CLINICAL INVESTIGATION

EDITORIAL

Clin. Invest. (2011) 1(5), 611–614

Assessing clinical efficacy of drugs in cancer patients: are we on the right track?

Ciro Gallo[†1] & Francesco Perrone[2]

"...calls for more clinically substantive differences in end points that directly reflect benefit (mainly overall survival or quality of life) have been made..."

[1]Medical Statistics Unit, Department of Medicine & Public Health, Second University of Napoli, Napoli, Italy
[2]Clinical Trials Unit, National Cancer Institute, Napoli, Italy
[†]Author for correspondence:
Tel: +39 081 566 7726
Fax: +39 081 566 6021
E-mail: ciro.gallo@unina2.it

The history of the relationship between medicine and statistics in cancer treatment looks like a lucky and happy marriage with some inevitable crises. The happy side of the marriage is that application of statistics and methodology to clinical trials allowed the evolution from anecdotal and narrative to evidence-based medicine. However, more recently, some concerns have arisen about credibility and meaningfulness of the trial findings for actually improving the treatment of cancer patients in standard clinical practice [1–4].

Since the 1970s, the quality of design and analysis of randomized clinical trials (RCTs) has largely improved and more trials were claimed 'positive' by the authors, but the relative benefits of the new drugs versus the older ones have remained stable over time [5]. RCTs have become increasingly larger leading to a greater proportion of statistically significant results with only marginal benefit and increasing costs [6]. For example, in RCTs of breast and colorectal cancer treatment, absolute benefits of experimental drugs have decreased over time, or have not changed at all in metastatic disease. On the contrary, costs have increased dramatically, even by 100-fold [7]. Furthermore, small but statistically significant improvements in early end points, such as progression-free survival (PFS), did not often translate into benefits of more clinically meaningful outcomes, such as overall survival (OS) [8]. Therefore, calls for more clinically substantive differences in end points that directly reflect benefit (mainly OS or quality of life) have been made [1,4].

Clinical relevance is more than statistical significance

A side effect of statistical implementation in clinical trials is the undue reliance on statistical significance. A p-value of <0.05 is the Holy Grail to be pursued for asserting that a finding is statistically significant and that a 'positive' result is attained. However, the p-value is largely misunderstood. It represents the probability that the observed result (or a more extreme one) occurred by chance given that the null hypothesis (H_0, usually the absence of effect) was true. It is not the probability that a null hypothesis is true or false, since it is calculated under the assumption that H_0 is true. A small p-value by itself does not always translate into evidence in favor of the alternative hypothesis and gives no information about the magnitude of the change [9], since it results from both sample size and magnitude of effect. $p = 0.05$ greatly overstates the evidence against the null hypothesis and is, at best, only a promising result [9]. However, p-value remains popular because it does not require alternative hypotheses be defined and very low p-values (e.g., <0.001) are unlikely to mislead [10].

The p-value is strongly affected by the size of the study sample. Demands for trials large enough to provide adequate statistical power (80% or more) have been repeatedly solicited in the past and large trials did what they were asked for – detecting marginal outcomes with statistical validity. As a consequence the rate

ISSN 2041-6792 10.4155/CLI.11.36

of statistically significant results gradually increased over the last three decades [5], and trivial benefits were reported as impressive improvements of clinical practice [11]. Minimum detectable effects were targeted rather than clinically worthwhile benefits [12], while observed differences even smaller than that anticipated in the protocol might eventually result in statistically significant results and be used for US FDA/EMA registration [4].

What is clinical relevance?

Clearly, it is critical to define which size of benefit can be deemed as clinically worthwhile. It might vary according to the different stakeholders. Pharmaceutical companies might prefer to look for small benefits, with very large trials that increase the chance of statistically significant results; third-payers might prefer trials looking for large improvements, in order to better spend their budget; patients and their doctors might stay in the middle, hoping for large improvements but not rejecting smaller benefits, particularly when therapeutic alternatives do not exist.

To assess the clinical benefit of treatments, greater emphasis should be given to the absolute measures of efficacy (absolute difference, number-needed-to-treat [NNT]) rather than the relative ones (hazard ratio [HR], relative risk reduction), both in designing trials and reporting results. Clearly, it is more impressive to say that the relative risk of death was 15% less in the experimental arm than the control than saying that median survival increased just 5 weeks. In addition, few cancer trialists are acquainted with the NNT, that is the reciprocal of the absolute difference between the control and the experimental arm, and represents the number of patients you expect to treat in order to see one additional success (e.g., using survival estimates at a given time [13]) – the larger the NNT, the less appealing the drug. Which absolute benefits are worthwhile largely depend on the prognosis of the study patients: a 3-month increase for patients with advanced metastatic solid tumors [4], with a median survival of 6 months in the control group, leads to a HR of 0.67, while a 6-month increase from a median survival of 3 years corresponds to a HR of 0.86.

