Therapeutic Trials, Statistical Significance, and Clinical Relevance

Large therapeutic clinical trials are the best scientific sustenance of evidence-based medicine. The article published in 1983, which could be called today the manifesto of the mega-trials (*Why do we need some large, simple randomized trials?*),[1] focused on the most common health problems, marked a shift in paradigms about the way to design and carry out scientific trials.

In brief, what this publication proposed was to develop trials on thousands or tens of thousands of patients to answer relevant questions on common diseases. This would allow evaluation of treatments from which moderate beneficial effects were expected (20-30% mortality reduction), and which –if successful– would have epidemiological impact. It is no coincidence then that the great vanguard in this regard have been cardiovascular trials, and particularly myocardial infarction, through the ISIS (Oxford) and GISSI (Italy) networks, and their later expansion to the large business and scientific networks of today. That led cardiologists to get familiar –perhaps sooner than other medical specialists– with the new nomenclature and ways of interpreting data.

Little by little, we learned what a confidence interval and a significant P expressed, and then a number of concepts such as relative risk, absolute and relative risk reduction, odds ratio, and number needed to treat. Meta-analysis also became a common tool in the practice of cardiology.

Statistical concepts have been evolving both for the design and the interpretation of results, but their spread and understanding on the part of doctors is variable. Over the past years, there has been a growing discussion as a result of a different historical moment.

What are the differences between these two periods, and how they are impacting on the interpretation of clinical trials?

I know the reader will understand when I say that this is an outlined description focused only on major trends. In the 1980s, we had no effective treatments for myocardial infarction –except when facing some complications–, and very few resources in heart failure. There were no validated routine behaviors for secondary prevention either. The emergence of the new form of research and its extraordinary success led to include a valuable therapeutic arsenal, and to markedly modify the natural evolution in different areas of cardiovascular disease. It was easy and ethical to conduct studies against placebo with different interventions, and an extraordinary community enthusiasm to participate in research networks was generated. So local designed trials in Argentina were conducted, like EMERAS[2], GEMICA[3], GESICA[4], ENAI[5], with thousands of patients and no –or minimum– compensation for researchers. The unquestionable endpoint was mortality or –at that time– a hard event, like evolving heart attack (pain, ECG changes and increased enzyme levels). Finding significant differences and expressing them in the new nomenclature allowed to rapidly put into practice many of the measures under research, with a remarkable impact on quality of life and survival improvement. Given the community motivation, the nature of the problems studied and the independent structures, regulatory guidelines and time spent on paperwork or stories were minimal.

At present, there are multiple validated therapies, so it is difficult to conduct studies against placebo, and also try to impact on mortality above an established treatment with similar outcomes. In turn, catheter intervention has become a routine therapy in coronary artery disease, and an important aspect of medical expenses is related to its outcomes and complications. The description refers to cardiology, but similar problems cover almost all the areas in medicine.

Table 1

	1980s	2010
Cardiovascular disease Treatments with impact on mortality	Few	Multiple
Networks for research	Communitary and collaborative	High cost business and scientific networks
Endpoints	Hard: mortality	Soft: combined (Different definitions of relevant myocardial infarctions, major bleeding, revascularization, etc.)
Statistical significance	Close to clinical relevance	Clinical relevance questioned

So, studies raise questions that do not have the revolutionary weight of the 80s:

Can a patient who already receives beta-blockers, aspirin, angiotensin-converting enzyme and statins after a myocardial infarction benefit from a new drug that acts on a different anti-atherosclerotic mechanism?

During angioplasty, what antithrombotic and antiplatelet therapy is more effective in terms of periprocedural infarction and bleeding?

Is a certain therapy equivalent, and not inferior to the previous validated treatment: angiotensin receptor blockers vs. inhibitors, different beta-blockers, different antihypertensive groups, different hypoglycemic agents?

Since most new treatments cannot reduce mortality with respect to the previous ones, surrogate or combined endpoints are assessed. Thus, there is a classification into cardiovascular and non cardiovascular mortality to enhance the power of the study, myocardial infarction is defined with criteria that increase its incidence and that have not been standardized yet, or benefits are obtained for some points, but not for other relevant ones, for which impact is low. With very large numbers of patients, results are statistically significant, but with a questionable eventual clinical impact, or on events of little clinical significance, such as the late-loss in studies with coronary stents.

