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, Rolf H.H. Groenwold, Mirjam J. Knol, Arno W. Hoes, Mirjam Nielens, Kit C.B. Roes, Anthonius de Boer, Olaf H. Klungel

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, The Netherlands
Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, The Netherlands
Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, The Netherlands
RIVM, National Institute for Public Health and the Environment Centre for Infectious Disease Control Netherlands (CIb) Epidemiology and Surveillance Unit (EPI), The Netherlands

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 ≤ 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...
Key findings

- Results from both the review and the simulation study showed that power was limited in small sample sizes (e.g., ≤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 (e.g., 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 (e.g., 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.

What is new?

- 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 (i.e., X and S) and a dichotomous outcome. In such settings, subjects can be in one of four possible exposure categories, indicated by i = 0 or 1 if exposure to X is absent or present, and j = 0 or 1 depending on the absence or presence of exposure to S. The corresponding outcome probabilities are indicated by P_{ij}.

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. Initially, six scenarios (A–F, see Table 1) were created,
Table 1. Six simulation scenarios used to evaluate interaction test on the risk difference and odds ratios scale

<table>
<thead>
<tr>
<th>Scenario (tests used)</th>
<th>P00</th>
<th>P01</th>
<th>P10</th>
<th>P11</th>
<th>RD interaction</th>
<th>OR interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (RD tests)</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>0.50</td>
<td>0.20</td>
<td>2.33</td>
</tr>
<tr>
<td>B (RD tests)</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>0.10</td>
<td>−0.20</td>
<td>0.26</td>
</tr>
<tr>
<td>C (OR tests)</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.40</td>
<td>0.15</td>
<td>2.00</td>
</tr>
<tr>
<td>D (OR tests)</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.14</td>
<td>−0.11</td>
<td>0.49</td>
</tr>
<tr>
<td>E (OR and RD tests)</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>F (OR and RD tests)</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Abbreviations: OR, odds ratio; RD, risk difference.

The expected number of events and non-events can be calculated as follows: number of events = \( P_{0ij} \times N \times F_{ij} \); number of non-events = \((1 - P_{0ij}) \times N \times F_{ij}\).

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

<table>
<thead>
<tr>
<th>Simulation parameter</th>
<th>Total sample size</th>
<th>Event probabilities</th>
<th>Exposure prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( N = 1,000 )</td>
<td>( P_{ij} = 0.30, 0.30, 0.30, 0.10 )</td>
<td>( f_{ij} = 0.25, 0.25, 0.25, 0.25 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X exposure status</th>
<th>S exposure status</th>
<th>Events</th>
<th>Non-events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>25 = 0.10 \times \frac{100}{100}</td>
<td>225 = (1 - 0.10) \times \frac{100}{100}</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>75 = 0.30 \times \frac{100}{100}</td>
<td>175 = (1 - 0.30) \times \frac{100}{100}</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>75 = 0.30 \times \frac{100}{100}</td>
<td>175 = (1 - 0.30) \times \frac{100}{100}</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>75 = 0.30 \times \frac{100}{100}</td>
<td>175 = (1 - 0.30) \times \frac{100}{100}</td>
</tr>
</tbody>
</table>

Abbreviations: OR, odds ratio; RD, risk difference.
In the generated scenarios, the previously mentioned interaction tests were compared. The LR test was applied as a model comparison test; comparing a model with a product term to a model without a product term. On the OR scale, logistic models were compared. On the RD scale, the LR test was based on Poisson models with an identity link and robust variance estimators, specifically the heteroscedasticity-consistent covariance estimator 0, ie, HC0 [13]. Two RERI-based tests were evaluated, the first using the delta (RERI_delta) estimator [18] and the second using the bootstrap percentile [18] variance estimator (RERI_bs). For the RERI_delta test, a Poisson model with log link and robust variance estimators was used, the RERI_bs test was estimated from the crude 2-by-2-by-2 table and used 1,000 bootstraps.