Still to be considered is that therapies with only marginal benefits can be rather harmful because the majority of patients have no benefit while most of them suffer side-effects [6]. Presenting results as a risk–benefit profile might be useful; for example, using measures like the NNT:NNH ratio, where NNH is the number-needed-to-harm, that is the number of patients you would have to treat in order to see one additional adverse effect. The smaller the NNH, the higher the number of extra side-effects with the new therapy. For example, NNT:NNH equal to two would indicate that you need to treat twice the number of subjects to have an additional success, rather than to have an additional side effect (or, reciprocally, you would expect two additional side-effects for every additional efficacy success).

Surrogate end points

Another issue that is overly critical in clinical trials with patients affected by metastatic cancer is the use of surrogate end points. The reasonable aim of treatment of metastatic cancer is to improve the length of, or the quality of, survival. However, many recent drugs have been registered without OS as the primary end point [14]. As for quality of life, it is seldom used as the primary end point and interpretation of results may be further complicated by methodological problems (e.g., missing data are usually informative since patients with deteriorated clinical conditions are more likely to miss questionnaires rather than those who are in a state of well-being).

> "...it is critical to define which size of benefit can be deemed as clinically worthwhile."

Surrogate end points are frequently used upon the assumption that they should be able to replace survival, the clinically relevant end point, and to speed up the approval process. However, it is not possible to determine which proportion of the effect on survival is accounted for by the effect on a surrogate end point, and often the observed effects on PFS or TTP translate into smaller, if any, effects on survival. In addition, proper statistical validation of a candidate surrogate end point is a difficult task and needs to turn to meta-analyses of RCTs where both end points have been assessed [15]; thus, confirmations of true surrogacy in cancer are scarce.

Demonstration of true surrogacy is context and treatment specific. This means that we cannot automatically extrapolate to metastatic disease, a conclusion that was derived in trials of adjuvant treatment (or vice versa) and that we cannot apply to a new class of drugs what was demonstrated with another one. The latter statement is particularly relevant here. Indeed, while some demonstration of surrogacy for PFS or TTP was produced with chemotherapeutic agents, such end points have been commonly used in registrative trials of target-based agents, that is, a new generation of drugs considered highly innovative and profoundly different from old chemotherapy. Consequently, the use of surrogate end points for new drugs is decidedly not validated and we might expect that advantages seen in PFS will not necessarily translate into survival gain. A critical example is bevacizumab that was registered

by the US FDA to be used in metastatic breast cancer combined with paclitaxel, thanks to a doubled PFS. Later, it became clear that the outstanding improvement in PFS did not translate into a survival gain and, recently, the FDA has withdrawn the authorization in light of a little but significant incidence of fatal adverse events [16,101].

Furthermore, (presumed) surrogate measures, such as PFS or TTP, are available only at certain times and are estimated with error, while survival is a continuous outcome, that is assessed without error.

In a metastatic cancer with a very dismal prognosis, another problem arises; PFS can be sensitive to detect very small benefits thanks to the application of highly intensive restaging procedures. For example, 4–5 week benefits in median PFS were targeted in two recent studies on trastuzumab beyond progression [12,17]. However, such a small advantage can only be demonstrated if radiological progression is detected through a monthly restaging, and will inevitably be obscured in clinical practice, where less frequent radiological restaging is performed and more emphasis is given to clinical assessment of treatment outcome.

Unfortunately, pharmaceutical companies driving registrative trials prefer the use of PFS (sometimes even changing the end point while the trial is ongoing [8]).

Shifting to confidence intervals

Knowledge is improved by shifting from testing to estimation, that is, from p-values to confidence intervals (CIs). CIs identify a range of plausible values of the true effect compatible with the observed result. A 95% CI is not a range of values within which the unknown true value lies with 95% probability (i.e., it does not have a 95% probability of including the unknown parameter); rather we set *a priori* a 95% probability that our final CI will contain the parameter value (i.e., the true effect we are looking for).

Unfortunately, the 95% rule (i.e., the complement of the familiar 5% significance level) has encouraged the mechanistic and reductive interpretation of CIs only in terms of statistical significance. If the CI contains the no effect value (e.g., HR = 1), then the observed difference is statistically significant, irrespective of any clinical interpretation of the observed change.

With CIs, we quantify our uncertainty. Drawing conclusions from the observed treatment effect is inadequate because of the uncertainty inherently associated to our estimate. The estimates least influenced by the play of chance, that is more statistically stable results, are not those with low p-values, but those with narrow CIs.

As a simple solution, entirely within the framework of the usual frequentist approach, we propose to define a clinically worthwhile threshold and to accept a new drug as clinically useful only if the upper limit of the CI is below that threshold. If it is not, the drug may still be useful, but the effect should be better ascertained with further research, and would be included in the 'limbo level' proposed by Sobrero and Bruzzi [1]. Thus, clinical needs would be integrated in the study conclusions.

Our proposal mimics, for superiority studies, what already happens for noninferiority studies, where a noninferiority margin is defined as the smallest value that would represent a clinically important effect [18]. In noninferiority trials, the CI approach is preferred in design, analysis and reporting as it is more informative and the same should be true for superiority trials.