With this background, there have been proposals to focus on outcomes and eventually the design of clinical trials from a different perspective, oriented towards reaching a consensus on measures of statistical significance, in addition to some criteria about clinical relevance. In this letter, I will review some basic concepts of statistical interpretation of clinical trials, and the characteristics of these new proposals.

P-LEVEL AND STATISTICAL SIGNIFICANCE

The introduction of statistics in the medical thinking has been difficult, partly due to an essential and unsolved conflict: the underlying need –in the medical act– to collaborate in the care of the patient who comes to consultation today, as a result of scientific tests based on trials grounded on concepts of probability of the large numbers, but say very little about this individual patient.

The first complex concept is to understand the p-level and the meaning of statistical significance.

Comparative treatment research on clinical events is formulated statistically as a hypothesis testing. We state a null hypothesis, which expresses that there will be no differences between the treatments to be assessed in the study, and an alternative hypothesis, which –by rejecting the first one– suggests that treatments have different effects.

Let's suppose that treatments are a drug A vs. placebo on mortality.

Null hypothesis:
mortality A = mortality P;
it can also be expressed as
mortality A - mortality of P = 0.

Alternative hypothesis:
A<>P or A-P<>0.

Clinical trials work with samples, and statistical evaluations try to limit the effect of chance on the results, assuming an uncertainty margin that is typical of probabilistic thinking. Errors in the interpretation are assumed to occur, and the idea is to limit those errors to a minimum, or to levels accepted by the community.

Alpha error

The alpha error is set, in practice the p-level to be considered significant for the trial, usually <0.05 or < 0.01. If, as a result of the statistical comparison, the p-value is lower than the alpha error set a priori, we reject the null hypothesis and embrace the alternative hypothesis: there are 'significant' differences between treatments.

Alpha error expresses the probability of being wrong when rejecting the null hypothesis, that is, of considering the study a false positive.

In other words, if the p-value was < 0.05, we are confirming that we reject the null hypothesis, that we accept that there are significant differences between the treatments, and that the error we can incur when confirming is less than 5%; there is a probability of less than 5% that a difference of this magnitude could happen by chance.

The p-value does not express something simple or intuitive, and is influenced by the number of the sample: large differences in the outcome of treatments may result in a signification of p < 0.05 or not significant in small studies; in turn, small differences in very large studies may result in significant differences of p < 0.01. *There is no direct relationship between p-value and clinical relevance.*

The choice of p-level < 0.05 as cut-off point for statistical significance is absolutely conventional, and has received several technical and conceptual criticisms. However, the advantage is that it is easy to calculate through statistical programs, and enjoys consensus in the literature and regulatory bodies: if p is 0.048, the treatment was useful, but if it is 0.052, it was not useful. According to what we have commented above, the difference between the two results is only a variation of 4 per thousand in the possibility of committing a statistical error, conceptually trivial but accepted as dichotomous (success – failure) in the body of current beliefs.

P-value or confidence intervals

In 1980s, several authors spread the limitations of the conventional p-level, its lack of relationship with clinical relevance, and its weakness to turn complex and uncertain problems into successes and failures through a conventional parameter. They proposed to report the outcomes in terms of confidence intervals, and even to ignore –when possible– the p-value, which was immediately accepted by many medical journals. [4]

In Table 2, which was taken from the software Evicardio®, we calculate the impact on mortality of streptokinase versus placebo in myocardial infarction in the GISSI I study.

Confidence intervals allow to estimate the effect size that is closer to clinical thinking; from a healthcare approach, they allow to estimate the eventual cost of treatment by number needed to treat: in this case, 45 patients should be treated in order to save a life.

The confidence interval has a clear statistical basis: the outcomes of the studies always have a degree of uncertainty and possibility of error.

The basic conceptual translation of the confidence interval is that there is a 95% probability that the true effect is covered by those values. In this case, we have a confidence interval of 95% or a margin of error of 5% when stating that the true reduction in mortality with streptokinase is between 1.1 and 3.4 deaths per 100 treated patients. There is a probability of less than 2.5% that the reduction be < 1.1, and also a probability of less than 2.5 that it be > 3.4.

