Split decision - When can dichotomizing continuous endpoints be the best approach?

Sophie O'Shaughnessy & Clare Newall
Veramed Limited, Twickenham, UK

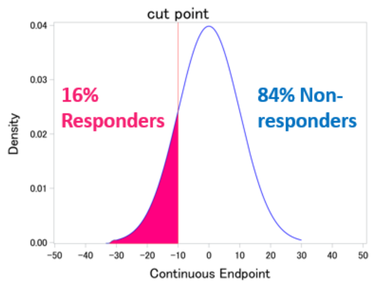
1. Introduction

Continuous endpoints such as patient reported outcome scores or physiological measures like blood pressure, can be analysed in 2 key ways:

Mean Analysis - compare group means e.g. with t-test or linear regression

Responder Analysis

- dichotomize subjects into 'responders' and 'non-responders' using a predetermined cut point
- compare proportion of responders in each group e.g. with chi-squared test, logistic regression



A mean analysis is commonly preferred by statisticians BUT when can a responder analysis be the better approach?

2. Pros and Cons of Responder Analyses

(as compared to mean analyses)

PROS

- ✓ Less prone to yield a statistically significant result for a treatment group difference which is not clinically meaningful
- ✓ More intuitive clinical relevance & easier for non-statisticians to understand
- ✓ May be the established endpoint in a therapeutic area
- ✓ The nonoccurrence of an intercurrent event (e.g. use of rescue medication) can easily be added as a criterion for a binary 'response' endpoint – incorporation into a continuous endpoint is less straight forward
- ✓ May be more appropriate if there isn't a linear relationship between the continuous endpoint and clinical efficacy
- ✓ Requires fewer assumptions, particularly about the distribution of the continuous endpoint
- ✓ Enables elucidation of factors predicting 'response' in subpopulations

CONS

- × Reduced efficiency (though power is maximised if you choose a cut point half way between the population means of the two groups)
- × Choice of cut point can be difficult/arbitrary and can affect results
- × Cruder approach - doesn't account for the fact various magnitudes of response may be clinically relevant
- × The term 'responder' can be misleading as endpoint may change due to natural history of the disease or measurement error, not treatment

4. Recommendations

Snappin & Jiang (2007) [2]- Sequential approach:

1) Establish **statistical significance** based on the most powerful approach (usually a comparison of means).

2) Assess **clinical relevance** based on examination of the mean difference between groups and on response rates.

It is also important to examine the **cumulative distribution functions** to help assess the consistency of response among subjects.

EMEA [3]

"In some situations, the 'responder' criterion may be the primary endpoint advised in guidelines [and]…should be used to provide the main test of the null hypothesis."

"If a 'responder' analysis is used to allow a judgement on clinical relevance, once a statistically significant treatment effect on the mean level of the primary variable(s) has been established, **the results of the 'responder' analysis need not be statistically significant** but the difference in the proportions of responders should support a statement that the investigated treatment induces clinically relevant effects."

"A 'responder' analysis cannot rescue otherwise disappointing results."

FDA [4]

"Because statistical significance can sometimes be achieved for small changes in PRO measures that may not be clinically meaningful (i.e., do not indicate treatment benefit), we encourage sponsors to avoid proposing labelling claims based on statistical significance alone.

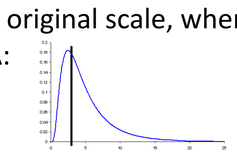
To demonstrate treatment benefit, we find it informative to examine the cumulative distribution function (CDF) of responses between treatment groups to characterize the treatment effect and examine the possibility that the mean improvement reflects different responses in patient subsets."

3. Comparing Power

Are there any situations where a responder analysis is more powerful than a mean analysis?

Uryniak et al (2011) [1] compared the power of t-tests with tests of equality of proportions while varying: 1) distributions of 2 treatment groups, 2) distribution parameters, 3) cut points, 4) size of treatment effects, 5) sample sizes.

