Subgroup analysis in randomized controlled trials appeared to be dependent on whether relative or absolute effect measures were used

Roderick P. Venekamp, Maroeska M. Rovers, Arno W. Hoes, Mirjam J. Knol

Abstract

Objectives: To assess whether relative or absolute effect measures were used in subgroup analyses of randomized controlled trials (RCTs) and study whether conclusions would change if subgroup effects were calculated on a different scale than reported.

Study Design and Setting: We studied all 327 RCTs published in 2010 in five major medical journals. For trials with a dichotomous primary outcome, we extracted reported main and subgroup effect measures. If crude subgrouping data were reported, we calculated the subgroup effects on both relative and absolute scales.

Results: Of the 229 RCTs with a dichotomous primary outcome, 120 (52%) performed subgroup analyses. In 106 of these 120 (88%) RCTs, relative effect measures were used for subgroup analyses, whereas an absolute scale was used in 9 (8%) trials. Two (2%) RCTs reported both relative and absolute subgroup effects. Crude data of the subgroups could be extracted in 41 of the 120 (34%) RCTs. Calculating subgroup effects on a different scale than reported lead to a change in conclusion in 17% of the 41 trials.

Conclusion: Almost all RCTs used relative effect measures for subgroup analyses. Interpretation of subgroup effects, however, appeared to be dependent on whether relative or absolute effect measures were used.

Keywords: Subgroup analysis; Randomized controlled trials; Treatment effects; Relative risk reduction; Absolute risk reduction; Epidemiology

1. Introduction

Randomized controlled trials (RCTs) are widely regarded as providing the most reliable evidence on the benefits and harms of interventions. In addition to main analyses, RCTs frequently perform subgroup analyses to identify specific subgroups of patients who do (or do not) benefit from the intervention [1–3]. Clinical guidelines often incorporate results of subgroup analyses, and such findings can therefore influence clinical decisions considerably.

Previous studies demonstrated that interpretation of trial results may be influenced by the use of either relative [eg, relative risk (RR), odds ratio (OR), hazard ratio (HR)] or absolute [eg, risk difference (RD)] effect measures in outcome reporting as benefits of interventions are often perceived larger if outcomes were reported with relative effect measures than if the same trial results were presented with absolute effect measures [4–8]. Consequently, reporting both relative and absolute effect measures for primary and secondary outcomes in RCTs is, nowadays, strongly recommended by the Consolidated Standards of Reporting Trials (CONSORT) statement [9]. Opposite to these explicit recommendations for the main analyses, the current CONSORT statement does not include clear recommendations on the use of specific effect measures for subgroup analyses. This, however, is remarkable as it has been acknowledged that subgroup analyses can lead to different results and conclusions with regard to statistical significance depending on whether relative or absolute effect measures are used [10]. To illustrate this phenomenon, we provide numerical examples based on RCTs performed by Dondorp et al. [11] (Appendix A at www.jclinepi.com) and Decousus et al. [12] (Appendix B at www.jclinepi.com).

As far as we are aware, no previous studies have been performed to investigate whether subgroup analyses are reported with relative or absolute effect measures and what the impact of such choices may be. We therefore systematically reviewed RCTs that were published in five major
general medical journals to assess whether relative or absolute effect measures were used in subgroup analyses and whether these subgroup effect measures differed from the main effect measures. We also studied whether conclusions would change if subgroup effects were calculated on a different scale than reported.

2. Methods

2.1. Selection of trials

We included all RCTs that were published in 2010 in five major general medical journals: *Annals of Internal Medicine (AIM)*, *British Medical Journal (BMJ)*, *Journal of the American Medical Association (JAMA)*, *Lancet*, and *New England Journal of Medicine (NEJM)*. These RCTs were retrieved using a search filter for PubMed that combined the journal names with publication date [pd] “2010” and publication type [pt] “randomized controlled trials” (Fig. 1). We included all RCTs irrespective of design (eg, parallel, factorial, crossover), study type (eg, superiority, equivalence, noninferiority), method of randomization, or sample size. Trials that were published online in 2010 but in article in 2011 were excluded. We also excluded research letters, cost-effectiveness analyses, diagnostic accuracy studies, studies that were not RCTs, and secondary analyses of RCTs.

