REVIEW

To test or to estimate? *P*-values versus effect sizes

Daniela Dunkler[1], Maria Haller[1,2], Rainer Oberbauer[3] & Georg Heinze[1]

1 Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria
2 Ordensklinikum Linz, Elisabethinen, Nephrology, Linz, Austria
3 Department of Medicine III, Medical University of Vienna, Vienna, Austria

Correspondence
Georg Heinze, Center for Medical Statistics, Informatics, and Intelligent Systems, Section for Clinical Biometrics, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria.
Tel.: +43-1-40400-66890;
Fax: +43(1)4040066870;
e-mail:
georg.heinze@meduniwien.ac.at

SUMMARY

Most research in transplant medicine includes statistical analysis of observed data. Too often authors solely rely on *P*-values derived by statistical tests to answer their research questions. A *P*-value smaller than 0.05 is typically used to declare "statistical significance" and hence, "proves" that, for example, an intervention has an effect on the outcome of interest. Especially in observational studies, such an approach is highly problematic and can lead to false conclusions. Instead, adequate estimates of the observed size of the effect, for example, expressed as the risk difference, the relative risk or the hazard ratio, should be reported. These effect size measures have to be accompanied with an estimate of their precision, like a 95% confidence interval. Such a duo of effect size measure and confidence interval can then be used to answer the important question of clinical relevance.

Transplant International 2020; 33: 50–55

Key words
clinical significance, effect size measure, statistical inference, statistical significance, statistical tests

Received: 8 July 2019; Revision requested: 24 July 2019; Accepted: 23 September 2019; Published online: 21 October 2019

The weather will change significantly in the next days (P = 0.04). In a weather report, a comment like this is would be inconceivable, but such statements can be found in many scientific articles. The statement does not contain the information that a recipient is actually interested in: will it get warmer or colder? How much change in weather do we have to expect? The *P*-value derived by a statistical test does not answer these questions. It leaves only a vague feeling that the weather may not stay the same. However, these questions can be answered satisfactorily by reporting an adequate measure of the size of the effect, in this example, the expected change in temperature.

Current practice in transplant research

In order to evaluate the current practice of reporting *P*-values and effect sizes in transplant research, we reviewed all manuscripts published in Transplant International in 2018 in the category "clinical research" (Table 1). Among 68 summaries of retrospective studies, in 27 (40%) *P*-values and measures of effect size were reported, while in 30 (44%) only *P*-values were given.

P-values and effect size measures

In transplant research, as in any other scientific discipline, using a *P*-value as a measure of "difference" is entirely uninformative and can even be misleading [1–8]. (For definitions on the concepts of statistical testing and estimation, see Table 2.) As an example, consider a study comparing the 5-year survival after kidney transplantation between two interventions or exposures. Generally, a *P*-value depends on two quantities: the observed difference between the groups and the sample size. With a sample size of 100 in each group, assuming 5-year survival probabilities of 85% and 90% in the two

doi:10.1111/tri.13535

Table 1. Review of all manuscripts published in Transplant International in the category "clinical research" in 2018

N = 83 manuscripts published in Transplant International in the category 'clinical research' in 2018.

N = 4 excluded,
because the study included no statistical analysis, or
if a prediction model was developed.

	Retrospective studies N = 68 (86%) incl. observational (n=56), pilot (n=2), registry (n=10) studies	**Prospective studies** N = 11 (14%) incl. observational (n=5), pilot (n=1), single-arm trial (n=1) survey (n=1), RCT (n=3) studies
Median sample size (min, max)	278 (8, 166776)	133 (40, 500)
Sample size/power calculation reported?	3 (4%)	3 (27%)

In the Summary: Reporting of …

P-value or statistical significance	Effect size &CI		
No	No	8 (12%)	3 (28%)
No	Yes	3 (4%)	0 (0%)
Yes	No	30 (44%)	4 (36%)
Yes	Yes	27 (40%)	4 (36%)
Clinical relevance		1 (1%)	0 (0%)

Random sample of 10 retrospective studies
incl. observational (n=8), registry (n=2).