Findings: Mean analysis was usually more powerful, often substantially so. For distributions other than normal and beta, responder analyses could occasionally be more powerful → **when there was a large proportion of observations around the cut point for dichotomization, i.e. a small change in mean (horizontal shift of the distribution) corresponded to a very large change in the proportion of responders.** Examples were comparison of 2 lognormal distributions on the original scale, when the cut point was at the lower end (Fig A) and normal vs. mixture normal distributions. However, as total sample size increased (from 100 to 400), the responder advantage diminished.

Fig A: 

To illustrate this more clearly, we present several theoretical scenarios in the table below, as well as distribution plots of scenarios not covered by Uryniak et al where the responder analysis is more powerful. We used SAS's CDF function, proc power and simulation (for the skewed normal distribution) to calculate % of responders and sample sizes.

Endpoint: Change from baseline (CFB) in St George's Respiratory Questionnaire (SGRQ) Total Score. Possible range: 0 to 100 (higher score=worse health) Individual 'Response': CFB <= -4 (min. clinically important difference (MCID)). 1:1 ratio of Control & Active Treatment Groups. Control Group: was always assumed to be drawn from a normal distribution N~(0, 10).

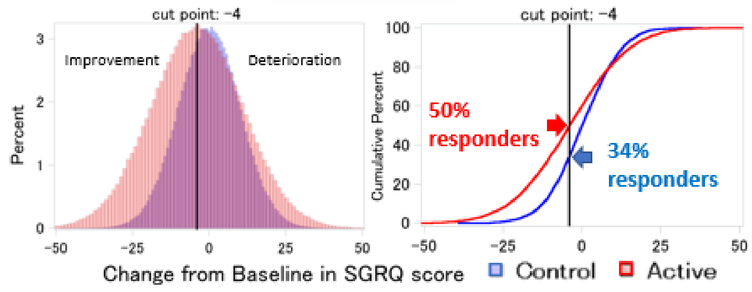
Yellow highlighted scenarios are where the responder analysis requires a smaller sample size than a mean analysis.

			% of Responders			Total Sample Size*	
Scenario	Population Distribution of Active Treatment Group	Difference in Trt Group Pop Means	Control Population	Active Population	Difference in Trt Group Pop % of Responders	Mean Analysis	Responder Analysis
1A	Normal, SD 10	-1	34%	38%	4%	3142	5160
1B	Normal, SD 10	-4	34%	50%	16%	200	316
1C	Normal, SD 10	-8	34%	66%	32%	52	80
2A	Normal, **SD 15**	-1	34%	42%	8%	5104	1278
2B	Normal, **SD 15**	-4	34%	50%	16%	322	316
2C	Normal, **SD 15**	-8	34%	61%	27%	84	114
3A	**Positively Skewed Normal+** , SD 10	-1	34%	43%	9%	3142	1042
3B	**Positively Skewed Normal+**, SD 10	-4	34%	55%	21%	200	180
3C	**Positively Skewed Normal+**, SD 10	-8	34%	69%	35%	52	64

*Sample size to reject a null hypothesis of a) no difference in means or b) no difference in proportions with 80% power and type 1 error of 5%. + Shape parameter λ = 3.

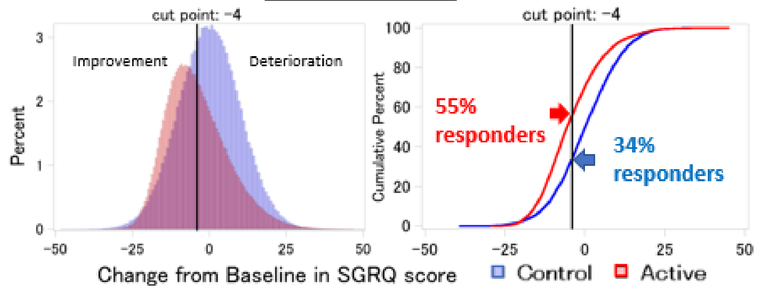
PLOTS: Left= Distribution Plot, Right= Cumulative Distribution Function (CDF) Plot

SCENARIO 2B



1) Variability of the endpoint (i.e. CFB score) is much greater in active than control group, and the mean endpoint value is better by the magnitude of the MCID (i.e. 4 points lower) in the active than the control group.
2) In the active group the extremes of negative CFBs (improvement) but also, to a lesser extent, positive CFBs (deterioration), observed are high compared to control.
3) Note the crossing lines on CDF plot suggest the active treatment may be worse than control for subjects with large deterioration (high positive CFBs). (Parallel lines indicate a consistent treatment effect across the endpoint range.)