Graphically, the topic has been simplified, and we have got used to recognizing quickly whether or not the outcome crossess the tie line: RR = 1 when working with relative risk, or ARR = 0 with absolute risk reduction.

STATISTICAL SIGNIFICANCE AND CLINICAL RELEVANCE

P-value and confidence interval limitations

In the 1990s, Alvan Feinstein and other authors made some critical remarks about p-value and confidence interval limitations, which have currently turned into a different way of dealing with trials results. [5] In one of their most representative papers, they argued that both p-value and confidence interval referred to the uncertainty of rejecting the null hypothesis, but that they were just two sides of the same coin. None of them was oriented to determine whether the result was medically relevant, or whether it was large enough to influence in a therapeutic behavior in the individual patient. While the concepts of evidence-based medicine and measures of effect are more clinical than the p-value, they do not tell us much about the relevance of the introduction of treatment in practice.

Considering what was discussed in the introduction, a historical moment in which there are several works with combined endpoints of different relevance, or large non-inferiority studies, with small –but statistically significant– differences, the issue becomes even more important.

Table 2. Mortality with streptokinase vs. placebo in the GISSI I study

Intervention	Event +	Total	Percentage
STREPTOKINASE	628	5860	10.7%
CONTROL	758	5852	13.0%
Global	1386	11712	11.8%
	p < 0.001	**CI 95%**	
ARR	2.2%	1.1%	3.4%
NNT	44.7	29.4	93.7
RR	0.83	0.75	0.91
RRR	17.3%	8.6%	25.1%
OR	0.81	0.72	0.90

Very significant p-value. Expressed in terms of measures of treatment effect and its confidence intervals, we have a more precise idea of the magnitude of the clinical impact: Mortality is reduced 17% globally (95% CI 8.6-25.1), every 100 treated patients we can reduce deaths to 2.2 (95% CI 1.1 3-3.4).

The Bayesian proposal

A contribution from the Bayesian thinking has been projected to the analysis of therapeutic trials. [6]. The argument is that p-values and confidence interval give a false sense of security against the logical uncertainty of therapeutic trials. The idea is to convey the physician that margin of uncertainty through new conceptual tools.

The Bayesian diagnostic thinking tells us that given a certain test with its sensitivity and specificity, we can estimate how much it will help us improve interpretation when applied on different populations with different prevalences: The so-called pre-test probability and post-test probability, when the result

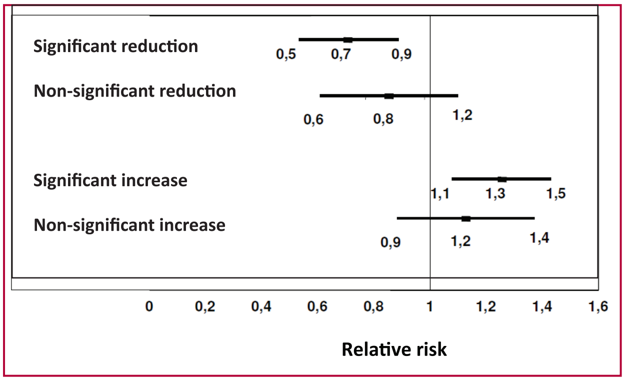


Fig 1. Graphic representation of the comparative outcomes of a treatment against placebo with respect to an adverse event. The central rectangle indicates the relative risk RR observed, and the arms indicate the confidence interval of 95% (95% CI). If treatment is superior to placebo, it will be expressed as a RR lower than 1, and if it is harmful, as a RR greater than 1. If the 95% CI does not cross 1, then the beneficial or harmful effect is significant (we can ensure that the treatment is beneficial or harmful). If the 95% CI crosses 1, we cannot confirm with an error < 5% whether the treatment is beneficial or harmful.

is positive or negative. For the interpretation of therapeutic trials, the proposal is to explain different scenarios and turn the information into a continuum of percentage possibilities. One of the major problems of this thinking process is to establish the 'prior probability', simple in diagnostic studies but –so far– very difficult in therapeutic trials. The attempts of the Bayesian groups for the American Food and Drug Administration (FDA) to accept their arguments as the basis for the validity of clinical trials have advanced significantly in recent years, and an official guidance has been recently released for use by the industry and the FDA.[7]. It has also led to a conceptual attempt by non-Bayesian statisticians, called *frequentists* in this case, to express the same concepts with the classical analysis tools.