2.2. Data extraction

We used a standardized data extraction form to assess the RCTs. This data extraction form was designed based on the five RCTs that were published in article in 2011. Two reviewers (R.P.V. and M.J.K.) independently extracted data from the included trials. Discrepancies between the reviewers were resolved by discussion. For trials with a dichotomous primary outcome, we extracted the reported effect measure for the main effect [RR, OR, HR, incidence rate ratio (IRR), RD, and incidence rate difference (IRD)], and determined whether results were statistically significant ($P \leq 0.05$). Additionally, we assessed whether these RCTs performed subgroup analysis by reviewing the methods and results sections (including tables and supplementary appendix) of these trials. If so, we investigated the number of subgroup analyses performed and whether relative or absolute effect measures (or both) were used. In addition, we assessed whether these trials used the appropriate statistical method to test whether treatment effect varies across the subgroup of interest, that is, whether tests for interaction were performed [2,13]. If possible, we extracted the crude data of the different subgroups to determine whether results and conclusions would change.

2.3. Sample size and data analysis

The decision to include all RCTs of 2010 was based on pragmatic considerations rather than formal sample size calculations. Frequencies and summary statistics of the extracted items were calculated. We used SPSS version 17 (SPSS Inc., Chicago, IL, USA) for these analyses.

If crude data of subgroups with two categories were reported, we calculated subgroup effects on both relative (ratio of RRs or ratio of IRRs across strata and 95% confidence interval (CI) and $P$-value) and absolute (difference of RDs or difference of IRDs across the subgroup strata and 95% CI and $P$-value) scales [14]. For further explanation of these calculations, see numerical example based on the study by Dondorp et al. [11] (Appendix A at www.jclinepi.com). For trials that used HRs as effect measure for the subgroup analyses and which reported only events and absolute numbers of patients across subgroups with two categories (ie, they did not report person-time of follow-up across the subgroups), we calculated the RR and the RD of both subgroup strata. Additionally, we calculated both the ratio of RRs and the difference of RDs across strata with their 95% CIs and $P$-values. For further explanation of these calculations, see numerical example based on the study by Decousus et al. [12] (Appendix B at www.jclinepi.com). For subgroups with more than two categories, we used Rothman Episheet version June 11, 2008 (http://www.druepi.org/dope-downloads/#Episheet) to derive the
The Mantel–Haenszel test for homogeneity for both relative and absolute scales. A change in conclusion was defined as a difference between relative and absolute subgroup effects with regard to statistical significance ($P \leq 0.05$) in one or more subgroups of the included RCTs.

3. Results

3.1. Characteristics of included trials

We retrieved a total of 361 records from our initial search, of which 327 were eligible for our analyses (Fig. 1). Most of the 327 RCTs were published in NEJM ($n = 124; 38\%$), followed by Lancet ($n = 84; 26\%$), BMJ ($n = 49; 15\%$), JAMA ($n = 47; 14\%$), and AIM ($n = 23; 7\%$). Most RCTs investigated the effect of medication ($n = 198; 61\%$), followed by surgical ($n = 32; 10\%$) and behavioral interventions ($n = 30; 9\%$). The median sample size of the RCTs was 499, ranging from 13 to 207,781 participants. A dichotomous primary outcome was reported in 229 of the 327 (70\%) trials. Of these 229 RCTs, 109 (48\%) trials were excluded because they did not perform subgroup analysis, leaving 120 RCTs for further analysis (Fig. 1).

3.2. Subgroup analysis

Subgroup analyses were more likely to be reported in trials without statistically significant main effects compared with trials with statistically significant main effects ($P = 0.04$). Subgroup effects were reported on a relative scale in 106 (88\%) of the 120 RCTs, whereas 9 (8\%) trials presented the subgroup effects on an absolute scale (Table 1). Two (2\%) of the 120 trials used both relative and absolute effect measures for subgroup analyses. The majority of the 120 RCTs ($n = 89; 74\%$) reported the same effect measure for the main and subgroup effects. However, eight (8/120; 7\%) trials that used an absolute effect measure (RD) for the main effect used a relative scale for their subgroup effects (Table 1). Most trials in which the effect measures for the main and subgroup effects differed used

---

Fig. 1. Flowchart.
ORs to report their subgroup effects (Table 1). Tests for interaction were performed in 82 (68%) of the 120 trials, in which subgroup analyses were reported.

### 3.3. Absolute vs. relative scale

In 59 (49%) of the 120 trials, no crude data on their subgroup analyses were presented, for example, trials only reported on the association measures for subgroup strata and/or the interaction terms in the results section without providing subgroup data in more detail (Table 2). Crude subgrouping data could be extracted in 41 (34%) of the 120 trials, whereas 20 (17%) RCTs that used HRs as measure of effect for subgroup analyses reported only events and absolute number of patients across subgroups (ie, they did not report person-time of follow-up across the subgroups).