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

3. Results

3.1. Systematic review

We identified 15 studies that evaluated the performance of interaction tests (Fig. 1). Of these, six focused on interaction tests on the RD scale [14,15,17–20], seven focused on interaction tests on the OR scale [6,7,9–11,21,22] and two addressed both [8,23]. The results of these studies are summarized in (Appendix Table 1 at www.jclinepi.com).

The nine studies that studied the OR scale, explored a large number of scenarios, ranging from scenarios with expected cell counts of 1 to scenarios with expected cell counts of 55 or more [9,10]. Studies that explored the LR tests showed that in most scenarios its type 1 error rate was larger than 0.05. In the same scenarios, the Tarone, BD, and Score tests showed error rates closer to 0.05 [6,7,9,10]. In extreme settings with expected cell counts of 1 to 5, the LR type 1 error rate could be as high as 0.97, whereas in the same settings, the BD and Tarone tests had error rates of 0.44. In three other studies [21–23], the LR had type 1 error rates < 0.05, however, these only explored large sample situations. The Wald and Q statistics also showed type 1 error rates close to 0.05 in relatively nonsparse data settings (eg, 75 cases and 150 controls) [8,21]. Generally, power did not differ much between the tests studied, differing usually not more than 5%, except in very small sample sizes.

The eight studies that evaluated RD interaction tests performance used a great number of scenarios, similar to the OR scenarios, including scenarios with expected cell counts of 1 to simulations with 1,000 cases and controls [18,23]. Two studies explored the LR tests performance for the RD scale and showed type 1 error rates close to 0.05 [17] or lower [23], the first study used a small amount of replications (maximum 600) [14] the latter used 5,000 repetitions with at least 500 subjects [23]. The Q tests on the RD scale was evaluated by three studies [18–20]. All three studies showed that in sparse data, the type 1 error rate was often higher than 0.05, however, generally not larger than 0.06. In scenarios with expected cell counts of 1 and a large number of subgroups, the type 1 error rate was seriously inflated to 0.60 [19]. A single study explored the Wald test for the RD scale and showed type 1 error rates below 0.05 [8]. RERI-based interaction tests were evaluated by three studies, which showed that the type 1 error rate (or the 95% CI coverage) was below 0.05 (or the coverage rate was above 95%), for example, a type 1 error rate of

Fig. 1. Flow of studies in the Medline search for simulation studies on the performance of interaction test.
0.025 or a coverage rate of 97% [14,15,18]. All of the simulations studies that evaluated the RERI-based tests used the OR as an approximation of the RR [15,18]. Power did not differ much between RD interaction tests, and was mostly driven by sparseness of the data and interaction magnitude.

3.2. Simulation study

Simulations on the OR scale showed that in small sample sizes ($N \leq 250$ or when one of the exposure groups contributed $\leq 0.10$ to the overall $N$) the Score and LR test had error rates above 0.05, sometimes as high as 0.10 (Fig. 2). The Q and Wald tests displayed type 1 error rates below 0.05, whereas the BD and Tarone tests had type 1 error rates closest to 0.05. Power did not markedly differ between tests (Appendix III, Fig. a at www.jclinepi.com), for example, the maximum difference in power for scenario C using 50 subjects was 0.07. Fig. 3 shows that in samples of 100 subjects, a high power, for example 80%, was only reached when the interaction effects were very large; interaction OR of 0.05 or 19. Note that the symmetry around 1 is expected because the interaction effects in positive and negative settings only differ regarding their sign, e.g., $1/19 = 0.0526$. Because of empty cell counts in scenarios of 50 subjects the Q tests

![Fig. 2. Type 1 error rates of odds ratio (OR) interaction tests evaluated in simulation scenarios E and F. Note that in the upper parts of scenarios E and F the sample size was increased from 50 observations to 2,000. In the bottom part, sample size was fixed at 250 observations and the fraction of $F_{11}$ was increased from 0.05 to 0.25. Thus the relative number of subjects that were exposed to both factors increased from 5% to 25% of the total sample size of 250. All simulations were repeated 10,000 times.](image-url)
did not converge in 25% of the cases and the BD, Tarone, and Score tests failed up to 10% of the replications. In these settings, the Wald and LR tests did converge because these were implemented using generalized linear models.

