PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2016; **25**: 355–362
Published online 15 February 2016 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/pds.3965

REVIEW

Tailoring treatments using treatment effect modification†

A. F. Schmidt[1,2,3,4]*, O. H. Klungel[1,2], M. Nielen[3], A. de Boer[2], R. H. H. Groenwold[1,2] and A. W. Hoes[1]

[1] *Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands*
[2] *Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht, the Netherlands*
[3] *Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, Utrecht, the Netherlands*
[4] *Institute of Cardiovascular Science, Faculty of Population Health, University College London, London, UK*

ABSTRACT

Background and objective Applying results from clinical studies to individual patients can be a difficult process. Using the concept of treatment effect modification (also referred to as interaction), defined as a difference in treatment response between patient groups, we discuss whether and how treatment effects can be tailored to better meet patients' needs.

Results First we argue that contrary to how most studies are designed, treatment effect modification should be expected. Second, given this expected heterogeneity, a small number of clinically relevant subgroups should be a priori selected, depending on the expected magnitude of effect modification, and prevalence of the patient type. Third, by defining generalizability as the absence of treatment effect modification we show that generalizability can be evaluated within the usual statistical framework of equivalence testing. Fourth, when equivalence cannot be confirmed, we address the need for further analyses and studies tailoring treatment towards groups of patients with similar response to treatment. Fifth, we argue that to properly frame, the entire body of evidence on effect modification should be quantified in a prior probability.

KEY WORDS—randomized controlled trial; nonrandomized study design; observational study design; statistics; effect modification; interaction; generalizability; pharmacoepidemiology

Received 29 June 2015; Revised 16 December 2015; Accepted 17 December 2015

BACKGROUND

Before launching a new treatment on the market, medical interventions and most notably drugs are typically evaluated in randomized clinical trials (RCTs), which primarily focus on the intended effects of interventions. Sometimes, RCTs can also provide information on relatively common 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. Consider a hypothetical trial (Table 1) that includes patients with diabetes (40%) and patients without (60%). The risk ratio (RR) of the intervention effect on the 5-year incidence of stroke differs between patients with and without diabetes: e.g. RR = 0.75 among patients with diabetes and RR = 0.63 among patients without diabetes. The observed (average) intervention effect is a weighted average of the effects among patients with and patients without diabetes: RR = 0.68. In this example, the intervention effect differs between subgroups based on diabetes status; i.e. there is effect modification by diabetes. Patients may be treated suboptimally when effect modification is not recognized.

Throughout this paper, we will use the term *effect modification*, *interaction* and *heterogeneity* interchangeably. Some reserve the term *interaction* for the

*Correspondence to: A. F. Schmidt, Institute of Cardiovascular Science, Faculty of Population Health, University College London, London WC1E 6BT, UK. E-mail: amand.schmidt@ucl.ac.uk

†Prior postings and presentations: This study and its results have not been previously published, neither has it been presented at conferences.

Table 1. Stroke risk by exposure (X) and baseline diabetes (D) status and their interaction on different measurement scales

	$D=0$	$D=1$
$X=0$	0.80	0.89
$X=1$	0.50	0.67
Risk difference interaction	Risk ratio interaction	Odds ratio interaction
$0.67-0.50-0.89+0.80=0.08$	$\frac{0.67\times0.80}{0.50\times0.89}=1.20$	$\frac{\frac{0.67}{0.33}\times\frac{0.80}{0.20}}{\frac{0.89}{0.11}\times\frac{0.50}{0.50}}=1.00$

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 not be made here either. Also, it has been recognized that the presence of effect modification depends on the effect measure chosen.[9–11] In the example RCT (Table 1) there is interaction on the RR [and on the risk difference (RD)] scale; however, using the odds ratio the effect of treatment is 0.25, in both diabetic and non-diabetic patients.[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.

In this paper we build upon work by others[4,13–15] and use the concept of treatment effect modification to discuss how generalizability of treatment effects can evaluated, and if generalizability is absent, how to tailor effects to patients with a more homogenous treatment response.

SHOULD TREATMENT EFFECT MODIFICATION BE EXPECTED?

