Null Hypothesis Significance Tests, a Misleading Approach to Scientific Knowledge: Some Implications for Eating Disorders Research

Roser Granero[1,2], Janet Treasure[10], Laurence Claes[3,4], Angela Favaro[5], Susana Jiménez-Murcia[2,6,7], Andreas Karwautz[8], Daniel Le Grange[9], Kate Tchanturia[10], Fernando Fernández-Aranda*[2,6,7]

Affiliations

1. Department of Psychobiology and Methodology, Autonomous University of Barcelona, Barcelona 08193, Spain.
2. CIBER Fisiopatología Obesidad y Nutrición (CIBEROBN), Instituto Salud Carlos III, Madrid, Spain
3. Faculty of Psychology and Educational Sciences, University of Leuven, Belgium.
4. Faculty of Medicine and Health Sciences, University of Antwerp, Belgium
5. Department of Neuroscience, University of Padua and Neuroscience Center (PNC), University of Padua, Padua, Italy.
6. Department of Psychiatry, University Hospital of Bellvitge-IDIBELL, Barcelona, Spain
7. Department of Clinical Sciences, School of Medicine and Health Sciences, University of Barcelona, Spain
8. Eating Disorders Unit, Department of Child and Adolescent Psychiatry, Medical University of Vienna, Vienna, Austria.
9. Eating Disorders Program, Department of Psychiatry, University of California, San Francisco, CA 94143, USA.
10. King's College London, Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience (IoPPN), London SE5 8AF, UK.

***Corresponding author:**

Fernando Fernandez-Aranda. Eating Disorders Unit, Department of Psychiatry, University Hospital of Bellvitge-IDIBELL and CIBEROBN, Feixa Llarga s/n 08907 Hospitalet del Llobregat (Barcelona, Spain). Tel. +34-93-2607227, Fax. +34-93-2607193. e-mail: ffernandez@bellvitgehospital.cat

Background

The application of the quantitative scientific method to the research of eating disorders (ED) uses statistical inference as its inductive analytical procedure of reference to obtain knowledge about the target populations based on the empirical evidence observed in specific samples. The validity of the studies published in different scientific dissemination forums (journals, congresses, seminars and scientific meetings) depends on different questions: formulation of relevant empirical hypothesis, adequate planning of the research, the selection and use of appropriate statistical techniques, and the adequate interpretation of the numerical results obtained with these analytical procedures.

The most commonly formulated problems in the ED research area are: estimation of population parameters and hypothesis testing. Studies focusing on the estimation of population parameters face the challenge of deducing the value of a parameter (or parameters) that characterize the frequency distribution within a population, often through confidence intervals. Parameter estimation is the objective of epidemiological studies conducted to find out the frequency of an event in a certain population, for example studies aiming at assessing the prevalence (also risk or rate) of disorders, symptoms or exposure to specific risk factors. In the ED area, epidemiological studies have been designed to solve different estimation problems, such as: a) determining the prevalence of eating problems in clinical or community populations [such as the study by Bagaric and colleagues among a community sample of South Australia looking at the lifetime prevalence of Bulimia Nervosa and Binge Eating Disorder (Bagaric, Touyz, Heriseanu, Conti, & Hay, 2020) or the study by Riberio and colleagues aiming at estimating the presence of the Binge Eating Disorder in Portuguese students (Ribeiro, Conceição, Vaz, & Machado, 2014)]; and b) finding out the frequency of eating symptoms/problems within specific segments of populations characterized by high vulnerability [such as the study by Aoun and colleagues among a sample of Syrian refugees (Aoun, Joundi, & El Gerges, 2019). These primary research publications can later be included in epidemiological systematic reviews or meta-analyses, which are based on structuring and synthesizing the available empirical evidence in order to answer a specific research question. For example, the publication that compiles previously published results for the association of disordered eating behaviours and autistic traits in nonclinical populations (Christensen, Bentz, Clemmensen, Strandberg-Larsen, & Olsen, 2019), or the study measuring the longitudinal evolution of ED prevalence from 2000 to 2018 (Galmiche, Déchelotte, Lambert, & Tavolacci, 2019).

On the other hand, hypothesis testing studies face the challenge of assessing the likelihood of an empirical hypothesis (also called working hypothesis or research hypothesis), which usually contains the supposed sense and/or level of the association/s between variables. Hypothesis testing studies analyze the empirical evidence obtained in a specific sample with different purposes: a) to

identify risk factors and underlying mechanisms that enable a better understanding of the etiology and the phenotypes of disorders [for example the study by Mallorquí-Bagué and colleagues aiming at investigating clinical and electrophysiological correlates of emotion regulation and craving regulation in AN (Mallorquí-Bagué et al., 2020)]; b) to assess the therapeutic efficacy of treatments [such as the study by Fernández-Aranda and colleagues analyzing the benefits of a serious video game as a complementary program to enhance the general functioning of BN patients (Fernandez-Aranda et al., 2015)]; and c) to find out the evolution over time of different disorders and their possible correlated factors [such as the work by Svedlund and colleagues, which assessed whether the efficiency of a medium-term intervention in women with ED may be due to ADHD symptoms (Svedlund, Norring, Ginsberg, & von Hausswolff-Juhlin, 2018), or the randomized clinical trial (RCT) by Quadflieg et al. aimed at assessing the efficacy of a video-based skills training program designed to reduce burden and distress in caregivers of female ED treated inpatients (Quadflieg, Schädler, Naab, & Fichter, 2017)].

