# In Search of p < 0.05: The Problem of Multiple Comparisons

María E. Knorre, Mariano Bergier y Arturo M. Cagide
Departament of Cardiology, Hospital Italiano. Buenos Aires, Argentina

## ABSTRACT

This article examines the role of chance and statistics in medical research, focusing on statistical significance and the p-value. Although randomization in clinical trials helps control randomness, the risk of false positive findings persists when multiple comparisons are performed without appropriate statistical adjustment.

This issue is illustrated through an analogy involving a winning casino bet, in which a successful outcome is reported without disclosing previous failed attempts, a situation comparable to uncorrected post hoc analyses. Strategies to mitigate this bias are discussed, including the a priori definition of hypotheses and endpoints, the application of statistical corrections for multiple testing, and the validation of findings in independent cohorts. Finally, the article emphasizes that artificial intelligence is also susceptible to this problem when data are not properly managed.

**Keywords:** statistics as topic, statistical significance, analysis of variance, clinical trials as topic, artificial intelligence.

## En búsqueda de una p < 0,05: el problema de las comparaciones múltiples

## RESUMEN

El artículo analiza el papel del azar y la estadística en la investigación médica a partir del concepto de significación estadística y el valor de *p*. Si bien la aleatorización en los ensayos clínicos permite controlar el azar, el riesgo de obtener resultados falsamente positivos persiste cuando se realizan múltiples comparaciones sin los ajustes estadísticos correspondientes.

Esta problemática se ilustra mediante una analogía con una apuesta ganadora en el casino, en la que se presenta un resultado exitoso sin considerar los intentos fallidos previos, situación comparable al análisis *post hoc* sin corrección. Se describen distintas estrategias para reducir este sesgo, como la definición *a priori* de hipótesis y puntos finales, la aplicación de correcciones estadísticas por comparaciones múltiples y la validación de los hallazgos en cohortes independientes. Finalmente, se destaca que la inteligencia artificial tampoco está exenta de este problema si no se aplican criterios estadísticos adecuados.

**Palabras clave:** análisis estadístico, significación estadística, análisis de varianza, investigación clínica, inteligencia artificial.

Corresponding author: eugenia.knorre@hospitalitaliano.org.ar, Knorre ME.

Received: 12/1/2025 | Accepted: 04/28/2025 | Published: 09/09/2026

DOI: http://doi.org/10.51987/rev.hosp.ital.b.aires.v46i3.1301

How to cite: Knorre ME, Bergier M, Cagide AM. In Search of p < 0.05: The Problem of Multiple Comparisons. Rev Hosp Ital B.Aires. 2026;46(3):e0001301



ISSN (en línea) 2314-3312

## INTRODUCTION

Are there any points of contact between chance and medicine? At first glance, the answer is no.

In medicine, decisions have a logical basis grounded in evidence and the clinician's experience. Chance comes into play only in randomized experimental research, when participants are assigned to the treatment or control groups being compared. The underlying concept is that randomization distributes variables that could affect outcomes equally between groups, so the only difference between them is the treatment under study. Under this premise, a $p < 0.05$ indicates that the probability of a false-positive conclusion is only 5 in 100.

Let us now contrast the above considerations with the following real-life situation. *A group of friends is on vacation, and one of them tells the others that the previous night he went to a casino and won a substantial amount of money by placing a straight-up bet on number 18. Surprised, his friends ask for details. The lucky winner replies, "The thing is..."*

The anecdote, though entirely plausible, hinges on a subtle detail that was not mentioned at first.

## DEVELOPMENT

### On Lottery Drums and Probabilities (Statistical Concepts)

Suppose that a lottery drum contains 1,000 balls, each labeled with a given difference, either positive or negative, and whose *frequency distribution* follows a bell-shaped curve, such that the mean value of these differences is zero (Fig. 1). The shaded area represents 5% of the extreme differences (50 balls): 2.5 % with negative values (25 balls) and 2.5% with positive values (25 balls).