Of the 41 trials from whom subgroup data could be extracted, 34 (83%) revealed similar subgroup effects on a relative and absolute scales regarding statistical significance (Table 2). In the other seven (17%) trials, the subgroup effects differed regarding statistical significance when using a relative or absolute effect measure. For five of these seven (71%) RCTs, the calculations revealed a nonstatistical significant subgroup effect on a relative scale and a statistically significant result on an absolute scale.

In 15 (75%) of the 20 trials that used HR as measure of effect for subgroup analysis and whom did not report person-time of follow-up across subgroups, relative and absolute subgroup effects were similar with regard to statistical significance (Table 2). In the other five (25%) trials, the subgroup effects revealed a statistically significant result on an absolute scale but not on a relative scale.

### 4. Discussion

We found that almost all RCTs used relative effect measures to report subgroup analyses even in the minority of trials in which main effects were presented with an absolute effect measure. Most trials reported an OR or HR as measure of subgroup effect. This may be explained by the fact that most researchers are familiar with logistic regression or Cox proportional hazard regression from which subgroup effects can be derived by putting the interaction term in the model. Especially, the frequent use of ORs as effect measure in RCTs is remarkable as the use of OR may lead to an overestimation of the effect compared with RR [15]. There are several methods available to estimate subgroup effects with RR as measure of effect, such as adding an

### Table 1. Reported effect measures for main and subgroup effects

<table>
<thead>
<tr>
<th>Effect measure</th>
<th>Main effect (%)</th>
<th>Subgroup effect (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk difference</td>
<td>14 (12)</td>
<td>9 (8)</td>
</tr>
<tr>
<td>Incidence rate difference</td>
<td>13 (11)</td>
<td>8 (7)</td>
</tr>
<tr>
<td>Relative scale</td>
<td>84 (70)</td>
<td>106 (88)</td>
</tr>
<tr>
<td>Relative risk</td>
<td>19 (16)</td>
<td>19 (16)</td>
</tr>
<tr>
<td>Odds ratio</td>
<td>11 (9)</td>
<td>24 (20)</td>
</tr>
<tr>
<td>Hazard ratio</td>
<td>53 (44)</td>
<td>62 (52)</td>
</tr>
<tr>
<td>Relative probability</td>
<td>1 (1)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Both relative and absolute scale</td>
<td>18 (15)</td>
<td>2 (2)</td>
</tr>
</tbody>
</table>

### Table 2. Subgroup analyses on relative and absolute scales based on crude data of 120 trials

<table>
<thead>
<tr>
<th>Subgroup effects calculated on both relative and absolute scale</th>
<th>No data available (%)</th>
<th>Crude data (%)</th>
<th>HR (%)a</th>
<th>Crude data and HR (%)b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change in conclusionc</td>
<td>n/a</td>
<td>7 (17)</td>
<td>5 (25)</td>
<td>12 (20)</td>
</tr>
<tr>
<td>Stat. sign. relative scale and non-Stat. sign. absolute scale</td>
<td>n/a</td>
<td>2 (5)</td>
<td>0 (0)</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Non-stat. sign. relative scale and Stat. sign. absolute scale</td>
<td>n/a</td>
<td>5 (12)</td>
<td>5 (25)</td>
<td>10 (16)</td>
</tr>
<tr>
<td>No difference</td>
<td>n/a</td>
<td>34 (83)</td>
<td>15 (75)</td>
<td>49 (80)</td>
</tr>
<tr>
<td>Total</td>
<td>59 (100)</td>
<td>41 (100)</td>
<td>20 (100)</td>
<td>61 (100)</td>
</tr>
</tbody>
</table>

Abbreviations: HR, hazard ratio; n/a, not available; Stat. sign., statistical significance; RCT, randomized controlled trial.

a HR: These RCTs used HRs as measure of effect for subgroup analyses and reported only events and absolute number of patients across subgroups (ie, they did not report person-time of follow-up across the subgroups). For these trials, we calculated relative subgroup effect as the ratio of relative risks (RRs) and absolute subgroup effect as the difference of risk differences (RDs) (see numerical example based on the study Decousus et al. [12]; Appendix B at www.jclinepi.com).

b Percentages do not sum to 100 because of rounding.

c Change in conclusion: it depends on the scale of subgroup analysis whether there is statistical significant heterogeneity of treatment effect among the subgroup (Appendices A and B at www.jclinepi.com). Appendix A, at www.jclinepi.com, describes a numerical example of calculations with crude data of the RCT of Dondorp et al. [11], in which there is no statistically significance on a relative scale (ratio odds ratios $P = 0.21$ and ratio RRs $P = 0.27$), whereas there is statistically significance on absolute scale (difference of RDs $P = 0.02$).
interactions in a log-binomial regression [16] or by calculating the ratio of RRs [14].