Median sample size (min, max)	885 (41, 62961)
Median number of p-values reported (min, max)	50 (9, 272)
…in the Summary	3 (0, 12)
…in the Results	14 (0, 27)
…in the Tables & Figures	34 (0, 247)

Results are given separately for retrospective and prospective studies. We evaluated if the summary of a manuscript mentioned at least one measure of effect size with a measure of variability (usually the 95% confidence interval) and if at least one *P*-value or a mention of statistical (in)significance was given. For a random sample of ten retrospective studies, we counted the number of reported *P*-values in the Summary, the Results and in the Tables and Figures. CI, confidence interval; RCT, randomized clinical trial.

groups, and with complete follow-up, the *P*-value results as 0.39. The statistically nonsignificant result with $n = 100$ could be falsely interpreted as evidence for lack of a difference between the groups "proving" the null hypothesis that the intervention makes no difference in survival. Researchers more aware of the fallacies in interpreting *P*-values would—more cautiously and correctly—verbalize the result as "we could not find any evidence suggesting a difference between the groups." However, including 1000 patients per group into the study and observing the same survival probabilities in both groups, one would compute a *P*-value of 0.0009 and conclude the opposite: strong evidence against the null hypothesis of no difference between the groups. Still, in both examples, the observed 5-year mortality risks in the two groups are 15% and 10% and thus the relative risk is 1.5 (=0.15/0.10), which is usually considered a strong effect. In the medical field, a relative risk of 1.5 of a hard outcome such as mortality or graft loss is almost never achievable with a single intervention. Therefore, prospective clinical trials are planned expecting much smaller differences in outcome such as relative risks of 1.25. In this example, solely the sample size will determine conclusions about the effect of an intervention and lead to contradictory conclusions if those are based only on the strict dichotomy of statistical significance or nonsignificance; but consider that the 100 observations could just be a subsample of the 1000.

The null hypothesis implies that the effect is exactly zero. In observational studies, this is hardly ever the case, and even in prospective controlled trials, exact equality of outcomes is unlikely. If there is even only a small difference which will typically have no clinical relevance, then a large enough sample will detect it and reject the null hypothesis (see most cardiovascular trials where the sample size is usually several thousand). In reaction to this unfortunate paradox, it has been proposed to avoid the notion "statistically significant" or "statistical significance" in observational research completely [1]. Statements on statistical significance for the primary outcome should be confined to randomized clinical trials, where the sample size is prespecified to detect a minimal clinically relevant difference.

Generally, *reporting the effect size is much more informative than a statement on statistical significance*. In the example, one could report the mortality risk difference or the relative risk. With both sample sizes, the same conclusions would then be drawn: The absolute mortality risk difference is 5%, and the relative risk is 1.5 [9].

Quantifying uncertainty of estimates

To address the uncertainty that is attached to these estimates of effect size, they should always be accompanied by 95% confidence intervals. The intervals would clearly reflect that a more precise estimate can be obtained with more data. In our example, the 95% confidence intervals for the mortality risk difference for sample sizes 100 and 1000 range from -4.1% to 14.1% and from 2.1% to 7.9%, respectively. These confidence intervals do not contradict each other. By contrast, the confidence interval obtained from $n = 100$ entirely includes the interval for $n = 1000$, that is, what already can be concluded from $n = 100$ can be made more precise with the larger sample.

Emphasizing clinical relevance

Effect size measures and confidence intervals put the emphasis on aspects of *clinical relevance* (also denoted by "clinical significance") compared to the concept of statistical significance assessed by *P*-values. Clinical relevance is determined depending on whether a difference is "real and noticeable" by individuals [10]. Effect size measures like the absolute or relative risk difference address this issue directly. A seeming drawback of the concept of clinical relevance is that its threshold must be determined on a case-by-case basis, as it depends on values and preferences patients and healthcare professionals attach to the outcome under investigation, or on existence of other effective treatments. Unlike statistical significance, clinical relevance is never determined by the data. Suppose that a mortality risk difference of 2.5% would constitute the minimal clinically relevant difference, corresponding to a number needed to treat of 40 (=1/0.025). While the analysis with $n = 1000$ was adequately powered to detect a small difference in mortality risk and excluded parity, it cannot exclude a difference less than 2.5%, as this value is covered by the confidence interval. Hence, such an analysis could not "prove" clinical relevance.

