REVIEW

Understanding of interaction (subgroup) analysis in clinical trials

Milos Brankovic[1,2] (iD) | Isabella Kardys[1] | Ewout W. Steyerberg[3] (iD) |
Stanley Lemeshow[4] | Maja Markovic[5] | Dimitris Rizopoulos[6] | Eric Boersma[1]

[1]Clinical Epidemiology Unit, Department of Cardiology, Erasmus Medical Center, Rotterdam, The Netherlands

[2]School of Medicine, University of Belgrade, Belgrade, Serbia

[3]Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands

[4]Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, Ohio

[5]Department of Child and Adolescent Psychiatry, Erasmus Medical Center, Rotterdam, The Netherlands

[6]Department of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands

Correspondence
Eric Boersma, Erasmus MC, Erasmus University Rotterdam, office: Ba-042 PO Box 2040, 3000 CA Rotterdam, The Netherlands.
Email: h.boersma@erasmusmc.nl

Abstract

Background: When the treatment effect on the outcome of interest is influenced by a baseline/demographic factor, investigators say that an interaction is present. In randomized clinical trials (RCTs), this type of analysis is typically referred to as subgroup analysis. Although interaction (or subgroup) analyses are usually stated as a secondary study objective, it is not uncommon that these results lead to changes in treatment protocols or even modify public health policies. Nonetheless, recent reviews have indicated that their proper assessment, interpretation and reporting remain challenging.

Results: Therefore, this article provides an overview of these challenges, to help investigators find the best strategy for application of interaction analyses on binary outcomes in RCTs. Specifically, we discuss the key points of formal interaction testing, including the estimation of both additive and multiplicative interaction effects. We also provide recommendations that, if adhered to, could increase the clarity and the completeness of reports of RCTs.

Conclusion: Altogether, this article provides a brief non-statistical guide for clinical investigators on how to perform, interpret and report interaction (subgroup) analyses in RCTs.

KEYWORDS
effect modification, heterogeneity, interaction, randomized clinical trial, stratification, subgroup analysis, trial

1 | INTRODUCTION

When the treatment effect on the outcome of interest differs according to the presence (or absence) of a baseline/demographic factor, investigators say that a (statistical) interaction is present. In randomized clinical trials (RCTs), statistical analysis of such a phenomenon is typically referred to as a subgroup analysis. The reason that motivates interaction (or subgroup) analysis is to learn how to use the treatment most effectively by identifying subgroups of patients who would and those who would not benefit from treatment, or to learn whether treatment would be harmful in specific subgroups defined by the baseline/demographic factor.[1] Although interaction analysis in RCTs is usually stated as the secondary study objective, if incorrectly tested or misinterpreted, it may lead to unnecessary withholding of treatment, ineffective or even harmful treatment effects.[2]

Although the concept of statistical interaction is not new, it still poses problems for clinical investigators. In 2000, Assmann et al[3] reviewed 50 RCTs in high-impact journals and found that 70% of trials tested interactions, but only 43% of the studies testing interaction reported the test they used,

and 37% of them reported *P*-values only. In 2006, Hernandez et al[4] reported similar results after investigating published cardiovascular RCTs. In 2007, Wang et al[1] evaluated 97 RCTs of which 61% tested interactions, but in 68% of the studies testing interaction, it was unclear whether analyses were prespecified or post hoc and only 27% of them reported formal testing. In 2017, Wallach et al[5] demonstrated that 61% of RCTs that claimed subgroup heterogeneity already in their abstracts (assuming these were most credible) were not supported by their results. Therefore, previous reports have tried to address this important topic including the issue of multiple testing and the importance of prespecifying the subgroup-treatment interaction.[1-3,6,7] These reviews were informative but did not consider certain statistical aspects which are important for analysis and interpretation of the results. To date, a few reports[8,9] have addressed some of these aspects but they were mainly intended for an epidemiological audience.

This article provides an overview of the key aspects of interaction testing to assist clinical investigators to appropriately apply statistical interaction analyses for binary outcomes and categorical covariates. In the following sections, we start by explaining how to analyse an interaction, then describe how to interpret, and finally report the results.

2 | ASSESSMENT OF STATISTICAL INTERACTION

A statistical interaction can be assessed in two ways: by *stratification*—when treatment effects are assessed across subgroups defined by a baseline/demographic factor; or by *interaction modelling*—when the treatment and the baseline/demographic factor are included together with an interaction term into a statistical model (treatment + baseline factor + treatment × baseline factor).[10]

Of note is that an interaction does not have a consistent meaning across statistical models. This is because different models estimate different effect measures (eg risk difference [RD], risk ratio [RR], odds ratio [OR], hazard ratio [HR]). Consequently, some statistical models are constructed as linear models (eg a linear regression model) and others as exponential models (eg logistic and Cox regression models). In a linear regression model, the β coefficient for an interaction term estimates a deviation from the *sum* of treatment subgroup effects. This implies that a linear regression model utilizes an additive scale for interaction testing. In logistic and Cox regression models, a ratio for an interaction term estimates a deviation from the *product* of treatment subgroup effects. This implies that these exponential models utilize a multiplicative scale for interaction testing.