Significance level is not truly significant

A large number of conclusions published for hypothesis testing in clinical scientific research are based on statistical significance tests [known as the "null hypothesis significance test" (NHST)], developed by Ronald Almer Fisher in the 1920s under the frequency statistical approach (Fisher, 1925). NHST provide the well-known index called *"significance level"* (*p-value*), which is considered by most researchers to be the (only) criterion to decide whether there is (or is not) a statistically significant relationship between the variables. The general decision rule is as simple as possible: $p \leq 0.05$ is interpreted as a statistically significant result (considered in practice to be strong evidence for the expected effect or association), while $p > 0.05$ is considered to be a statistically non-significant result (which for most researchers means that no effect is observed in the empirical data). But despite the popularity of the significance level, misuse of *p-values* is very common, mainly because a large number of researchers do not know how to properly interpret these indexes.

A frequent analytical procedure in ED research is to calculate the *p-value* provided by NHST and then use a decision rule based on the theory of hypothesis testing developed by Jerzy Neyman and Egon Pearson (Neyman & Pearson, 1933). The Neyman-Pearson approach has provided researchers with important and valuable tools to accept (confirm) or reject (refute) a contrasted empirical hypothesis, such as the definition of Type-I and Type-II errors (α and β), the statistical power ($1-\beta$), the critical regions within the decision rule, or the basis for calculating the minimum sample size needed to get a specific effect. But since the algorithmic approach offered by Neyman-Pearson is different to (and largely incompatible with) the Fisher method, its result has been in historical conflict with the making of statistical judgments based on error rates that are well-

known among mathematicians-statisticians and highly unknown among researchers (S. N. Goodman, 1993). And this regular misunderstanding of the rationales of both the Fisher and Neyman-Pearson theories has contributed even more to the uncertainty regarding the fundamentals of statistical procedures in medical research leading to unreliable conclusions (Griffiths & Needleman, 2019; Savitz, Tolo, & Poole, 1994; Smith, 2020; Wellek, 2017).

One of the most common misconceptions of the NHST is the consideration that *p-value* is the probability of the "null hypothesis" (denoted H_0) being true. But the interpretation of the significance level is not so simple. To approximate the true meaning of the *p-value*, it must be borne in mind that the rationale of NHST starts from the theoretical assumption that a certain statistical hypothesis is true. This is popularly known as the H_0, which is formulated by the absence of association between the variables (it is important to note that H_0 rarely corresponds to the empirical hypothesis that really interests researchers). And given the conditional assumption that the H_0 is true, a set of mathematical algorithms are developed to obtain a measure of the probability of discrepancies equal to or greater than those obtained in the empirical study being obtained by chance. This value is known as the *significance level* (the famous *p-value*), which is mathematically equivalent to the following conditional probability: $p\text{-}value=Pr(d \geq d_{study}|H_0)$. This statistical interpretation of the *p-value* is somewhat complex (it is not intuitive in clinical terms), and is therefore not the interpretation made by most researchers who base their final conclusions on the significance level (Lazzeroni & Ray, 2012). Many scientists simply (and wrongly) assume that *p-value* is the probability (understood as the credibility or likelihood) that empirical data attributes to H_0, which in mathematical terms would be equivalent to assuming that $p\text{-}value=pr(H_0)$. And this incorrect use of the *p-value* leads to the use of this index as a simple measure of the probability of success/error in the context of a simple decision between two mutually exclusive options (accept versus reject the H_0): if *p-value* is small (by consensus in the medical scientific community $p \leq 0.05$), the probability of success when choosing H_0 is considered low and therefore this hypothesis is rejected; conversely, if *p-value* is large ($p>0.05$), the probability of H_0 being true is high and therefore is not ruled out. This mistake when interpreting the significance level has meant that the identification of the associations between variables has led to an incessant decades-long search for covert "statistically significant results", which in turn led to mythologization of the *p-value* and its use as irrefutable proof of scientific evidence (the finding of small *p-values* has brought much joy to many of our colleagues, who have reported these values as unequivocal and irrefutable proof of the success of their research).

In recent decades, many examples have been published to draw attention to the key limitations of NHST and to the consequences of relying on statistical significance (Van Calster, Steyerberg, Collins, & Smits, 2018). We would also like to present here some illustrations of

problems of statistical inference based on probabilistic premises, which can lead to bizarre conclusions. In 1996, the prestigious publication *Nature* presented an example to prove that the change from absolute certainty to probability makes the syllogistic reasoning false under the statistical reasoning process (Beck-Bomholdt & Dubben, 1996). The authors numerically developed a single logical fallacy to obtain evidence regarding the possible non-human origin of the leader of the Roman Catholic Church (John Paul II, the Pope at the time). Under the title *"Is the Pope an alien?"* the surprising solution to this problem was that the Pontiff's human status was supported by an extremely low probability (p=0.00000000017!). And although the most dogmatic Catholic believers could have interpreted this value as irrefutable proof of the Pope's divine creation, atheists and practitioners of other religions could also interpret it as evidence of the Pontiff's extraterrestrial origin. Obviously, the most logical conclusion is to employ common sense and seriously doubt the interpretative deductive mathematical method used to solve the absurd problem regarding the Pope's nature and origin.

