Author’s response to reviews

Title: Responder Analyses and the Assessment of a Clinically Relevant Treatment Effect

Authors:

Steven M Snapinn (ssnapinn@amgen.com)
Qi Jiang (qjiang@amgen.com)

Version: 3 Date: 13 August 2007

Author’s response to reviews: see over
Referee 1

General

I approve entirely of your thesis. It is nothing less than a scandal that the FDA and EMEA and drug regulatory agencies continue to place emphasis on so-called responder analysis. The definition of response is naïve and the consequent rise in sample size simply adds to the cost of drug development and delays the introduction of treatment for patients. Judging by the frequency with which methodologically unsound papers on this topic are published in the medical press, it seems that the regulators' confusion is shared widely by researchers, editors and referees. I hope that your paper will promote clearer thinking on this issue but, fear, alas, that this nonsense will continue.

Minor Essential Revisions

There are two points where I disagree (slightly) with what you claim.

1. You imply at one point that individual response could be determined by a cross-over trial. In fact even a cross-over trial is inadequate for this purpose. You need a cross-over with repeated periods. See Senn, S. J. (2004), Individual response to treatment: is it a valid assumption? British Medical Journal, 329, 966-968.

   The point raised in the referenced paper is an interesting one, and we agree with the comment. We have removed the reference to crossover trials from the manuscript.

2. I think it is also not simple to judge response graphically as you imply. Certainly it is a necessary condition for response on the original scale to be additive for the distribution under active treatment and placebo to be identical but for location. However, it is not sufficient.

   We agree with the point. If the distributions for the active treatment and placebo differ only in location then it may be reasonable to assume that the effect of the treatment is roughly constant among subjects (which would make the assessment of response clear), but there are other possible explanations. We have softened the wording in the manuscript: “The consistent horizontal separation between the distribution functions (Figure 1B) suggests that the benefit on the continuous scale was consistent among subjects, although it should be recognized that other explanations are possible. If the assumption of a consistent benefit among subjects seems reasonable, the mean difference between groups would be an appropriate summary of the treatment benefit, and its magnitude should be used to help determine clinical relevance.”

Discretionary Revisions

You might consider citing the following papers.

Thank you for the suggestion. We have added these references.
Referee 2

Major Compulsory Revisions

The authors discuss the weaknesses of responder analyses and offer an alternative. They come up with some interesting arguments and suggestions. However, I have some major remarks:

The authors greatly appreciate the referee’s careful review and excellent comments.

I am not sure about the objective of the paper. Is it ‘to examine [responder analysis] in detail (page 6, or to ‘discuss various weaknesses’. In the latter case there may be no need to discuss the advantages of responder analysis in detail, but I have the feeling that the paper is more like a list of the disadvantages, without paying much attention to possible advantages. Would it not do a service to the reader if also the advantages were discussed? Or are there none? Of course, if the authors prefer to discuss the disadvantages only, they can do so. I would then suggest that they change the title. Even if they do so, most of the remarks below are still relevant.

The responder analysis has one well-known disadvantage (a loss of power relative to an analysis of the original continuous variable) and one purported advantage (to help ensure that any identified difference between groups is clinically meaningful). It was not our main purpose to list disadvantages, but rather to explain that this supposed advantage is illusory. In this case, assuming our arguments are convincing, the cost in statistical power comes with no real advantage.

There is actually one additional potential advantage of the responder analysis: Simplicity. That is, some researchers may feel more comfortable interpreting a difference in response rates rather than a mean difference in a continuous measure. However, this is a personal preference, not a scientific advantage, and we did not address it in this paper.

Many of the arguments mentioned in the paper have been discussed by several others, for example Doug Altman and Stephen Senn. That would not be a problem, because it can be quite interesting to discuss existing opinions again and to add new ideas and interpretations. However, some of the arguments seem somewhat flawed and do not stand scrutiny. I mention some aspects that would rethinking and reformulating.

1. The authors correctly remark that it not an easy task to define a cut-off value for response and use this as an argument to compare means. However, they state that in that case an estimate of the clinically relevant difference is required. They do not discuss how to come to a decision about what is a clinically relevant difference for the means. I am not sure that is an easy task either. They also mention that it makes no sense that a weight loss of 5.1% is a response, while 4.9% is not. Does not the same problem hold for
difference between the means? Might it even be that for responder analysis it does not matter that such small differences as 4.9 vs 5.1 lead to opposite classifications: You may misclassify some patients, but those misclassifications will cancel, because you are not considering individuals but the overall response over a group.

It’s our belief that the existence of a treatment effect should be based on a statistical test using the continuous variable, but that the clinically meaningfulness of this difference should be based on an examination of both the mean difference and the response rates: “The next step is to determine clinical importance by examination of the mean difference between groups, as well as by examination of response rates, possibly using various response definitions.” We agree that any yes/no decision about clinical meaningfulness is essentially arbitrary. However, applying this criterion to an individual subject, rather than to the group results, is what leads to the cost in power, regardless of the fact that misclassifications will tend to cancel.

2. On page 4 the authors suggest that when responder analysis is used, rejection of the null hypothesis $P_x \leq P_c$ leads to the conclusion that the difference in clinically relevant. I am not so sure about this. If $P_x=0.55$ and $P_c=0.50$, this may not always be a clinically relevant improvement, especially not if the side effect profiles or other parameters that affect clinical relevance differ.

