Nerea Fuertes Cayón

Statistical significance and Clinical relevance

1. Introduction:

Statistics guide researchers in gaining a clearer understanding of data and making conclusions from it. Thus, it is crucial to fully comprehend what statistical procedures and their results mean in the context of the study. In other words, in clinical research, it is essential to consider both statistical significance and clinical relevance when interpreting study findings.[1]

However, statistical significance and clinical relevance are concepts that are commonly mixed. Statistical significance measures the chance of study outcomes being random. Conversely, clinical relevance focuses on the size of the treatment effect, indicating whether the trial results are likely to have a substantial impact on current medial practices.[2] Consequently, even though there are accepted criteria for statistical significance, a similar framework is absent when it comes to assessing clinical significance. In many cases, it is the clinician's judgment, along with input from the patient, what determines whether a result holds clinical significance or not.[3]

2. p-value:

As mentioned, there is an accepted criterion to measure statistical significance. Statistical significance is usually evaluated using the *p*-value. This statistical measure quantifies the probability of obtaining observed results when the null hypothesis (H_0) is true. To understand *p*-value, first is crucial to know that H_0 denies the existence of a relationship between two variables, while the alternative hypothesis (H_a) states that these variables are correlated. Usually, it is interesting to reject H_0 to favour H_a, as it allows for assessing the correlation of variables to confirm theoretical expectations.[4]

Therefore, in the context of a null hypothesis, the p-value reflects the likelihood of the observed data being inconsistent with the assumptions made by a model. A smaller p-value indicates increased statistical inconsistency with the null hypothesis, given that the assumptions for calculating the p-value are valid. Also, *p*-value is not a measure that can state the truth of a null hypothesis or state the probability that the obtained data is due to random chance; it is a statement about data in connection with a suggested explanation, not about the explanation itself.

On one hand, reducing the analysis of the data to this value can result in mistaken beliefs and unwise decision-making. Investigators or analysts should consider the context of the study when drawing conclusions, such as the study's design, the quality of measurement, what other evidence exists for the study, and if the assumptions behind the data analysis are valid. And on the other hand, relying on

"statistical significance," often interpreted as p≤0.05, to support a scientific claim or imply truth significantly distorts the scientific process. So, it is evident that decisions regarding scientific conclusions should not rely solely on considering *p*-value or if the *p*-value surpasses or not a fixed value.

When presenting research findings, there is a risk of data dredging, also known as data snooping or p-hacking. Data dredging involves conducting multiple statistical tests until a statistically significant result is achieved, potentially resulting in false-positive findings. If multiple statistical tests are conducted without a fixed objective, the risk of finding patterns or relationships by chance increases, thereby increasing the risk of reporting inaccurate outcomes. Therefore, the presentation of positive and negative outcomes is crucial to maintain the integrity of research findings. Researchers must transparently report the number of hypotheses explored, document all decisions made during data collection, and provide details on every statistical analysis performed. In other words, to make valid scientific conclusions using p-values and related statistics, it is important to know how many analyses were performed, what they involved, and the criteria for choosing which results to report, including p-values.

P-value should not be used as a measure of the importance of a result. If *p*-value is below the commonly used threshold of 0.05, there is statistical evidence to reject the H_0. This implies that the observed results were not likely due to chance and that they demonstrate statistical significance. Although lower *p*-values do not mean that the result are more important than the ones with higher *p*-values. Small *p*-values can be attained by simply conducting the measurements in precisely, while larger *p*-values can result from factors such as using a small sample size. With a smaller sample size, there is a greater likelihood of committing bias, as some potential outcomes may not be adequately analysed.

By taking all the above mentioned into account, is clear that *p*-value is not enough to provide a good measure the strength of evidence for a model or hypothesis. Researchers should not base their conclusions only by analysing *p*-value, as by itself it provides limited information.

3. Alternatives to *p*-value:

As seems, *p*-value lacks to provide enough statistical evidence to reject a hypothesis. As a result, alternatives like the confidence interval (CI) or effect sizes (ES) should be employed.

Confidence intervals can be interpreted to measure a sample or a research quality. A confidence interval is identified by the range of its margins of error, a range that shifts according to the selected confidence level. Common confidence levels in biomedical literature range from 90% to 99%, and, to a lesser extent, 99.9%. A more

precise estimate is associated with narrower interval margins. The 95% confidence interval is conventionally preferred in literature, corresponding to the acknowledged statistical significance level of $P < 0.05$. A rule applies to samples of equal size: as the confidence level diminishes, estimate accuracy heightens.

