoo KJ. Impact of High-Dose Statin Pretreatment in Patients with Stable Angina during Off-Pump Coronary Artery Bypass. Korean J Thorac Cardiovasc Surg 2011 Jun;44(3):208-14.

#112 Yu LT, Zhu J, Tan HQ, Wang GG, Teo KK, Liu LS. Telmisartan, ramipril, or both in high-risk Chinese patients: analysis of ONTARGET China data. Chin Med J (Engl) 2011 Jun;124(12):1763-8.

#113 Zeymer U, Arntz HR, Mark B, Fichtlscherer S, Werner G, Scholler R, et al. Efficacy and safety of a high loading dose of clopidogrel administered prehospitally to improve primary percutaneous coronary intervention in acute myocardial infarction: the randomized CIPAMI trial. Clin Res Cardiol 2012 Apr;101(4):305-12.

Appendix IV Reported exclusion criteria in RCTs in secondary prevention of cardiovascular events stratified for low and high impact journals*.

	Trial publications in high impact journal (N=42)	Trial publications in low impact journal (N=70)
Self-explanatory criteria		
Contraindication to intervention	31 (69% 95CI 55; 83)	44 (63% 95CI 52; 74)
Any impaired renal condition	23 (55% 95CI 40; 70)	40 (57% 95CI 46; 69)
Any impaired liver condition	25 (60% 95CI 45; 74)	34 (49% 95CI 37; 60)
High risk of bleeding	22 (52% 95CI 37; 67)	34 (49% 95CI 37; 60)
Criteria requiring justification		
Age below 18	32 (76% 95CI 63; 89)	46 (66% 95CI 55; 77)
Other age restrictions	18 (43% 95CI 63; 58)	23 (33% 95CI 22; 44)
Pregnant and/or fertile	23 (55% 95CI 40; 70)	30 (43% 95CI 31; 54)
Lactating women	14 (33% 95CI 19; 48)	16 (23% 95CI 13; 33)
Female gender	0	1 (1% 95CI 0; 4)
Male gender	0	0
Any medication usage at baseline	31 (74% 95CI 61; 87)	49 (70% 95CI 59; 81)
Non-naïve for Intervention	19 (45% 95CI 30; 60)	21 (30% 95CI 19; 41)
Opinion of physician	12 (29% 95CI 15; 42)	16 (23% 95CI 13; 33)
Indication for either treatment arm	3 (7% 95CI 0; 15)	10 (14% 95CI 6; 22)
Likely to be lost to follow-up	4 (10% 95CI 1; 18)	5 (7% 95CI 1; 13)
Short life expectancy	22 (52% 95CI 37; 67)	23 (33% 95CI 22; 44)
Lack of cognition or mental impairment	9 (21% 95CI 9; 34)	7 (10% 95CI 3; 17)
Physical disability	5 (12% 95CI 2; 22)	7 (10% 95CI 3; 17)

** High impact factor journals were those journals with an impact factor higher than 5, whereas low impact journals had an impact factor equal or lower than 5. The 95% confidence intervals (95CI)_are based on the asymptotic Wald method and values below 0% and above 100% were truncated. One study did not report any exclusion criteria.*

Appendix V Treatment effect estimates per intervention type (beta-blocker, clopidogrel and statin therapy) stratified for exclusion criteria*.

Exclusion criteria	Beta-blocker RCTs		Clopidogrel RCTs		Statin RCTs	
	Study	RR 95%CI	Study	RR 95%CI	Study	RR 95%CI
Exclusion due to: any medication usage at baseline.	Dungen 2011 Funck 2011 Marzazzi 2011 White 2012	2.25 (0.70; 7.30) 0.94 (0.50; 1.80) 0.86 (0.50; 1.50) 0.90 (0.53; 1.50)	Ahn 2010 Aradi 2012 Bhatt 2010 Fernandez 2011 Good 2012 Khosravi 2011 Lee 2011 a Patti 2011 Ruff 2012 Steg 2010 Valgmigli 2012 Wang 2011	0.80 (0.34; 1.90) 0.12 (0.02; 0.95) 0.99 (0.42; 2.33) 0.83 (0.37; 1.88) 0.73 (0.57; 0.95) 0.36 (0.12; 1.13) 0.67 (0.24; 1.86) 0.40 (0.16; 0.99) 0.81 (0.52; 1.27) 0.84 (0.55; 1.28) 0.98 (0.75; 1.28) 0.47 (0.26; 0.85)	Baran 2011 Callahan 2011 Mora 2012 Nohara 2011 Sardella 2012 Truong 2011	0.50 (0.05; 5.22) 0.84 (0.48; 1.49) 0.82 (0.48; 1.40) 0.50 (0.13; 1.97) 0.37 (0.19; 0.71) 0.87 (0.78; 0.96)
Did not apply exclusion criterion	Ambrosio 2010	0.68 (0.32; 1.40)	Lee 2011 b	0.60 (0.34; 1.05)	Arimura 2012 Athyros 2010 Gerdts 2012 Kouvelos 2012 Veselka 2011 Youn 2011	1.50 (0.28; 8.12) 0.57 (0.35; 0.95) 0.83 (0.35; 1.99) 0.50 (0.23; 1.07) 0.92 (0.43; 1.98) 0.25 (0.03; 2.18)
Exclusion due to: non-naïve for Intervention.	Ambrosio 2010 Dungen 2011 Funck 2011	0.68 (0.32; 1.40) 2.25 (0.70; 7.30) 0.94 (0.50; 1.80)	Aradi 2012 Bhatt 2010 Fernandez 2011 Good 2012 Khosravi 2011 Lee 2011 a Patti 2011 Ruff 2012	0.12 (0.02; 0.95) 0.99 (0.42; 2.33) 0.83 (0.37; 1.88) 0.73 (0.57; 0.95) 0.36 (0.12; 1.13) 0.67 (0.24; 1.86) 0.40 (0.16; 0.99) 0.81 (0.52; 1.27)	Athyros 2010 Baran 2011 Callahan 2011 Gerdts 2012 Mora 2012 Nohara 2011 Truong 2011 Veselka 2011	0.57 (0.35; 0.95) 0.50 (0.05; 5.22) 0.84 (0.48; 1.49) 0.83 (0.35; 1.99) 0.82 (0.48; 1.40) 0.50 (0.13; 1.97) 0.87 (0.78; 0.96) 0.92 (0.43; 1.98)
Did not apply exclusion criterion	Marzazzi 2011 White 2012	0.86 (0.50; 1.50) 0.90 (0.53; 1.50)	Ahn 2010 Lee 2011 b Steg 2010 Valgmigli 2012 Wang 2011	0.80 (0.34; 1.90) 0.60 (0.34; 1.05) 0.84 (0.55; 1.28) 0.98 (0.75; 1.28) 0.47 (0.26; 0.85)	Arimura 2012 Kouvelos 2012 Sardella 2012 Youn 2011	1.50 (0.28; 8.12) 0.50 (0.23; 1.07) 0.37 (0.19; 0.71) 0.25 (0.03; 2.18)
Exclusion due to: opinion of physician	Ambrosio 2010 Funck 2011 White 2012	0.68 (0.32; 1.40) 0.94 (0.50; 1.80) 0.90 (0.53; 1.50)	Bhatt 2010 Ruff 2012 Steg 2010	0.99 (0.42; 2.33) 0.81 (0.52; 1.27) 0.84 (0.55; 1.28)	Callahan 2011 Nohara 2011	0.84 (0.48; 1.49) 0.50 (0.13; 1.97)