We agree that the conclusions from any clinical trial must weigh the results of all analyses, not just those from the primary efficacy variable. We also agree that a small significant difference in responder rates does not necessarily imply a clinically meaningful difference. However, the desire to draw exactly that conclusion seems to be the main goal behind the responder analysis. Certainly, the quotes provided from regulatory guidance documents imply as much. It is not clear to us why anyone would use a responder analysis unless there was a desire to demonstrate a clinically meaningful (not just a statistically significant) effect.

3. On page 7 the authors discuss the loss of power when responder analysis is used. Their table 1 provides a nice example of the possible advantages of responder analysis in some situations. If the cut-off is 2, the groups sizes differ enormously: 526 versus 4053. So responder analyses seems inappropriate. However, if it is the case that ONLY such a large results has any clinical meaning, an analysis on mean response if of no value. For example it may not be sufficient to decrease the average amount of infectious agents. They may have to be below a certain (very low) level for the patient to benefit.

We completely agree. We have modified existing text in the manuscript to read: “Clearly, there are some situations where the achievement of a certain value on a continuous scale has enormous clinical implications, such as when a test value is used as the basis for a decision on hospital admission or surgical intervention. In these cases, an analysis of the response rates would be highly relevant despite any cost in power. Alternatively, the clinical event itself (i.e., the hospitalization or the
surgery) would make an appropriate endpoint for determining the treatment effect.”

Later on, the authors suggest that there are two situations: the distributions in the groups only differ in location, or they (also) differ in variance. May this not be a simplification? Firstly, the distributions may differ in other aspects, but more importantly, is it not usually the case that things are not so clear and that you do not and cannot know what situation you are in? How to decide on the exact properties of the distributions? In any case, it may be difficult to decide whether the means approach is equivalent to the responder approach. It usually simply is impossible to determine whether two distributions are equivalent. The authors suggest to consider the empirical distributions, but is it not the aim of statistical analyses to have a more robust approach? If it is sufficient to look at raw data, why then use statistical methods at all?

We agree that we might have provided an oversimplified description of the value of examination of the distributions of outcomes on the continuous scale. We have modified the manuscript to highlight some the difficulties with this approach, as suggested. However, we feel that, despite these difficulties, examination of the distributions is of excellent value. First, it is a useful tool to illustrate the relationship between the analysis of means and the responder analysis. In addition, as we discuss in the manuscript, the empirical distribution functions provide the results of the responder analyses for any potential definition of response. This is clearly of more value than simply examining the results of a responder analysis for a single response definition.

To address these points we have added the following: “Of course, the distributions of responses may differ in considerably more complex ways than illustrated here. In addition, due to limited sample sizes, it may be difficult to tell whether the horizontal difference between two empirical distribution functions suggests a consistent effect from patient to patient. Despite these limitations, examination of the empirical distribution functions should provide valuable information that simple comparisons of means or response rates do not provide.”

4. The authors argue that responder analysis is no good, because it used in some disease areas and not in others. This is a remarkable argument. Do they also feel that survival analysis is no good because it is used in some areas and not in others? May it not be that different disease areas have a different way of looking at things, different interests or even different ideas about statistical methods?

The point we were trying to make is that the inconsistent application of the responder analysis is a symptom of the fact that there is no clear rationale for when it should or should not be used. Survival analysis is used when the efficacy variable is a time to event, because in these cases it provides clear benefit over other methods. In the case of the responder analysis the rationale appears to be the desire to ensure that the treatment effect is clinically meaningful. However, this rationale should apply in all cases. Do researchers who use a survival
analysis care nothing about clinical meaningfulness? To help clarify the point, we have changed the mention of the “problem” with the inconsistent application of the responder analysis to: “It is also interesting to consider the inconsistent application of the responder analysis.”

5. The authors also argue that responder analysis is no good, because the TERM responder is misleading. I am not sure this is an appropriate argument. The term error has many meanings in statistics but the term surely does not imply that errors have been made. So the term is misleading. Do the authors suggest that we should refrain from statistical models with error terms, because of this? The authors also argue that patients in the control group cannot be responders, because you do not know whether the response is due to the treatment. It would require a cross-over trial to properly define responders. I am not sure this is a proper argument either. If you use this type of reasoning, would survival then be a proper endpoint?

We do not believe that the term “responder” is misleading. Quite the contrary – we believe that it is an accurate indication of the intention of the analysis. That is, a “responder” is meant to indicate a patient who has benefited from the treatment in a meaningful way. It is our contention, however, that it does no such thing. A causal effect is defined as the difference between the outcome a patient experienced while taking the treatment and the hypothetical outcome the patient would have experienced in the absence of the treatment (i.e., the counterfactual). The response definition is meant to capture a causal effect, but does not.

6. May it not be that in some situations the size of the effect of the intervention is not so relevant. Could it possibly be that for blood pressure for example the final blood pressure is not so relevant but that it only matters whether the blood pressure is between certain limits?

This is an excellent point, and an additional complication. For example, blood pressure reductions in hypertensive patients beyond a certain level may provide no additional clinical benefit, and may even cause harm. In our experience, analyses of trials of antihypertensive drugs do not account for this fact. Situations like this might require two arbitrary cutpoints to define a “responder,” one on the upper end and one on the lower end. However, this situation would not alter our fundamental approach: Determine the existence of a treatment effect on an analysis of the continuous variable, and determine the clinical meaningfulness of that effect by examination of the mean difference and of the response rates for various response definitions.