Importantly, whether an interaction is present or in which direction it operates will depend on which of these two scales it is tested. Consider the following hypothetical example: a study finds that in women, 1% of participants receiving treatment and 3% of those receiving placebo reached the outcome, and in men, 2% of participants receiving treatment and 4% of those receiving placebo reached the outcome (Figure 1.1). The risk difference (RD) between the placebo and treatment arm is 2% (3%-1%) in women, and 2% (4%-2%) in men, suggesting no additive interaction between treatment and sex. The study also finds that the RR between the placebo and treatment arms in women is 3 ($\frac{3\%}{1\%}$), and in men is 2 ($\frac{4\%}{2\%}$), suggesting a multiplicative interaction between treatment and sex. Figure 1 illustrates that this situation, where additive and multiplicative interaction effects do not match, is not just a theoretical possibility, but even common, when analysing statistical interactions (Figure 1.1-4, 7, 8).

In RCTs, many statistical analyses are based on logistic and Cox regression models (ie binary outcomes are often analysed) which utilize the multiplicative scale.[11] Hence, these analyses will only test multiplicative, rather than additive interaction effects. At the same time, from the public health perspective, additive effects are favourable over multiplicative effects to increase the net benefit by allocating the treatment to the proper subgroup.[12,13] In addition, some authors have argued that showing an additive effect of a treatment across subgroups may also provide stronger evidence for an underlying biological interaction.[12,14] Therefore, it is reasonable that investigators assess the additive, apart from the multiplicative, interaction effects. Moreover, the confidence intervals (CIs) for both interaction effects should be calculated to assess the statistical strength for such inferences.

2.1 | Multiplicative interaction effect

For binary outcomes, logistic or Cox regression models can be applied to test for multiplicative interaction between treatment and a baseline/demographic factor (Table 1). From the model's output, a ratio with 95% CI for an interaction term indicates the magnitude of the interaction and the *P*-value indicates the significance level.

FIGURE 1 Statistical interactions on additive and multiplicative scales. "*a*" denotes the effect in the placebo arm in the subgroup where the baseline factor equals zero; "*b*" denotes the effect in the treatment arm in the subgroup where the baseline factor equals zero; "*c*" denotes the effect in the placebo arm in the subgroup where the baseline factor equals 1; "*d*" the effect in the treatment arm in the subgroup where the baseline factor equals 1. RD denotes risk difference, whereas RR denotes risk ratio within subgroups. *Y*–axes display numerical values (rates of outcome per 1000 patients) which can be used for calculation of RD_0, RD_1, RR_0, and RR_1 ($RD_0 = a-b$; $RD_1 = c-d$; $RR_0 = a/b$; $RR_1 = c/d$). Eight potential scenarios can be observed when a deviation exists from the sum of treatment subgroup effects (additive scale) or their product (multiplicative scale)