Did not apply exclusion criterion	Dungen 2011 Marzazzi 2011	2.25 (0.70; 7.30) 0.86 (0.50; 1.50)	Ahn 2010 Aradi 2012 Fernandez 2011 Good 2012 Khosravi 2011 Lee 2011 a Lee 2011 b Patti 2011 Valgmigli 2012 Wang 2011	0.80 (0.34; 1.90) 0.12 (0.02; 0.95) 0.83 (0.37; 1.88) 0.73 (0.57; 0.95) 0.36 (0.12; 1.13) 0.67 (0.24; 1.86) 0.60 (0.34; 1.05 0.40 (0.16; 0.99) 0.98 (0.75; 1.28) 0.47 (0.26; 0.85)	Arimura 2012 Athyros 2010 Baran 2011 Gerdts 2012 Kouvelos 2012 Mora 2012 Sardella 2012 Truong 2011 Veselka 2011 Youn 2011	1.50 (0.28; 8.12) 0.57 (0.35; 0.95) 0.50 (0.05; 5.22) 0.83 (0.35; 1.99) 0.50 (0.23; 1.07) 0.82 (0.48; 1.40) 0.37 (0.19; 0.71) 0.87 (0.78; 0.96) 0.92 (0.43; 1.98) 0.25 (0.03; 2.18)

**Numbers indicate risk ratios (RR) with 95% confidence intervals (95%CI)*

CHAPTER 8

Approaches to determine generalizability of treatment effects

Amand F Schmidt, Arno W Hoes and Rolf H H Groenwold

Based on: Comments on: "'The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias"
Statistics in Medicine 2014. Feb 33(3):536-537
doi: 10.1002/sim.5929.

It has been widely accepted that most randomized control trials (RCTs) include patient groups that are not a representative sample of the patients who will receive the intervention in daily practice [1-3]. This has raised concerns about the generalizability of RCT results. Recently Pressler and Kaizar [4] enriched the discussion by asserting that the bias that results from a lack of generalizability can be quantified. They define two populations: the first consisting of subjects fulfilling the inclusion criteria of a RCT (in their notation group I) and the second group (group E) comprising subjects not meeting the inclusion criteria. They propose to estimate the treatment effects in both groups ($\hat{\Delta}(I)$ and $\hat{\Delta}(E)$, respectively) using nonrandomized (i.e., observational) data. Assuming equal amounts of confounding in both groups $\hat{\beta} = \hat{\Delta}(I) - \hat{\Delta}(E)$ provides an unbiased estimate of how much treatment effect modification there exists between included and excluded subjects. If the interest is in estimating the "population average treatment effect " (PATE), weighing $\hat{\beta}$ by the proportion of E among the total population of interest $\hat{\pi} = \frac{n_E}{n}$ provides an estimate $\hat{\gamma} = \hat{\pi}\hat{\beta}$ of how much "generalizability bias" is created by relying on $\hat{\Delta}(I)$ to estimate the PATE. This weighing of $\hat{\beta}$ is necessary because if there is treatment effect modification between groups I and E, the PATE is dependent on the proportionate size of both groups. While we acknowledge the relevance of the approach suggested by Pressler and Kaizer, we wish to touch upon some concerns and discuss alternative strategies for exploring generalizability.

First, Pressler and Kaizer fail to address why one would be interested in the treatment effect in group E. For example, if we explore the effectiveness of a new antihypertensive drug and E comprises subjects without hypertension it seems illogic to try to estimate treatment effect modification between groups I and E.

Second, using nonrandomized data to estimate $\hat{\gamma}$ or $\hat{\beta}$ only results in an unbiased estimate if the amount of confounding is equal in both groups E and I. This assumption is not testable, as the authors acknowledge, and results in a problem encountered in virtually all nonrandomized studies; i.e., not knowing whether estimates are unbiased or biased by confounding.

Third, when $\hat{\gamma}$ is estimated an implicit assumption is made that the effects within groups I and E are homogenous, otherwise $\hat{\gamma}$ does not necessarily reflect a lack of generalizability due to excluding group E. Imagine a RCT in which only subjects younger than 50 years are enrolled and let there be treatment effect modification by diabetes status. Hence,

group E would consist of subjects older than 50 years. In that case, $\hat{\gamma}$ might deviate from 0 simply because age increases the number of diabetic subjects; i.e., the magnitude of $\hat{\gamma}$ becomes dependent on the proportion diabetics subjects. To show that this is not a lack of generlizability between excluded and included patients, note that while there is effect modification between diabetic and non-diabetic subjects $\left(\hat{\beta} = \hat{\Delta}_{DM} - \hat{\Delta}_{NDM} \neq 0\right)$ within diabetic and non-diabetic subgroups the difference between groups I and E equals 0; $\left(\hat{\gamma}_{DM} = \hat{\pi}_{DM} * \left[\hat{\Delta}(I)_{DM} - \hat{\Delta}(E)_{DM}\right] = 0\right) = \left(\hat{\gamma}_{NDM} = \hat{\pi}_{NDM} * \left[\hat{\Delta}(I)_{NDM} - \hat{\Delta}(E)_{NDM}\right] = 0\right)$.

This brings us to the final issue. If there is treatment effect modification between groups I and E (and assuming homogenous effects within groups I and E) the "population average treatment effect" (PATE) will depend on the proportion of excluded patients $\hat{\pi}$. For example, let there be treatment effect modification between group I and E caused by age. Specifically, the relative risk for the outcome under treatment is 0.4 in subjects younger than 50 years of age (group I) and 1 among subjects older than 50 (group E). Furthermore, let the proportion E differ from 0.1 to 0.9 between certain regions of a country. Consequently, the PATE will range from 0.44 to 0.91. Reporting numerous region specific PATE estimates is at the very least inefficient compared to the alternative of reporting two age specific estimates. Furthermore, in the presence of treatment effect modification, the PATE is not applicable to any (group of) subject(s) making this an inappropriate effect estimate. Instead, the age specific effect estimates are applicable to their respective group members. If estimating the PATE is inappropriate when there is treatment effect modification, it is equally inappropriate to interpret $\hat{\gamma}$ as the amount of "generalizability bias". Therefore, we suggest that instead of focussing on "generalizability bias" it is more helpful to simply indicate if treatment effect modification is present or absent (i.e., if $\hat{\beta} \neq 0$). This can be estimated between groups I and E, as Pressler and Kaizer advise, but also within subgroups of I or E.

Before discussing alternative strategies we want to recognize that in settings in which all trials exclude the same kind of subjects the approach suggested by Pressler and Kaizar to estimate $\hat{\beta}$ can indeed provide valuable insights on generalizability. However, at the potential cost of confounding bias because the effect estimates in E and I are based on nonrandomized data. Alternatively, if some RCTs include patients excluded by other RCTs, comparisons of the effect estimates between trials can also provide information on generalizability

[5]. While this latter approach prevents confounding bias within trials, estimates may still differ due to differences between studies, e.g. concomitant drug use, which could also affect conclusions regarding generalizability [6]. Ideally $\hat{\beta}$ should be estimated within RCTs, which guard both against bias due to study specific effects and confounding bias. Therefore, we suggest that researchers explore generalizability by focusing on treatment effect modification using individual patient data (IPD) from multiple RCTs, which allows $\hat{\beta}$ to be estimated within studies.

Reference List

1. Van Spall HG, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. JAMA 2007; 297(11):1233-1240.
2. Lee PY, Alexander KP, Hammill BG, Pasquali SK, Peterson ED. Representation of elderly persons and women in published randomized trials of acute coronary syndromes. JAMA 2001; 286(6):708-713.
3. Bugeja G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. BMJ 1997; 315(7115):1059.
4. Pressler TR, Kaizar EE. The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. Stat.Med. 2013.
5. Schmidt AF, Rovers MM, Klungel OH, Hoes AW, Knol MJ, Nielen M, de BA, Groenwold RH. Differences in interaction and subgroup-specific effects were observed between randomized and nonrandomized studies in three empirical examples. J.Clin.Epidemiol. 2013; 66(6):599-607.
6. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Stat.Med. 2002; 21(3):371-387.