And as there have been ongoing attempts to reconcile religion and science throughout history, we offer another provocative example that uses Bayes' conditional probability to obtain scientific arguments for the existence of God [the theorem was formulated by the Presbyterian minister Thomas Bayes in the 18th century (Bayes & Price, 1763)]. In fact, it is suspected that Reverend Bayes himself, along with his friend and fellow mathematician and minister Richard Price, were possibly *tempted* to find answers that went beyond faith to questions as philosophical as the existence of Deity (this could be reasonable, since it is known that in the 17th and 18th centuries, statistical theory was used to prove the existence of God). Based on Bayesian Decision Theory, the recent book by the physicist and risk scientist Stephen D. Unwin *revealed* how a math equation can be used to calculate the probability of God (Unwin, 2004). This top publication sparked heated international debate, since according to the mathematical reasoning of the Bayes Theorem, the hypothesis that the known universe was the result of God's creation achieved a probability p=0.62. But beyond calculation, how should this probability be interpreted? Does this suppose that the existence of God is *evident* at 62%? This result is not a great *revelation* to Religious Believers, who undoubtedly trust 100% in the existence of God (people of faith surely doubt the reliability of the mathematical calculation). The relevant question here would be: is p=62% an impressive and convincing result for non-believers? It would be unsurprising for agnostics and atheists to continue doubting the mystery of God, since an additional 38% of faith is ultimately required to complete the mathematical calculation.

One last example that shows the absurdity of some conclusions based on mere statistical inference reasoning is a prospective RCT designed to assess the potential positive therapeutic effects of intercessory prayer to the Judeo-Christian God. Based on a double-blind protocol, a

sample of n=393 hospitalized coronary patients were assigned to an intervention group (with participant Christians praying) or to a control group (Byrd, 1988). The authors' conclusion was, literally *"The intercessory prayer group subsequently had a significantly lower severity score based on the hospital course after entry (P less than .01). The control patients required ventilatory assistance, antibiotics, and diuretics more frequently than patients in the IP group. These data suggest that intercessory prayer to the Judeo-Christian God has a beneficial therapeutic effect in patients admitted to a coronary care unit"* [(Byrd, 1988) p.826)]. Again, what are the potential implications for this striking conclusion? Should hospitals increase their workforce by employing Judeo-Christians to pray for the patients? Should this new item be included in the social security budget? But the most disturbing conclusions are related with aims of the work itself: even assuming a positive effect of intercessory prayers, should the Judeo-Christian God receive a commission for his mediational divine healing? How can other Gods be encouraged to help with the treatment of unhealthy people? In short, these examples serve to understand statements as emphatic as that of Jacob Cohen, a passionate defender of the use of alternative and complementary approaches to the NHST (such as effect size measures), who around 30 years ago noted that the *"significance test has not only failed to support and advance Psychology as a science but also has seriously impeded it"* [(Cohen, 1994) p. 997].

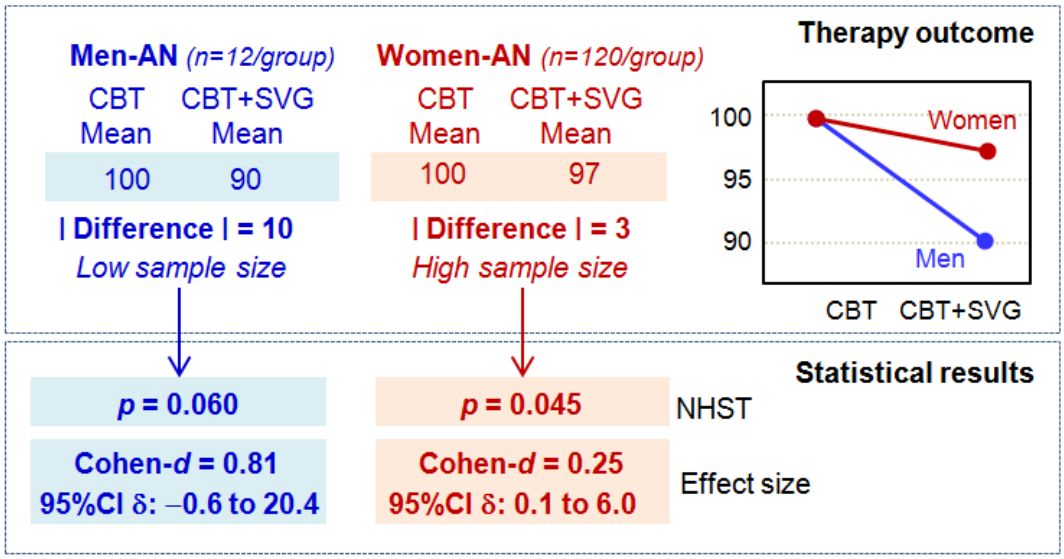
Sample size, significance level and effect size

