Journal of Clinical Epidemiology 134 (2021) 174–177

Journal of Clinical Epidemiology

KEY CONCEPTS SERIES

Effect Modifiers and Statistical Tests for Interaction in Randomized Trials

Robin Christensen[a,b], Martijn J.L. Bours[c], Sabrina M. Nielsen[a,b,*]

[a] *Section for Biostatistics and Evidence-Based Research, the Parker Institute, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark*
[b] *Research Unit of Rheumatology, Department of Clinical Research, University of Southern Denmark, Odense University Hospital, Denmark*
[c] *Department of Epidemiology, GROW – School for Oncology and Developmental Biology, Maastricht University, Maastricht, The Netherlands*

Received 1 March 2021; Received in revised form 8 March 2021; Accepted 10 March 2021

Abstract

Statistical analyses of randomized controlled trials (RCTs) yield a causally valid estimate of the overall treatment effect, which is the contrast between the outcomes in two randomized treatment groups commonly accompanied by a confidence interval. In addition, the trial investigators may want to examine whether the observed treatment effect varies across patient subgroups (also called 'heterogeneity of treatment effects'), i.e. whether the treatment effect is modified by the value of a variable assessed at baseline. The statistical approach for this evaluation of potential effect modifiers is a test for statistical interaction to evaluate whether the treatment effect varies across levels of the effect modifier. In this article, we provide a concise and nontechnical explanation of the use of simple statistical tests for interaction to identify effect modifiers in RCTs. We explain how to calculate the test of interaction by hand, applied to a dataset with simulated data on 1,000 imaginary participants for illustration.

1. Background

Randomized controlled trials (RCTs) are considered the gold standard when evaluating a treatment's effectiveness because of their high *internal validity* when appropriately conducted. The goal of randomization is to balance both observed and unobserved participant characteristics between two (or more) randomly allocated treatment groups. Thus, the RCT design allows causal effects of treatments to be estimated because confounding will generally not be an issue [1]. Usually, statistical analyses of RCTs yield an estimate of the *overall* treatment effect (say, E_{overall}), which is the contrast between the outcomes in two treatment groups commonly accompanied by a confidence interval.

RCTs can also have good *external validity* if they are based on real-life populations that are relevant for the intervention, treats the control group with an acceptable standard of care, and reports outcomes that are meaningful. An ideal trial in this regard enrolls patients with a broad range of background characteristics, for example, disease severity, age, sex, race, and prior therapies. Following the primary analyses estimating the overall treatment effect, E_{overall}, the trial investigators may want to examine whether the observed treatment effect varies across patient subgroups (also called 'heterogeneity of treatment effects'). In such cases we are interested in examining whether the treatment effect is modified by the value of another variable (i.e. the effect modifier) [2]. The statistical approach for evaluating potential effect modifiers is a test for statistical interaction [3].

Findings from investigating heterogeneity of treatment effects for an RCT are important for understanding, interpreting, and translating the findings, and consequently for determining whether there is an appropriate patient sub-population for treatment use. Evidence for effect modification therefore helps to delineate applicability of an intervention, showing in whom the treatment is most likely to work and thus indicative of an RCT's *external validity*. In this article, we provide a concise and nontechnical explanation of simple statistical tests for interaction to identify effect modifiers in RCTs.

* Corresponding author.
E-mail address: sabrina.mai.nielsen@regionh.dk (S.M. Nielsen).

https://doi.org/10.1016/j.jclinepi.2021.03.009

2. Definition

The statistical tests for interaction are often referred to as subgroup analyses, implying any comparison of effect between treatment groups (net benefit) across subsets (i.e. subgroups) of patients with specific characteristics that could be potentially relevant effect modifiers. Usually subgroup analyses investigate subgroups defined by a factor measured either before or at baseline, such as sex (males vs. females). Subgroup analyses can be misleading if they are based on data-driven hypotheses, employ inappropriate statistical methods, or fail to account for multiple testing [4]. As exemplified by Alosh et al [5], one should distinguish between three categories of subgroup analysis: (*i*) *exploratory analyses* search for differential responses from early clinical trial data or from clinical trials that failed to establish treatment efficacy in its intended population; (*ii*) *supportive analyses* aim at investigating the consistency of treatment effect across subgroups for a clinical trial that has established treatment efficacy in its intended overall population; and finally (*iii*) *inferential analyses* aim at establishing treatment efficacy in a pre-defined targeted subgroup and/or in the overall population.