Most clinical studies are not designed to detect treatment effect modification and usually assume homogeneity of treatment effects.[16] Because of this, power to detect interaction effects is generally low, and absence of significant interaction effects should not be seen as proof for the absence of treatment effect modification (a point we will revisit). Despite this expected low power, Poole, Shrier and VanderWeele[17] describe that between 34% and 47% of the meta-analyses reviewed by Engels et al.,[18] Deeks,[19] and Sterne and Egger,[20] rejected homogeneity of treatment effects. This, perhaps unexpected, high percentage of heterogeneity is likely not solely attributable to differences in treatment response but may also be explained by between study differences in dosage, adherence strategies, or co-medication (refer to Sun et al.[13,21] and Rothwell[22] for a more complete discussion).

Given the problems of empirical evidence for treatment effect heterogeneity we additionally refer to theoretical work of Greenland who showed that if both treatment and a potential effect modifier have an effect on the outcome, treatment effect modification must be present on at least one effect measure scale[23] (e.g. RD or RR). Given that most human diseases are complex in nature, multiple factors will be involved in a wide range of endpoints. Combining this with the tendency of more representative studies,[24,25] and therefore more heterogeneous patient samples, we feel that some degree of treatment effect modification should be expected in most studies in which treatment has an effect on an outcome. Whether this effect modification is relevant for clinical practice is a question, which should be explored case by case.

WHICH POTENTIAL EFFECT MODIFIERS TO PRE-SPECIFY

An essential question when designing a study is for whom we want to assess the effects of treatment, whether treatment effects may differ, and if so what defines the subgroups for which treatment effects may differ. To pre-specify potential effect modifier it seems sensible to take account of any prior knowledge of the biological mechanism, potential patient benefit, the frequency certain patients are encountered in practice, and the costs involved in measuring a patient characteristic. When for example comorbidity is a potential effect modifier, it seems reasonable to assess whether relatively common diseases, such as diabetes, modify the effect of treatment. Discussions on the choice of subgroups should focus on patients included but also certainly on patients not included in a (future) study.[26]

Too often, however, discussions on generalizability or the absence of treatment effect modification revolve around the question whether a patient sample is *representative* of the target population or the 'average' patient.[27] Representativeness, however, plays only a minor role in applying treatment effects to individuals.[14,25,28,29] 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, because of unequal subgroup sizes, a representative sample will more often than not preclude detection of treatment effect modification. Hence, representativeness often results in wrongfully assuming homogeneity of treatment effects and thus possibly in patients being treated suboptimally. A more fruitful approach when expecting treatment effect modification is to design a

DOI: 10.1002/pds

study to oversample the pre-specified patient subgroups to ensure sufficient power to detect interaction or its absence.

Even if one is interested in population average treatment effect[30] one should be aware that in the presence of treatment effect modification, small differences between populations can result in markedly different main treatment effects.[31] Assume, for example, that in a population aged 65, the main treatment effect is 1.00 (RR). 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. $e^{\ln(1.00)+\ln(0.95)*(70-65)} \approx 0.77$]. Hence, unless treatment effect modification is minimal, population average treatment effects are not expected to generalize to other settings.

Thus, when discussing generalizability or treatment effect modification, it is essential to define the patient group(s) of interest. Such subgroups should be chosen based on biological plausibility, potential patient benefit, subgroup frequency and measurement costs; this should, however, not be guided by the issue of representativeness.

WHEN ARE TREATMENT EFFECTS GENERALIZABLE?

When interaction effects can be quantified with sufficient precision to exclude clinically relevant treatment effect modification, the main (i.e. average) treatment effect equally applies to all subjects studied and—because there is no direct reason to believe the treatment acts differently in other subjects—this treatment effect is possibly generalizable to, and perhaps beyond, the population included in the study.[30,31] As stated previously non-significant interaction tests are not sufficient to claim generalizability, to quote Altman[32] 'absence of evidence is not evidence of absence'. Instead to 'prove' generalizability, so called equivalence tests should be used.