Why is the *p-value* misinterpreted in clinical scientific research? The most probable reason is that the significance level is a very slippery concept that requires a lot of background knowledge to understand (Badenes-Ribera, Frias-Navarro, Iotti, Bonilla-Campos, & Longobardi, 2018; Morris, 2020). When interpreting *p-values* it should be understood that NHST only provides a measure of the compatibility between the empirical data registered in a specific study with a theoretical statistical hypothesis of reference formulated for the target population, through a theory based on the principles of the frequentist inference. Therefore, a *p-value* should never be considered to be an estimate of the probability of the empirical research hypothesis being true or false, or of the discrepancies between the data and the H_0 having been produced by the effect of mere chance, mainly because different factors influence the *p-value*. First, the significance level is related to the discrepancies between the empirical data and the theoretical model of reference (the higher the differences the lower the *p-value*); second, the spread of the data also affects significance [the higher the precision of the measures (lower variance), the lower the *p-value*]; and third, the sample size, which is one of the main contributors to the *p-value* (the larger the sample the greater the likelihood of a lower significance level). The relationship between the effect size and *p-value* is more intuitive for researchers, but not the influence of the sample size on the NHST results. In

general, researchers understand the convenience of analyzing large samples, but they do not always know the implications of this preference. The basic reason lies in an important analytical concept: statistical power. Studies carried out with small samples are underpowered and have a low capacity to detect real effects (significance level easily tends to $p>0.05$). On the contrary, studies with large samples have a high capacity to identify real effects (and therefore, it is easier to achieve $p\leq0.05$). The problem, however, is that very large samples are also overpowered, with the risk of achieving very small *p-values* for irrelevant clinical effects. This leads to the paradoxical situation that two studies that observe identical effects obtain very different *p-values* depending solely on the size of the samples.

The next example will illustrate the paradox of the *p-value* and how the measures of the effect size help to obtain more realistic knowledge of the problem. Suppose that an RCT, with a low sample size for some groups, aims to assess the benefit of including a serious video game (SVG) program together with cognitive behavioral therapy (CBT) to improve emotion regulation in ED patients. Since the authors suppose that sex and diagnostic subtype could act as an interaction (moderation) variable, stratified analyses are carried out (separately according to gender and diagnosis) (Figure 1). In the subsample of men with anorexia nervosa (AN), the CBT+SVG group (n=12) obtains a final mean of 90 points on a global measure of emotion dysregulation, compared to a mean equal to 100 points in the control group (composed of n=12 men who only received CBT). Therefore, in this work, SVG in AN males is related to a decrease of 10 points on the emotion dysregulation scale. On the other hand, an emotion dysregulation mean score equal to 97 points is obtained for the CBT+SVG group of AN women (consisting of n=120 participants) compared to a mean score of 100 points in the control group (with n=120 women). In females, the SVG is associated to a decrease of 3 points in the emotion dysregulation score. With this empirical evidence, NHST achieves $p=0.060$ for men (statistically not significant) and $p=0.045$ for women (statistically significant). These significance levels suggest the lack of evidence against the H_0 within men, and this statement could lead many researchers to the conclusion that the SVG program has no benefits for emotion regulation in AN males (probably discouraging its future use). On the contrary, the existence of statistical evidence against H_0 for AN women could lead to the consideration that SVG is an effective intervention to reduce the emotion regulation severity of these patients, thus making its use advisable.

Figure 1. Benefits of the SVG intervention on male and female AN patients



However, the results obtained in the previous example seem confusing. How can a higher difference in emotion dysregulation equal to 10 points in AN men be non-significant, while a much lower difference of 3 points in AN women achieved a significant result? This is a simple question of statistical power: the subsample of men is very small (n=12 subjects per group), and therefore large (or even huge) differences are required to reach the threshold of $p \leq 0.05$. Conversely, when the samples are large (as in the female AN group), small differences can easily reach the threshold of statistical significance. But a more relevant question is whether a difference of only 3 points in the emotion dysregulation scale obtained within the female subsample should be considered solid scientific evidence to recommend complementing CBT with the SVG? The answer *is not evident*: the final clinical decision depends on multiple factors together with the statistical evidence, such as the cost of implementing the program, the clinicians' expertise and the patients' values and preferences.

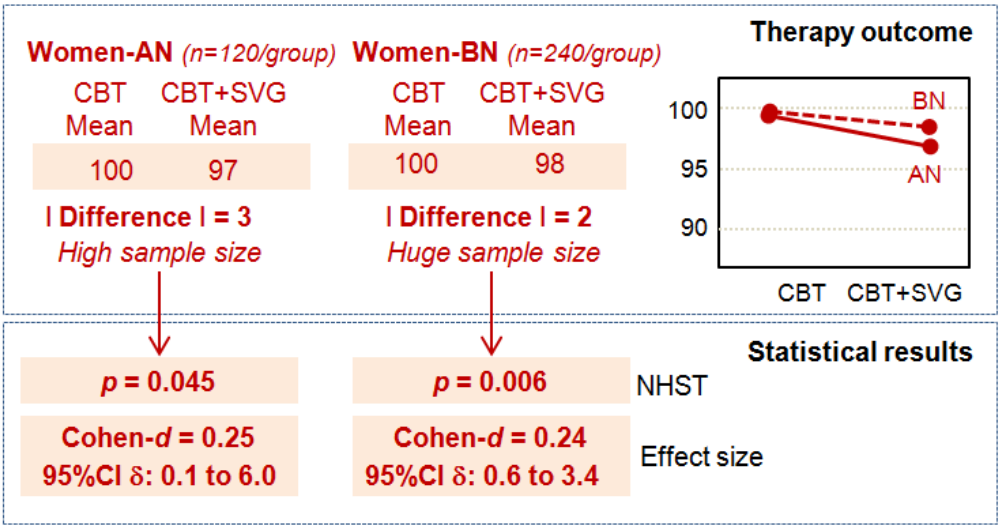
What's more, scientific knowledge in the ED area should never be built on the basis of generic and imprecise tests that simply state that two treatments differ: additional measures of the effect sizes are required. The key question is not whether two (or more) groups differ, but how much the groups differ. It is probably not relevant enough for clinicians to know that two groups differ. What they really need is a measure of the real difference or impact (Lee, 2016). The SVG program might have a real effect on emotion dysregulation in AN women, but it might be so irrelevant in clinical terms that the cost-benefit ratio is discouraging. So, what should be done? In research, *p-values* should *always* be complemented with measures that help expert clinicians assess effect sizes, to have stronger elements for formulating properly founded conclusions and making clinical decisions based on adequate empirical evidence. In the example of the SVG program, the