Part V

General discussion

Before launching a new treatment to the market, medical interventions and most notably drugs, are typically evaluated in randomized clinical trials (RCTs) that primarily focus on the intended effects of interventions. Sometimes, RCTs can also provide information on relatively common (type A) unintended (i.e. adverse) effects (1-3). After marketing, intervention effects (both intended and unintended) are often monitored using nonrandomized studies (e.g., case-control or cohort studies), supplemented by post-launch RCTs when needed. These studies are usually designed to provide information on the average intervention effect. Therefore, differences in treatment effects between a wide range of potential users will often remain undetected (4-7).

When treatment effects differ between patients, this is referred to as effect modification, interaction, or heterogeneity of treatment effects. For example, consider a hypothetical trial that includes patients with (20%) and without (80%) diabetes. The risk ratio (RR) of the intervention effect on the 5-years incidence of stroke differs, between patients with and without diabetes: say RR= 0.40 among patients with diabetes and RR = 1.00 among patients without diabetes. The observed (average) intervention effect is a weighted average of the effect among patients with and patients without diabetes: RR = 0.83. In this example, the intervention effect differs between subgroups based on diabetes status, i.e., there is effect modification by diabetes. In the presence of effect modification, average treatment effects are non-informative: the RR = 0.83 applies neither to patients with diabetes, nor to patients without diabetes. Instead, subgroup-specific effect estimates are more meaningful for future patients.

Throughout this thesis, we used the term *effect modification, interaction* and *heterogeneity* interchangeably. Some reserve the term *interaction* for the specific situation of heterogeneity of treatment effect when a factor biologically interacts with the treatment and *effect modific*ation for the situation where it does not (8). This distinction can usually not be determined analytically and will therefore not be made here. Also, it has been recognized that the presence of effect modification depends on the effect measure chosen (e.g., risk difference, risk ratio or odds ratio) [**Chapter 1**] (9-11). For example, if the odds ratio is constant across patient subgroups this does not preclude heterogeneity of the risk ratio (12). Effect modification is therefore also referred to as effect measure modification. Here, we consider situations where the effect measure was selected a priori and thus only consider effect modification of the particular effect measure chosen.

When study results do not suggest any effect modification, the main (i.e. average) treatment effect found in a study is likely generalizable beyond the population included in the study; because there is no direct reason to believe the treatment acts differently in other subjects [**Chapters 7 and 8**] (13;14). However, most clinical studies are not designed to detect treatment effect modification and often assume homogeneity of treatment effects(15). As the aforementioned example showed, undetected effect modification can result in wrongfully applying the main treatment effect to patients. In our example, patients without diabetes would be treated despite not having any benefit from treatment. On the other hand, the effect in patients with diabetes was underestimated; affecting, for example, the willingness to prescribe or take the medication, the cost-effectiveness ratio and possibly resulting in negative decisions regarding reimbursement. Overall, patients are treated suboptimally when effect modification is not recognized.

In this chapter, we will consider how study results can be translated to individual patients using the concept of treatment effect modification. First, strategies to detect treatment effect modification are addressed. Second, we present a structured approach to assess to whom estimated treatment effects are generalizable and for whom a more tailored treatment effect estimate is needed.

Detecting treatment effect modification

Detecting treatment effect modification is essential in translating study results to individual patients. Effect modification can be detected by testing whether the interaction effect, e.g. in a regression model, differs from zero (16;17). However, such interaction tests are renowned for their lack of power (i.e., the probability of correctly concluding that interaction exists) combined with large type 1 errors (i.e., the probability of falsely concluding that interaction exists) [**Chapter 1**] (10;18-23). While this might seem counterintuitive, this is due to the erratic behaviour of interaction tests (as any other test) in sparse data settings. In sparse data settings, interaction tests may show large outliers (away from zero) under the null-hypothesis, but also large outliers away from the expected value (thus toward zero), under the alternative hypothesis. Often this performance is viewed as inevitable, but proper sample size calculations with oversampling of the relevant patients can prevent data sparseness. Thus, interaction tests do not inherently underperform, but rather they lack proper planning.

For more definitive conclusions on the absence or presence of treatment effect modification, the current approach to interaction testing needs improvement. The first step is to more actively share and pool individual patient data to increase the effective sample size and thus increase the power of interaction tests [**Chapters 1, 2, 3, 5 and 8**] (14;24;25). Increases in sample size can increase power (and decrease type 1 errors). Perhaps equally important, data sharing allows exploring consistency of effects across multiple studies.

A second improvement, might be to follow a more inclusive view on the body of evidence, allotting a more prominent role to nonrandomized intervention studies (26-29). Exploring interactions in RCTs as well as nonrandomized studies can provide information on generalizability to less controlled settings, i.e. daily clinical practice. Recently, empirical studies on various clinical topics showed comparable results in RCTs and nonrandomized studies [**Chapters 2, 3 and 7**] (24;30-33). At the same time we recognize that randomized and nonrandomized studies differ in their likelihood of bias (notably confounding) and consequently the strength of evidence. A possible way to incorporate this strength of evidence component is to use Bayesian methods. Bayesian methods provide an intuitive way to incorporate this [**Chapter 3**] (34-39); for example by reweighting poor quality studies. At the very least, this reweighting of nonrandomized studies is more promising than acting as if nonrandomized studies provide no evidence at all; a practice commonly applied in most systematic reviews and meta-analyses on intended treatment effects (29). Similarly, when RCT and randomized studies contradict each other, focusing on RCT results alone, too easily puts the blame on nonrandomized studies. Instead researchers should try to explore why results differed between different designs [**chapter 6**] (40).

Third, for interaction tests to be anything but exploratory, interaction tests should be prespecified including proper sample size calculations. Sample size calculations and sampling strategies (e.g., equally sized subgroups) can ensure appropriate power and type 1 error rates. One attractive idea is to incorporate interaction tests using adaptive trial designs (41-43). For example, design a trial to show presence of a main effect in a homogenous group of patients. When during interim analysis there is enough evidence to expect that the treatment is effective (i.e. there is a beneficial average effect), the second study period (the period following the interim analysis) can be used to enrich the patient sample to explore heterogeneity

between important patient subgroups. We recognize that this contrasts with the more usual approach of focusing on a single promising subgroup after interim (41;44). Here we actually reverse the usual approach: we start with a subgroup where we expect treatment to be most beneficial and in the second stage (after interim) explore consistency of this treatment effect across important subgroups.

Finally, while focusing on detecting generalizability of treatment effects, one should be aware that a non-significant interaction test can never be interpreted as proof for the absence of treatment effect modification. To quote Altman (45): “absence of evidence is not evidence of absence”. Instead to ‘prove’ equivalence, so called equivalence tests should be used. Recognizing that the strict null-hypothesis (i.e., H_0: $\mu_0 = 0$) never holds, tests of equivalence determine margins between which differences in treatment effect estimates are small enough to be deemed clinically irrelevant (46;47). When the treatment effect estimate and its confidence interval fall between these margins equivalence is ‘proven’ (**Figure 1)**. This approach has been frequently applied to main effects (48). Equivalence tests can be extended to interaction tests by determining a margin around the neutral interaction effect, which is sufficiently small for the subgroup-specific estimates to be considered equivalent. As with any test, equivalence interaction tests require proper planning to ensure sufficient power and sample size. One approach could be to use the adaptive trial design suggested previously and in the second stage include enough patients to show equivalence for subgroups of major interest.

Treating individuals

Probably there is not a single treatment that is equally effective in every patient, so there will always be a need for subgroup-specific (or even patient specific) treatment effect estimates; i.e., identification of subsets of patients for whom treatment effects are more or less similar. In Box 1, a scheme is presented to assess effects of interventions that are applicable to an individual rather than apply to a population that comprises a wide variety of patients. In this scheme it is suggested to first explore whether generalizability can be rejected and if so, confirmatory analyses are suggested to estimate specific treatment effects for individuals.