Of course, alternative limits should be used. If in noninferiority studies it is common to set HRs of 1.2 or 1.25 as upper limits of CIs, their reciprocals (0.83 or 0.8) could be used in superiority studies. It might be noted that with these limits, many of the trials recently used for registration of new target-based anticancer drugs would not be fully convincing [4]. The predefined superiority margin would be a matter of clinical judgment [1,4], as it is for noninferiority studies. What we claim is that the clinically worthwhile effect should definitely be the same.

Conclusion

In conclusion, some simple suggestions seem appropriate in order to be smarter about oncology clinical trials [19]:

- Trials should look for substantial effects on important clinical end points and not for marginal effects on surrogate outcomes;
- More concern should be given to safety and longer follow-up time, particularly with drugs candidated for long-term use (i.e., many recent target-based drugs);
- Identifying patients who would most benefit (or be harmed) by treatment should be pursued by early validation of reliable biomarkers;
- CIs should be given more reliance than significance testing when assessing the clinical evidence of benefit;
- Results should be reported in a clearer fashion, not encouraging (even inadvertently) overly optimistic interpretations.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Bibliography

1 Sobrero A, Bruzzi P. Incremental advance or seismic shift? The need to raise the bar of efficacy for drug approval. *J. Clin. Oncol.* 27(35), 5868–5873 (2009).

2 Stewart DJ, Kurzrock R. Cancer: the road to Amiens. *J. Clin. Oncol.* 27(3), 328–333 (2009).

3 LoRusso PM, Schnipper LE, Stewart DJ, Boerner SA, Averbuch SD, Wolf W. Translating clinical trials into meaningful outcomes. *Clin. Cancer Res.* 16, 5951–5955 (2010).

4 Ocana A, Tannock IF. When are positive clinical trials in oncology truly positive? *J. Natl Cancer Inst.* 103(1), 16–20 (2011).

5 Booth CM, Cescon DW, Wang L, Tannock IF, Krzyzanowska MK. Evolution of the randomized controlled trial in oncology over three decades. *J. Clin. Oncol.* 26, 5458–5464 (2008).

6 Fojo T, Parkinson DR. Biologically targeted cancer therapy and marginal benefits: are we making too much of too little or are we achieving too little by giving too much? *Clin. Cancer Res.* 16, 5972–5980 (2010).

7 Seruga B, Hertz PC, Wang L *et al.* Absolute benefits of medical therapies in Phase III clinical trials for breast and colorectal cancer. *Ann. Oncol.* 21, 1411–1418 (2010).

8 Ocana A, Amir E, Vera F, Eisenhauer EA, Tannock IF. Addition of bevacizumab to chemotherapy for treatment of solid tumors: similar results but different conclusions. *J. Clin. Oncol.* 29(3), 254–256 (2011).

9 Goodman SN. Of p-values and Bayes: a modest proposal. *Epidemiology* 12(3), 295–297 (2001).

10 Katki HA. Invited commentary: evidence-based evaluation of p values and Bayes factors. *Am. J. Epidemiol.* 168(4), 384–388 (2008).

11 Seruga B, Tannock IF. Up-front use of aromatase inhibitors as adjuvant therapy for breast cancer: the emperor has no clothes. *J. Clin. Oncol.* 27(6), 840–842 (2009).

12 Blackwell KL, Burstein HJ, Storniolo AM *et al.* Randomized study of lapatinib alone or in combination with trastuzumab in women with Erbb2-positive, trastuzumab-refractory metastatic breast cancer. *J. Clin. Oncol.* 28(7), 1124–1130 (2010).

13 Altman DG, Andersen PK. Calculating the number needed to treat for trials where the outcome is time to an event. *Br. Med. J.* 319, 1492–1495 (1999).

14 Apolone G, Joppi R, Bertele' V, Garattini S. Ten years of marketing approvals of anticancer drugs in Europe: regulatory policy and guidance documents need to find a balance between different pressures. *Br. J. Cancer* 93, 504–509 (2005).

15 Buyse M, Sargent DJ, Grothey A, Matheson A, de Gramont A. Biomarkers and surrogate end points – the challenge of statistical validation. *Nat. Rev. Clin. Oncol.* 7, 309–317 (2010).

16 Ranpura V, Hapani S, Wu S. Treatment-related mortality with bevacizumab in cancer patients. A meta-analysis. *J. Am. Med. Ass.* 305(5), 487–494 (2011).

17 von Minckwitz G, du Bois A, Schmidt M *et al.* Trastuzumab beyond progression in human epidermal growth factor receptor 2-positive advanced breast cancer: a german breast group 26/breast international group 03–05 study. *J. Clin. Oncol.* 27(12), 1999–2006 (2009).

18 Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJW for the CONSORT Group. Reporting of noninferiority and equivalence randomized trials. An Extension of the CONSORT Statement *J. Am. Med. Assoc.* 295, 1152–1160 (2006).

19 Kramer BS, Wilentz J, Alexander D *et al.* Getting it right: being smarter about clinical trials. *PLoS Med.* 3(6), e144 (2006).

Website

101 US FDA begins process to remove breast cancer indication from Avastin® label www.fda.gov/newsevents/newsroom/pressannouncements/ucm237172.htm (Accessed 1 March 2011)