Cohen's-*d* coefficient [a standardized measure of the differences between means (Cohen, 1988)] could be obtained, whose value |*d*|=0.81 obtained for men is interpreted as a possible high-large effect size in practical terms, while the value |*d*|=0.25 achieved for women is interpreted as a poor effect size. Another method to assess clinical impact is to obtain the confidence intervals for the mean differences (95%CI-δ): -0.6 to 20.4 points for men, compared to 0.1 to 6.0 for women. What do these intervals indicate? For men, the interval is too wide (not very informative), but the upper limit assumes that the SVG program could obtain decreases in emotion dysregulation of up to 20 points. For women, the interval is narrow (highly informative), and this indicates that the differences could be practically nil or reach a maximum of 6 points on the emotion dysregulation scale.

Another frequent mistake is to interpret a non-significant result for an NHST ($p>0.05$) as evidence for the null hypothesis H_0 being proven (S. Goodman, 2008). Based on the example above (Figure 1), the reality can be quite different. In statistical terms, a non-significant result only suggests that the empirical data do not provide sufficient evidence to rule out the likelihood of H_0. But this finding does not guarantee that H_0 is false, and a new study carried out with higher statistical power could detect the relationship between the variables expected by researchers. Therefore, a non-significant result is only an indication of a “not found” (or “not evidenced”) relationship.

On the contrary, it is wrong to suppose that a significant result necessarily implies a relationship between the variables, and still less to assume the existence of a good-large impact (Steyerberg & Van Calster, 2020; Sullivan & Feinn, 2012). In scientific research, (very) small *p-values* (highly significant in statistical terms) could be associated with poor effects in overpowered studies carried out with very large samples. For example, imagine that in the RCT carried out to assess the benefits of the SVG program in ED patients (Figure 1), sample size for BN women was n=240 for both the CBT+SVG and the control groups and the mean difference was only 2 points for the emotion dysregulation measure (Figure 2). This small difference has achieved a $p=0.006$ in the NHST, which cannot be interpreted as a great evidence for the benefit of the SVG program. On the contrary, effect size measures suggest a poor benefit (Cohen-*d*=0.24) that in clinical terms could suppose a decrease in the emotion dysregulation scale of between 0.6 to 3.4 points. On the basis of this new result, one could suppose that the achievement of low *p-values* (and therefore highlighting the hypothetical relationship between variables) is a matter of time, patience and having the necessary resources to recruit large samples (Boukrina, Kucukboyaci, & Dobryakova, 2020).