Recognizing that the strict null-hypothesis (i.e. H_0: $\mu_0 = \text{null}$) probably never holds, tests of equivalence determine margins between which differences in treatment effect estimates are small enough to be deemed clinically irrelevant.[33,34] When the treatment effect estimate and its confidence interval fall between these margins, equivalence is 'proven' (Figure 1). Equivalence tests can be applied to interactions effects by determining a margin around the neutral interaction effect or around the subgroup specific effects, and test if both the point estimates and their confidence intervals fall within this margin. For example, let d be the predefined margin of equivalence, δ_i the effect for the ith subgroup, when $i = \{0, 1\}$ the interaction effects equals $\theta = \delta_0 - \delta_1$, σ_i the subgroup specific

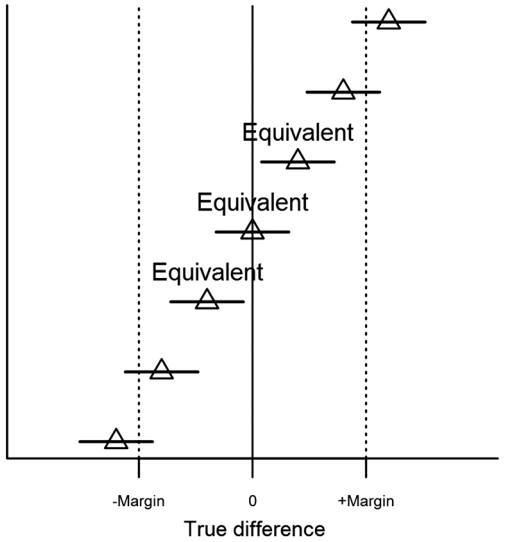

Figure 1. Examples of equivalence testing using confidence intervals. Based on Jones et al. [33]

standard errors of δ_i, and σ_θ the standard error of the interaction effect. In this case an interaction effect is sufficiently absent when $\left(\hat{\delta}_i - z^*\sigma_i, \hat{\delta}_i + z^*\sigma_i\right) \subset (-d, d)$ is true for all subgroup effects or $\left(\hat{\theta} - z^*\sigma_\theta, \hat{\theta} + z^*\sigma_\theta\right) \subset (-d, d)$. Where $z = \varphi^{-1}\left(1 - \frac{\alpha}{2k}\right)$, typically $\alpha = 0.05$ and k equals the number of subgroups or the number of interaction effects and e equals the expected effect under homogeneity; most intuitively this can be set to the estimated main effect of treatment. These procedures test against the following null-hypotheses $H_0: |\delta_i| < d + \text{e}$ and $H_0: |\theta| < d$. Intuitively both procedures should results in the same results, however, because the subgroup specific margins are scaled on the estimated main effect, sampling variation may decrease power in small margin settings (Figure 2). A clear benefit of the interaction equivalence tests over its subgroup specific counter part is that it straightforwardly extends to linear effect modifiers (e.g., age), preventing arbitrary categorizations.

DETECTING TREATMENT EFFECT MODIFICATION

Effect modification can be detected by testing whether the interaction effect differs from zero.[35,36] However, such interaction tests are renowned for their lack of power (i.e. the probability of correctly concluding that an interaction exists), which may be compounded by large type 1 errors (i.e. the probability of falsely concluding that an interaction exists) when the data are sparse.[10,37–43] Note that data sparseness is intuitively

DOI: 10.1002/pds

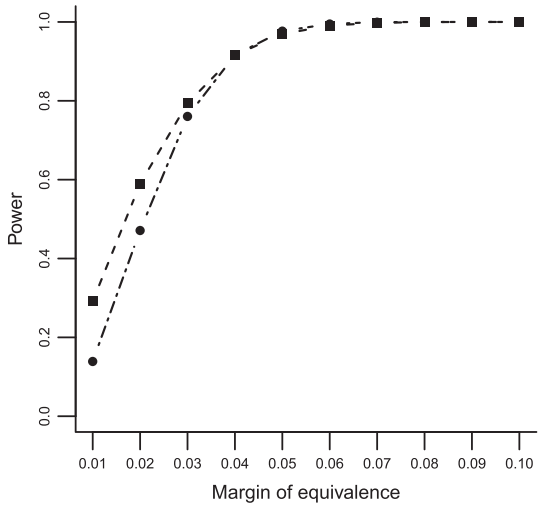

Figure 2. The dashed line with a square symbol indicates power for an equivalence test using an interaction effect, the dashed dotted line with the circle symbol indicates power for the equivalence test based on subgroup specific effects. Simulated results were based on a scenario (10,00 replications) with subjects treated or untreated $j = \{1, 0\}$ and exposed or unexposed to a potential effect modifier $i = \{1, 0\}$, with the endpoint incidence equaling $r_{ij} = \{0.20, 0.15, 0.15, 0.10\}$, $\theta = 0$, and each group of ij subjects occurring 1,000 times.