Simulations on the RD scale showed that in sample sizes of 250 subjects or less (or when the prevalence of combined exposure subjects was $\leq 0.15$) the type 1 error rates were typically above 0.05 (Fig. 4). In such settings, the LR and the RERI$\Delta$ tests had type 1 error rates closest to 0.05, but in some scenarios, type 1 error rates above 0.05 were also shown. Results of scenarios A and B showed that, excluding the RERI-based tests, there was little difference in power between the RD tests (Appendix III, Fig. b at www.jclinepi.com). Fig. 3 shows that the RERI-based tests were relatively underpowered (compared with the other interaction tests on the RD scale) when the interaction effect was negative. For example, when the interaction RD $= -0.3$ the power of the Wald test $= 0.36$ and the power of the RERI$\Delta$ test $= 0.11$. This difference decreased as sample size increased to 1,000 (Appendix III, Figures b and c at www.jclinepi.com). Because of empty cell counts in scenarios of 50 subjects the LR tests failed up to 15% of the times, the other RD tests failed in less than 5% of the replications.

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

4. Discussion

Our simulation study showed that in small sample sizes (on the OR scale) the BD and Tarone tests had type 1 error rates closest to 0.05 whereas the LR test had type 1 error rates as high as 0.10. On the RD scale, simulation results revealed that RD interaction tests frequently had type 1 error rates $> 0.05$. Of all the RD tests evaluated, the LR and RERI$\Delta$ tests had type 1 error rate closest to 0.05. Additionally, our simulations showed that the RERI-based tests were relatively underpowered (as compared with the other RD interaction tests) in the presence of negative interaction effects, this difference decreased as sample size increased to 1,000.

Results of our simulation study were generally supported by the review results. For example, both the review and simulation study showed that on the OR scale the BD and Tarone tests had type 1 error rates closest to 0.05. However, there are also differences between the review and simulations results. The most important difference is that our simulation showed a relative lack of power for the RERI-based tests, as compared with the other RD tests, in negative interaction settings with less than 1,000 subjects. Studies included in the review did not show such results. This difference can be explained by the scenarios that were
considered. Note that the RERI uses relative measures, such as the RR, to calculate the RD interaction. In our simulations, negative interaction settings were based on scenarios where, compared with the unexposed, exposure prevented the outcome; i.e., by taking the unexposed group as the default reference group, sometimes the RR < 1. The studies included in the review never used scenarios where (combined) exposure was protective. Instead, these studies created negative interaction scenarios by setting RR_{11} < RR_{10} and/or RR_{01}, but at the same time ensuring that all RRs > 1. Previously, it has been recognized that the RERI should only be used when all types of exposure increase the outcome risk [24]. To achieve this, it is recommended to recode the exposure ensuring that the reference category is always the group with the smallest risk [24]. We showed that given sufficient sample size, the RERI-based tests performed similar to the other RD interaction tests even when exposure prevents the outcome. Note that in case-control studies, the RERI is often the only test available to explore RD interaction, making recoding an important consideration.

Several limitations and strengths warrant discussion. First, although we searched systematically, we concede that we may have missed studies. Our review and simulation study showed comparable results. Therefore, it seems unlikely that the overall conclusions would materially change.