The probability of drawing a ball with a negative or positive difference corresponding to those in the shaded area is only 5 %. We can return the ball to the drum and draw again, repeating the experiment 100 times, after which we would have drawn 5 balls with differences corresponding to those in the shaded area of the curve.

#### *From the Lottery Drum to Applied Inferential Statistics*

Let us leave games of chance behind and turn to clinical research.[1-4]

Suppose that an observational study seeks to estimate the value of a diagnostic or prognostic indicator or the outcome of a treatment, or that a randomized trial seeks to estimate the effect of a particular intervention. In both cases, two samples that differ with respect to the variable under study (endpoint) are compared to determine whether there is a "statistically significant difference."

Starting from the assumption that there are no differences (null hypothesis), the possible differences found in the endpoint between the two samples would have a distribution (frequency distribution) such as that illustrated in Figure 2 (similar to Fig. 1), with a mean value (mean) of zero.

If the trial results show that the difference found lies between −1.5 and +1.5 (both tails), as indicated by the arrows in Figure 2, the study was negative because

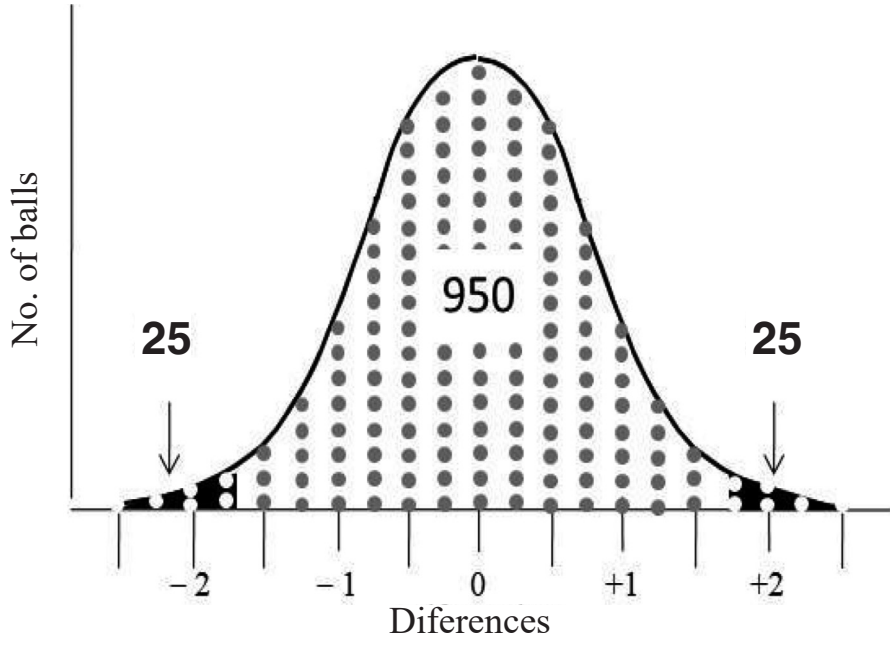


**Figure 1.** This figure illustrates the probability of drawing a ball labeled with a given difference (positive or negative) from a lottery drum: if the drum contains 1,000 balls, 950 with differences ranging from −1.5 to +1.5 and 50 with extreme differences—25 with −1.5 and 25 with +1.5—the probability of drawing one of these balls is 5 %. This probability increases with repeated trials; after 100 attempts, we would expect to draw 5 balls with these extreme differences (alternatively, we could express this relationship as 20 to 1). Source: Authors' own work.