Moreover, we demonstrated that results and conclusions changed with regard to statistical significance in 17% of the 41 trials when subgroup effects were calculated on a different scale than reported. Subgroup analyses that reveal statistically significant relative but nonstatistically significant absolute effects could be accompanied by only small absolute treatment effects and may therefore have limited clinical relevance. For trials with nonstatistically significant subgroup effects on a relative scale and statistical significant results on an absolute scale, we found both small and large absolute treatment effects among subgroups (Appendices A and B at www.jclinepi.com) [11,12]. The clinical impact of such subgroup findings are, however, also highly dependent on (the severity of) the primary outcome of interest. This phenomenon is illustrated by the numerical example of the study by Dondorp et al. (Appendix A at www.jclinepi.com). Although the absolute effect size might be judged modest (difference of RDs, 4.5%; 95% CI: 0.9, 8.1), this finding may be clinically relevant because of the severity of the primary outcome, that is, in-hospital mortality. This subgroup finding would not have been detected when results were only presented with relative subgroup effects. As a consequence, reporting of both relative and absolute subgroup effect measures is crucial for determining the clinical impact of subgroup results. Absolute subgroup effects (difference of RDs) can easily be derived using Rothman Episheet version June 11, 2008 (http://www.drugepi.org/dope-downloads/#Episheet).

To our knowledge, this is the first study to investigate whether subgroups are reported as relative or absolute effect measures. Moreover, we studied the impact of such choices by calculating subgroup effects on a different scale than reported. To enhance validity, we performed a systematic literature search and included all RCTs that were published in five major general medical journals in 2010.

To appreciate our results, some potential limitations should also be discussed. First, inclusion of RCTs in our study was restricted to trials that were published in 2010 in five major medical journals, and our results may therefore not be generalizable to trials published in less prominent journals. Second, we might have missed some RCTs with our PubMed search syntax. It is, however, unlikely that this would have affected our results because these RCTs are likely to be missing at random. Third, we have included data from trials that reported HRs as effect measure for subgroup analysis and which reported only events and absolute numbers of patients across subgroups (ie, did not report person-time of follow-up across the subgroup). Results of these calculations cannot be directly linked to the results as presented in the articles. Our aim, however, was to investigate whether results would change if subgroup effects were calculated on both relative and absolute scales. Because we were able to derive both relative (ratio of RRs) and absolute (difference of RDs) subgroup effects, we do not consider this as a drawback. Finally, we pragmatically used a difference in statistical significance ($P \leq 0.05$) to conclude whether relative and absolute subgroup effects differed. Although frequently used in medical science, the use of a $P$-value of $\leq 0.05$ for statistical significance is arbitrary. Furthermore, not only the statistical significance but also the magnitude of the reported effect size and the severity of the primary outcome of interest are of crucial importance when translating research findings into practice as illustrated by the numerical examples (Appendices A and B at www.jclinepi.com).

5. Conclusion

Almost all RCTs that are currently published in high impact journals use relative effect measures for reporting subgroup analyses. In 17% of the 41 trials, conclusions changed when subgroup effects were calculated on a different scale than reported. Because of the potential for relative estimates to lead to misinterpretations of the absolute value of benefit, a strong argument could be made that the reporting of RR reduction should always be accompanied by presentation of absolute risk reductions. We therefore recommend to incorporate such recommendations in the CONSORT statement not only for primary and secondary outcomes but also for subgroup analyses.

Acknowledgments

Authors’ contribution: All authors had full access to all the data (including statistical reports and tables) in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design was contributed by R.P.V., M.M.R., and M.J.K.; data extraction and acquisition of data was contributed by R.P.V. and M.J.K.; analysis and interpretation of data was contributed by R.P.V., M.M.R., A.W.H., and M.J.K.; drafting the manuscript was done by R.P.V.; critical revision of the manuscript for intellectual content was done by M.M.R., A.W.H., M.J.K.; and study supervision was performed by M.M.R., A.W.H., and M.J.K. All the authors provided final approval for submission.

Appendix

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2013.11.003

References