defined as a small expected cell count but generalizes to continuous data with low densities (or frequencies) at certain values. Often this underperformance of interaction tests is viewed as inevitable; however, this underperformance is merely a result of a lack of a proper design to detect interaction effect, often resulting in sparse data.

For more definitive conclusions on the absence or presence of treatment effect modification, the current approach to interaction testing needs improvement. A first step is to more actively share and pool individual patient data to increase the effective sample size, power, and by decreasing data sparseness, ensuring nominal type 1 error rates for interaction tests.[31,44,45]

Second, for interaction tests to be anything but exploratory, interaction tests should not only be pre-specified but also include proper sample size calculations and sampling strategies (e.g. equally sized subgroups), ensuring appropriate power and type 1 error rates. One attractive idea is to incorporate interaction tests using adaptive trial designs.[46–48] For example, consider an RCT of a particular treatment, conducted within a homogenous group of patients. If 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 pre-specified clinically important patient subgroups. We recognize that this contrasts with the more usual approach of focusing on a single promising subgroup after interim.[46,49] 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.

Third, to increase the interpretability of interaction tests (or any test) we feel that it is essential to a priori define the prior probability of rejecting a test. For example let us assume that data from multiple well-designed studies are available, ensuring sufficient power (let us say 70% or type 2 error rate of 0.30) to reject an interaction test using a statistical significance level (alpha) of 0.05. Suppose two different drug compounds are evaluated, for the first compound we know that for a similar drug 1% of interaction test were true positives, for the second compound this was 25%. In the first case the probability that a rejected interaction test reflects a true positive equals $1 - \frac{\alpha(1-prior)}{\alpha(1-prior)+prior(1-\beta)} = 1 - \frac{0.05(1-0.01)}{0.05(1-0.01)+0.01(1-0.30)} = 0.13$, while for the second compound this equals 0.82. Quantifying a prior probability is of course inherently subjective an issue, which we address later.

HOW TO PERSONALIZE TREATMENT EFFECTS

We suggest that after one identifies important potential effect modifiers (based on the criteria discussed), and quantifying the available prior knowledge, one explores if generalizability can be shown. If generalizability cannot be proven, we propose a thorough multivariable analyses to explore for which patients treatment needs to be modified.

To explore generalizability one first needs to define regions of equivalence as discussed in the preceding texts. After which pre-specified interaction tests can be compared against this region. In itself pre-specification does not significantly increase power to detect interactions unless proper design steps are taken (e.g. oversampling of subgroups).[50] We suggest that regardless of pre-specification or not, these interaction tests are deemed exploratory unless steps are made to quantify the prior evidence, ensure sufficient sample size, power and type 1 error levels.

After determining the amount of within study heterogeneity, and assuming multiple studies exist, between study heterogeneity should be explored, for example by comparing aggregated results from different studies.[44,51] However, attributing differences in treatment effects between studies to differences in baseline characteristics or study design, using for example meta-regression,

DOI: 10.1002/pds

may result in (ecological) bias. Therefore, significant interaction effects found in aggregated meta-analyses should always be confirmed using independent individual patient data.

If, after performing the previously mentioned analyses, absence of effect modification cannot be excluded with confidence, confirmatory analyses are needed, tailoring treatment effects towards groups or individuals. If treatment homogeneity is rejected one may be tempted to treat this as a true positive results. However, as with any discovery, replicating results is essential; hence, results on interaction effects should be independently confirmed. If the results are replicated, it seems sensible to finally combined data from both the confirmatory and exploratory steps to increase precision in the subgroup specific effect estimates of the treatment[52] and use these to tailor treatment (e.g. RR = 0. 75 for diabetes patients versus RR = 0.63 in patients without diabetes).