Some practical recommendations

In line with the statistical guidelines of this journal and currently effective reporting guidelines [11], we propose that *any study that compares outcomes between two groups should report an appropriate estimate of effect size and provide a 95% confidence interval as a measure of precision. These measures should be interpreted regarding*

Table 2. Simplified definitions of selected concepts of statistical testing and estimation

Some key ingredients to statistical testing	
Null hypothesis	States that two different interventions (or exposures) lead to the same outcome, that is, that the effect size is 0 if expressed as a difference, or 1 if expressed as a ratio.
Alternative hypothesis	States that two different interventions lead to different outcomes, that is, that the effect size is not equal to 0 if expressed as a difference, or not equal to 1 if expressed as a ratio.
Statistical test	Depending on the research question (e.g., scale of the outcome), various statistical tests, like a t-test, are available. A test statistic measures the "distance" between the data and the null hypothesis. A test is only valid if its underlying assumptions are met. These assumptions do not only encompass direct assumptions of the test, like approximate normal distribution in the case of a t-test, but also assumptions about the conduct of the study, like random selection of subjects and treatment or that no interim analyses were conducted.
P-value	To facilitate interpretation and comparison, a test statistic is usually transformed to a probability scale and expressed as a *P*-value. The *P*-value measures the compatibility of the observed data with the null hypothesis. Technically, it expresses the probability with which, given the null hypothesis was true, data with an effect size as extreme as the observed one or more extreme than the observed one can be obtained. The *P*-value cannot separate implausibility of the null hypothesis from implausibility of any of the assumptions: A small *P*-value gives evidence that the data are not compatible with the specified model—encompassing the null hypothesis and all assumptions. Hence, the *P*-value should be viewed as a continuous measure of compatibility of the data to the model ranging from 0 (complete incompatibility) to 1 (complete compatibility) [1]. Consequently, precise *P*-values should be presented (e.g., $P = 0.07$ and not P = NS or $P > 0.05$).
Key ingredients to estimation	
Effect size estimate	Expresses the expected difference or ratio in the outcome between two interventions.
Confidence interval	Expresses the imprecision of an estimate of effect size that arises from a limited sample size. Technically, when the study could be repeated very often and the confidence level is set to 95%, then 95% of the confidence intervals computed on the study repetitions will cover the true effect size.
Clinical relevance	Based on the effect size estimate and confidence interval in addition to subject matter knowledge and other published results, a researcher can finally answer the question of clinical relevance "Are observed differences between the two study groups large enough to be of clinical significance?"

For methodologically correct definitions, we refer to Greenland *et al*. [2]. For information on statistical testing, we refer to textbooks on statistics, for example, Agresti *et al*. [14].

their clinical relevance and should be compared to reported effect size measures stemming from similar studies. Table 3 describes such effect size measures for group comparisons of outcomes typically investigated in transplantation research. An adequate effect size measure is selected depending on the scale of measurement of the outcome, that is, whether it is a continuous, binary, or time-to-event variable. For all presented effect size measures, 95% confidence intervals can be estimated. To conveniently obtain effect size measures with accompanying 95% confidence intervals for two proportions or two survival rates at a specific time point, our online calculator https://biometrician.shinyapps.io/effectsizeci/ can be used.

If more than two groups should be compared, separate effect size measures, each comparing one group with the reference group, can be computed. Typically, the reference group is chosen as the group receiving the standard therapy, or the group of patients with the most common characteristic. Regression models can be used to obtain effect size measures that are generalized to continuous variables. Consider, for example, that we would like to express the excess mortality risk that is associated with a 10-year increase in donor age. A Cox regression model could be used to estimate the association of donor age with survival and may come to the conclusion that the hazard ratio per 10-year increase in donor age is 1.2. In this example, a hazard ratio of 1.2 per 10 years is preferred to a hazard ratio of 1.02 per year. Effect sizes for categorical variables should be unambiguously reported, for example, as a hazard ratio of 1.3 for males vs. females, not just as a hazard ratio of

Table 3. Some commonly used effect size measures to compare two interventions in transplantation research