PURE MULTIPLICATIVE INTERACTION

1. SUB-MULTIPLICATIVE

$RD_0 = RD_1$

$RR_0 > RR_1$

2. SUPER-MULTIPLICATIVE

$RD_0 = RD_1$

$RR_0 < RR_1$

PURE ADDITIVE INTERACTION

3. SUB-ADDITIVE

$RD_0 > RD_1$

$RR_0 = RR_1$

4. SUPER-ADDITIVE

$RD_0 < RD_1$

$RR_0 = RR_1$

MULTIPLICATIVE & ADDITIVE INTERACTION

5. SUPER-MULTIPLICATIVE & SUPER-ADDITIVE

$RD_0 < RD_1$

$RR_0 < RR_1$

$RD_0 < RD_1$

$RR_0 < RR_1$

6. SUB-MULTIPLICATIVE & SUB-ADDITIVE

$RD_0 > RD_1$

$RR_0 > RR_1$

$RD_0 > RD_1$

$RR_0 > RR_1$

7. SUB-ADDITIVE & SUPER-MULTIPLICATIVE

$RD_0 > RD_1$
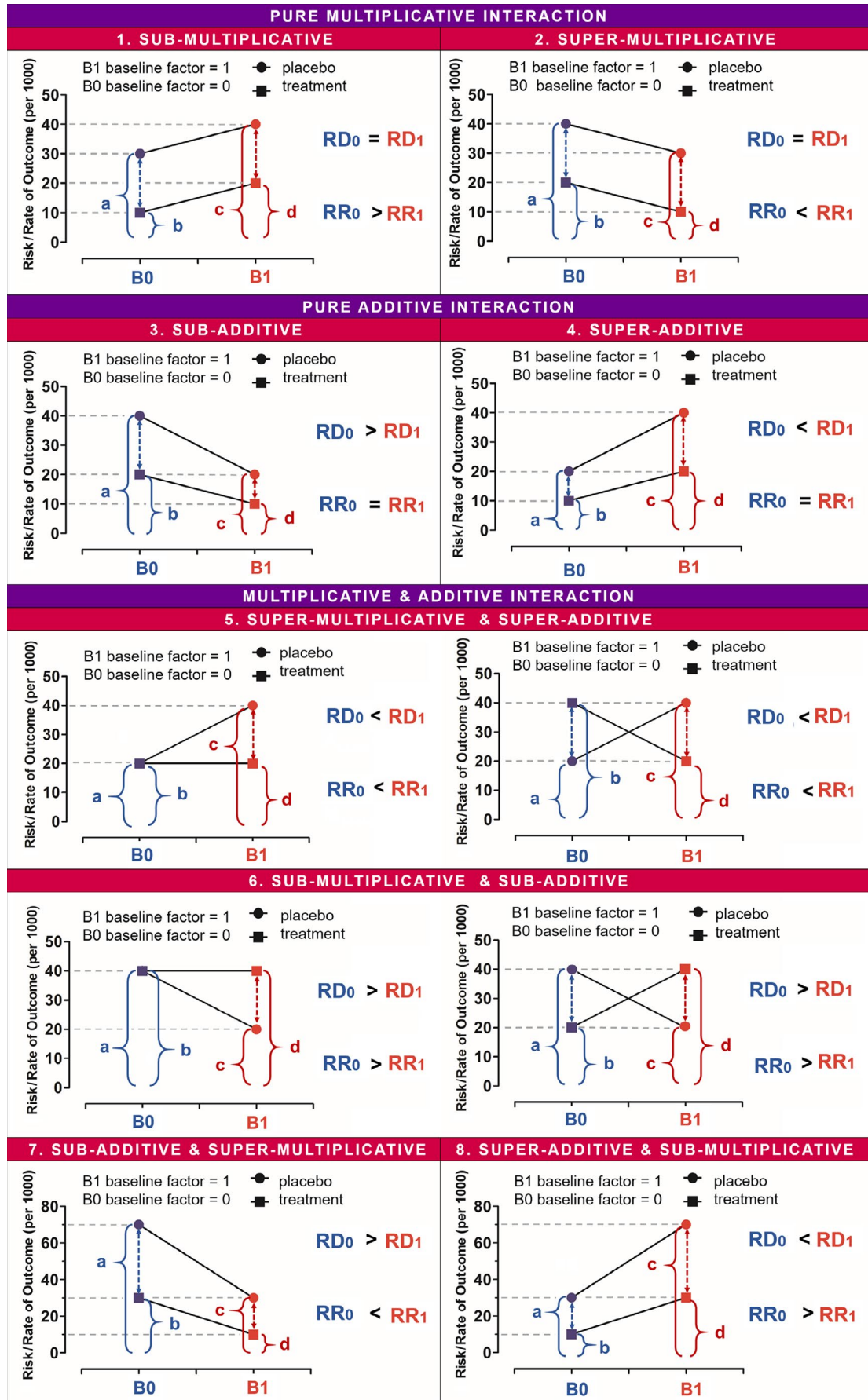
$RR_0 < RR_1$

8. SUPER-ADDITIVE & SUB-MULTIPLICATIVE

$RD_0 < RD_1$

$RR_0 > RR_1$

TABLE 1 Multiplicative interaction effects

Relative risk ratio due to interaction (stratification)	Eq.
Formula (RR, OR, HR):	
$\dfrac{RR_{T+,B+}}{RR_{T+,B-} \times RR_{T-,B+}}$	(1)
Description:	
T, treatment;	
B, baseline factor;	
$\dfrac{RR_{T+,B+}}{RR_{T+,B-} \times RR_{T-,B+}}$ equals to the ratio for the interaction term in the regression model	
Logistic regression model (interaction modelling)	
Formula:	
$\mathrm{Ln}\left[\dfrac{\mathrm{Pr}_{Y=1}}{(1-\mathrm{Pr}_{Y=1})}\right] = \beta 0 + \beta 1\,(T) + \beta 2\,(B) + \beta 3\,(T \times B)$ (exponentiation of both sides of the equation will eliminate the logarithm)	
$\dfrac{\mathrm{Pr}_{Y=1}}{(1-\mathrm{Pr}_{Y=1})} = e^{\beta 0} \times e^{\beta 1(T)} \times e^{\beta 2(B)} \times e^{\beta 3(T \times B)}$ (this can also be rewritten as)	
$\mathrm{Odds} = O_0 \times \mathrm{OR}_T \times \mathrm{OR}_B \times \mathrm{OR}_{T \times B}$	(2)
Description:	
$\mathrm{Pr}_{Y=1}$, probability of outcome $Y = 1$ (eg a patient dies)	
O_0, odds of outcome $Y = 1$ in the subgroup receiving placebo without the effect of the baseline factor ($T-,B-$); this is a background risk because it is not defined by treatment or baseline factor	
OR_T, odds ratio between the subgroup receiving treatment without the effect of the baseline factor ($T+,B-$) and the subgroup in which both treatment and baseline factor are absent ($T-,B-$)	
OR_B, odds ratio between the subgroup receiving placebo with the effect of the baseline factor ($T-,B+$) and the subgroup in which both treatment and the baseline factor are absent	
$\mathrm{OR}_T \times \mathrm{OR}_B \times \mathrm{OR}_{T \times B}$, odds ratio between the subgroup receiving treatment with the effect of the baseline factor ($T+,B+$) and the subgroup in which both treatment and baseline factor are absent	
$\mathrm{OR}_{T \times B}$, odds ratio for the interaction term quantifies the multiplicative interaction effect	
Cox regression model (interaction modelling)	
Formula:	
$\mathrm{Ln}\,[H\,(t)] = \beta 0 + \beta 1\,(T) + \beta 2\,(B) + \beta 3\,(T \times B)$	
$H\,(t) = e^{\beta 0} \times e^{\beta 1(T)} \times e^{\beta 2(B)} \times e^{\beta 3(T \times B)}$ (this can also be rewritten as)	
$H\,(t) = H_0\,(t) \times \mathrm{HR}_T \times \mathrm{HR}_B \times \mathrm{HR}_{T \times B}$	(3)
Description:	