SCENARIO 3B



1) Overall variability of the endpoint (i.e. CFB score) is the same in the active as the control group. But in the active group the distribution of CFB scores is positively skewed so that the bulk of the active distribution is shifted towards the improvement end of the scale and the active mean is better than control by the magnitude of the MCID. Therefore 55% of active subjects are 'responders' vs. 34% of controls.
2) Note however, the most extreme negative CFBs (improvement) in the active group are not as extreme as in the control group.

Key Points to Note from the Table:

Scenarios 1x: When both distributions are non-skewed normal, a mean analysis is always more powerful [1]. Approx. 60% larger sample is required for a responder analysis than a mean analysis.

Scenarios 2x & 3x: If the mean difference is small (2A & 3A) then the responder analysis has a much greater advantage over the mean analysis. However, an absolute mean difference <4 would not be considered clinically important so these scenarios are actually of limited real-life relevance. Once the mean difference becomes sizeable at 8 points (2C & 3C), even though there is a very large group difference in % of responders, the mean analysis regains its advantage.

Summary:

There are few situations where a responder analysis is substantially more powerful than a mean analysis. These:

- are unusual
- and can be complicated - if the active vs. control treatment effect is qualitatively different across different parts of the distribution (i.e. CDF lines cross), then really neither a mean nor a responder rate are satisfactory summary measures as both can mask this heterogeneity of effect.

5. New Methods

Lin (2016) [5] proposes applying Ganju et al (2013)'s **'MinP' approach** to responder analysis, which **allows multiple pre-specified responder cut points to be explored** while **improving power/efficiency** relative to a standard responder analysis and controlling the type 1 error rate. Thus it is of particular interest for an endpoint where there is no consensus on the most appropriate cut point. A test is carried out for each cut point and the minimum of the p-values is used for formal inference with the critical value for hypothesis testing coming from permutation.

Zhang et al (2016) [6] have demonstrated the implementation of a **responder analysis without *a priori* dichotomization** of the data which may be **useful when treatment 'success' is defined by multiple criteria or over a longitudinal time period**. The proportion of responders for each treatment is derived from a model based on the original data collected for the specified variables, and estimated by substituting maximum likelihood estimators of the model parameters. The authors also showed that this model-based approach is both **more efficient** and **more effective at handling missing data** than the usual approach of analysing dichotomized data.

6. Conclusions

- A comparison of means is more powerful than a responder analysis in nearly all situations.
- Exceptions are unusual and tend to be when the active treatment (vs. control) causes more of a change in distribution shape than shift in location.
- However, responder analyses have key advantages in some situations, particularly in terms of clinical relevance and when there is a non-linear relationship between the endpoint and clinical efficacy.
- Advice for most typical scenarios is to use a mean analysis to establish statistical significance plus a secondary responder analysis to support clinical relevance.
- Ideally present CDFs to show the consistency of the treatment effect across the whole range of endpoint values.

References: [1] Uryniak, T., Chan, I. S. F., Fedorov, V. V., Jiang, Q., Oppenheimer, L., Snapinn, S. M., Teng, C.-H., Zhang, J. (2011). Responder analyses – a PhRMA position paper. Statistics in Biopharmaceutical Research 3:476–487. [2] Snapinn, S. M., Jiang, Q. (2007). Responder analyses and the assessment of a clinically relevant treatment effect. Trials 8:article 31. [3] EMA (2002) Points to Consider on Multiplicity Issues in Clinical Trials [4] FDA (Dec 2009) Guidance for Industry [on] Patient Reported Outcome Measures: Use in Medical Product Development to Support Labelling Claims. [5] Lin (2016) Robust inference for responder analysis: Innovative clinical trial design using a minimum p-value approach. Contemporary Clinical Trials Communications 3 65-69. [6] Zhiwei Zhang, Jianxiong Chu, Dewi Rahardja, Hui Zhang & Li Tang (2016) Responder analysis without dichotomization, Journal of Biopharmaceutical Statistics, 26:6, 1125-1135.