To start with, one should decide for which subgroups the presence or absence of treatment effect modification should be determined e.g., based on age,, gender, comorbidity, etc. Often subgroups are chosen based on prior knowledge of the biological. However, it also seems

important to take into account how frequently certain patients are encountered in practice. When, for example comorbidity is a potential effect modifier, it seems more reasonable to assess whether relatively common diseases, such as diabetes, modify the effect than rare diseases. Focussing on common subgroups will obviously result in more patients benefitting. Furthermore, the costs of measuring the patient characteristic (49) should also be considered e.g., age, a common effect modifier, is easily measurable, whereas determining a genetic marker is quite expensive. Discussions on the choice of subgroups should focus on patients included but also certainly on patients not included in a (future) study (50). Often such discussions revolve around the question whether the patient sample was *representative* of the target population or the "average" patient (51). However, representativeness plays only a minor role in generalizing treatment effects to individuals, instead treatment effect modification play a more important role (29;52;53). In the absence of effect modification the same treatment effect applies to every patient subgroup, and thus, representativeness is irrelevant. In the presence of treatment effect modification, a representative sample will more often than not actually preclude detection of treatment effect modification (due to unequal subgroup sizes). Hence, representativeness often results in wrongfully assuming generalizability of treatment effects and in patients being treated suboptimally. Furthermore, even if one is interested in population average treatment effect (13), in the presence of interaction small differences between populations can result in markedly different main treatment effects [**Chapter 8**] (14). Assume, for example that the main treatment effect is 1.00 (RR) in a population aged 65. In the presence of an interaction effect of 0.95 (RR) per year, the treatment effect in a population aged 70 will be 0.77 (RR) [i.e, $e^{ln(1.00)+ln(0.95)*(70-65)} \approx 0.77$].Thus, discussing which patients might respond differently to treatment is essential, however, this should not be guided by the issue of representativeness.

In a second step, internal homogeneity of a study should be explored, preferably by selecting subgroups based on the results of step 1. Often this entails performing multiple interaction tests which inflate the overall type 1 error rate (54). In an attempt to decrease the number of false positive findings pre-specification of interaction tests has often been advocated (55-57). However, pre-specification does not necessarily decrease the number of tests applied, nor will it prevent an increase in the overall type 1 error rate. Furthermore, pre-specification does not significantly increase power to detect interactions unless proper design steps are taken (e.g., oversampling of subgroups) (58). Therefore, we suggest that these interaction tests are

deemed exploratory unless pre-specification coincides with steps ensuring sufficient sample size, power and type 1 error levels.

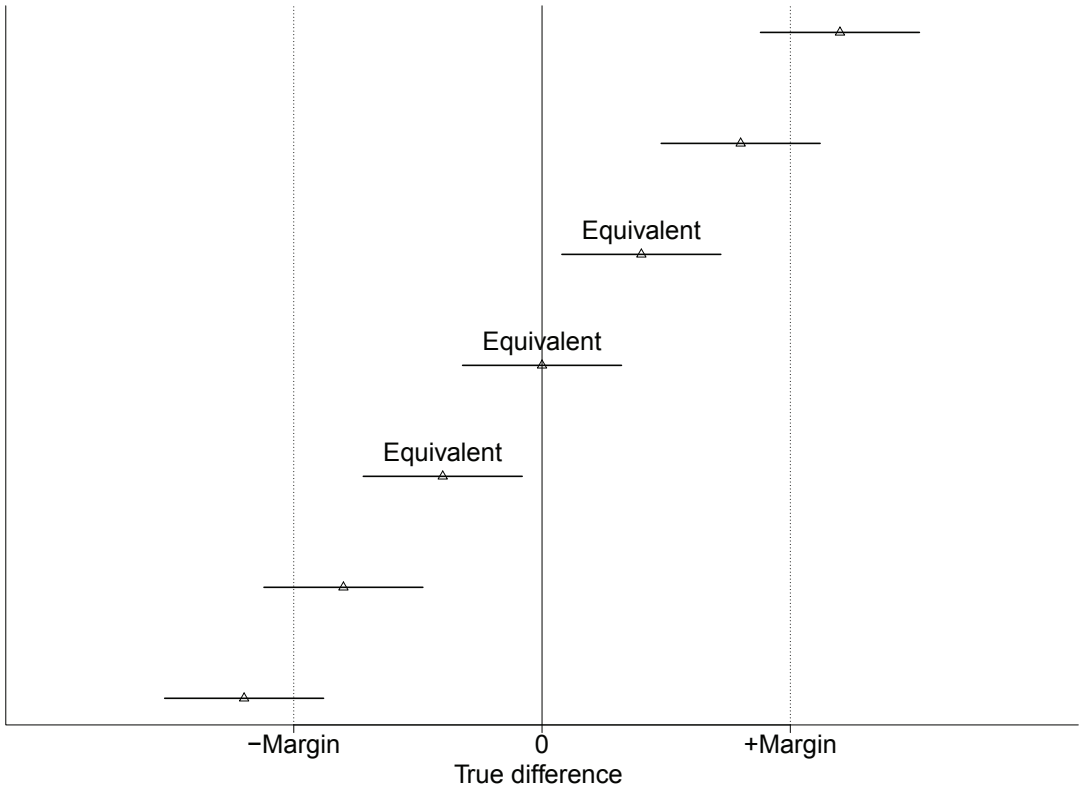
In a third step heterogeneity across studies should be explored; again incorporating information from the previous steps. The most basic approach is to compare aggregated results from different studies (24;59). This might also be an opportunity to explore whether similar estimates where discovered in less controlled settings, e.g., comparing results from RCTs and nonrandomized studies [**Chapters 2, 3 and 6**] (24;59). However, attributing differences in treatment effects between studies to differences in baseline characteristics, using for example meta-regression, may result in (ecological) bias. To prevent this bias, it is recommended to acquire access to the individual patient data from multiple studies and to explore if differences can be explained by (multiple) interactions terms [**Chapters 4,5 and 6**] (14;25;60).

If, after performing the above discussed exploratory analyses, absence of effect modification cannot be excluded with confidence, one should estimate more specific treatment effects. The most common way is to estimate subgroup-specific treatment effects, such as the diabetes specific estimates in our example study (RR = 0.40 versus RR = 1.00 in patients without diabetes).

Recently, such subgroup-specific estimates based on a single variable (i.e., univariable interaction tests) have been criticized (61-65). Among other reasons, critics recognized that patients likely differ on more than one characteristic (i.e., there is unexplained treatment effect modification). A straightforward solution is to include multiple interaction tests, for example exploring whether treatment effects differ by diabetes, gender, and age. However, exploring higher order interactions inevitably increases data sparseness, which dramatically reduces power and increases type 1 error rates [**Chapters 1 and 3**] (10;20-23;66-71). To solve this, a two-step multivariable method has been suggested. First, a multivariable risk prediction model is developed, predicting the risk of the outcome if no treatment is applied [**Chapter 4**] (72;73). Second, the predicted absolute risk is multiplied by a relative treatment effect estimate (e.g. a risk ratio) (65). Assume, for example, that in our previous trial the multivariable predicted 5-year stroke risk is 0.10, for a particular patient with diabetes. Treating this patient will then result in a predicted 5 year risk of 0.04 (i.e., 0.40 * 0.10 = 0.04).

While this multivariable approach to subgroup analysis is indeed an improvement, there are some remaining challenges. First, in the above described multivariable approach one implicitly assumes that the main relative treatment effect estimate (i.e., the RR) is homogeneous but that the effect of treatment is heterogeneous on an absolute scale (i.e., there is effect modification on the risk difference scale). While this is possible, it seems advisable to explore whether this is the case by adding a treatment by predicted risk interaction term to the statistical model [**Chapter 5**]. Second, the described multivariable approach induces heterogeneity on the absolute scale, which depends on the magnitude of the treatment effect estimate on the relative scale and the range of the baseline risk across different subgroups. However, it is unclear whether this reflects true heterogeneity of the absolute risk and currently no statistical tests are available to explore this. Finally, this approach ignores factors unrelated to the outcome (such as genetic variants that are related to drug metabolism (74)).