![Fig. 4. Type 1 error rates of risk difference (RD) interaction tests evaluated in simulation scenarios E and F. Note that in the upper parts of scenarios E and F the sample size was increased from 50 observations to 2,000. In the bottom part, sample size was fixed at 250 observations, and the fraction of $F_{11}$ was increased from 0.05 to 0.25. Thus, the relative number of subjects that were exposed to both factors increased from 5% to 25% of the total sample size of 250. All simulations were repeated 10,000 times.](image)
by including additional studies. Second, we recognize that instead of using asymptotic tests, exact tests perform better with respect to type I error control. However, power of exact tests may be lower in a situation where by design there, usually, already is a lack of power. Third, we concede that in this study the number of simulation scenarios was limited and that it would be interesting to explore scenarios with larger disbalance between subgroups and different event probabilities. Furthermore, we simulated scenarios with expected cell counts as low as 1.25 which might seem unrealistic. However, when adjusting for multiple (co)variables these scenarios might occur more often than initially expected, making these simulations potentially very relevant. Fourth, some might question the use of Poisson or binomial models, because these are known to provide impossible estimates (eg, probabilities outside the range 0–1) or to not converge at all [25]. Obviously, as with all models, it is advisable to check the plausibility of derived estimates. In settings where estimates are (expected to be) implausible, methods based on calculating the “marginal probabilities of success” using logistic regression models might be preferable [25–27]. Fifth, in sparse data settings (eg, N = 50), some tests failed. Often, tests based on generalized linear models (GLMs) did converge and tests based on 2-by-2-by-2 tables failed, because of empty cells. For comparison, the Wald test was estimated based on GLMs and 2-by-2-by-2 tables, where the latter did indeed fail more often. Despite this increase in failed tests, results were equal up to two decimals (results not shown). Thus, it seems unlikely that differences in failure rate can explain our results. Obviously, performing interaction testing (or any testing at all) is problematic when empty cells exist and researchers should generally reconsider testing in such settings. On the other hand, empty or sparse cell counts could also be because of large (interaction) effects, which are important to report. Finally, an often-heard comment on interaction testing in small sample sizes is that “when a significant result is found, despite low power, this interaction effect must therefore be present”. However, this study showed that, depending on the test used and to some extent the scale chosen, the type I error rate can be high thus invalidating the previous comment.

Based on our results, we recommend the following. First, when sufficient sample size is available (eg, 500–1,000 subjects) all interaction tests perform similarly; hence the choice of interaction tests is irrelevant. Second, in smaller sample sizes, power is limited (unless a large interaction effect is present) and type I error rates are high, hence exploring interactions in such settings might not be appropriate. Furthermore, when deciding whether sample size is sufficient, researchers should also consider the distribution across exposure categories and across other potentially relevant (confounding) variables. Fourth, in sparse data settings, the Tarone and BD tests on the OR scale and the LR or RERI\textsubscript{delta} tests on the RD scale should be preferred because these tests have type I error rates closest to 0.05. Finally, users of the RERI-based tests should be aware of its behavior when exposure is protective and should consider recoding the statistic or use one of the other RD tests.

5. Conclusion

In small sample sizes (eg, N < 1,000), the Tarone and BD tests are preferred when assessing interaction on the OR scale. On the RD scale, the LR and RERI\textsubscript{delta} are the preferred tests for interaction. However, when exposure is preventative for the outcome, RERI-based tests are relatively underpowered compared with other interaction tests unless sample size is large. Recoding the exposure so that the RERI interaction effect becomes positive will resolve this problem.

Acknowledgments

AFS, RHHG, AWH, MJK, MN, KCBR, AdB, and OHK contributed to the conception and design of the study. AFS performed the literature search, analyses, and drafted the manuscript. RHHG, AWH, MJK, MN, KCBR, AdB, and OHK provided guidance during initial planning of the article and during critical revision. AFS had full access to all the data and takes responsibility for the integrity of the data presented. This work was presented at the 2013 Dutch WEON conference.

Appendix

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2014.02.008.
References


[26] Austin PC. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. J Clin Epidemiol 2010;63:2–6.