Recently, subgroup-specific estimates based on a single variable (i.e. univariable interactions) have been criticized.[15,53–56] 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, depending on the number of subgroups (and type, e.g. binary or not), exploring higher order interactions will increase data sparseness, which may dramatically reduce power and increase type 1 error rates.[10,39–42,57–62]

To (partially) 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 a subjects is not treated.[63,64] For example, using a logistic model, the predicted risk equals

$$logit(\hat{p}_i) = logit(Prob[Y = 1|Z]) = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j z_{ij}$$

[Equation (1)], where **Z** equals a *n by k* matrix and **Y** a *n by* 1 column matrix. In the second step, the predicted risk is multiplied by a relative treatment effect estimate (e.g. a RR). [56] Assume, for example, that in our previous trial the multivariable 5-year predicted risk of stroke without treatment equals $\frac{1}{1+e^{-logit(\hat{p}_i)}} = \frac{1}{1+e^{2.20}} = 0.10$, for a particular patient with diabetes. Based on the RR of 0.75, treating this patient will result in a predicted 5 year risk of 0.075 (i.e. $0.75 \times 0.10 = 0.075$) and in an individualized RD of $0.100 - 0.075 = 0.025$; in the general case the RD can be individualized using Equation (2): $RD_i = logit(\hat{p}_i)^{-1} * RR$.

While this multivariable approach to subgroup analysis is indeed an improvement, one should be aware that this approach is only valid if the relative treatment effect measure (e.g. the RR) is homogenous across different levels of the predicted risk. If unknown to the researcher, the relative treatment effect measure is in fact heterogeneous, applying the previously mentioned approach may falsely induce treatment effect modification on the RD scale. We propose that when the previously mentioned approach is applied, this should be combined with sensitivity analysis exploring (1) if the relative effect measure is heterogeneous across the range of predicted risks and (2) to what extend the RD is truly heterogeneous across the predicted risk. Following the risk stratification approach by Kent and Hayward [15] and others we suggest to subdivide the sample based on quantiles of the predicted risk [Equation (1)] and estimate quantile specific treatment effects (e.g. RR and RD)[15] to explore if the treatment effects changes with increasing risk. Using this approach one can judge if the relative effect measure is fairly homogenous across the predicted risk and if individualized treatment effects on the RD scale [Equation (2)] agree with the quantile specific treatment effects on the RD scale. As with any testing procedure one should be careful not to over interpret non-significant interaction tests results, because, as addressed before, this does not imply homogeneity. To increase power, and only if quantile specific treatment estimates linearly change, one could use the predicted risk as a linear term in a statistical model and include a treatment by predicted risk interaction term.[65] Depending on the amount of linearity such a model is expected to be more precise and powerful than the quantile specific approach. A remaining issue with the previously described approaches is that, typically, the predicted risk is treated as if it was observed without error, possibly erroneously decreasing the standard error of any test.[66] A second more general comment is that all the discussed multivariable approaches only allow for individualized treatment effect estimates in so far as variables are related to the outcome. A strategy to include variables unrelated with the outcome in a multivariable interaction test is to use unsupervised cluster analysis to identify multivariable patient clusters, and test if treatment effectiveness differs across cluster memberships.[67–69]

QUANTIFYING THE PRIOR PROBABILITY OF TREATMENT EFFECT MODIFICATION

Throughout the previous sections we emphasized the need to quantify the prior probability for the presence

DOI: 10.1002/pds

of an interaction. Here we detail what to base this prior probability on.

As stated previously, RCTs are the gold standard in intervention research. Despite this, we feel strongly against a priori deciding to quantifying the prior probability solely on RCT results. RCTs are not initiated at random. Instead, RCTs are initiated based on information from basic experiments, genetic studies, nonrandomized studies and/or previous RCTs; therefore, to properly quantify the prior knowledge these sources should all be considered. Depending, however, on the potential risk of bias, taking account of the endpoints of interest, and the general potential risk of an intervention, these multiple source of prior knowledge should be reweighted. In some cases, for example, when exploring the intended effect of statins on a myocardial infraction, one may choose to weight non-RCT data by zero. This reweighing or elimination of data should obviously be clearly presented and justified. We appreciate that this introduces a certain amount of subjectivity in analyses that may seem otherwise objective. However, this is no different than excluding RCTs at perceived high risk of bias from a meta-analysis, a thing which is customarily (although not without discussion) performed in, for example, Cochrane Reviews.