Figure 1. Examples of equivalence testing using confidence intervals.



Based on Jones et al. (66).

Recommendations and conclusions

In the present commentary we have argued that detecting treatment effect modification is essential to bridge the gap between the results from clinical studies and treating individuals

in daily practice. We addressed strategies to detect effect modification and used these in a framework to estimate more individualized effects of treatment where needed (**Box 1**).

Box 1 Proposed strategy to estimate differential treatment effects.

1. **Explore whether generalizability of treatment effects can be rejected:**
   a. Discuss for which patients treatment effects are expected to differ. This should be guided by biological plausibility, prevalence of the patient characteristic and cost-effectiveness of determining the patient characteristics and subsequent tailored treatment.
   b. Use results of 1.a to assess internal homogeneity of treatment effects by performing multiple exploratory interaction tests.
   c. Use results of 1.a and 1.b to explore whether treatment effects differ between studies and if this can be explained by differences in included patients.
2. **Perform confirmatory studies to estimate differential treatment effects based on the results from step 1:**
   a. Use sample size calculations and oversampling strategies of patient subgroups to ensure appropriate levels of power and type 1 error.
   b. Perform multivariable interaction tests to determine and individual's reaction to treatment.

We conclude with the following recommendations. First, treatment effect modification should be formally assessed using interaction tests. Second, for interaction tests to be anything but exploratory, these should not only be prespecified, but proper sample size calculations and sampling strategies should ensure appropriate levels of power and type 1 error rates. Third, prespecified subgroups should be selected based on biological plausibility, prevalence of the patient type and cost-effectiveness of determining the patient characteristic and subsequent tailored treatment. Fourth, to obtain sufficient sample size, researchers should collaborate and pool individual patient data. This includes combining data from randomized and non-randomized studies, possibly by means of Bayesian statistical methods. Fifth, because it is important to assess whether effect modification is absent (and thus judge the generalizability of the results), equivalence testing can be considered. Finally, the above focuses very much on the analyses of study results within and between studies. However, discussing potential patients not included in clinical studies also seems as essential. Such discussions should be aimed at potential effect modifiers, however, and not on representativeness.

In conclusion, generalizability and treatment effect modification are strongly interlinked. In the presence of effect modification, estimates of average (overall) treatment effects are non-informative and only subgroup-specific effect estimates can be generalized. Generalizability of these specific treatment effect estimates should subsequently be assessed, resulting in a continuing cycle of ever improving knowledge on how individual patients will react to treatment.

Reference List

1. Grobbee DE, Hoes AW. Intervention Research: Unintended Effects. Clinical Epidemiology: Principles, Methods and Applications for Clinical Research. 2 ed. Burlington: Jones and Bartlett Learning; 2015. p. 181-214.
2. Vandenbroucke JP. When are observational studies as credible as randomised trials? Lancet 2004 May 22;363(9422):1728-31.
3. Vandenbroucke JP. What is the best evidence for determining harms of medical treatment? CMAJ 2006 Feb 28;174(5):645-6.
4. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005 Jan 1;365(9453):82-93.
5. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005 Jan 8;365(9454):176-86.
6. Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. Lancet 2005 Jan 15;365(9455):256-65.
7. Bugeja G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. BMJ 1997 Oct 25;315(7115):1059.
8. VanderWeele TJ. On the distinction between interaction and effect modification. Epidemiology 2009 Nov;20(6):863-71.
9. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. Am J Epidemiol 1980 Oct;112(4):467-70.
10. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. Stat Med 1983 Apr;2(2):243-51.
11. White IR, Elbourne D. Assessing subgroup effects with binary data: can the use of different effect measures lead to different conclusions? BMC Med Res Methodol 2005;5:15.
12. Morabia A, Ten HT, Landis JR. Interaction fallacy. J Clin Epidemiol 1997 Jul;50(7):809-12.
13. Pressler TR, Kaizar EE. The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. Stat Med 2013 Apr 1.
14. Schmidt AF, Hoes AW, Groenwold RH. Comments on 'The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias' by Taylor R. Pressler and Eloise E. Kaizar, Statistics in Medicine 2013. Stat Med 2014 Feb 10;33(3):536-7.
15. Hernan MA. A definition of causal effect for epidemiological research. J Epidemiol Community Health 2004 Apr;58(4):265-71.
16. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. BMJ 2003 Jan 25;326(7382):219.
17. Matthews JN, Altman DG. Statistics notes. Interaction 2: Compare effect sizes not P values. BMJ 1996 Sep 28;313(7060):808.
18. Bagheri Z, Ayatollahi SM, Jafari P. Comparison of three tests of homogeneity of odds ratios in multicenter trials with unequal sample sizes within and among centers. BMC Med Res Methodol 2011;11:58.
19. Lui KJ. Testing homogeneity of the risk ratio in stratified noncompliance randomized trials. Contemp Clin Trials 2007 Sep;28(5):614-25.
20. Lui KJ. A simple test of the homogeneity of risk difference in sparse data: an application to a multicenter study. Biom J 2005 Oct;47(5):654-61.
21. O'Gorman TW, Woolson RF, Jones MP, Lemke JH. Statistical analysis of K 2 x 2 tables: a comparative study of estimators/test statistics for association and homogeneity. Environ Health Perspect 1990 Jul;87:103-7.
22. Paul SR, Donner A. Small sample performance of tests of homogeneity of odds ratios in K 2 x 2 tables. Stat Med 1992 Jan 30;11(2):159-65.
23. Zhang L, Yang H, Cho I. Test homogeneity of risk difference across subgroups in clinical trials. J

Biopharm Stat 2009 Jan 7;(1520-5711 (Electronic)).
24. Schmidt AF, Rovers MM, Klungel OH, Hoes AW, Knol MJ, Nielen M, et al. Differences in interaction and subgroup-specific effects were observed between randomized and nonrandomized studies in three empirical examples. J Clin Epidemiol 2013 Jun;66(6):599-607.
25. Koopman L, van der Heijden GJ, Hoes AW, Grobbee DE, Rovers MM. Empirical comparison of subgroup effects in conventional and individual patient data meta-analyses. Int J Technol Assess Health Care 2008;24(3):358-61.
26. Vandenbroucke JP. Why do the results of randomised and observational studies differ? BMJ 2011;343:d7020.
27. Vandenbroucke JP. Commentary: the HRT story: vindication of old epidemiological theory. Int J Epidemiol 2004 Jun;33(3):456-7.
28. Lawlor DA, Davey SG, Ebrahim S. Commentary: the hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology? Int J Epidemiol 2004 Jun;33(3):464-7.
29. Rothman KJ. Six Persistent Research Misconceptions. J Gen Intern Med 2014 Jan 23.
30. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000 Jun 22;342(25):1887-92.
31. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001 Aug 15;286(7):821-30.
32. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. Am J Ophthalmol 2000 Nov;130(5):688.
33. Shikata S, Nakayama T, Noguchi Y, Taji Y, Yamagishi H. Comparison of effects in randomized controlled trials with observational studies in digestive surgery. Ann Surg 2006 Nov;244(5):668-76.
34. Abrams K, Ashby D, Errington D. Simple Bayesian analysis in clinical trials: a tutorial. Control Clin Trials 1994 Oct;15(5):349-59.
35. Jones HE, Ohlssen DI, Neuenschwander B, Racine A, Branson M. Bayesian models for subgroup analysis in clinical trials. Clin Trials 2011 Apr;8(2):129-43.
36. Simon R. Bayesian subset analysis: application to studying treatment-by-gender interactions. Stat Med 2002 Oct 15;21(19):2909-16.
37. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. Stat Med 1995 Dec 30;14(24):2685-99.
38. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. Health Technol Assess 2000;4(38):1-130.
39. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian Approaches to Randomized Trials. Journal of the Royal Statistical Society Series A (Statistics in Society) 1994 Jan 1;157(3):357-416.
40. Hernan MA, Alonso A, Logan R, Grodstein FMKB, Stampfer MJ, Willet WC, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology 2008;19(6):766-79.
41. Boessen R, van der BF, Groenwold R, Egberts A, Klungel O, Grobbee D, et al. Optimizing trial design in pharmacogenetics research: comparing a fixed parallel group, group sequential, and adaptive selection design on sample size requirements. Pharm Stat 2013 Nov;12(6):366-74.
42. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. Stat Med 2009 Apr 15;28(8):1181-217.
43. van der Baan FH, Knol MJ, Klungel OH, Egberts AC, Grobbee DE, Roes KC. Potential of adaptive clinical trial designs in pharmacogenetic research. Pharmacogenomics 2012 Apr;13(5):571-8.
44. Tanniou J, Tweel vd T, Teerenstra S, Kit CBR. Level of evidence for promising subgroup findings in an overall non-significant trial. Statistical Methods in Medical Research 2014.
45. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995 Aug 19;311(7003):485.

46. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. BMJ 1996 Jul 6;313(7048):36-9.
47. Fleming TR. Design and interpretation of equivalence trials. Am Heart J 2000 Apr;139(4):S171-S176.
48. Home PD, Pocock SJ, Beck-Nielsen H, Curtis PS, Gomis R, Hanefeld M, et al. Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial. Lancet 2009 Jun 20;373(9681):2125-35.
49. Veenstra DL, Higashi MK, Phillips KA. Assessing the cost-effectiveness of pharmacogenomics. AAPS PharmSci 2000;2(3):E29.
50. Graaf vdR, Groenwold RHH, Kalkman S, Grobbee DE, Delden JJM. From Justifying Inclusion to Justifying Exclusion of Study Populations: Strengths and Limitations. World Medical Journal 2013;59(5):192-7.
51. Dekkers OM, von EE, Algra A, Romijn JA, Vandenbroucke JP. How to assess the external validity of therapeutic trials: a conceptual approach. Int J Epidemiol 2010 Feb;39(1):89-94.
52. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. Int J Epidemiol 2013 Aug;42(4):1012-4.
53. Rothman KJ, Gallacher JE, Hatch EE. Rebuttal: When it comes to scientific inference, sometimes a cigar is just a cigar. Int J Epidemiol 2013 Aug;42(4):1026-8.
54. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995 Jan 21;310(6973):170.
55. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005 Jan 8;365(9454):176-86.
56. Barraclough H, Govindan R. Biostatistics primer: what a clinician ought to know: subgroup analyses. J Thorac Oncol 2010 May;5(5):741-6.
57. Fletcher J. Subgroup analyses: how to avoid being misled. BMJ 2007 Jul 14;335(7610):96-7.
58. Peterson B, George SL. Sample size requirements and length of study for testing interaction in a 2 x k factorial design when time-to-failure is the outcome [corrected]. Control Clin Trials 1993 Dec;14(6):511-22.
59. Rovers MM, Straatman H, Ingels K, van der Wilt GJ, van den Broek P., Zielhuis GA. Generalizability of trial results based on randomized versus nonrandomized allocation of OME infants to ventilation tubes or watchful waiting. J Clin Epidemiol 2001 Aug;54(8):789-94.
60. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Stat Med 2002 Feb 15;21(3):371-87.
61. Kent DM, Ruthazer R, Selker HP. Are some patients likely to benefit from recombinant tissue-type plasminogen activator for acute ischemic stroke even beyond 3 hours from symptom onset? Stroke 2003 Feb;34(2):464-7.
62. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Medical Research Methodology 2006;6:18.
63. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. Journal of the American Medical Association 2007 Sep 12;298(10):1209-12.
64. Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 2010;11:85.
65. Dorresteijn JA, Visseren FL, Ridker PM, Wassink AM, Paynter NP, Steyerberg EW, et al. Estimating treatment effects for individual patients based on the results of randomised clinical trials. BMJ 2011;343:d5888.
66. Jones MP, O'Gorman TW, Lemke JH, Woolson RF. A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. Biometrics 1989 Mar;45(1):171-

81.
67. Liang KY, Self SG. Tests for Homogeneity of Odds Ratio When the Data are Sparse 3. Biometrika 1985 Aug 1;72(2):353-8.
68. Lipsitz SR, Dear KB, Laird NM, Molenberghs G. Tests for homogeneity of the risk difference when data are sparse. Biometrics 1998 Mar;54(1):148-60.
69. Lui KJ, Kelly C. Tests for homogeneity of the risk ratio in a series of 2x2 tables. Stat Med 2000 Nov 15;19(21):2919-32.
70. Lui KJ, Chang KC. Test homogeneity of odds ratio in a randomized clinical trial with noncompliance. J Biopharm Stat 2009 Sep;19(5):916-32.
71. Reis IM, Hirji KF, Afifi AA. Exact and asymptotic tests for homogeneity in several 2 x 2 tables. Stat Med 1999 Apr 30;18(8):893-906.
72. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. Heart 2012 May;98(9):683-90.
73. Schmidt AF, Nielen M, Klungel OH, Hoes AW, de Boer A, Groenwold RH, et al. Prognostic factors of early metastasis and mortality in dogs with appendicular osteosarcoma after receiving surgery: An individual patient data meta-analysis. Preventive Veterinary Medicine 2013 Nov 1;112(3-4):414-22.
74. Schelleman H, Stricker BH, de BA, Kroon AA, Verschuren MW, Van Duijn CM, et al. Drug-gene interactions between genetic polymorphisms and antihypertensive therapy. Drugs 2004;64(16):1801-16.

Het is doorgaans moeilijk om resultaten uit klinische studies toe te passen bij de behandeling van individuele patiënten. De meeste klinische studies zijn ontworpen om informatie te verstrekken over het gemiddelde effect van een interventie. Hierdoor blijven de potentieel verschillende reacties van patiënten op interventies vaak onopgemerkt.

Wanneer behandeleffect(en) verschillen tussen patiënten, spreekt men van effectmodificatie, interactie of heterogeniteit van de behandelingseffecten. Met behulp van het concept effect modificatie is in dit proefschrift beschreven hoe studieresultaten kunnen worden vertaald naar behandeleffecten die relevant zijn voor individuele patiënten.

Het detecteren van effect modificatie

Zoals beschreven is het detecteren van factoren die behandeleffecten modificeren essentieel om individuele patiënten optimaal te behandelen. Doorgaans wordt dit gedaan door middel van zogenaamde interactietoetsen. In hoofdstuk 1 is de prestatie van een aantal veel gebruikte interactietoesten geëvalueerd. Wanneer het aantal patiënten in een onderzoek groot is (bv. 1000 patiënten) is de prestatie van de verschillende testen gelijkwaardig. In kleinere steekproeven is, afhankelijk van de test, de kans om een interactieeffect te detecteren (als dat er werkelijk is) aanzienlijk kleiner dan het gebruikelijke niveau van 80%. Tegelijkertijd neemt de kans op fout-positieve bevindingen toe: van 5% naar 10%. De volgende testen presteerden het beste en worden daarom aanbevolen: Tarone, Breslow-Day, Likelihood Ratio en de testen gebaseerd op het Relative Excess Risk due to Interaction (RERI).