The subgroups of interest are defined, preferably *a priori*, and the baseline variable under consideration needs to precede treatment in time. In the simplest case, our baseline factor is a covariate with only two levels (e.g. male vs. female subjects), leading to two subgroups (e.g. subgroup 1: males, subgroup 2: females). If we want to compare the treatment effects observed in the two subgroups, a first step is to estimate the treatment effects (i.e. net benefit) *within each subgroup* in separate analyses (E_1 and E_2, respectively). Next, a test for statistical interaction comparing the two subgroups can be calculated by hand based on the subgroup treatment effects (E_1 and E_2) and their corresponding standard errors (SE_{E1} and SE_{E2}) [3]:

Difference between subgroup effects, $\boldsymbol{d} = \boldsymbol{E}_1 - \boldsymbol{E}_2$

Standard error for d, $\boldsymbol{SE_d} = \sqrt{\boldsymbol{SE}^2_{\boldsymbol{E1}} + \boldsymbol{SE}^2_{\boldsymbol{E2}}}$

Test statistics for the z-test, $\boldsymbol{z\ value} = \frac{\boldsymbol{d}}{\boldsymbol{SE_d}}$

The p-value can be found by using the absolute (non-negative) z value which gives a test of the null hypothesis that in the population the difference between subgroups (d) is zero, by comparing the value of z to the standard normal distribution. For effect measures on a multiplicative scale (such as risk ratio, hazard ratio, or odds ratio) as opposed to the additive scale (such as risk differences), the analyses should be performed using the log-transformation and with the corresponding standard errors [3]. Importantly, effect modification may be present on one scale but not on another, and conflicting opinions exist on which scale to use [6]. The European Medicines Agency (EMA) recommends using the scale on which the endpoint is commonly analyzed, and to present supplementary analyses on the complementary scale where inconsistency is observed [7].

3. Application

For presenting the results of subgroup analyses graphically, forest plots are useful. Preferably, the plots should include a bold vertical line at the overall treatment effect (i.e., E_{overall}) rather than at the null (i.e., 'no effect') to guide correct interpretation regarding heterogeneity of treatment effects across subgroups. Fig. 1 illustrates an example based on a simulated dataset on 1,000 imaginary participants (randomized 1:1); the data was generated to reveal a standardized mean difference corresponding to a statistically significant moderate overall treatment effect of $E_{\text{overall}} = 5.00$ (95%CI: 3.73 to 6.27) units. To this dataset we deliberately generated a contextual factor (CF 1) that would create two separate subgroups with different magnitudes of treatment effects (E_1: 8.00 and E_2: 2.00 units, respectively). The standard errors can be calculated from the confidence intervals shown in the figure, SE_{E1}: (9.75-8.00)/1.96 = 0.89 and SE_{E2}: (3.78-2.00)/1.96 = 0.91, respectively. From these values we can test the interaction and estimate the difference between the subgroups (with confidence interval). The test of interaction:

$$\boldsymbol{d} = \ 8.00 - 2.00 = 6.00$$

$$\boldsymbol{SE_d} = \sqrt{0.89^2 + 0.91^2} = 1.273$$

$$\boldsymbol{z\ value} = \ \frac{6.00}{1.273} = 4.71$$

A z-value of 4.71 gives $p<0.001$ when we refer it to a table of the normal distribution. The estimated interaction effect is $d = 6.00$ units; the corresponding 95% confidence interval is $6.00 \pm 1.96^{*}1.273$ (i.e., 95%CI 3.50 to 8.50). The data thus provide evidence for effect modification, indicating that the treatment effect is significantly stronger in CF 1-positive than CF 1-negative trial participants. The other presented contextual factors shown in Fig. 1 were computer-generated completely at random, and thus any apparent effect modification across CF 2, CF 3, …., and CF 7 reflect purely chance findings (a well-known caveat to multiple testing without an *a priori* hypothesis).