Use of confidence interval in the Bayesian approach

We have discussed that 95% confidence interval expresses the values between which we believe the true effect is, with 95% confidence. Another way of interpreting confidence interval is to asssume, on the basis of results, the probability of different percentage effects; it has been proposed by Sackett in recent years. [8]

Returning to the example of GISSI:

We have observed a reduction in mortality of around 17% (9% to 25%). The confidence interval is calculated according to the Gaussian distribution, by estimating the standard error of mortality reduction, and with the formula:

95% CI = % of reduction observed ± 1.96 * Standard error reduction.

In this case, we might outline 1.96 as 2, and assume that the confidence interval is symmetric: 8% above and below 17%, which would indicate a standard error of about 8%/2= 4%.

Being aware of the Gaussian distribution, we can estimate different percentage points of effect, and their probability. For example, if we recall that ± 1 standard error covers 68% of the probability (34% below and 34% above the average percentage), we may argue that it is 16% less likely that the effect be > 21 or < 13% (calculating 17% ± 4%, ie, observed percentage ± 1 standard error).

CLINICAL RELEVANCE AND STATISTICAL SIGNIFICANCE

In February 2010, Sanjay Kaul and George Diamond proposed a new form of graphication of results, trying to approach to the concept of clinical relevance . It is clear that the importance or relevance of a therapeutic intervention is strongly influenced by the context of the problem, its costs, therapeutic alternatives, and risks involved, so it is impossible to set a 'logical' or desirable percentage. The creators of the paradigm of large trials spoke about moderate effects, 20-30%, on highly relevant events. A 30% reduction in mortality is not the same as the periprocedural infarction diagnosed by minor increases in troponin, or the need for a late intervention that involves no mortality.

Despite these limitations, the pattern is valid as a conceptual tool. As proposed by the authors, let's suppose that for a problem in particular, we set a minimum level of impact that we will consider important, in this case, a 15% reduction of an event. The relative risk considered clinically important (minimum clinically important difference, MCID) is 0.85. In Figure 1, Part A, of the work cited (13), the authors propose the following pattern, which we have reproduced here, in Figure 2.

Starting at the top in Figure 2, the first study did not achieve a statistically significant reduction, ie, it could not discard the null hypothesis expressed by the RR = 1 in this case. The arm of its confidence interval is also separated from the minimum clinically important difference –MCID–, set as RR 0.85, ie, a 15% reduction of the event.

The second study does not reject the null hypothesis, but its confidence interval includes the MCID. The need for an evaluation with a larger number of patients would be recommendable.

The third study has a narrow confidence interval, and while the null hypothesis is rejected, its benefit on the event is reduced and is not close to the MCID that we have set as relevant. Chances of achieving a greater reduction than the MCID are less than 2.5%, because it is outside one of the two tails of the 95% confidence interval.

The fourth study rejects the null hypothesis and includes the MCID; therefore, it may be relevant according to our clinical criteria. In this case, we can estimate that the probability of achieving a reduction greater than the MCID is 50% (central effect coincides with 0.85). As you see, by knowing the results of the study and the magnitude of the desired reduction, we can calculate its probability.

The fifth study not only rejects the null hypothesis but also ensures the reduction of the event over our MCID, ie, it provides a statistically significant and clinically relevant response to our criterion, established a priori.