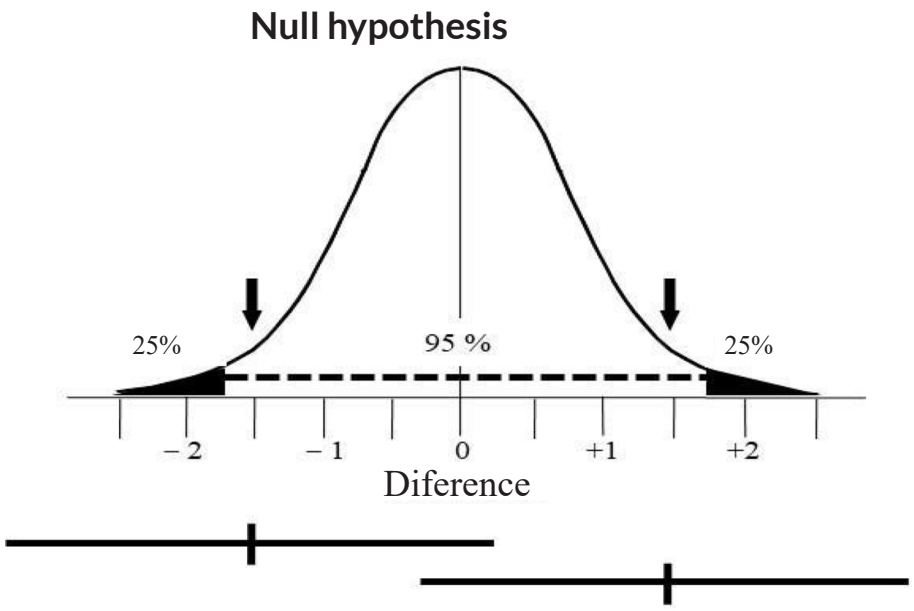


**Figure 2.** Frequency distribution curve (Gaussian curve) under the null hypothesis. Both arrows indicate a trial result in which the difference between the samples (positive or negative) did not reach statistical significance, as it falls within the area of the curve corresponding to 5 % in both tails. (The dotted line delimits 95 % of the area under the curve). The bottom panel shows another way to express the result: both solid lines represent the confidence intervals constructed from the differences, indicating that the study was negative because the intervals do not include zero. Source: Authors' own work.

statistical significance was not "reached," as the difference "falls" within the area corresponding to p ≥ 95 % (it does not "reach" < 2.5%).

In clinical research, a 95% confidence interval (CI) is usually constructed for the difference found, −/+ 1.5, which –because it crosses zero on the null hypothesis curve– again indicates that the difference is not statistically significant (Fig. 2).

In the previous example, we used the difference in percentages (risk difference [RD]) between, for example, the intervention and control groups. This could also be expressed as a ratio between the two, that is, as relative risk (RR). In this case, the mean of the curve will be 1, and the null hypothesis is rejected when the CI reaches or crosses 1.

Thus, the null hypothesis is accepted, and it is concluded that the endpoint (diagnostic/prognostic criterion or intervention) does not differ between the samples being compared.

However, if the study were repeated 100 times under the null hypothesis, about 95 trials would yield a nonsignificant result. In contrast, approximately 5 could yield a statistically significant result by chance and thus be falsely considered positive.

## Multiple Comparisons: Statistical Foundations

As shown above, the probability of finding a significant difference increases when you repeat comparisons. The p-value obtained after multiple comparisons is often referred to as the *nominal p-value*, meaning a numerical value that is not interpreted in the context of the actual probability. Several statistical procedures are available to calculate a *p-value adjusted* for the number of comparisons performed.

One approach is to adjust the *p*-value for the number of comparisons performed using multiple-comparison adjustment procedures.

The Bonferroni correction was initially applied, in which the corrected significance level ($\alpha c$) is obtained by dividing the nominal $\alpha$ level by the number of comparisons ($\kappa$), according to the following expression: $\alpha c = \alpha/\kappa$.

The required significance level according to the number of comparisons was also calculated using the following expression: $\alpha c = 1 - (1 - \alpha)\kappa$, where $\alpha$ represents the nominal p-value and $\kappa$ the number of comparisons performed.