4. Pointers

Altman and Bland originally presented this simple approach as an "interaction revisited" statistics notes, in the BMJ back in 2003 [3]. This approach is transparent and feasible when we want to compare two estimated quantities, such as means (Fig. 1) or proportions (Fig. 2), each with its standard error.

Although highly feasible, investigating subgroup effects should be done with great care and interpreted cautiously. Most trials are not powered to detect subgroup differences but reporting the results anyway will allow future meta-analyses to investigate this based on several trials thereby

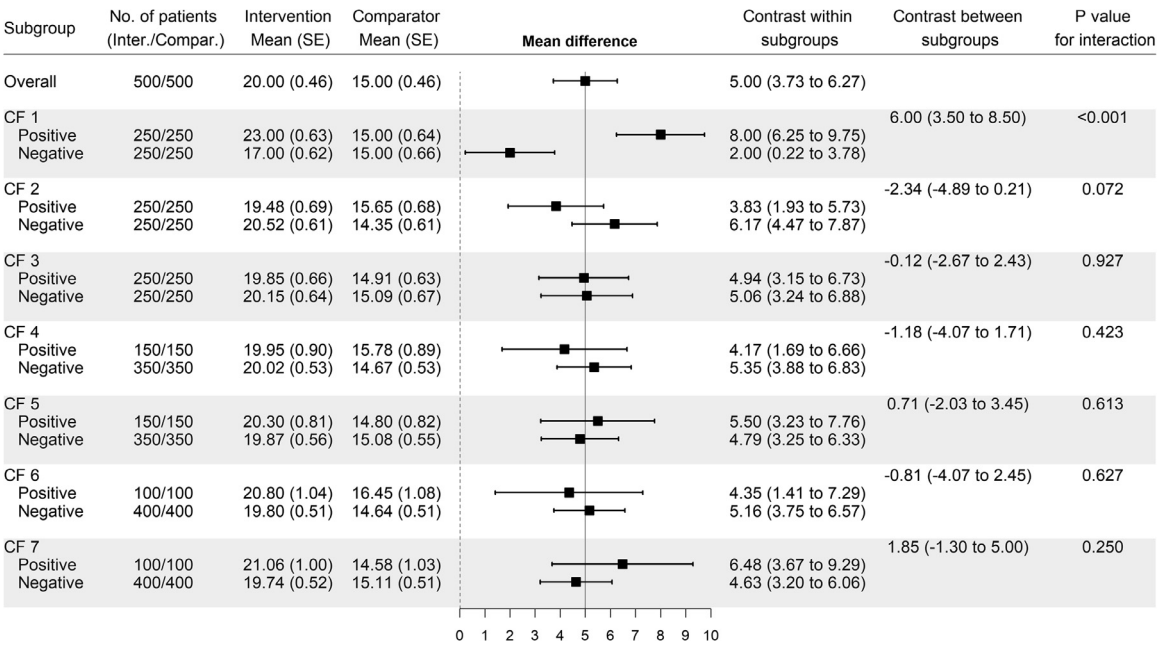

Fig. 1. Forest plot showing the results of subgroup analyses based on a simulated dataset on 1,000 imaginary participants. The outcome is based on continuous data. The bold vertical line indicates the overall treatment effect, and the dashed line indicates no effect.

