Cancer Immunol Immunother (2010) 59:1137–1140
DOI 10.1007/s00262-010-0859-4

EDITORIAL

The tyranny of statistics in medicine: a critique of unthinking adherence to an arbitrary *p* value

Malcolm S. Mitchell · Mimi C. Yu · Theresa L. Whiteside

Received: 5 April 2010 / Accepted: 22 April 2010 / Published online: 9 May 2010

There is no question in anyone's mind, and certainly not in ours, that statistics are essential to ensure that clinical trials are properly designed, with sufficient numbers of patients and observations to lead to a meaningful evaluation of the outcome. From that evaluation, also based on appropriate statistical analysis, emanates the design of follow-up trials either with the same treatment (if the outcome is "positive"), or a significantly modified therapy or entirely different approach (if the results are "negative"). It is this binary emphasis on "positive" and "negative" results that this commentary will focus upon, questioning what these terms really mean and how they affect our perception of "therapeutic success". We will specifically consider the possibility that potentially successful treatments have been overlooked or abandoned because of an inappropriate emphasis on an arbitrarily fixed level of "type 1" or "α" error (namely, declaring the existence of a treatment effect when none exists). We also will consider the case of insufficient emphasis on the possibility of declaring an absence of treatment effect when it actually exists ("type 2" or "β" error). The biological or medical importance of the treatment often seems to be given a secondary role to the statistics imposed on the analysis of the study, which is exactly opposite to what should be the order of importance[1].

A key element of the issue we will discuss is the role of the p value, more specifically the (in)famous "$p < 0.05$". In the early development of statistics, by such luminaries as R.A. Fisher—who was, not incidentally, a renowned geneticist besides being a statistician—it was important to decide what level of confidence investigators should reasonably have in the significance of their observations [2]. One could be somewhat liberal in accepting differences between treatment and control, choosing a level where the differences could have occurred by chance 10 or 15 times in 100 ($p = 0.10$ or 0.15), or choose a much more rigorous level of p, such as $p = 0.01$ or less). Fisher and others suggested that an intermediate level, for example $p < 0.05$, might be sufficient for most investigations, to give one an idea of whether there was a difference between the sets of observations. Thus, "$p < 0.05$" was an arbitrary value selected more as a convenience than a rule. It was assumed that the investigators would pursue additional studies designed to test the difference in other ways; to affirm or refute the results of the first, and that the one trial would not end the exploration of the question (often a treatment). In addition, the co-equal importance of the study's Expected Power (or $1 - \beta$, often expressed as a %) was also stressed, to avoid overlooking a potentially important treatment; perhaps one that showed its usefulness not in all subjects, but in a subgroup that could be further defined. This level was often set, again arbitrarily, at 0.8 or 0.9.

The choice of $p < 0.05$, rather than stating the exact value of probability of the differences, was based in part on

This study has not been presented elsewhere.

M. S. Mitchell (✉)
University of Texas at El Paso,
500 West University Drive, El Paso, TX 79968, USA
e-mail: malcolmsmitchell@yahoo.com

M. C. Yu
The Masonic Cancer Center, University of Minnesota
School of Medicine, Minneapolis, MN, USA

T. L. Whiteside
Pittsburgh Cancer Institute, University of Pittsburgh
School of Medicine, Pittsburgh, PA, USA

[1] Reference [1], Chap. 7, p. 161 ff, considers several of these points in a succinct, insightful exposition.

the simplicity of keeping in mind a single value level of statistical "significance", a round number of 0.05 [1]. Most parametric statistical analyses rely on the assumption that data follow a normal distribution. When the normality assumption is satisfied, a probability value of 0.05 corresponds to the sum of the two-tailed areas under the normal distribution curve, which are approximately two standard deviations (1.96) from the mean. Further, it was based on the level where one's tolerance for the role of chance begins to "wear thin" [1]. For example, as Ingelfinger et al. [1] point out, if one throws a fair (unweighted) coin five times and either heads or tails come up all five times, the probability of each set of occurrences is 1/32, and the probability of one or the other set of occurrences is 1/16 or approximately 0.05. It is not that at slightly more than $p = 0.05$ the difference is not real; it is just that it is a little more likely to have occurred by chance than if the probability is less than 0.05.