Decimal values are expressed using commas rather than periods.

Table 3 shows, for a nominal p-value of 0.05, the actual p-values and the resulting adjusted values obtained by applying the formula described above, which should be used in a study when more than one comparison is performed.

The possibility that 5 out of every 100 comparisons may be false positives might seem unlikely in clinical research. However, expressing the same probability as 1 in 20 makes its significance easier to appreciate. Returning

**Table 1.** Nominal p-values and adjusted values according to the number of comparisons, calculated using the Bonferroni correction and the significance-level adjustment method.

| $\alpha$ | $\kappa$ | $\alpha_c$ | $\alpha / \kappa$ |
|---|---|---|---|
| 0.05 | 1 | 0.05 | 0.05 |
| 0.05 | 2 | 0.10 | 0.025 |
| 0.05 | 3 | 0.14 | 0.017 |
| 0.05 | 4 | 0.19 | 0.013 |
| 0.05 | 5 | 0.22 | 0.010 |

$\alpha$: nominal p-value; $\alpha$c: unadjusted p-value; k: number of comparisons; $\alpha$/k: adjusted p-value.

Source: Authors' own work.

to Table 1, note that with only two or three comparisons, the p-value required to achieve statistical significance changes substantially.

## Multiple Comparisons in Clinical Research

In clinical research, multiple comparisons may arise in different scenarios. In clinical trials, researchers usually define an overall hypothesis and a primary objective, either to evaluate an intervention's effect or the prognostic value of a given factor. However, when the initial comparison does not show a statistically significant association, there may be a temptation to explore alternative objectives or modify the endpoint, repeating the analyses until a favorable result is obtained.[1-5]

Even without modifying the primary endpoint, multiple comparisons may arise when that endpoint is assessed at different follow-up times, for example at 3, 6, or 12 months, with subsequent selection of the time point showing a statistically significant result.[3]

Another common source of multiple comparisons is analyzing population subgroups. When the overall result is negative, researchers may divide the original sample to explore associations in different subgroups, retaining only those with favorable results.[7]

Likewise, researchers may apply different statistical tests to the same data and report only the test that yields a significant p-value. These strategies, whether used separately or in combination, substantially increase the probability of obtaining spurious associations.[8]

## Correction Strategies in Clinical Research

The main strategy for ensuring the credibility of clinical research findings is to precisely define, *a priori*, the hypothesis to be tested, the study objective, and the primary endpoint, as well as the statistical test to be used and the required sample size.[1] Additional findings should be presented as secondary endpoints, explicitly stating that they are based on a post hoc analysis and that the *p*-value has not been adjusted for multiple comparisons.[5,6]

For secondary analyses, adjust p-values according to the number of comparisons using appropriate statistical procedures, such as the Bonferroni correction or other adjustment methods.[4,6]

Another alternative is to establish a hierarchical analysis of secondary endpoints. In this approach, once statistical significance has been reached for the primary endpoint, a prespecified hierarchical order is established for the secondary endpoints. When a nonsignificant *p*-value is obtained, subsequent analyses should be interpreted as yielding nominal values, even when they are below 0.05.[1]

Finally, in observational studies, findings from multiple comparisons may be considered exploratory and presented as results from a derivation cohort, requiring validation in an independent cohort.[2,4]

## Broadening the Scope

Researchers address multiple comparisons using different criteria depending on the research setting.

In *randomized controlled intervention trials* with a predefined primary endpoint, researchers usually handle secondary endpoints by applying different adjustment methods or stating that they have not adjusted for multiple comparisons.

*Observational studies based on large databases* that assess prognosis or interventions (Phase IV) vary in methodological rigor, as they sometimes lack the necessary corrections for the multiple analyses commonly performed in this type of research.

Finally, in *single-center studies*, unreported *post hoc* analyses are common and may explore—within the general framework of a hypothesis—multiple objectives or endpoints without appropriate adjustment for multiple comparisons.