Estimation of clinical relevance and IT tools

In their article, the authors use the example of the TACTICS-TIMI 18 study[10] to analyze the effects of reducing the incidence of the combined event. In the protocol, an expected reduction of 25% had been estimated for the calculation of the sample. The real reduction was 18% (95% CI 2-32%), statistically significant with a p of 0.028. It was calculated that, according to results, the probability to achieve a 25% reduction –which, a priori, was considered valid– was only 17%. One more element to judge the relevance of the trial, which does not arise directly from the observation of the confidence interval. The authors

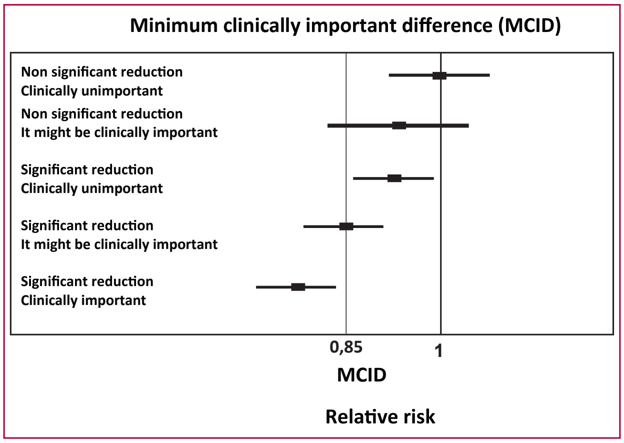


Fig 2. Five clinical trials have been graphicated, with their 95% CI and relative risk. In addition to the vertical line for RR = 1, a line was drawn in the RR 0.85, which is the minimum clinically important difference considered acceptable here. Explanation in the text. Modified from cite 13.

make this calculation on the basis of the Bayesian reasoning.

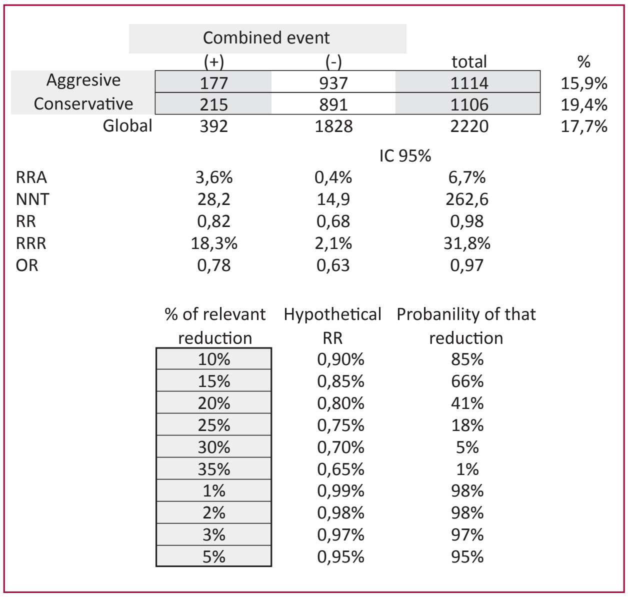
The same can be obtained from the conventional ('frequentist') statistics. I have built a spreadsheet that will be available as an attachment to this work, and can be downloaded from http://www.sac.org.ar/web/es/revista-argentina-de-cardiologia. Here is an example of its use. Figure 3.

We have entered the data from the TACTICS-TIMI 18 study in the shaded boxes of the 2x2 table. The combined event was presented in 177/1114 patients with the initial aggressive treatment, and in 215/1106 patients from the conservative group. The spreadsheet calculates the percentages, measures of effect and confidence interval, and, at the bottom, the probability in this study to achieve a reduction of different percentages of the event, which can be changed. In this case, 10-15-20%, etc. were used. Clearly, the probability of an estimated 25% reduction was 18% (17.6% with a decimal place), very similar to the 17% calculated by the authors using Bayesian methods.

CONCLUSIONS AND FINAL COMMENTS

Conceptually, evidence-based medicine provides tools to avoid measures proved ineffective, and help us choose the therapies for which there is better evidence. Despite the imaginary of Sackett et al., it is virtually impossible to address the search for evidences on an individual basis, due to time and increased complexity in the interpretation of literature.

The physician has to resort to community consensus or guidelines that should evaluate the information in a serious and responsible way, but that, in practice, are strongly influenced by conflicts of interest that have resulted in many recommendations being debatable or based on weak evidence.