Figure 2. RCT to assess the benefits of the SVG intervention on AN and BN women

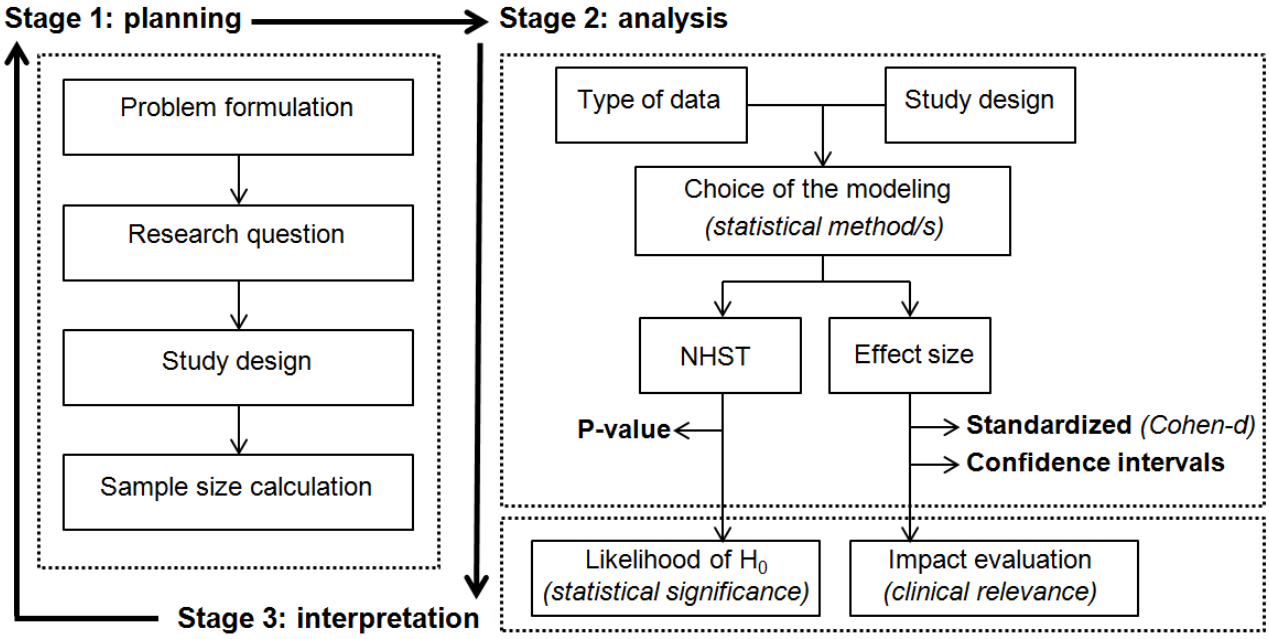


Searching the evidence in ED

Considering the benefits and difficulties of the current statistical approaches in medical scientific research, what should be considered the most appropriate procedure for the contrast of hypothesis in the ED area? Despite the difficulties, the Fisher and Neyman-Pearson theories have been key elements of the statistical methodology for the last century. It is undeniable that NHST and hypothesis testing have provided indispensable tools for clinical studies, and continue to be the framework for basic and applied research. And while the drawbacks of NHST have been detailed in endless forums, it seems that other alternatives proposed to replace or complement *p-values* have not been successful. But now is the time to recognize the value of alternative paradigms for supplementing and enhancing the methods of data analysis, such as the new-Bayesian theory [a number of significant Bayesian factors and effect sizes measures exist (Jeon & De Boeck, 2017; Kelter, 2020; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017)] or other suitable statistics (Krueger & Heck, 2017; Lovell, 2020; Wilson, Harris, & Wixted, 2020).

At present, an increasing number of scientific journals in Medicine and Health, such as the European Eating Disorders Review, are publishing studies that (complementarily to NHST) provide researchers with the tools required to assess the clinical relevance of the empirical evidence: effect size measures. This editorial decision agrees with the recommendations of the American Statistical Association [https://www.amstat.org/], which warns that *p-values* should never be interpreted in isolation from other additional evidence observed in research studies (Wasserstein & Lazar, 2016). The Publication Manual for academic and scientific documents of the American Psychological Association [https://www.apa.org/ (American Psychological Association (APA), 2019)], which contains the standards for a large number of papers published on Social and Behavioral Sciences, also indicates that an adequate interpretation of the empirical results should be based on other elements that complement the NHST, mainly the calculation of effect sizes.

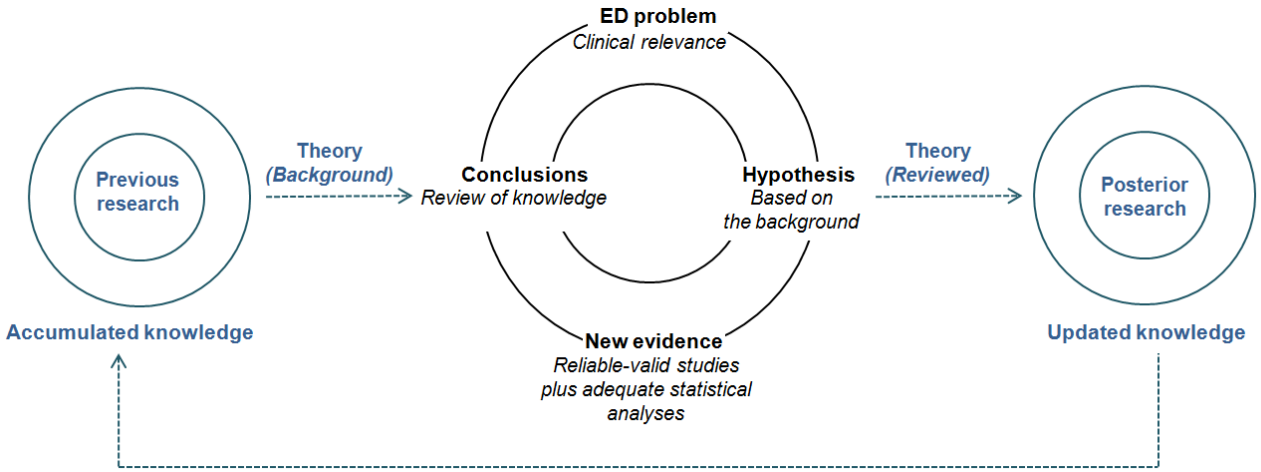
Figure 3. The process of the study



Our recommendation is to follow the process shown in Figure 3. Proper statistical analytical practice involves *always* complementing the *p-value* obtained through NHST with other tools that can assess the clinical relevance of the effect (effect size measures and graphics are useful). These measures of the effect size play a fundamental role because they can offer a more complete, detailed and realistic view of the phenomenon (problem) under study than conclusions based only on the *p-values* (which are also often subject to misinterpretation and over-valuations). Complete numerical and graphic results obtained in the analytical plan should be logically integrated within the theoretical context, since only clinically consistent results can lead to progress in scientific reasoning. This is a key concept of evidence based medicine EBM (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996), which promotes the integration of clinical knowledge with the best available empirical evidence in order to make proficient decisions about the care of patients. The principles of EBM have represented a relevant step toward the implementation of valuable tools in ED clinical practice (Bulik, 2016; Hilbert, Hoek, & Schmidt, 2017; Stice, Johnson, & Turgon, 2019), with a growing body of literature including well-designed, well-analyzed and well-interpreted studies that constitute the basis for offering clinically useful, reliable and updated guidance.