## From Artificial Intelligence to Multiple Comparisons

Does artificial intelligence (AI) solve the problem of potentially spurious associations resulting from multiple comparisons? The answer is no.

On the contrary, AI's ability to explore thousands of associations across millions of data points may amplify the risk of identifying spurious relationships if researchers do not apply appropriate statistical controls.[9]

# CONCLUSION

Returning to the anecdote with which this article began:

"...the thing is, I had been betting all night and losing repeatedly on every bet until I finally won with 18" (we do not know whether our character recovered all his money, but in any event, this does not change the story).

That is precisely where the problem lies: in not telling the whole story and presenting the result as the original objective rather than as one reached after several attempts, which inevitably alter the p-value.

Simply put, saying "I found what I was looking for" is not the same as saying "Look what I found." A statement this simple encapsulates the central message of this article.

In this context, there is nothing more fitting than to recall: "If you torture your data long enough they will tell you whatever you want to hear"; "...the perpetrator simply pores over the data until a 'significant' association is found between variables and then devises a biologically plausible hypothesis to fit the association."[8,10]

Or perhaps, "a p-value is no substitute for a brain" (*A p-value is no substitute for a brain*. Anonymous), as has been noted in contemporary discussions on the use and interpretation of the *p*-value.[11] An even more fitting reflection, however, is that of physicist Richard Feynman, who concluded: "It is much more interesting to live with uncertainty than to live with answers that might be wrong."[12]

In medicine, this last statement requires a slight qualification: even in the face of uncertainty, we must always make decisions, whether through action or inaction. The challenge is to remain aware of the constraints that uncertainty imposes.

**Author Contributions:** Conceptualization, Supervision, Validation (AMC). Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software (MEK, MB, AMC).

**Conflicts of Interest:** The authors declare no conflicts of interest related to the content of this work

**Funding:** The authors declare that this study received no funding from external sources

## REFERENCES

1. Pocock SJ, Stone GW. The primary outcome fails - what next? N Engl J Med. 2016;375(9):861-870. https://doi.org/10.1056/NEJMra1510064.
2. Sugitani T, Bretz F, Maurer W. A simple and flexible graphical approach for adaptive group-sequential clinical trials. J Biopharm Stat. 2016;26(2):202-216. https://doi.org/10.1080/10543406.2014.972509.
3. Freemantle N, Calvert MJ. Interpreting composite outcomes in trials. BMJ. 2010;341:c3529. https://doi.org/10.1136/bmj.c3529.
4. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol. 1995;57(1):289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
5. Bauer P. Multiple testing in clinical trials. Stat Med. 1991;10(6):871-889; discussion 889-890. https://doi.org/10.1002/sim.4780100609.
6. Perneger TV. What's wrong with Bonferroni adjustments. BMJ. 1998;316(7139):1236-1238. https://doi.org/10.1136/bmj.316.7139.1236.
7. Pocock SJ, Assmann SE, Enos LE, et al. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. Statist. Med.2002;21:2917-2930. https://doi.org/10.1002/sim.12968.
8. Bender R, Lange S. Adjusting for multiple testing--when and how? J Clin Epidemiol. 2001;54(4):343-349. https://doi.org/10.1016/s0895-4356(00)00314-0.
9. Westphal M, Zapf A, Brannath W. A multiple testing framework for diagnostic accuracy studies with co-primary endpoints. Stat Med. 2022;41(10):1848-1866. doi:10.1002/sim.9308.
10. Mills JL. Data torturing. N Engl J Med. 1993;329(16):1196-1199. https://doi.org/10.1056/NEJM199310143291613.
11. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05". Am Stat. 2019;73(Suppl 1):1-19. https://doi.org/10.1080/00031305.2019.1583913.
12. Feynman RP. The pleasure of finding things out. Cambridge (MA): Perseus Books; 1999. 270 p.