$\mathrm{HR}_{T \times B}$, hazard ratio for the interaction term quantifies multiplicative interaction effect	

$\beta 0, \beta 1, \beta 2, \beta 3$, coefficients in a regression model.

The benefits of interaction testing in regression models include the following: multivariable adjustment, testing interactions between >2 factors and continuous factors, and testing treatment effects across subgroups defined by a risk model. Note that when testing an interaction with a continuous factor, the amount of change of the interaction coefficient will depend on the chosen unit (or unit interval) of the continuous factor (for details see Knol et al[15]). For continuous factors, a non-linear interaction should also be considered because the interaction may not be uniform across the entire range of the continuous factor. In such cases, the choice could be to categorize the continuous factor.

2.2 | Additive interaction effect

For binary outcomes, the additive interaction can be expressed as the *absolute excess risk due to interaction* (AERI). The AERI can only be calculated if absolute risks are known, and under the assumption that the risks are unbiased (ie without confounding). An AERI >0 will indicate super-additive interaction (ie joint effect is higher than the sum of individual effects), whereas AERI <0 will indicate sub-additive interaction (ie joint effect is lower than the sum of individual effects). Of note, to further define a direction of an interaction as super- or sub-, one needs to

specify the exact subgroups on which this particular notation is based.

Consider the study by Head et al,[16] who investigated the effects of primary coronary intervention (PCI) and coronary artery bypass grafting (CABG) on 5-year mortality among patients with complex coronary artery disease (CAD) using pooled data from eleven RCTs. The investigators found a significantly higher 5-year mortality in patients treated with PCI compared to those treated with CABG only in the subgroup of diabetic patients. They applied a Cox regression model implying that the interaction was analysed on the multiplicative scale. To examine whether this interaction also exists on the additive scale, we can calculate AERI using the numbers provided in their Table 2.[16] In their study, the 5-year mortality risk was 15.7% in patients with diabetes treated with PCI, 8.4% in patients without diabetes treated with CABG, 10.7% in patients with diabetes treated with CABG, and 8.7% in patients without diabetes treated with PCI. We calculate AERI using equation 4 from Table 2 as $15.7 + 8.4 - 10.7 - 8.7 = 4.7\%$, suggesting a super-additive interaction between diabetes and PCI. Assuming this AERI of 4.7% is unbiased, then the direction

alone (AERI >0), rather than its magnitude (4.7%), is important to answer the question whether diabetic patients should be treated with CABG over PCI. Yet, in certain situations it may also be relevant to consider the magnitude of the interaction itself, which will be discussed later.