SUMMARY

In the present commentary we have argued that detecting treatment effect modification is essential to bridge the gap between results from clinical studies and treating individuals in daily practice. We addressed strategies to detect effect modification and used these in a framework to assess if there is a need for more individualized treatment effects, and estimate this in confirmatory analyses.

We conclude with the following recommendations. First, treatment effect modification should be formally assessed using interaction tests. Second, pre-specified subgroups should be selected based on biological plausibility, prevalence of the patient type and cost-effectiveness of determining the patient characteristic. Third, before tailoring treatment effects to patient subgroups one should first consider if generalizability or the absence of treatment effect modification can be proven, using, e.g. an equivalence test. Fourth, for interaction tests to be anything but exploratory, these should not only be pre-specified, but include a quantification of the prior knowledge, use proper sample size calculations and sampling strategies to ensure appropriate levels of power and type 1 error rates (taking account of possible multiple testing). Finally, if after careful consideration and sufficient replication, subgroup effects are found to be consistent across different studies, this should have an impact in daily clinical practice. What is sufficient evidence, however, should be determined on a case by case basis and depends, amongst other things, on the disease, intervention related risks and the magnitude of interaction.

CONFLICT OF INTEREST

None of the authors of this paper have a financial or personal relationship with other people or organizations that could inappropriately influence or bias the content of the paper.

KEY POINTS

- Clinical studies are designed to provide evidence on average treatment effects.
- To tailor treatment towards individual patients, the presence or absence of treatment effect modification needs to be systematically elucidated.
- Generalizability of treatment effects can be tested within the framework of equivalence testing.
- The type of patient, the presence, the magnitude and the number of effect modifiers determine whether no further analyses, univariable subgroup analyses or multivariable subgroup analyses may need to be performed.

ETHICS STATEMENT

This study did not include data from human, animal or plant subjects. Therefore no ethical approval was sought.

ACKNOWLEDGEMENTS

This work was supported by Research Focus Areas funding of the Utrecht University, which is a collaboration between the faculties of medicine, science and veterinary medicine. The funding body had no role in decisions on the design, writing or submission of the manuscript. We want to acknowledge and thank the two anonymous reviewers for their helpful suggestions which markedly improved the manuscript.

AUTHOR CONTRIBUTIONS

AFS drafted the manuscript. OHK, MN, AB, AWH and RHHG provided guidance during initial planning of the paper and during critical revision.