More importantly, an arbitrary cutoff at any level ignores what the biological importance of the treatment may be, and what failing to recognize that might mean for the future. That is true in both directions: "significant" or "insignificant." Let us say that a very toxic treatment appears to be effective at $p < 0.05$, perhaps at $p = 0.045$, and leads to approval and licensing of the treatment, and its subsequent widespread use. The actual number of patients who are beneficially affected, with prolonged lifespan, might turn out to be very few, with a far larger number having to tolerate the toxicity in vain. Moreover, repeated trials of the treatment might prove to be "insignificantly" different from either the current or historical controls. Conversely, in a properly designed prospective trial comprising all eligible patients, a treatment might exhibit a minimal effect overall, but demonstrate a highly significant effect within a subgroup of patients who have some biological basis for responding. We recognize the perils of performing a posteriori analyses of subgroups, which constitute statistical testing of hypotheses that were not part of the aims of the original study. These retrospective, multiple analyses are prone to false positive findings, with a collective probability of type 1 error that can be many times higher than the 5% conventional rule. However, if a particular subgroup analysis has a high degree of biological plausibility, these findings should be treated as hypothesis-generating data, and be followed up with appropriately designed second generation studies.

Those conducting biological laboratory experiments or clinical trials, should first decide what *biological* difference would be meaningful to them, and then design their experiments to demonstrate the desired outcome. Likewise, in a clinical trial one must first decide what percentage of difference would be *clinically* significant. Usually a 10% or more difference in a parameter such as overall survival, disease-free survival, or the like is sought. Then one determines the number of patients (a value dependent upon the postulated rate of events in the study population) needed to demonstrate this difference with a high probability of yielding a statistically significant set of results when an effect truly exists (the expected power of the study) and a low probability of yielding statistically significant results when no effect actually exists (type 1 or α error). If the difference is suspected to be large, a relatively small number of patients will suffice. However, if it is a small (but clinically important) difference, perhaps 1%, a large number of patients will be required.

There are numerous examples in the literature of misguided interpretation of statistical significance (type 1 error) or expected power (type 2 error) in clinical trials. These include examples of dismissing a good treatment inappropriately or of accepting a marginally useful treatment based upon one "statistically significant" trial, often in a situation where the control group had an unexpected and irreproducibly long survival in that first Phase III trial. One poignant example of the first situation is the randomized controlled trial performed by Hersey et al. [3], an extensive and careful study of his vaccinia viral oncolysate melanoma vaccine. 700 patients with resected Stage IIB or Stage III melanoma were randomized to receive either the vaccine or no treatment followed for an average (median) of 8 years. To quote these authors: "Analysis on the basis of all eligible, randomized patients (n = 675) found, after a median follow-up period of *8 years*, a median OS of *88* months in the control versus *151* months in the treated group (hazard ratio [HR], *0.81*; 95% confidence interval [CI], 0.64 to 1.02; $p = 0.068$ by stratified univariate Cox analysis). At 5 and 10 yrs, survival rates for control and treated patients were *54.8%* vs. *60.6%* and *41%* vs. *53.4%,* respectively. Median RFS [relapse-free survival] was *43* months in the control group compared with *83* months in the treated group (HR, 0.86; 95% CI, 0.7 to 1.07; $p = 0.17$)." (emphases, ours). Yet, because the 95% confidence intervals of HR (the range in which the probability of occurrence of the true mean is 95% on repeated sampling) barely included 1.0, the authors were obliged to conclude that the treatment was ineffective. Nonetheless, they maintained—correctly we believe—that the statistics did not rule out "important gains from such treatment". The 7.3-year survival of the controls was an unpredictable and improbable longevity; one would reasonably have expected a 3–4 year median survival for the latter [4]. Nonetheless, the hazard ratio indicated the possibility of nearly a 20% improvement in overall survival. The infeasibility of performing another trial of comparable magnitude, in light of the "negative" findings of this study, led to the disuse of the vaccinia oncolysate after many years of investigation.

It is also useful to note the opposite danger: of accepting a biologically marginal result that was statistically "significant." In a single large cooperative group trial of high-dose IFN-α in resected Stage III melanoma the control, untreated melanoma patients had somewhat shorter than usual relapse-free and overall median survivals [5]. If the expected median survivals for untreated controls had been observed in that first Phase III trial, namely approximately 18-month relapse-free survival and 36-month overall survival [4], high-dose IFN-α would not have been approved. Instead, a 12-month median time to relapse and an 18-month median overall survival were found. The 7 months of improved survival in the high-dose IFN-α group versus untreated controls was maintained to the end of the observation period in that study, but was not reproduced in subsequent trials with the same regimen. Indeed, a letter to the editor of the *Journal of Clinical Oncology* by Lens and Dawes [6] concludes that their statistical analysis of all results found no evidence for the effectiveness of IFN-α in this setting. Despite many caveats then, melanoma patients after undergoing resection of their lymph nodes or involved skin, who would otherwise remain asymptomatic for several years, are now being treated routinely with toxic high doses of IFN-α after resection of their lymph node(s) hoping to achieve a 7-month increase in their overall survival.