When absolute risks are not reported, when treatment effects are derived from multivariable models (ie treatment effects adjusted for other covariates), or when interaction between treatment and a continuous factor is considered, additive interaction can be assessed using relative excess risk due to interaction (RERI) and synergy index.[17-19] Because these indices operate with ratios (derived from logistic or Cox regression models) instead of absolute risks, they can only be used to assess the direction, and not the magnitude, of additive interaction for absolute risks, as AERI can. Moreover, since ratios are asymmetrically distributed (ie preventive effects range from 0 to 1 and hazardous effects range from 1 to ∞), the subgroup effects should be recoded before calculation. Otherwise, these indices can differ if preventive and hazardous effects are combined in the equation. The easiest way to recode the effects is to use the subgroup with the lowest risk as the reference when

TABLE 2 Additive interaction effects

Absolute excess risk due to interaction (AERI)	Eq. n.
Formula (absolute risks):	
$\text{AERI} = R_{T+,B+} + R_{T-,B-} - R_{T+,B-} - R_{T-,B+}$	(4)
Description:	
T, treatment	
B, baseline factor	
$R_{T+,B+}$, risk in the subgroup receiving treatment with the effect of the baseline factor	
$R_{T-,B-}$, risk in the subgroup receiving placebo without the effect of the baseline factor	
$R_{T+,B-}$, risk in the subgroup receiving treatment without the effect of the baseline factor	
$R_{T-,B+}$, risk in the subgroup receiving placebo with the effect of the baseline factor	
Relative excess risk due to interaction (RERI)	
Formula (RR, OR, HR):	
$\text{RERI} = RR_{T+,B+} - RR_{T+,B-} - RR_{T-,B+} + 1$ (stratification)	(5)
$\text{RERI} = e^{\beta1+\beta2+\beta3} - e^{\beta1(T)} - e^{\beta1(B)} + 1$ (interaction modelling)	(6)
(this can also be rewritten as)	
$\text{RERI} = OR_T \times OR_B \times OR_{T \times B} - OR_T - OR_B + 1$	(7)
Description:	
Note that $OR_{T+,B+}$ is not provided in the regression model's output using the interaction term	
$OR_T \times OR_B \times OR_{T \times B}$ equals to $OR_{T+,B+}$	
Attributable proportion of joint effect due to interaction (modified AP)	
Formula (absolute risks)	
$\text{modified AP} = \frac{\text{AERI}}{R_{T+,B+} - R_{T-,B-}}$	(8)
Formula (RR, OR, HR):	
$\text{modified AP} = \frac{\text{RERI}}{RR_{T+,B+} - 1}$	(9)

$\beta0$, $\beta1$, $\beta2$, $\beta3$, coefficients in a regression model.

the treatment and the baseline factor are jointly considered (note that some statistical packages perform recoding automatically,[20] to do this manually see Knol et al[21]).

2.2.1 | Relative excess risk due to interaction (RERI)

The RERI (synonym: interaction contrast ratio [ICR]) is the difference between the joint effect of treatment and a demographic/baseline factor and their effects considered individually (Table 2).[13] The RERI ranges from $-\infty$ to $+\infty$ and can indicate super-additive (RERI >0) or sub-additive (RERI <0) interaction effects.[21] The 95% CI for RERI can be calculated using the delta method[22] or using the first percentile Bootstrap method.[23] The latter is more suitable for continuous factors.[15] If additional covariates are included into the model, RERI may vary across levels of those covariates.[24] The codes for calculating RERI with 95% CI are available for SAS,[17,25] STATA,[18] R,[20,26] and using excel sheets.[8,15]

Consider another RCT by Andrews et al,[27] who found that treatment based on a sepsis protocol has paradoxically increased in-hospital mortality compared to usual care in septic patients with hypotension. They reported that this was only the case in the subgroup of patients with normal Glasgow coma score (GCS ≥13) at baseline. From their Figure 3, we re-calculated RR as 3.55 in patients with GCS <13 treated using the sepsis protocol, 3.09 in patients with GCS <13 receiving usual care, and 1.91 in patients with GCS ≥13 treated using the sepsis protocol, as compared to the subgroup of patients with GCS ≥13 receiving usual care. To illustrate how additivity can be assessed using ratio measures, we calculated RERI using equation 5 from Table 2 as 3.55−1.91−3.09 + 1 = −0.45. The RERI suggested a sub-additive interaction (RERI <0) between the sepsis-protocol treatment and the lower GCS score. This can be explained by the fact that patients with lower GSC score at baseline had a poorer health condition than those who did not (ie the GSC score was a proxy for patient health condition), which on its part altered the effect of the sepsis-protocol treatment on patient outcome. Note that the direction, and not the magnitude, of RERI is relevant for drawing this conclusion.