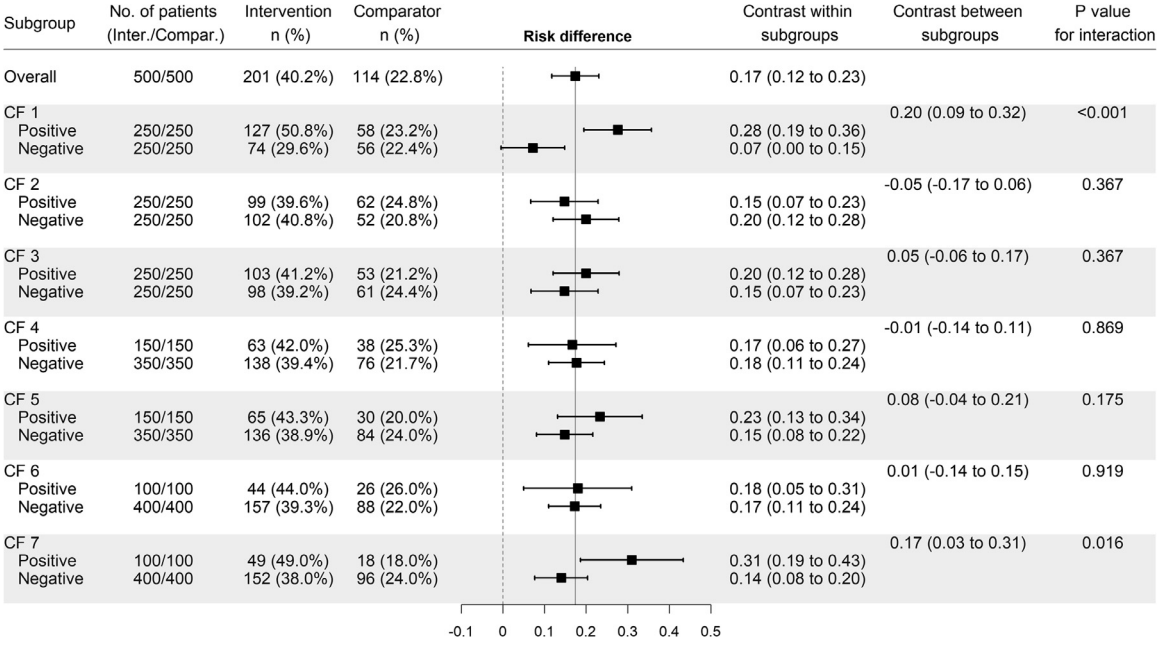

Fig. 2. Forest plot showing the results of subgroup analyses based on a simulated dataset on 1,000 imaginary participants. The outcome is based on dichotomous data. The bold vertical line indicates the overall treatment effect, and the dashed line indicates no effect.

achieving the sufficient power. Currently, there exist no explicit/standard list of factors to be investigated for effect modification in trials. However, one may initially be inspired by the U.S. Food and Drug Administration (FDA) requiring effectiveness data to be analyzed by sex, age, and racial subgroups.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2021.03.009.

References

[1] Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. Annu Rev Public Health 2000;21:121–45.

[2] Kent DM, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Ann Intern Med 2020;172(1):35–45.

[3] Altman DG, Bland JM. Interaction revisited: the difference between two estimates. Bmj 2003;326(7382):219.

[4] Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. Cmaj 2020;192(32):E901–6.

[5] Alosh M, Huque MF, Bretz F, D'Agostino RB Sr. Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. Stat Med 2017;36(8):1334–60.

[6] Doi SA, Furuya-Kanamori L, Xu C, Lin L, Chivese T, Thalib L. Questionable utility of the relative risk in clinical research: a call for change to practice. J Clin Epidemiol 2020.

[7] European Medicines Agency (EMA) Guideline on the investigation of subgroups in confirmatory clinical trials. London, United Kingdom: European Medicines Agency (EMA); 2019.

Further reading

Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. Annu Rev Public Health 2000;21:121–45 *Note: How to make inferences about causal effects of treatments or exposures*.

Kent DM, Paulus JK, van Klaveren D, D'Agostino R, Goodman S, Hayward R, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Ann Intern Med 2020;172(1):35–45 *Note: Guidance on an alternative (advanced) approach to subgroup analysis, namely predictive heterogeneity of treatment effects analysis, taking into account all relevant patient attributes (contextual factors) simultaneously*.

Altman DG, Bland JM. Interaction revisited: the difference between two estimates. Bmj 2003;326(7382):219 *Note: A short hands-on description of how to estimate and test interactions by hand*.

Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. Cmaj 2020;192(32):E901–6. *Note: An instrument for assessing the credibility of subgroup analyses. The full instrument* is *provided in their appendix 5 (for RCTs) and in appendix 6 (for meta-analyses of RCTs), with a detailed manual in their appendix 3; all available here:* Available at: https://www.cmaj.ca/content/192/32/E901/tab-related-content.

Alosh M, Huque MF, Bretz F, D'Agostino RB Sr, et al. Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. Stat Med 2017;36(8):1334–60 *Note: Technical and elaborate description of subgroup analysis in confirmatory clinical trials*.