The European Journal of General Practice

ISSN: 1381-4788 (Print) 1751-1402 (Online) Journal homepage: https://www.tandfonline.com/loi/igen20

The importance of effect sizes

Sil Aarts, Marjan van den Akker & Bjorn Winkens

To cite this article: Sil Aarts, Marjan van den Akker & Bjorn Winkens (2014) The importance of effect sizes, The European Journal of General Practice, 20:1, 61-64, DOI: 10.3109/13814788.2013.818655

To link to this article: https://doi.org/10.3109/13814788.2013.818655

Published online: 30 Aug 2013.

European Journal of General Practice, 2014; 20: 61–64

informa
healthcare

Methodological Paper

The importance of effect sizes

Sil Aarts[1], *Marjan van den Akker*[2,3] *& Bjorn Winkens*[4]

[1]Department of Allied Health Professions, Fontys University of Applied Sciences, Eindhoven, The Netherlands, [2]Department of Family Practice, School for Public Health and Primary Care: Caphri, Maastricht University, The Netherlands, [3]Department of Family Practice, Katholieke Universiteit Leuven, Belgium, and [4]Department of Methodology and Statistics, School for Public Health and Primary Care: Caphri, Maastricht University, The Netherlands

KEY MESSAGE:

- Statistical significance testing alone is not the most adequate manner to evaluate if there is indeed a clinically relevant effect.
- Effect sizes should be added to significance testing.
- Effect sizes facilitate the decision whether a clinically relevant effect is found, helps determining the sample size for future studies, and facilitates comparison between scientific studies.

Keywords: standardized effect sizes, unstandardized effect sizes, statistical testing, significance

INTRODUCTION

In a previous article in this journal, we discussed the use and misuse of statistical testing in scientific medical research articles (1), stating that significance testing should only be used when generalising from a sample to the population, i.e. the sample data is used to generalize conclusions for all members of the population under study. More importantly, significance tests conducted with either too few or too many participants can be misleading (2,3). That is, studies that include too few participants, lack statistical power to detect a clinically relevant effect. In contrast, in studies with large sample sizes, even small effects are likely to be evaluated as statistically significant while these effects might lack any clinical relevance. These disadvantages pinpoint that testing statistical significance alone is not the most adequate manner to evaluate if there is indeed a clinically relevant effect. Significance testing conveys little information on the *size* of an observed effect, i.e. 'how large is the obtained effect?' For example, statistical testing and corresponding *P*-values make it impossible to evaluate if an effect obtained in study A is smaller or larger than an effect evaluated in study B.

Editorial boards of peer-reviewed journals have tried to encourage researchers to complement statements regarding statistical significance with more clinically meaningful results, i.e. results independent of sample size and/or measurement scale (4). Specifically, so-called effect sizes, a manner to quantify the magnitude of an effect (2,5), are recommended in addition to significance testing. Hence, the question that medical researchers should focus on is not so much 'is there a treatment effect?' but rather 'how large is the treatment effect?'

The current article is aimed at informing medical researchers on the concept of 'effect size.' This article is not intended to provide an exhaustive overview of all effect sizes, but is merely aimed at discussing the advantages and disadvantages of this concept and the most frequently used types thereof.

WHY EFFECT SIZES?

Since statistical testing and, more specifically, *P*-values are affected by the sample size of a study, the magnitude of an effect cannot be estimated using statistical hypothesis testing alone (i.e. 'significant' or 'not significant').

Correspondence: S. Aarts, Department of Allied Health Professions, Fontys University of Applied Sciences, PO Box 347, 5600 AH Eindhoven, The Netherlands. E-mail: s.aarts@fontys.nl

(Received 19 April 2013; accepted 2 June 2013)

ISSN 1381-4788 print/ISSN 1751-1402 online
DOI: 10.3109/13814788.2013.818655

Consider two fictitious studies in which therapies A and B for depression are compared to a placebo group. Both studies conclude that in 80% of the included patients depression was absent after the therapy (i.e. outcome measure: "depression absent" vs. "depression present"), whereas in both studies "only" 20% of the placebo group showed absence of depression. Hence, the two therapies show the same clinical effect, i.e. both therapies show a relative risk of therapy versus placebo of 4. However, the first study, evaluating therapy A, included 100 patients and 100 controls while the second study, evaluating therapy B, included only 20 individuals in each group. Since *P*-values heavily depend on the sample size used (1,6,7), the *P*-value of the first study will be much smaller than that of the second study. One could then easily conclude that therapy B is inferior to therapy A, but, as shown from this example, this conclusion would be utterly false. The addition of "effect sizes" can be used to overcome this problem.