2.2.2 | Attributable proportion of joint effect due to interaction

As noted above, in certain situations it is relevant to consider the magnitude of the interaction, that is to what extent the treatment effect is changed due to a certain baseline factor. The motivation behind this is to test the robustness of the interaction by assessing its magnitude and limits of confidence interval. Another motivation can be that investigators may consider a future intervention on that baseline factor to improve the treatment effect. Alternatively, if intervening on the primary exposure is impossible, investigators can try to target other factors that interact with the primary exposure to eliminate most of its effects. For this purpose, investigators could assess attributing proportions of interaction effect to identify the most relevant baseline factors. Further reading on this topic is provided elsewhere.[28]

Attributable proportion of joint effect due to interaction, called here modified AP, indicates the proportion of the joint effect of the treatment and a baseline/demographic factor that is due to the interaction itself (Table 2).[29] It ranges from (−)100% to (+)100% and indicates super-additive (modified AP >0) or sub-additive (modified AP <0) interaction effects. It can be calculated using either absolute risks or ratios (Table 2). It is independent of covariate adjustment.[29] The codes for calculating modified AP with 95% CI are available in SAS,[28] STATA[28] and R.[20,26] In the study by Head et al, modified AP can be calculated using equation 8 from Table 2 as $\frac{4.7}{15.7-8.4} = 0.64$ suggesting a super-additive interaction. It also indicates that 64% of the joint effect is due to the interaction itself between diabetes and PCI (the rest of 36% is the sum of the proportions of their effects considered individually).

3 | CLINICAL INTERPRETATION AND REPORTING

In previous sections, we explained that a presence, and even direction, of the interaction can change with the choice of the statistical model. We also discussed arguments for preferring additive over multiplicative interaction effects for binary outcomes. The following section discusses the interpretation of interaction analyses in RCTs accompanied by relevant recommendations (Table 3).

Statistical interaction between the treatment and a baseline/demographic factor can be interpreted as *effect-measure modification* or as *causal interaction*. When treatment effects vary across the subgroups of baseline/demographic factor, this can be interpreted as effect-measure modification.[30] For effect-measure modification, this baseline/demographic factor does not need to affect the outcome directly, but only needs to correlate with another factor that does.[18] As a consequence, investigators cannot attribute treatment subgroup effects to the baseline factor itself. Therefore, some authors refer to it simply as effect heterogeneity.[13,31] The clinical motivation behind effect modification (or heterogeneity) can be to identify the subgroups wherein treatment is most effective (or perhaps harmful). However, the interaction can be interpreted as causal only if both the treatment and the baseline factor directly affect the outcome.[30,32] For example, the clinical motivation behind assessing causal interaction could be to intervene on the baseline factor to improve the effect of treatment.

TABLE 3 Recommendations on the use of the interaction analysis in RCTs

Methods
1. Specify whether effect-measure modification or causal interaction is in view
2. Describe whether an interaction analysis is prespecified or post hoc
3. Describe how confounding was controlled for (eg randomization, multivariable adjustment)
a For effect–measure modification, additional adjustment is generally not needed because the treatment is randomized
b Consider further adjustment If multiple treatment subgroup modifications are found to be significant in order to identify the most relevant subgroups
c Report which relation is controlled for (eg "treatment" – "outcome of interest" and/or "baseline/demographic factor" – "outcome of interest") and the set of relevant confounders
d For causal interaction, confounding between the "baseline/demographic factor" and the "outcome of interest" must be taken into consideration

Results
1. Report the number of patients with and without the "outcome of interest" in treatment and placebo arms per each subgroup defined by the baseline/demographic factor
2. Report the treatment effect (eg RR/OR/HR) per each subgroup defined by the baseline/demographic factor using the subgroup with the lowest risk as the reference category
3. Report both multiplicative and additive interaction effects with 95% confidence intervals
4. To define a direction of an interaction (positive or negative), specify the subgroups on which this particular notation is based

In RCTs, investigators could claim that treatment directly affects the outcome even across subgroups of the baseline factor due to randomization of the treatment (assuming also adequate sample size, adherence to the study protocol, and no differential loss to follow-up).[33] However, claiming that the baseline factor itself is responsible for the subgroup effects is not immediately possible if confounding of the baseline factor on the outcome was not controlled for. This is because randomization accounts for unbiased comparability of treatment arms, but does not account for imbalances between the subgroups themselves that affect the outcome. Consider again the study by Head et al,[16] who found that PCI was associated with higher mortality than CABG in the subgroup of diabetic patients. The subgroup analysis would validly indicate that CABG is more effective than PCI in diabetic patients. However, concluding that diabetes itself is responsible for the subgroup effects is only possible if the investigators had controlled for other baseline factors that affect patient survival and are unequally distributed between the subgroups. For example, it could be that diabetic patients were treated less proactively with PCI than non-diabetic patients (eg they waited longer for PCI) which on its part affected patient survival, instead of diabetes itself.