In hoofdstuk 2 is onderzocht of interactie-effecten gebaseerd op gerandomiseerd onderzoek (RCTs) verschillen van de interactie-effecten verkregen uit niet-gerandomiseerd onderzoek. Doordat behandeling in RCTs willekeurig (at random) wordt toebedeeld aan patiënten is theoretisch de kans op verstoring kleiner en daarom heeft deze onderzoeksopzet de voorkeur boven niet-gerandomiseerde studies. Daarentegen kunnen niet-gerandomiseerde studies doorgaans meer patiënten includeren, wat de kans vergroot om interactie-effecten te detecteren. Om te bepalen of het mogelijk is om resultaten uit beide type onderzoeken te combineren is empirische data vergeleken. Specifiek zijn de effecten van statine, een bypass operatie en mammografiescreening systematisch vergeleken voor gerandomiseerd en niet-gerandomiseerd onderzoek. De vergelijkbaarheid van interactie-effecten tussen verschillende onderzoeksopzetten was afhankelijk van de onderzochte interventie. Ondanks vergelijkbare

gemiddelde effecten, verschilden interactie effecten tussen RCTs en niet-gerandomiseerde studies.

In hoofdstuk 3 is bestudeerd of het combineren van resultaten uit RCTs en niet-gerandomiseerde studies mogelijk is met Bayesiaanse statistiek. Als voorbeeld is gebruik gemaakt van het geslachtsspecifieke effect van rosiglitazone op de incidentie van heupfracturen. De resultaten van dit onderzoek tonen aan dat in de meeste gevallen de kans om interactie-effecten te detecteren toe nam bij het gebruik van Bayesiaanse statistiek ten opzichten van frequentistische statistiek Dit, zonder een onacceptabel hoge kans op fout-positieve bevindingen. Echter, wanneer RCTs en niet-gerandomiseerde studies resultaten in tegenovergestelde richtingen lieten zien, werden liep het risico op fout-positieve en fout-negatieve bevindingen op tot wel 100%.

Het behandelen van individuele patiënten

In de hiervoor besproken en geëvalueerde methoden lag de focus voornamelijk op effect modificatie door een enkele factor, bijvoorbeeld geslacht. Uiteraard, verschillen patiënten op meer dan één factor van elkaar. Idealiter worden daarom meer variabelen gelijktijdig gebruikt om te bepalen hoe een patiënt reageert op een behandeling. Een manier om multivariabele effectmodificatie te bestuderen is om eerst een predictie regel te construeren om bijvoorbeeld sterfte te voorspellen. Vervolgens kan worden bestudeerd of de reactie op behandeling af hangt van deze voorspelde kans op sterfte. Een voorbeeld hiervan is beschreven in de hoofdstukken 4 en 5. In hoofdstuk 4 is een voorspelmodel gemaakt om de kans te voorspellen op sterfte of het ontwikkelen van een metastase bij honden met operatief behandelde botkanker. Onafhankelijk van andere risicofactoren bleken serum alkalische fosfatase, gewicht, tumor locatie en leeftijd gerelateerd te zijn aan sterfte en/of het ontwikkelen van een metastase. Vervolgens is in hoofdstuk 5 bepaald of de reactie van honden op additionele chemotherapie afhangt van het voorspelde risico op sterfte en/of metastase. De resultaten suggereerden dat voornamelijk honden met een relatief laag risico op sterfte baat hadden van deze additionele chemotherapie.

Generaliseerbaarheid van behandeleffecten

Wanneer effect modificatie afwezig is, is het waarschijnlijker dat patiënten niet verschillend reageren op de behandeling. In zulke gevallen zijn behandeleffecten ‘generaliseerbaar’.

Generaliseerbaarheid van behandeleffecten wordt soms in twijfel getrokken. Getwijfeld wordt er met name aan de generaliseerbaarheid van behandeleffecten die gebaseerd zijn op RCTs, omdat RCTs vaak slechts een zeer selecte groep patiënten includeren. Daarentegen worden in niet-gerandomiseerde onderzoeken vaak patiënten geïncludeerd die een betere afspiegeling zijn van de klinische praktijk. Als een empirisch voorbeeld zijn, in hoofdstuk 6, de RCT resultaten van de effecten van atenolol en propranolol op het voorkomen van een hartinfarct vergeleken met resultaten uit niet-gerandomiseerde onderzoeken. Er was weinig verschil tussen de effecten geschat in de RCTs en de effecten geschat in de niet-gerandomiseerde studies. Dit impliceert dat, in dit voorbeeld, de resultaten van de RCT generaliseerbaar zijn. In hoofdstuk 7 zijn de resultaten gepresenteerd van een systematische review van RCTs over secundaire cardiovasculaire preventie. In deze review is bestudeerd of RCTs verschillende typen patienten includeerden en of dit verschil leidde tot andere effectschattingen van bètablokkers, clopidogrel en statines. Hoewel een verschil niet helemaal kon worden uitgesloten, leken de resultaten aan te geven dat er, ondanks het includeren van verschillende patienten, geen systematisch verschil was tussen de RCTs. In hoofdstuk 8 is de relatie tussen effect modificatie en generaliseerbaarheid formeel gedefinieerd. Daarnaast, wordt er in dit hoofdstuk aangetoond dat, in de aanwezigheid van effect modificatie, het gebruiken van subgroep-specifieke effect schatters te verkiezen is boven het gemiddelde effect van behandeling.

In de general discussion wordt op basis van de voorgaande hoofstukken een stappenplan gepresenteerd om resultaten van klinische studies toe te passen bij het behandelen van individuele patiënten.

Applying results from clinical studies to individual patients is challenging. Clinical studies are usually designed to provide information on the average intervention effect. Therefore, differences in treatment effects between a wide range of patients will often remain undetected.

When treatment effects differ between patients, this is referred to as effect modification, interaction, or heterogeneity of treatment effects. In this thesis, we considered how study results can be translated to individual patients using the concept of treatment effect modification.

Detecting effect modification of interventions

Before results from clinical studies can be translated to individual patients it is essential to determine whether results differ between patients. To detect this potential treatment effect modification a plethora of tests is available. In chapter 1, the performance of these tests was evaluated by simulating clinical studies of different sizes and with different interaction effects. When the number of patients was large (e.g., 1,000 subjects) all tests performed equally. In smaller sample sizes, depending on the tests chosen, the probability to detect treatment effect modification when it was present, was well below the customary level of 0.80. At the same time the probability of falsely concluding interaction could be as high as 0.10. In these small sample size settings the Tarone, Breslow-Day, Likelihood Ratio and Relative Excess Risk due to Interaction (RERI) based tests performed best and are recommended.

In chapter 2 we explored whether interaction effects observed in randomized controlled trials (RCTs) were comparable to nonrandomized studies, using a review of empirical studies on statin therapy, bypass surgery and mammography screening. While nonrandomized studies are more prone to bias (mainly due to the lack of randomization), they also have the potential to include more subjects, increasing the probability of detecting treatment effect modification. This review showed that comparability of interaction effects across study designs was topic specific. Despite comparable main effect estimates, interaction effects could still considerably differ between RCTs and nonrandomized studies.

In chapter 3 the utility of Bayesian methods to incorporate nonrandomized study results in the analysis of an RCT was studied. As an example, we focused on the effect of rosiglitazone on bone fracture incidence and modification by gender. In most settings, the probability to detect an interaction effect increased in Bayesian analysis compared to frequentist analysis,

without increasing the false positive rate too severely. In settings where results from nonrandomized and RCT studies were in opposing directions, false positive and negative rates as high as 100% were observed.