Although the *p*-value and the confidence interval describe the same thing, they do it diverse ways and complement each other. While the *p*-value indicates the likelihood of the hypothesis occurring by chance, the confidence interval establishes ranges of uncertainty within which one can anticipate the value of that hypothesis. For example, to know if the difference in the concentration of 0.46 µg/g in the concentration of lindane in two soils is statistically significant or the difference is simply by chance, the *p*-value calculated by t-test and the CI should be considered.

If the confidence interval for the difference of two samples does not include zero, it aligns with rejecting the null hypothesis that there is no difference. With this example is observed what was mentioned before: CI and *p*-value give the same information in different ways.[5] However, it must be kept in mind that CI also allows researchers to assess clinical relevance as it provides information about the difference between two populations, or treatment groups, that can be statistically significant but not clinically significant. In other words, CI allows to know if a treatment show a statistically significant improvement in a certain outcome, and if the magnitude of improvement is big enough to have practical or meaningful implications.[6]

Another parameter that can be use as an alternative, or in combination, to *p*-value is the effect size. For any reader, it may be complicated to understand the outcome of, for example, the use of a treatment if the article only says that the results are statistically significant. Consequently, the most convenient way to make the understanding of result easier is to establish a standardized way of measuring the treatment effect, the effect size.

There are numerous ways to calculate the effect size. The Cohen's d if often used to assess the standardized difference between two means, and it is determined by dividing the difference between the means by the pooled standard deviation. The Pearson's r measures the strength and direction of a linear regression between two variables, using the r value to indicate the size of the effect. The odds ratio quantifies the strength and direction of the association between two binary variables. Variance-accounted (η^2) is used in analysis of variances like ANOVA or ANCOVA to represent the proportion of variance in the dependent variable attributable to the independent variable. In epidemiology, the relative risk is frequently used, expressing the ratio of the probability of an event occurring in the exposed group to the probability in the unexposed group.[7]

4. Sample Study:

In clinical studies the main objectives are to analyse the origins of diseases, assess diagnostic methods, forecast prognosis, and determine interventions that can prevent adverse health outcomes. Although, many argue that this is simply impossible, since it is never possible to study all the possible outcomes, and therefore nothing guarantees that in the following clinical trial or experiment the same result are going to be achieved. Consequently, to defend the already observed outcomes it should be necessary to develop alternative theories and gather evidence to challenge them. The more a theory can manage efforts to prove it wrong, the more its trueness is maintained. Thus, to be sure that a theory is valid is essential to demonstrate that other theories are wrong to be more confident in the chosen one.

After comprehending the information detailed in the previous sections, one can now understand the following example about clinical and statistical relevance. The example is focused on the administration of renin-angiotensin enzyme inhibitors (ACEi) and angiotensin receptor blocker (ARBs) to patients suffering from chronic renal insufficiency. The hypothesis in this randomized controlled trial was that these drugs combined with intravenous iodine administration could produce damage to the kidneys. Therefore, it was studied if withdrawing these drugs, the chances of having a contrast-induce nephropathy (CIN) decreased. So, the H_0 in this study was: the likelihood of CIN is the same in patients who the ACEi/ARB treatment continued and those who stopped. In other words, that there was not difference between the groups.

The obtained *p*-value in this study for the difference in the frequency of CIN between groups was 0.16. As explained before, this value exceeds the threshold of 0.05, and consequently there is not statistical evidence to reject H_0. If only *p*-value is considered, statistically there is not a decrease in the likelihood of CIN in patients who stopped receiving ACEi/ARB. Although, it has been concluded that in clinical studies statistical evidence dictated by *p*-value is not enough to support or refuse a hypothesis. Accordingly, CI must be also considered; in Figure 1 the *p*-value function can be observed. The *p*-value is displayed on the left y-axis, while the confidence level is represented on the right y-axis. Examining both y-axes allows us to observe the reciprocal relationship between these two magnitudes.

