ELSEVIER

Journal of Clinical Epidemiology ■ (2014) ■

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

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

Amand F. Schmidt[a,b,c,*], Rolf H.H. Groenwold[a,b], Mirjam J. Knol[a,d], Arno W. Hoes[a], Mirjam Nielen[c], Kit C.B. Roes[a], Anthonius de Boer[b], Olaf H. Klungel[a,b]

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, P.O. Box 85500, 3508 GA Utrecht, The Netherlands*
[b]*Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, P.O. Box 80082, 3508 TB Utrecht, The Netherlands*
[c]*Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, Yalelaan 107, 3584 CL Utrecht, The Netherlands*
[d]*RIVM, National Institute for Public Health and the Environment Centre for Infectious Disease Control Netherlands (CIb) Epidemiology and Surveillance Unit (EPI), P.O. Box 1, 3720 BA Bilthoven, The Netherlands*

Accepted 10 February 2014; Published online xxxx

Abstract

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

Methods: A literature review and simulation study was performed evaluating interaction tests on the odds ratio (OR) or the risk difference (RD) scales: Cochran Q (Q), Breslow−Day (BD), Tarone, unconditional score, likelihood ratio (LR), Wald, and relative excess risk due to interaction (RERI)-based tests.

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

Conclusion: In small samples, interaction tests differed. On the OR scale, the Tarone and BD tests are recommended. On the RD scale, the LR and RERI-based tests performed best. However, RERI-based tests are underpowered compared with other tests, when exposure prevents the outcome, and sample size is limited.

Keywords: Statistics; Review; Simulation; Epidemiologic methods; Interaction; Effect modification; Subgroups; Odds ratio; Risk ratio; Relative excess risk due to interaction

1. Introduction

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

The presence of interaction depends on the type of effect measure that quantifies the relation between treatment and outcome [4,5]. For example, in case of a binary outcome (eg, myocardial infarction), an interaction can be present on the OR (multiplicative) scale but absent on the RD (additive) scale, or vice versa.

Previously, the performance of interaction tests was assessed using simulation studies [6−11]. Most studies focused on interaction tests using ORs, and no single study compared all the commonly used interaction tests together

Funding: This work was supported by Research Focus Areas funding of the Utrecht University and was a collaboration between the faculties of medicine, science, and veterinary medicine.

Conflict of interest: None of the authors of this article has a financial or personal relationship with other people or organizations that could inappropriately influence or bias the content of the article.

* Corresponding author. Tel.: +31-88-7550-461; fax: +31-88-7568-099.

E-mail address: a.f.schmidt@umcutrecht.nl (A.F. Schmidt).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2014.02.008

What is new?

Key findings

- Results from both the review and the simulation study showed that power was limited in small sample sizes (eg, ≤250 subjects) and that type 1 error rates could be relatively high. On the risk difference (RD) scale, when exposure was protective, the relative excess risk due to interaction (RERI)-based tests were relatively underpowered compared to other interaction tests. Given sufficient sample size (asymptotically) and independent of exposure being a risk factor or not. All interaction tests performed equal.

What this adds to what was known?

- Up till now, a comprehensive overview including all readily available and frequently used interaction tests was lacking. Compared with the other odds ratio (OR) interaction tests, the Tarone and Breslow−Day (BD) tests had type 1 error rates closest to 0.05. Among the RD tests, the likelihood ratio (LR) and RERI-based tests had type 1 error rates closest to 0.05. Previous research showed that the RERI should be recoded when exposure prevents the outcome. The present study revealed that recoding is unnecessary when exposure is protective and sample size is sufficiently large (eg, 1,000 subjects). Performance of all tests was equal in such settings.

What is the implication and what should change now?

- When comparing subgroup-specific effect using interaction test, researchers should be aware of the following: (1) When sample size is sufficiently large (eg, 500−1,000 subjects) and the choice of interaction test is irrelevant, they all performed equally. (2) In small sample sizes, depending on the tests chosen, type 1 error can be as high as 0.10. Therefore, exploring interactions in such settings might not be appropriate. If interaction testing is pursued in such settings, on the OR scale, the Tarone or BD test, and on the RD scale, the LR or RERI-based test should be used. (3) Users of the RERI-based tests should be aware of its behavior when exposure is protective and should consider recoding the statistic or use one of the other RD tests when sample size is limited.

in one scenario. We aimed to provide a comprehensive comparison of commonly applied test on the OR scale and the RD scale (specifically the Cochran Q [Q], BD, Tarone, unconditional score [Score], LR, and Wald test and tests based on the RERI). First, a systematic review was conducted providing an overview of previous simulations studies. Obviously, each simulation study used different simulation scenarios which could potentially explain any dissimilarity in performance between interaction tests. Therefore, in a second part we conducted a simulation study to compare all of the previously mentioned interaction tests under equal simulation conditions.

2. Methods

The review and subsequent simulation study evaluated the following asymptotic interaction tests: on the OR scale the Q, BD, Tarone, Score, LR, and the Wald test were compared. For the RD scale we compared the Q, LR, and the Wald test and tests based on the RERI. To our knowledge no variance estimator is available for the BD, Tarone, and Score tests using the RD scale, therefore these tests were not assessed for the RD scale. Similarly, the RERI is specifically proposed for estimating interaction on an RD scale using risk ratios (RRs) and, therefore, was only evaluated on the RD scale. For the formulae of these interaction tests we refer to Appendix I at www.jclinepi.com. In both the review and the subsequent simulation study we focused on sparse data scenarios because asymptotic tests differ in such settings. In small sample sizes, power is often limited therefore, while exploring both power and type 1 error rates, we focus on the latter.

2.1. Systematic review

Using the following search terms in title or abstract, Medline was searched (date: 5/24/13): (homogeneity OR modification OR interaction OR synergism OR antagonism) AND (simulation OR "monte carlo") AND (effect OR test OR statistic OR power OR significance)

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

2.2. Simulation study

A simulation study was performed to assess the statistical performance of the previously mentioned interaction tests. Most evaluated tests are only applicable to categorical data and, therefore, all simulations were based on scenarios with two dichotomous exposures (ie, X and S) and a dichotomous outcome. In such settings, subjects can be in one of four possible exposure categories, indicated by $i = 0$ or 1 if exposure to X is absent or present, and $j = 0$ or 1 depending on the absence or presence of exposure to S. The corresponding outcome probabilities are indicated by P_{ij}. Initially, six scenarios (A−F, see Table 1) were created,