Bridging the gap between clinical studies and individual patient care

The previously discussed methods predominantly focused on whether treatment effects differ between patients using a single baseline variable, for example gender. Obviously, patients differ on more than one variable. Ideally, treatment effect modification should include multiple variables. One approach to multivariable treatment effect modification is to first construct a rule based on multiple variables to predict a subject's future probability of developing a certain health outcome. By combining multiple variables into a single number (e.g., the probability of a future event), treatment effect modification can be explored in the usual manner only now including multiple predictors for the outcome. An application of such an approach is provided in chapters 4 and 5. In chapter 4 a prediction rule for early mortality and early metastasis was developed for canines who were surgically treated for osteosarcoma. This study described an individual patient data meta-analysis using data from 20 studies, which showed that independent of other risk factors serum alkaline phosphatase, weight, tumor location and age were associated with early mortality and/or early metastasis. The follow-up study presented in chapter 5 showed that dogs with different predicted risk of 5-month mortality responded differently to additional chemotherapy treatment. Result suggested that dogs with a relatively low risk of 5-month mortality benefitted most from additional chemotherapy.

Generalizability of the effects of interventions

In the absence of treatment effect modification, treatment effects are expected to be similar for every patient, in other words generalizable. In such settings, translating the results from clinical studies to individual patients becomes less complicated because the average (main) treatment effect is applicable to every patient. However, generalizability of RCT results is often questioned because it is well known that patients included in RCTs may differ from those included in nonrandomized studies. The latter include "real life" patients, whereas RCTs may include highly selected patient populations, as a result of strict in- and exclusion criteria. In chapter 6 the generalizability of RCT results on atenolol or propranolol compared to diuretics in preventing non-fatal myocardial infraction (MI) was assessed. Specifically, results of these RCTs were compared to results from two nonrandomized studies. There was evidence that

atenolol affected MI incidence differently compared to propranolol. However, the effect estimates of propranolol and atenolol were comparable across the study designs (RCTs vs nonrandomized studies), suggesting generalizability. Results from a systematic review of secondary cardiovascular RCTs, presented in chapter 7, showed that despite inclusion of different types of patients the effects of beta-blocker, clopidogrel, and statin therapy did not systematically differ between studies. While absence of effect modification could not be proven, results seemed to indicate that clinical study results might be generalizable. In chapter 8 the relationship between generalizability and treatment effect modification was formally addressed and it was argued to use subgroup-specific estimates when treatment effect modification is present.

Based on the previous chapters a unifying approach is presented in the general discussion to guide clinicians, patients and other potential stakeholders in most optimally treating patients encountered in clinical practice. We argue that to truly estimate individualized treatment effects, studies should be designed to detect treatment effect modification (or its absence), for example using adaptive designs. Additionally, we argue that it is time for a more inclusive view on intervention research allowing for a more prominent role of nonrandomized studies, potentially using Bayesian statistics to account for the possible bias in nonrandomized studies.

During the 4 years that I have spent working on this thesis if have become indebted to a great number of people; a few I wish to mention especially.

Prof. dr. A.W. Hoes, dear Arno besides my love of boxing, which you so frequently referred to, I am also very fond of cooking. To every draft paper you added the final missing ingredient that brought everything together. Without your input, this thesis would have been very bland. Thank you.

Prof. dr. M. Nielen, dear Mirjam to some this collaboration between veterinary and 'human' medicine might seem strange. I however am very glad that you became one of my promoters. Always, you would confidently insist that I explained something until you understood it. All too often, this led me to the insight that I too did not fully grasp the topic. I hope one day I can be half as confident as you.

Dr. R.H.H. Groenwold, dear Rolf I am very grateful and perhaps lucky that you decided to be my supervisor on this project. Certainly, without you I would not have finished half the paper in this thesis. At the start of my project I did not understand much about epidemiology or statistics (although I was convinced I did). Despite this, you always trusted me with topics, that you knew I did not fully grasp at that time. Because of your guidance I travelled a very long distance.

Dr. O.H. Klungel, dear Olaf thank you for making the time to supervise me. Your knowledge on drugs and experience with drug development were indispensible. You opened my eyes to see that drug development is not only about benefit but also about cost and adverse events and that this requires a different methodology.

Prof. dr. A. de Boer, dear Ton you have been a co-author on nearly every paper I have written thus far. Your input was very valuable especially on the applied papers were you reminded me time and time again to more carefully state my conclusion given that evidence add hand. I will try to remember this.

Prof. dr. J. Kirpesteijn, dear Jolle we met during the first year of my thesis. I was in need of empirical data and you had a vast network of colleagues willing to share their data. Without

a doubt, my thesis would have been two precious chapters shorter had we not met. I wish to thank you, not only for our research together, but also for you enthusiasm, patients and trust.

Dear co-authors Prof. dr. F. Gueyffier, Prof. dr. K. C. B. Roes, Prof. dr. J.J.M van Delden, Prof. dr. M.R. Rovers, Prof. dr. D. Vail, Prof. dr. J. Berg, Dr. I. Klugkist, Dr. M Knol, Dr. R. van der Graaf, Dr. P. Amsellem, Dr. N. Bacon, Dr K. Kow, Dr. I. Kurzman, Dr. K. Maritato, Dr. A. Moore, Dr. E. Morello and Dr J. Sottnik thank you for your contributions and the opportunities offered.

I also which to thank the biostatistics department for allowing me to assist in numerous courses, give lectures and finally also offering me the opportunity to develop my own. Thank you Bert, Caroline, Rene, Esther, Cas, Rebecca, Ingeborg, Paul, Peter, Julien, Rutger, Putri, Stavros, Victor and Konstantinos.

To the members of the weekly methodology meeting I am grateful for first putting up with my confusion and later putting up with my comments. Sanni, Jammal, Grace, Marijn, Stavros, Rolf and Mirjam, our bi-weekly causality meetings have given me much food for thought. To Patrick and Mark, I am not only grateful for the data that you extracted for me, but also for your patients with me while I tried to grasp the data structure.

All the people at the Julius center and at the UMC. Coby, Henk en Monique, I really appreciated you helping me find my way. For all the coffee's, lunches and inspiration I have to thank, Julien, John, Marijn, Ruud, Frederieke, Manon, Noor, Sara, Lotje, Charlotte, Puspha, Maarten, Maaike, Loes, Kim, Ellen, Stan, Thomas, Welling, Chantal, Nanne, Lizelotte and Willemijn. Coffee would of course not be possible without the MiCaffe and Jurjen, Sabine and Monique.

John, Elle and Nico, I am very glad that we have stayed in touch since so long. John I hope you and Jolieke will be very happy in the United States.

Anoukh and Sophie thank you for being my friends, colleagues and reading bits and pieces of my drivel. I Hope you are happy with this paragraph. I am looking forward to our next party.

Aziz, Derk, Julian and Luuk you are my oldest friends. Without you, I would have few good stories, parties, vacations and other (mis)adventures. Derk I am happy that you will be there once I move to London, it’s going to be fun!

My dear family Henk, Martha, Liesbeth, Andre and Lauri thank you for your love, trust and instilling in me the confidence to follow my dreams. Josef, Louella, Bram and Han thank you for the fun times during the holidays, your company and advice.

Stavros and Julian I am so happy you both agreed to be my paranymps. Be assured you are already forgiven for the jokes and stories at my expense.

Elke, without you I would not have started, let alone finished this research. I would be lost without you.

Amand Floriaan Schmidt was born on 22 October, 1986 in Amsterdam. In 2004 Floriaan studied Nutrition and Dietetics at the applied university of Arnhem and Nijmegen (HAN). After obtaining a bachelor's degree in 2008 he continued his studies at the VU University. There, in 2010, he obtained his Master of Science in Health Sciences with a major in public health and infectious disease. That same year he started working as a PhD student at the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, under supervision of prof. dr A.W. Hoes, prof. dr. M. Nielen, dr. O.H. Klungel and dr. R.H.H. Groenwold, culminating in the present thesis. In 2013 Floriaan attained a Master of Science in Clinical Epidemiology at the Utrecht University. In the near future Floriaan will continue his career as a postdoctoral researcher in genetic epidemiology with the Institute of Cardiovascular Science, University College London.