Although randomization accounts for comparability between treatment arms even across subgroups, imbalances can still occur due to chance. Stratified randomization on known baseline factors that influence patient outcome prevents these imbalances to occur.[34] Yet, it does not control for imbalances between subgroups of baseline factors other than the treatment. Stratified randomization only helps to obtain comparable numbers of participants in both treatment arms within each subgroup.[34] However, other covariates can still be unevenly distributed among the subgroups which could affect the outcome. Alternatively, if randomization of the baseline factor is possible, investigators can apply a factorial design to control for confounding of treatment and the secondary intervention on that baseline factor. Another approach could be to adjust for relevant factors by including them into the statistical model. Using this approach however, one can never be completely sure from trial data that unknown confounding does not exist.

For effect-measure modification, controlling for confounding is generally unnecessary but can be helpful in some instances. First, imbalances that occur even with randomization could be adjusted for. Second, in stratified randomization the number of strata should be as low as possible (total number of strata is the product of the number of subgroups of each factor; eg if stratifying on sex and age using 3 categories, one will have $2 \times 3 = 6$ strata). With too many strata, one can end up with low numbers of participants per subgroup. Thus, stratifying on some factors and adjusting for others is an option. Third, if multiple significant subgroups exist, further adjustments could help narrowing the choice to the most relevant.

Randomized clinical trials are principally conducted assuming homogeneous effects of the treatment within subgroups. Based on this assumption, sample size is usually calculated by estimating only one (relative) effect that is supposed to hold for all eligible study participants. This is the main reason why RCTs are often underpowered to detect differences of treatment effects between subgroups even if they truly exist. Investigators should, therefore, plan a priori to analyse subgroups and incorporate these considerations into the sample size calculation. In this way, an adequate number of participants will be recruited for each subgroup. Moreover, the choice which interactions to test should be based on pathophysiological (and genetic) considerations and other relevant clinical implications (eg benefits of treatment based on disease stage, timing of treatment, comorbidities).[2] Such prespecified analyses would also help prevent bias that may arise when subgroup analysis is assessed after obtaining overall findings. A prespecified analysis (synonyms: "a priori," "preplanned," "planned," "previously suggested") is specified before obtaining data or as an attempt of corroboration (ie a trial performing an analysis similar to a previously reported trial).[5] If this is not the case, the analysis is post hoc (synonyms: "non-prespecified," "secondary," "explanatory," "preliminary").[5] Note that post hoc analyses may be data-driven

or motivated by overall null findings.[35] Investigators could try to systematically assess all possible statistical interactions to reduce the chance of spurious results[36] but then also correct for multiple testing. Finally, the best way to validate a statistical interaction is to replicate it in subsequent trials.

4 | CONCLUSION

This article describes challenges associated with assessment and interpretation of statistical interactions for binary outcomes in RCTs. It also provides information on publicly available excel sheets, SAS, STATA and R codes which can be used to assess different additive and multiplicative interaction effects, as well as recommendations to increase completeness and reliability of interaction analyses in future RCTs. Altogether, this article provides a brief non-statistical guide for clinical investigators on how to perform, interpret and report statistical interaction analyses in RCTs.

CONFLICT OF INTEREST

All authors declare no conflict of interest.

ORCID

Milos Brankovic https://orcid.org/0000-0002-3996-0813
Ewout W. Steyerberg https://orcid.org/0000-0002-7787-0122