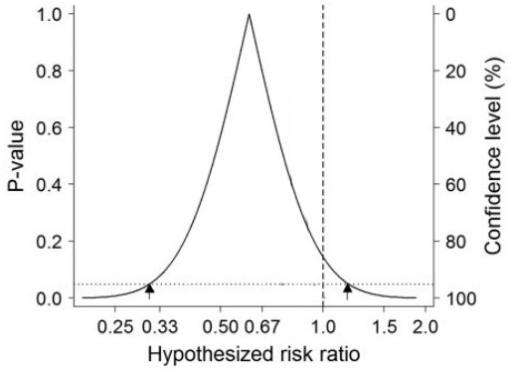

Figure 1. *P*-value function. Indicated by the peak appears the point for the ratio at 0.61. Indicated black arrows the 95% CI range, from 0.31 to 1.19, is observed.

The pick of the illustrate black line appears at 0.59, this value refers to the point estimate for the risk ratio. This ratio allows to compare the probability of an event occurring in one group compared to the another. Since it is lower than one, it suggests a decreased risk between the studied groups. Hence, the risk ratio is in accordance with the conclusion drawn from the *p*-value.

If a CI is reported around the point estimate, the corresponding confidence limits on the horizontal axis are obtained. With a confidence level set a 95%, reflecting a two-tailed *p*-value of 0.05, the resulting risk ratio has a lower limit of 0.31 and an upper limit in 1.19. Since the *p*-value falls into the interval it means that the observed results are consistent with the null hypothesis. Once again, thanks to this parament the information acquired from the *p*-value is sustained.

So, to statistically support the conclusion that can be drawn from the *p*-value different parameters must be considered. Although, it must not be forgotten that this is a clinical study in which the balance of benefit and harm is established by the size of the impact rather than its statistical significance.

In this study, for example, there was not a statistical significance in the decrease in CIN incidence when ACEi/ARBs treatment was retired, although is noteworthy that it was observed that only thirteen patients that stopped receiving the treatment were needed to prevent a single occurrence of CIN. This suggests that although the statistical significance was not reached, there might still be a clinically significant difference. Therefore, solely reporting that the hypothesis is not statistically significant without considering the practical effect could lead to overlooking meaningful clinical implications.

5. Conclusion:

In conclusion, even though statistical relevance is essential in the evaluation of a clinical study, it is important to emphasize the relevance of evaluating whether the observed difference in a study holds clinical significance, so many effects such as ratios of risk are proved. Is fundamental to present the evaluated effects whit a CI, which it provides a summary of both the effect size and the precision with which the

effect is estimated. In other words, presenting the evaluated effect along with a CI not only shows how big the effect is but also the degree of uncertainty linked to that measurement.

In summary, CI suggested a likely range for the risk ratio estimated between 0.30 and 1.19 with 95% certainty. Meanwhile, the *p*-value indicated that, assuming no actual difference, a similar or more extreme effect could occur in 16% of repeated experiments. The study finds that discontinuing ACEi/ARBs for a week requires treating only thirteen patients to prevent one contrast-induced nephropathy (CIN) event. Therefore, due to the relatively gentle impact of discontinuing ACEi/ARBs, the small number of patients needed to prevent CIN, and the potential severity of CIN suggest that the intervention's effect is substantial, supporting the need for consideration beyond statistical measures.

Bibliography

1. Andrade, C. (2019). The P value and statistical significance: misunderstandings, explanations, challenges, and alternatives. Indian journal of psychological medicine, 41(3), 210-215.
2. Ranganathan, P., Pramesh, C. S., & Buyse, M. (2015). Common pitfalls in statistical analysis: Clinical versus statistical significance. Perspectives in clinical research, 6(3), 169–170.
3. Fethney J. (2010). Statistical and clinical significance, and how to use confidence intervals to help interpret both. Australian critical care: official journal of the Confederation of Australian Critical Care Nurses, 23(2), 93–97.
4. Figueiredo Filho, D. B., Paranhos, R., Rocha, E. C. D., Batista, M., Silva Jr, J. A. D., Santos, M. L. W. D., & Marino, J. G. (2013). When is statistical significance not significant?. Brazilian Political Science Review, 7, 31-55.
5. Simundic, A. M. (2008). Confidence interval. Biochemia Medica, 18(2), 154-161.
6. O'Brien, S. F., & Yi, Q. L. (2016). How do I interpret a confidence interval?. Transfusion, 56(7), 1680-1683.
7. Lee, D. K. (2016). Alternatives to P value: confidence interval and effect size. Korean journal of anaesthesiology, 69(6), 555-562.