DOI: 10.1002/pds

REFERENCES

1. Vandenbroucke JP. When are observational studies as credible as randomised trials? *Lancet* 2004; **363**: 1728–1731.
2. Vandenbroucke JP. What is the best evidence for determining harms of medical treatment? *CMAJ* 2006; **174**: 645–646.
3. Grobbee DE, Hoes AW. Intervention Research: Unintended Effects, in Clinical Epidemiology: Principles, Methods and Applications for Clinical Research, chap 6. Burlington: Jones and Bartlett Learning, 2015; 181–214.
4. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet* 2005; **365**: 82–93.
5. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005; **365**: 176–186.
6. Rothwell PM, Mehta Z, Howard SC, *et al*. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. *Lancet* 2005; **365**: 256–265.
7. Bugeja G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. *BMJ* 1997; **315**: 1059.
8. VanderWeele TJ. On the distinction between interaction and effect modification. *Epidemiology* 2009; **20**: 863–871.
9. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *Am J Epidemiol* 1980; **112**: 467–470.
10. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. *Stat Med* 1983; **2**: 243–251.
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; **50**: 809–812.
13. Sun X, Briel M, Busse JW, *et al*. The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review. *BMJ* 2011; **342**: d1569.
14. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol* 2013; **42**: 1012–1014.
15. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA* 2007; **298**: 1209–1212.
16. Hernan MA. A definition of causal effect for epidemiological research. *J Epidemiol Community Health* 2004; **58**: 265–271.
17. Poole C, Shrier IF, VanderWeele TJ: Is the risk difference really a more heterogeneous measure? *Epidemiology* 2015; **26**(5): 714–718.
18. Engels EA, Schmid CH, Terrin N, *et al*. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med* 2000; **19**: 1707–1728.
19. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med* 2002; **21**: 1575–1600.
20. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol* 2001; **54**: 1046–1055.
21. Sun X, Briel M, Walter SD, *et al*. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ* 2010; **340**: c117.
22. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005; **365**: 176–186.
23. Greenland S, Lash TL, Rothman K: Concepts of interaction, in Rothman K, Greenland S, Lash TL (eds): Modern Epidemiology 3rd ed. Philadelpha: Lippincott Williams and Wilkins, 2008; 71–86.
24. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. *J Clin Epidemiol* 2009; **62**: 499–505.
25. Schmidt AF, Groenwold RH, van Delden JJ, *et al*. Justification of exclusion criteria was underreported in a review of cardiovascular trials. *J Clin Epidemiol* 2014; **67**(6): 635–644.
26. van der Graaf R, Groenwold RHH, Kalkman S, *et al*. From justifying inclusion to justifying exclusion of study populations: strengths and limitations. *World Med J* 2013; **59**: 192–197.
27. Dekkers OM, von Elm E, Algra A, *et al*. How to assess the external validity of therapeutic trials: a conceptual approach. *Int J Epidemiol* 2010; **39**: 89–94.
28. Rothman KJ, Gallacher JE, Hatch EE. Rebuttal: when it comes to scientific inference, sometimes a cigar is just a cigar. *Int J Epidemiol* 2013; **42**: 1026–1028.
29. Rothman KJ. Six persistent research misconceptions. *J Gen Intern Med* 2014; **29**(7): 1060–1064.
30. Pressler TR, Kaizar EE. The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. *Stat Med* 2013; **32**(20): 3552–3568.
31. 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; **33**: 536–537.
32. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995; **311**: 485.
33. Jones B, Jarvis P, Lewis JA, *et al*. Trials to assess equivalence: the importance of rigorous methods. *BMJ* 1996; **313**: 36–39.
34. Fleming TR. Design and interpretation of equivalence trials. *Am Heart J* 2000; **139**: S171–S176.
35. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *BMJ* 2003; **326**: 219.
36. Matthews JN, Altman DG. Statistics notes. Interaction 2: Compare effect sizes not P values. *BMJ* 1996; **313**: 808.
37. 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.
38. Lui KJ. Testing homogeneity of the risk ratio in stratified noncompliance randomized trials. *Contemp Clin Trials* 2007; **28**: 614–625.
39. Lui KJ. A simple test of the homogeneity of risk difference in sparse data: an application to a multicenter study. *Biom J* 2005; **47**: 654–661.
40. O'Gorman TW, Woolson RF, Jones MP, *et al*. Statistical analysis of K 2 × 2 tables: a comparative study of estimators/test statistics for association and homogeneity. *Environ Health Perspect* 1990; **87**: 103–107.
41. Paul SR, Donner A. Small sample performance of tests of homogeneity of odds ratios in K 2 x 2 tables. *Stat Med* 1992; **11**: 159–165.
42. Zhang L, Yang H, Cho I. Test homogeneity of risk difference across subgroups in clinical trials. *J Biopharm Stat* 2009; **19**(1): 67–76.
43. Schmidt AF, Groenwold RH, Knol MJ, *et al*. Exploring interaction effects in small samples increases rates of false-positive and false-negative findings: results from a systematic review and simulation study. *J Clin Epidemiol* 2014; **67**: 821–829.
44. Schmidt AF, Rovers MM, Klungel OH, *et al*. Differences in interaction and subgroup-specific effects were observed between randomized and nonrandomized studies in three empirical examples. *J Clin Epidemiol* 2013; **66**: 599–607.
45. Koopman L, van der Heijden GJ, Hoes AW, *et al*. Empirical comparison of subgroup effects in conventional and individual patient data meta-analyses. *Int J Technol Assess Health Care* 2008; **24**: 358–361.
46. Boessen R, van der Baan F, Groenwold R, *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; **12**: 366–374.
47. Bretz F, Koenig F, Brannath W, *et al*. Adaptive designs for confirmatory clinical trials. *Stat Med* 2009; **28**: 1181–1217.
48. van der Baan FH, Knol MJ, Klungel OH, *et al*. Potential of adaptive clinical trial designs in pharmacogenetic research. *Pharmacogenomics* 2012; **13**: 571–578.
49. Tanniou J, Tweel v T, Teerenstra S, *et al*. Level of evidence for promising subgroup findings in an overall non-significant trial. *Stat Methods Med Res* 2014.
50. Peterson B, George SL. Sample size requirements and length of study for testing interaction in a 2 × k factorial design when time-to-failure is the outcome [corrected]. *Control Clin Trials* 1993; **14**: 511–522.
51. Rovers MM, Straatman H, Ingels K, *et al*. Generalizability of trial results based on randomized versus nonrandomized allocation of OME infants to ventilation tubes or watchful waiting. *J Clin Epidemiol* 2001; **54**: 789–794.
52. Bowden J, Dudbridge F. Unbiased estimation of odds ratios: combining genomewide association scans with replication studies. *Genet Epidemiol* 2009; **33**: 406–418.
53. 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; **34**: 464–467.
54. Hayward RA, Kent DM, Vijan S, *et al*. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Med Res Methodol* 2006; **6**: 18.
55. Kent DM, Rothwell PM, Ioannidis JP, *et al*. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials* 2010; **11**: 85.
56. Dorresteijn JA, Visseren FL, Ridker PM, *et al*. Estimating treatment effects for individual patients based on the results of randomised clinical trials. *BMJ* 2011; **343**: d5888.
57. Jones MP, O'Gorman TW, Lemke JH, *et al*. A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. *Biometrics* 1989; **45**: 171–181.
58. Liang KY, Self SG. Tests for homogeneity of odds ratio when the data are Sparse 3. *Biometrika* 1985; **72**: 353–358.
59. Lipsitz SR, Dear KB, Laird NM, *et al*. Tests for homogeneity of the risk difference when data are sparse. *Biometrics* 1998; **54**: 148–160.
60. Lui KJ, Kelly C. Tests for homogeneity of the risk ratio in a series of 2x2 tables. *Stat Med* 2000; **19**: 2919–2932.
61. Lui KJ, Chang KC. Test homogeneity of odds ratio in a randomized clinical trial with noncompliance. *J Biopharm Stat* 2009; **19**: 916–932.
62. Reis IM, Hirji KF, Afifi AA. Exact and asymptotic tests for homogeneity in several 2 × 2 tables. *Stat Med* 1999; **18**: 893–906.
63. Moons KG, Kengne AP, Woodward M, *et al*. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart* 2012; **98**: 683–690.