An effect size refers to the magnitude of a result. There are several methods to calculate the size of an effect (8). The term effect size can refer to unstandardized effect sizes (e.g. the difference between group means, relative risk or odds ratio) or standardized effect sizes (such as 'correlation' or 'Cohen's d').

UNSTANDARDIZED EFFECT SIZES: RELATIVE RISK AND ODDS RATIO

Many research articles aim at describing the possible benefits of a (new) treatment or therapy on a binary outcome variable, i.e. presence or absence of disease. These articles often report their results ('effects') in terms of Relative Risk (RR) or Odds Ratio (OR). Since there is some tendency to think that these two effect sizes are similar, they are often, incorrectly, used interchangeably (9). Before describing RR and OR, we first need to comprehend the term 'absolute risk.' Absolute risk is a risk that is stated without comparison to any other risk, it is merely a probability that a certain event occurs, e.g. 'the risk of an individual developing depressive disorder is 5%.'

Relative risk

Relative risk (RR) is a comparison between different risk levels and is often used in prospective studies (e.g. cohort studies and experimental studies) (10,11). That is, RR is the ratio of the probability of some event to occur in one group to the probability of that event occurring in another group (11). For example, if the relative risk for depression for women compared to men is 1.14, this means that women are 14% more likely to develop depression than men. A way of defining this is

$$RR = \frac{risk_1}{risk_2} = \frac{\frac{a_1}{n_1}}{\frac{a_2}{n_2}},$$

where *RR* is the relative risk of group 1 versus group 2, a_1 and a_2 are the number of events within group 1 and 2, and n_1 and n_2 are the total number of subjects within group 1 and 2, respectively. Hence, RR is the ratio of the proportion of those exposed who did develop the outcome or condition to the proportion of those not exposed who did develop the outcome or condition.

Odds ratio

The odds ratio (OR) is a ratio of two odds. An 'odds' is calculated as the number of events (e.g. diseased) divided by the number of non-events (e.g. healthy) within a group. The OR is then calculated by dividing the odds of the one group (e.g. the treatment group) by the odds of the other group (e.g. the control group), as shown in the following formula;

$$OR = \frac{odds_1}{odds_2} = \frac{\frac{a_1}{b_1}}{\frac{a_2}{b_2}},$$

where *OR* is the odds ratio of group 1 versus group 2, a_1 and a_2 are the number of events ('disease') within group 1 and 2, respectively, and b_1 and b_2 are the number of non-events ('healthy') within group 1 and 2. Often, ORs are used in retrospective studies to provide an approximation of the relative risk.

Odds ratios might be hard to comprehend intuitively (i.e. an OR of six does not mean 'six times more likely to experience the outcome at hand'). Consequently, odds ratios are often interpreted as being equivalent to the relative risk. However, the OR is only similar to the RR when the initial risk of an event (i.e. the prevalence of the outcome at hand) is low. If the prevalence of a certain condition or event increases, a larger difference between the RR and OR becomes apparent.

Consider the following hypothetical situation: women develop depression in 80% of the time while men develop depression 70% of the time. The odds is therefore 4 for women (80/20) and 2.3 for men (70/30). Hence, the odds ratio of women compared to men will be 1.7 (4/2.3). However, the conclusion that women are 70% more likely to develop depression would be overestimating the risk of developing depression for women, since the relative risk is "only" 1.14 (80/70).

This example shows that interpreting an OR as if it was a RR is false when it results in concluding that an

effect size is bigger than it actually is. This does not mean that RRs are superior to ORs or vice versa. Both risks calculations are measured on a different 'scale,' with both having their own positive and negative elements and are often used in different study designs.

STANDARDIZED EFFECT SIZES: CORRELATION AND COHEN'S D

Standardized effect sizes are preferred over unstandardized effect sizes when studies using different measurement scales, are being compared; standardized effect sizes facilitate the comparison between studies (e.g. in meta-analysis) (12).

Correlation

The effect size that medical researchers might be most familiar with is correlation. Correlation determines the extent to which two numerical variables are 'proportional' to each other show a linear dependency. For example, cholesterol level (mmol/l) correlates with age, i.e. cholesterol generally increases when people get older. There are several correlation coefficients, of which the Pearson correlation coefficient ('r') is the most common one (13). This Pearson correlation coefficient is obtained by dividing the covariance (s_{xy}) of two variables by the product of their standard deviations (s_x and s_y):

$$r = \frac{S_{xy}}{S_x S_y}.$$

Correlation ranges from 'no linear relation' ($r = 0$) to a 'perfect linear relation' ($r = 1$ indicating a perfect positive linear relation and $r = -1$ indicating a perfect negative linear relation). In other words, if you plot the values of the two variables against each other (scatterplot), a correlation is high if it is possible to draw an ellipse around the points that can be approximated by either an upward or a downward straight line. A significant correlation between two variables does not necessarily imply a strong association, e.g. if the sample size is 50, Pearson's correlations of 0.28 or higher are found to be statistically significant while a correlation < 0.30 is rarely considered a strong association (13,14).