REFERENCES

1. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine–reporting of subgroup analyses in clinical trials. *N Engl J Med*. 2007;357(21):2189-2194.
2. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*. 2005;365(9454):176-186.
3. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. 2000;355(9209):1064-1069.
4. Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? *Am Heart J*. 2006;151(2):257-264.
5. Wallach JD, Sullivan PG, Trepanowski JF, Sainani KL, Steyerberg EW, Ioannidis JP. Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. *JAMA Intern Med*. 2017;177(4):554-560.
6. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med*. 2002;21(19):2917-2930.
7. Lagakos SW. The challenge of subgroup analyses–reporting without distorting. *N Engl J Med*. 2006;354(16):1667-1669.
8. Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. *Int J Epidemiol*. 2012;41(2):514-520.
9. Boffetta P, Winn DM, Ioannidis JP, et al. Recommendations and proposed guidelines for assessing the cumulative evidence on joint effects of genes and environments on cancer occurrence in humans. *Int J Epidemiol*. 2012;41(3):686-704.
10. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*, Vol 398, 3rd edn. Hoboken, NJ: John Wiley & Sons; 2013.
11. Gosho M, Sato Y, Nagashima K, Takahashi S. Trends in study design and the statistical methods employed in a leading general medicine journal. *J Clin Pharm Ther*. 2018;43(1):36-44.
12. Greenland S. Interactions in epidemiology: relevance, identification, and estimation. *Epidemiology*. 2009;20(1):14-17.
13. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*, 3rd edn. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
14. VanderWeele TJ, Robins JM. The identification of synergism in the sufficient-component-cause framework. *Epidemiology*. 2007;18(3):329-339.
15. Knol MJ, van der Tweel I, Grobbee DE, Numans ME, Geerlings MI. Estimating interaction on an additive scale between continuous determinants in a logistic regression model. *Int J Epidemiol*. 2007;36(5):1111-1118.
16. Head SJ, Milojevic M, Daemen J, et al. Mortality after coronary artery bypass grafting versus percutaneous coronary intervention with stenting for coronary artery disease: a pooled analysis of individual patient data. *Lancet*. 2018;391(10124):939-948.
17. Li R, Chambless L. Test for additive interaction in proportional hazards models. *Ann Epidemiol*. 2007;17(3):227-236.
18. VanderWeele T, Knol MJ. A tutorial on interaction. *Epidemiologic Methods*. 2014;3:33-72.
19. VanderWeele TJ. Causal interactions in the proportional hazards model. *Epidemiology*. 2011;22(5):713-717.
20. Mathur MB, VanderWeele TJ. R function for additive interaction measures. *Epidemiology*. 2018;29(1):e5-e6.
21. 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;26(6):433-438.
22. Hosmer DW, Lemeshow S. Confidence interval estimation of interaction. *Epidemiology*. 1992;3(5):452-456.
23. Assmann SF, Hosmer DW, Lemeshow S, Mundt KA. Confidence intervals for measures of interaction. *Epidemiology*. 1996;7(3):286-290.
24. Skrondal A. Interaction as departure from additivity in case-control studies: a cautionary note. *Am J Epidemiol*. 2003;158(3):251-258.
25. Lundberg M, Fredlund P, Hallqvist J, Diderichsen F. A SAS program calculating three measures of interaction with confidence intervals. *Epidemiology*. 1996;7(6):655-656.
26. Mark Stevenson with contributions from Telmo Nunes CH, Jonathon Marshall, Javier Sanchez, Ron Thornton, Jeno, Reiczigel

JR-C, Paola Sebastiani, Peter Solymos, Kazuki Yoshida, Geoff Jones, Sarah Pirikahu, Simon, Firestone RK, Johann Popp and Mathew Jay. epiR: Tools for the Analysis of Epidemiological Data. R, 2017; https://CRAN.R-project.org/package=epiR. Accessed May 6, 2019.

27. Andrews B, Semler MW, Muchemwa L, et al. Effect of an early resuscitation protocol on in-hospital mortality among adults with sepsis and hypotension: a randomized clinical trial. *JAMA*. 2017;318(13):1233-1240.

28. VanderWeele TJ, Tchetgen Tchetgen EJ. Attributing effects to interactions. *Epidemiology*. 2014;25(5):711-722.

29. VanderWeele TJ. Reconsidering the denominator of the attributable proportion for interaction. *Eur J Epidemiol*. 2013;28(10):779-784.

30. VanderWeele TJ. On the distinction between interaction and effect modification. *Epidemiology*. 2009;20(6):863-871.

31. Szklo M, Nieto FJ. *Epidemiology Beyond the Basics*, 3rd edn. Burlington, MA: Jones & Barlett Learning; 2014.

32. Vander Weele TJ. Confounding and effect modification: distribution and measure. *Epidemiologic Methods*. 2012;1(1):55-82.

33. Lachin J, Bautista O. Stratified-adjusted versus unstratified assessment of sample size and power for analyses of proportions. In: Thall P, ed. *Recent Advances in Clinical Trial Design and Analysis*. Boston, MA: Kluwer; 1995:258.

34. Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. *J Clin Epidemiol*. 1999;52(1):19-26.

35. 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.

36. Patel CJ, Chen R, Kodama K, Ioannidis JP, Butte AJ. Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus. *Hum Genet*. 2013;132(5):495-508.