64. Schmidt AF, Nielen M, Klungel OH, *et al*. Prognostic factors of early metastasis and mortality in dogs with appendicular osteosarcoma after receiving surgery: an individual patient data meta-analysis. *Prev Vet Med* 2013; **112**: 414–422.
65. Farooq V, van Klaveren D, Steyerberg EW, *et al*. Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score II. *Lancet* 2013; **381**: 639–650.
66. Schmidt AF, Groenwold RHH, Amsellem P, *et al*. Which dogs with appendicular osteosarcoma benefit most from chemotherapy after surgery? Results from an individual patient data meta-analysis. *Prev Vet Med* 2016. doi: 10.1016/j.prevetmed.2015.10.016.
67. van Giessen A, Moons KG, de Wit GA, *et al*. Tailoring the implementation of new biomarkers based on their added predictive value in subgroups of individuals. *PLoS One* 2015; **10**: e0114020.
68. van Giessen A, de Wit GA, Smit HA, *et al*. Patient selection for cardiac surgery: time to consider subgroups within risk categories? *Int J Cardiol* 2015; **203**: 1103–1108.
69. Everitt B, Hothorn T: Cluster Analysis, in Gentleman R, Hornik K, Parmigiani G (eds): An Introduction to Applied Multivariate Analysis with R, chap 6. New York, Springer, 2011, pp 163–200