Cohen's d

'Cohen's d' can be used when comparing the mean value of a numerical variable between two groups. For example, the mean cholesterol level (mmol/l) of patients with diabetes versus patients without diabetes or the mean score on a depression questionnaire of patients with cardiovascular disease versus without cardiovascular disease) (5). It indicates the standardized difference between two means, $\bar{y}_1 - \bar{y}_2$, expressed in standard deviation units, i.e. divided by the (pooled) within-group standard deviation of the data at hand, s_y. The formula for calculating Cohen's d is:

$$d = \frac{\bar{y}_1 - \bar{y}_2}{s_y}.$$

Of course, the interpretation of the size of Cohen's d needs to occur within the context of the study at hand, but it has been suggested that a value of 0.2 or less should be considered a small effect, a value between 0.2 and 0.5 as a medium effect size, and a value of 0.8 or larger as a large effect (4,5). This implies that, although an observed effect might be statistical significant, it might still be trivial when linked to a Cohen's d of 0.1 (5). Cohen's d is also frequently used in sample size calculations, where a lower Cohen's d indicates the need for a larger sample size (10).

DISCUSSION AND RECOMMENDATION

As stated earlier, a statement 'this is a significant effect' can be very misleading since statistical significance depends on the size of the effect, the number of participants in the sample, the research design and the statistical test being employed (2–4). Hence, even trivial effects can become statistically significant, and vice versa, i.e. clinically relevant effects need not be statistically significant. Apart from statements regarding statistical significance, researchers also need to report the obtained effect sizes.

Effect sizes, including the abovementioned examples thereof, enable medical researchers to measure the strength of a relationship between variables. In contrast to significance testing, (standardized) effect size gives medical researchers the opportunity to compare treatments or therapies reported in, for example, various randomized controlled trials (2). However, researchers should keep in mind that comparing effect sizes (e.g. between studies evaluating various therapies) is only useful if the manner in which these effects sizes are calculated, are comparable. That is, studies that differ substantially regarding design or used methodology could complicate the comparison of effect sizes and could, therefore, easily lead to false conclusions.

Given that reporting effect sizes facilitates the interpretation of the results of medical studies, medical researchers are highly encouraged to present effect sizes to provide an answer to the question 'how large is the effect?.' With the knowledge and expertise of medical professionals regarding their conducted research and area of expertise, the researcher can then determine, using these effect sizes, whether an observed effect is clinically relevant or not.

In conclusion, (standardized) effect sizes should be added to significance testing to be able to decide whether a clinically relevant effect is found, to help determine the sample size for a possible future study, and to facilitate comparison between studies in meta-analyses.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

REFERENCES

1. Aarts S, Winkens B, van den Akker M. The insignificance of statistical significance. Eur J Gen Pract. 2012;18:50–2.
2. Coe R. It's the effect size, stupid: What effect size is and why it is important. Annual Conference of the British Educational Research Association; University of Exeter, England; 2002.
3. Kirk RE. Practical significance: A concept whose time has come. Educ Psychol Meas. 1996;56:746–59.
4. Cohen J. The earth is round ($p < .05$). Am Psychologist 2004;49: 997–1003.
5. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, New Jersey: Lawrence Erlbaum Associates; 1988.
6. Estes WK. Significance testing in psychological research: Some persisting issues. Psychol Sci. 1997;8:18–20.
7. Johnson DH. The insignificance of statistical significance testing. J Wildl Manage. 1999;63:763–72.
8. Valentine JC, Cooper H. Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes. Washington, DC: What Works Clearinghouse; 2003.
9. Schmidt CO, Kohlmann T. When to use the odds ratio or the relative risk? Int J Public Health. 2008;53:165–7.
10. Imbos TJ, Berger MPF, Janssen MPE. Methodologie en statistiek I. Universiteit Pers Maastricht FdG, editor. Maastricht: Datawyse; 1996.
11. McHugh ML. Scientific inquiry: Clinical statistics for primary care practitioners: Part II-absolute risk reduction, relative risk, relative risk reduction, and number needed to treat. J Spec Pediatr Nurs. 2008;13:135–8.
12. Kline RB. Beyond significance testing: Reforming data analysis methods in behavioral research. Washington DC: American Psychological Association; 2004.
13. Rodgers JL, Nicewander WA. Thirteen ways to look at the correlation coefficient. Am Stat. 1988;42:59–66.
14. Wikipedia. Correlation and dependence. Available at: http://en.wikipedia.org/wiki/Correlation_and_dependence (accessed 18 February 2013).