Lastly, a final thought about the way knowledge is built in the ED area. The solving of a clinical problem involves research activities using the circular scientific method (Figure 4), and any point of the process could lead to many possible next steps. Within this iterative progression, adequate statistical analysis carried out in a well-designed study could lead to expected or surprising evidence, but should always contribute to better planning of posterior research.

Figure 4. The iterative process of the scientific knowledge



Acknowledgment

We thank CERCA Programme/Generalitat de Catalunya or institutional support. This work was partially supported by Plan Nacional sobre Drogas (project 2019I47), Instituto de Salud Carlos III (PI17/01167) and Generalitat de Catalunya (PERIS/SLT006/17/00246). CIBEROobn is an initiative of ISCIII Spain.

References

American Psychological Association (APA). (2019). Publication manual of the American Psychological Association (7th ed.). In *Publication manual of the American Psychological Association (7th ed.)*. https://doi.org/10.1037/0000165-000

Aoun, A, Joundi, J, El Gerges, N. (2019). Prevalence and correlates of a positive screen for eating disorders among Syrian refugees. *Eur Eat Disorders Rev*., 27: 263–273. https://doi.org/10.1002/erv.2660

Badenes-Ribera, L., Frias-Navarro, D., Iotti, N. O., Bonilla-Campos, A., & Longobardi, C. (2018). Perceived Statistical Knowledge Level and Self-Reported Statistical Practice Among Academic Psychologists. *Frontiers in Psychology*, *9*, 996. https://doi.org/10.3389/fpsyg.2018.00996

Bagaric, M, Touyz, S, Heriseanu, A, Conti, J, Hay, P. Are bulimia nervosa and binge eating disorder increasing? Results of a population-based study of lifetime prevalence and lifetime prevalence by age in South Australia. Eur Eat Disorders Rev. 2020; 28: 260– 268. https://doi.org/10.1002/erv.2726

Bayes, T., & Price, R. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418. https://doi.org/10.1098/rstl.1763.0053

Beck-Bomholdt, H. P., & Dubben, H. H. (1996). Is the Pope an alien? *Nature*, *381*, 730. https://doi.org/10.1038/381730d0

Boukrina, O., Kucukboyaci, N. E., & Dobryakova, E. (2020). Considerations of power and sample size in rehabilitation research. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, *154*, 6–14. https://doi.org/10.1016/j.ijpsycho.2019.08.009

Bulik, C. M. (2016). Towards a science of eating disorders: Replacing myths with realities: The fourth Birgit Olsson lecture. *Nordic Journal of Psychiatry*, *70*(3), 224–230.

https://doi.org/10.3109/08039488.2015.1074284

Byrd, R. C. (1988). Positive therapeutic effects of intercessory prayer in a coronary care unit population. *Southern Medical Journal*, *81*(7), 826–829. https://doi.org/10.1097/00007611-198807000-00005

Christensen, SS, Bentz, M, Clemmensen, L, Strandberg-Larsen, K, Olsen, EM. (2019). Disordered eating behaviours and autistic traits—Are there any associations in nonclinical populations? A systematic review. Eur Eat Disorders Rev., *27*, 8– 23. https://doi.org/10.1002/erv.2627

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences Hill (2nd Edition)*. https://doi.org/10.1111/1467-8721.ep10768783

Cohen, J. (1994). The Earth Is Round (p < .05). *American Psychologist*, *49*, 997–1003. https://doi.org/10.1037//0003-066X.49.12.997

Fernandez-Aranda, F., Jimenez-Murcia, S., Santamaría, J. J., Giner-Bartolomé, C., Mestre-Bach, G., Granero, R., … Menchón, J. M. (2015). The Use of Videogames as Complementary Therapeutic Tool for Cognitive Behavioral Therapy in Bulimia Nervosa Patients. *Cyberpsychology, Behavior, and Social Networking*, *18*(12), 1-8. https://doi.org/10.1089/cyber.2015.0265

Fisher, R. (1925). Statistical methods for research workers. In *Biological monographs and manuals*. Edinburgh: Oliver and Boyd.

Galmiche, M., Déchelotte, P., Lambert, G., & Tavolacci, M. P. (2019). Prevalence of eating disorders over the 2000-2018 period: a systematic literature review. *The American Journal of Clinical Nutrition*, *109*(5), 1402–1413. https://doi.org/10.1093/ajcn/nqy342

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Goodman, S. N. (1993). p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, *137*(5), 485–501. https://doi.org/10.1093/oxfordjournals.aje.a116700

Griffiths, P., & Needleman, J. (2019). Statistical significance testing and p-values: Defending the indefensible? A discussion paper and position statement. *International Journal of Nursing Studies*, *99*, 103384. https://doi.org/10.1016/j.ijnurstu.2019.07.001

Hilbert, A., Hoek, H. W., & Schmidt, R. (2017). Evidence-based clinical guidelines for eating disorders: international comparison. *Current Opinion in Psychiatry*, *30*(6), 423–437. https://doi.org/10.1097/YCO.0000000000000360

Jeon, M., & De Boeck, P. (2017). Decision qualities of Bayes factor and p value-based hypothesis testing. *Psychological Methods*, *22*(2), 340–360. https://doi.org/10.1037/met0000140

Kelter, R. (2020). Analysis of Bayesian posterior significance and effect size indices for the two-

sample t-test to support reproducible medical research. *BMC Medical Research Methodology*, *20*(1), 88. https://doi.org/10.1186/s12874-020-00968-2