Scale of the outcome	Examples	Effect size measures	Example of interpretation "If intervention 1 is compared to intervention 2, . . ."*	Statistical method for generalization
Continuous	Glomerular filtration rate, glucose level	Difference of means	". . .the expected difference in glomerular filtration rate is 5 mL/min/1.73 m^2."	General linear model (linear regression, ANCOVA)
Binary within a fixed, fully observable time frame	Complications during transplantation, delayed graft function	Risk difference p_1-p_2	". . . in 5% of all people the occurrence of a complication during transplantation could be avoided."	Risk prediction after logistic regression
		Relative risk (RR) p_1/p_2	". . . the probability of the occurrence of a complication during transplantation multiplies by 1.25."	Poisson regression (for rare outcome events)
		Odds ratio (OR) $Odds_1/Odds_2$	". . . the odds of the occurrence of a complication during transplantation multiples by 1.5." [$Odds_1 = p_1/(1-p_1)$]	Logistic regression
Binary within varying follow-up time	Incidence of acute rejection episodes	Incidence rate ratio	". . . the expected number of acute rejection episodes per patient year multiplies by 1.15."	Poisson regression
Time-to-event	Patient survival, graft survival	Survival difference at t years, $S_2(t)-S_1(t)$	". . . in 7% of all people graft loss within the first two years could be avoided."	Survival estimation after Cox regression
		Hazard ratio (HR)	". . . the instantaneous mortality multiplies by 1.2."	Cox regression

The choice of effect size measure depends on the scale of the outcome. Examples of correct interpretations for comparison of two interventions are given. Statistical methods to generalize the analysis for adjustment for potential confounders or continuous exposure variables are presented. p_1, p_2, the observed event rates after intervention 1 or 2; $S_1(t)$, $S_2(t)$, the observed survival proportions at t years after interventions 1 and 2.

*For comparing exposures, change to "If exposed individuals are compared to unexposed individuals, . . ."

1.3 for gender. Regression models can also be used to adjust the effect size measure for potential confounders. A confounder is a variable that is associated with the risk factors (e.g., donor age) and causally related to the outcome. In our example, the estimated glomerular filtration rate of the donor could assume the role of a confounder. Adjustment for confounders is especially important in observational studies where patients are not randomized and hence their characteristics likely have influenced the treatment decision [12]. It should always be stated if effect size estimates obtained from a regression model are unadjusted or adjusted. If they are adjusted, the adjustment variables should be stated. With continuous outcome variables, the scale of their measurement has to be reported. Generally, the International System of Units (SI units) should be preferred over other systems.

Detailed guidance on reporting of results from observational and randomized studies and other study types can be found on the website of the "Enhancing the Quality and Transparency of Health Research (EQUATOR) network" (http://www.equator-network.org, accessed 08 July 2019) [11,13]. The EQUATOR network website links to published reporting guidelines for any type of study in any medical discipline. Summarizing, *instead of solely relying on P-values to answer their research questions, authors are encouraged to present adequate effect size measures accompanied with 95% confidence intervals*.

Authorship

DD, GH: wrote the paper. MH, RO: critically revised the paper.

Transplant International 2020; 33: 50–55

Funding

The authors have declared no funding.

Conflicts of interest

The authors have declared no conflicts of interest.

REFERENCES

1. Amrhein V, Greenland S, McShane B. Retire statistical significance. *Nature* 2019; **567**: 305.
2. Greenland S, Senn SJ, Rothman KJ, *et al*. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016; **31**: 337.
3. Greenland S, Poole C. Living with statistics in observational research. *Epidemiology* 2013; **24**: 73.
4. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p<0.05". *Am Stat* 2019; **73**: 1.
5. Goodman S. A dirty dozen: twelve p-value misconceptions. *Semin Hematol* 2008; **45**: 135.
6. Greenland S. Null misinterpretation in statistical testing and its impact on health risk assessment. *Prev Med* 2011; **53**: 225.
7. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat* 2016; **70**: 129.
8. Chakkera HA, Schold JD, Kaplan B. P Value: Significance Is Not All Black and White. *Transplantation* 2016; **100**: 1607.
9. Noordzij M, van Diepen M, Caskey FC, Jager KJ. Relative risk versus absolute risk: one cannot be interpreted without the other. *Nephrol Dial Transplant* 2017; **32**: ii13.
10. Kendall PC. Clinical significance. *J Consult Clin Psychol* 1999; **67**: 283.
11. Vandenbroucke JP, von Elm E, Altman DG, *et al*. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Med* 2007; **4**: e297.
12. Jager KJ, Zoccali C, MacLeod A, Dekker FW. Confounding: What it is and how to deal with it. *Kidney Int* 2007; **73**: 256.
13. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *PLoS Med* 2010; **7**: e1000251.
14. Agresti A, Franklin CA, Klingenberg B. *Statistics: The Art and Science of Learning from Data*, 4th ed. Upper Saddle River, New Jersey: Pearson, 2017.