	Combined event			
	(+)	(-)	total	%
Aggressive	177	937	1114	15,9%
Conservative	215	891	1106	19,4%
Global	392	1828	2220	17,7%

		IC 95%	
RRA	3,6%	0,4%	6,7%
NNT	28,2	14,9	262,6
RR	0,82	0,68	0,98
RRR	18,3%	2,1%	31,8%
OR	0,78	0,63	0,97

% of relevant reduction	Hypothetical RR	Probanility of that reduction
10%	0,90%	85%
15%	0,85%	66%
20%	0,80%	41%
25%	0,75%	18%
30%	0,70%	5%
35%	0,65%	1%
1%	0,99%	98%
2%	0,98%	98%
3%	0,97%	97%
5%	0,95%	95%

Fig 3. Table for statistical analysis of a clinical trial. By entering the data in the four shaded boxes at the top, it calculates automatically the statistical significance of the difference in the effects of treatment, the measures of effect and their confidence intervals. In the shaded boxes at the bottom, different percentages of reduction of the event –like the MCID– can be set, and the spreadsheet indicates the probability of that reduction.

Overcoming these limitations is surely a collective task as well. The new proposals to analyze the information from the perspective of the relevance and clinical significance are a contribution in this regard, with necessarily subjective criteria, but closer to the physician's decisions about the patient.

This letter has intended to approach to this line of thought, which is now in its beginnings but will certainly be very fruitful, and to provide an IT tool that I hope will facilitate the implementation of this concept to the reading of clinical trials.

To establish the relevance and clinical significance is a community task, which may condition the design of trials in the future, which will be more patient-oriented than focused on intervention or drugs.

Carlos Daniel Tajer, M.D.
Director of the Argentine Society of Cardiology

BIBLIOGRAPHY

1. Yusuf S, Collins R, Peto R (1984) *Why do we need some large, simple randomized trials?* Stat Med 3: 409–422.

2. EMERAS (Estudio multicéntrico estreptoquinasa Repúblicas de América del Sur) Collaborative group. *Randomized trial of late thrombolysis in patients with suspected actue myocardial infarction.* Lancet 1993; 342:767-772.

3. Elizari M, Martinez JM, Belziti C y col. *Morbidity and mortality* *following early administration of amiodarone in acute myocardial infarction.* GEMICA study investigators, GEMA Group, Buenos Aires, Argentina. Grupo de Estudios Multicéntricos en Argentina. Eur Heart J. 2000 Feb; 21(3):198-205.

4. Doval H, Nul D, Grancelli H y col., for the Grupo de Estudio de la Sobrevida en la Insuficencia Cardíaca en Argentina (GESICA) Investigators. *Randomised trial of low dose amiodarone in severe congestive heart failure*. Lancet 1994; 344:493-498.

5. Tajer C, Grancelli H, Hirschson Prado A, et al. *Enalapril in unstable angina: a randomized double blind multicentre trial*. Eur Heart J 1995;16:259. *Enalapril en la angina inestable. Estudio multicéntrico* ENAI. Rev Argent Cardiol 1996; 64:31-47.

6. Lang J, Rothman K, Cann C. *The confounded p value. Epidemiology* 1998: 9:7-8

7. Ioannidis J. *Why most published research findings are false*. PLoS Med 2005; 2:696-701

8. Gardner MJ, Altman DG. *Confidence intervals rather than P values: estimation rather than hypothesis testing*. Br Med J 1986; 292:746–750

9. Feinstein A. *P-Values and Confidence Intervals:Two Sides of the Same Unsatisfactory Coin*. J Clin Epidemiol 1998: 51; 355–360

10. Diamond GA, Kaul S. *Bayesian approaches to the analysis and interpretation of clinical megatrials*. J Am Coll Cardiol 2004;43: 1929 –39

11. Campbell G. Guidance for Industry and FDA Staff Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. 2010. It can be downloaded at: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf

12. Sackett DL. *Superiority, equivalence and noninferiority trials*. In: Haynes RB, Sackett DL, Guyatt GH, Tugwell P, editors. Clinical Epidemiology. Third edition. Lippincott Williams & Wilkins, 2006:

13. Kaul S; Diamond G. *Trial and Error How to Avoid Commonly Encountered Limitations of Published Clinical Trials* J Am Coll Cardiol 2010;55:415–27

14. Wiviott SD, Braunwald E, McCabe CH, et al., for the TRITON TIMI 38 Investigators. *Prasugrel versus clopidogrel in patients with acute coronary syndromes*. N Engl J Med 2007; 357:2001–15