It is important to recognize that the mean is always ± a random component, such that if one does many trials, some control groups will be outside the 95% confidence limits. Meta-analyses were designed exactly for the purpose of looking at many similar trials and eliminating the bias of anomalous controls in any one of them. Repetition ultimately leads to a reliable conclusion. Notably, in epidemiology the association of smoking with lung cancer was the result of the analysis of many trials. Of course repetition of clinical trials may not be possible because of the expense involved in following large numbers of patients for a long period of time, which a regulatory agency should recognize, but sometimes does not in its demands for conclusive proof of efficacy.

As a final example, consider a genetically defined subgroup of patients who respond to a treatment with a high degree of statistical significance, but where analysis of the group as a whole does not produce a "significant" result. A trial of an allogeneic melanoma lysate vaccine elicited prolonged survival in patients with resected Stage IIB melanoma who had specific HLA haplotypes versus untreated controls. 97 vaccine-treated patients who had at least two of the five HLA antigens previously noted to have an influence on response to the vaccine had an improved relapse-free survival compared with the 78 observation patients (5-year relapse-free survival, 83% vs. 59%; $p = 0.0002$) [7]. Analysis of the overall group of 684 patients showed an "insignificant" difference in relapse-free survival, with a hazard ratio of 0.92. Relapses were fewer than expected in the control group as well as the treatment group, (as with the study mentioned earlier [3]), which made assessment of *overall survival* impossible until an unspecified later date. Moreover, to quote the investigators: "the power to detect a small but clinically significant difference was low", and they acknowledged the probable effect in the HLA subgroups [8]. Nevertheless, this extensive, carefully documented vaccine trial was deemed "negative", even though those who directed and interpreted the trial have remained convinced that efficacy was demonstrated, with minimal toxicity, in a distinct subgroup of patients who could be prospectively identified. Stewart and Kurzrock [9] make essentially the same point for patients who have the appropriate molecular markers for sensitivity to new biological agents such as monoclonal antibodies: they are a small subgroup of the whole, but are eminently detectable in advance of the trial by modern in vitro techniques.

There are other examples of similar occurrences, where instead of using statistics to evaluate an outcome and then proceeding to confirmation and extension of biologically interesting results, the failure to attain an arbitrary p value or a study underpowered to detect an important subgroup of patients has led to abandoning a useful treatment. Reporting the actual p value and leaving the interpretation to the investigators as well as to the scientific audience would be a much more useful approach in any case.

Rather than belabor the point, we feel it important simply to re-emphasize the perspective of the "founding fathers" of biostatistics, which has been stressed many times since, but apparently forgotten just as often: that statistics is simply a tool useful in research, and should not be its master. All statistics should be interpreted judiciously, by the scientists and clinicians performing the trials, and the readers of their papers, who may then decide which results are worth pursuing further and which are best viewed as less worthwhile from the standpoint of benefits versus risks (toxicity). The binary decisions often forced on investigators by external agencies, or imposed by the financial pressures on the sponsors (manufacturers), may be a fact of life at the moment, but they are surely not in the best interests of science or the public.

References

1. Ingelfinger JA, Mosteller F, Thibodeau LA, Ware JH (1995) Biostatistics in clinical medicine, 3rd edn. McGraw–Hill, New York
2. Fisher RA (1925) Statistical methods for research workers. Oliver and Boyd, Edinburgh

3. Hersey P, Coates AS, McCarthy WH et al (2002) Adjuvant immunotherapy of patients with high-risk melanoma using vaccinia viral lysates of melanoma: results of a randomized trial. J Clin Oncol 20:4181–4190
4. Balch CM, Soon S-J, Gershenwald JE et al (2001) Prognostic factors analysis of 17, 600 melanoma patients: validation of the american joint committee on cancer melanoma staging system. J Clin Oncol 19:3622–3634
5. Kirkwood JM, Strawderman MH, Ernstoff MS et al (1996) Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the Eastern Cooperative Oncology Group Trial EST 1684. J Clin Oncol 14:7–17
6. Lens MB, Dawes M (2002) Interferon alfa therapy for malignant melanoma: a systematic review of randomized controlled trials. Letter to the Editor. J Clin Oncol 20:1818–1825
7. Sosman JA, Unger JM, Liu PY et al (2002) Southwest Oncology Group. Adjuvant immunotherapy of resected, intermediate-thickness, node-negative melanoma with an allogeneic tumor vaccine: impact of HLA class I antigen expression on outcome. J Clin Oncol 20:2067–2075
8. Sondak VK, Liu PY, Tuthill RJ et al (2002) Adjuvant immunotherapy of resected, intermediate-thickness, node-negative melanoma with an allogeneic tumor vaccine: overall results of a randomized trial of the Southwest Oncology Group. J Clin Oncol 20:2058–2066
9. Stewart DJ, Kurzrock R (2009) Cancer: the road to Amiens. Commentary J Clin Oncol 27:328–333 (Epub 2008 Dec 8)