Krueger, J. I., & Heck, P. R. (2017). The Heuristic Value of p in Inductive Statistical Inference. *Frontiers in Psychology*, *8*, 908. https://doi.org/10.3389/fpsyg.2017.00908

Lazzeroni, L. C., & Ray, A. (2012). The cost of large numbers of hypothesis tests on power, effect size and sample size. *Molecular Psychiatry*, *17*(1), 108–114. https://doi.org/10.1038/mp.2010.117

Lee, D. K. (2016). Alternatives to P value: confidence interval and effect size. *Korean Journal of Anesthesiology*, *69*(6), 555–562. https://doi.org/10.4097/kjae.2016.69.6.555

Lie, S. Ø., Rø, Ø., & Bang, L. (2019). Is bullying and teasing associated with eating disorders? A systematic review and meta-analysis. *The International Journal of Eating Disorders*, *52*(5), 497–514. https://doi.org/10.1002/eat.23035

Lovell, D. P. (2020). Null hypothesis significance testing and effect sizes: can we "effect" everything … or … anything? *Current Opinion in Pharmacology*. https://doi.org/10.1016/j.coph.2019.12.001

Mallorquí-Bagué, N., Lozano-Madrid, M., Testa, G., Vintró-Alcaraz, C., Sánchez, I., Riesco, N., … Fernández-Aranda, F. (2020). Clinical and Neurophysiological Correlates of Emotion and Food Craving Regulation in Patients with Anorexia Nervosa. *Journal of Clinical Medicine*, *9*(4), e:960. https://doi.org/10.3390/jcm9040960

Morris, P. H. (2020). Misunderstandings and omissions in textbook accounts of effect sizes. *British Journal of Psychology (London, England : 1953)*, *111*(2), 395–410. https://doi.org/10.1111/bjop.12401

Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*, 289–337. https://doi.org/10.1098/rsta.1933.0009

Quadflieg, N., Schädler, D., Naab, S., & Fichter, M. M. (2017). RCT of a Video-based Intervention Program for Caregivers of Patients with an Eating Disorder. *European Eating Disorders Review : The Journal of the Eating Disorders Association*, *25*(4), 283–292. https://doi.org/10.1002/erv.2521

Ribeiro, M., Conceição, E., Vaz, A. R., & Machado, P. P. P. (2014). The prevalence of binge eating disorder in a sample of college students in the north of Portugal. *European Eating Disorders Review : The Journal of the Eating Disorders Association*, *22*(3), 185–190. https://doi.org/10.1002/erv.2283

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996).

Evidence based medicine: what it is and what it isn't. *BMJ (Clinical Research Ed.)*, Vol. 312, pp. 71–72. https://doi.org/10.1136/bmj.312.7023.71

Savitz, D. A., Tolo, K. A., & Poole, C. (1994). Statistical significance testing in the American Journal of Epidemiology, 1970-1990. *American Journal of Epidemiology*, *139*(10), 1047–1052. https://doi.org/10.1093/oxfordjournals.aje.a116944

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. https://doi.org/10.1037/met0000061

Smith, R. J. (2020). P > .05: The incorrect interpretation of "not significant" results is a significant problem. *American Journal of Physical Anthropology*, e24092. https://doi.org/10.1002/ajpa.24092

Steyerberg, E. W., & Van Calster, B. (2020, May). Redefining significance and reproducibility for medical research: A plea for higher P-value thresholds for diagnostic and prognostic models. *European Journal of Clinical Investigation*, Vol. 50, p. e13229. https://doi.org/10.1111/eci.13229

Stice, E., Johnson, S., & Turgon, R. (2019). Eating Disorder Prevention. *The Psychiatric Clinics of North America*, *42*(2), 309–318. https://doi.org/10.1016/j.psc.2019.01.012

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size-or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Svedlund, N. E., Norring, C., Ginsberg, Y., & von Hausswolff-Juhlin, Y. (2018). Are treatment results for eating disorders affected by ADHD symptoms? A one-year follow-up of adult females. *European Eating Disorders Review : The Journal of the Eating Disorders Association*, *26*(4), 337–345. https://doi.org/10.1002/erv.2598

Udo, T., & Grilo, C. M. (2018). Prevalence and Correlates of DSM-5-Defined Eating Disorders in a Nationally Representative Sample of U.S. Adults. *Biological Psychiatry*, *84*(5), 345–354. https://doi.org/10.1016/j.biopsych.2018.03.014

Unwin, S. (2004). *The probability of God: a simple calculation that proves the ultimate truth*. New York: Three Rivers Press.

Van Calster, B., Steyerberg, E. W., Collins, G. S., & Smits, T. (2018). Consequences of relying on statistical significance: Some illustrations. *European Journal of Clinical Investigation*, *48*(5), e12912. https://doi.org/10.1111/eci.12912

Wasserstein, R. L., & Lazar, N. (2016). The ASA Statement on Statistical Significance and P-values. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wellek, S. (2017). A critical evaluation of the current "p-value controversy". *Biometrical Journal.*

Biometrische Zeitschrift, *59*(5), 854–872. https://doi.org/10.1002/bimj.201700001

Wilson, B. M., Harris, C. R., & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(11), 5559–5567. https://doi.org/10.1073/pnas.1914237117