Contemporary Clinical Trials Communications 3 (2016) 65–69

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Robust inference for responder analysis: Innovative clinical trial design using a minimum p-value approach

Yunzhi Lin

Takeda Development Center Americas, Inc., Deerfield, IL 60015, USA

ARTICLE INFO

Article history:
Received 6 November 2015
Received in revised form
31 March 2016
Accepted 4 April 2016
Available online 14 April 2016

Keywords:
Clinical trials
Responder analysis
Multiple testing
Robust inference

ABSTRACT

Responder analysis is in common use in clinical trials, and has been described and endorsed in regulatory guidance documents, especially in trials where "soft" clinical endpoints such as rating scales are used. The procedure is useful, because responder rates can be understood more intuitively than a difference in means of rating scales. However, two major issues arise: 1) such dichotomized outcomes are inefficient in terms of using the information available and can seriously reduce the power of the study; and 2) the results of clinical trials depend considerably on the response cutoff chosen, yet in many disease areas there is no consensus as to what is the most appropriate cutoff. This article addresses these two issues, offering a novel approach for responder analysis that could both improve the power of responder analysis and explore different responder cutoffs if an agreed-upon common cutoff is not present. Specifically, we propose a statistically rigorous clinical trial design that pre-specifies multiple tests of responder rates between treatment groups based on a range of pre-specified responder cutoffs, and uses the minimum of the p-values for formal inference. The critical value for hypothesis testing comes from permutation distributions. Simulation studies are carried out to examine the finite sample performance of the proposed method. We demonstrate that the new method substantially improves the power of responder analysis, and in certain cases, yields power that is approaching the analysis using the original continuous (or ordinal) measure.

1. Introduction

In many disease areas in which "hard" clinical endpoints such as mortality are not appropriate measures of efficacy, rating scales and other continuous measures are used for the evaluation of treatments. For instance, in schizophrenia clinical trials, the MATRICS Consensus Cognitive Battery (MCCB) or the Negative Symptom Assessment-16 (NSA-16) are frequently used instruments for evaluating psychopathology in study subjects. Other examples include the use of the Expanded Disability Status Scale (EDSS) in multiple sclerosis trials, the use of exercise tolerance (ET) measures in trials of heart failure therapies, and etc. In such studies, overall treatment effect has typically been tested by assessing the difference in mean change over time of the continuous (or ordinal) measure between the treatment and control group. Although such analyses are usually the primary outcomes, one problem is that the translation of the results into clinical practice is difficult. We might not know what, for example, a difference which is statistically significant but amounts to only 1 MCCB point in magnitude means from a clinical perspective. Such a problem can be addressed by using a responder analysis, in which each subject is classified as either a "responder" or a "non-responder", and the proportions of patients who benefit are quantified and compared between treatment groups. A common approach is to define a threshold for the change from baseline in the continuous (or ordinal) endpoint, and define a patient as a "responder" if his/her change value is above (or below) the threshold.

Responder analysis provides several benefits and hence is in many cases proposed or recommended by regulatory guidance or clinical communities to be used in clinical trials. For example, draft guidance from the FDA on patient-reported outcomes specifically endorsed the responder analysis as an alternative approach to assessing clinical relevance [1]. The procedure is useful, because responder rates can be understood more intuitively than a difference in means of rating scales. It also helps ensure that a reported statistically significant result represents a clinically meaningful benefit. However, two major issues arise from this procedure. First, it is well known that dichotomization tends to result in a loss of statistical power compared to an analysis of the original continuous variable. The procedure hence is inefficient in terms of using the information available and requires greater sample size in clinical

E-mail address: yzlinn@gmail.com.

http://dx.doi.org/10.1016/j.conctc.2016.04.001

trials as analyzed, for example, by Snapinn and Jiang [2]. The second issue with responder analysis is that the results of clinical trials depend considerably on the response cutoff chosen. Yet in many disease areas across different clinical trials, various definitions of response have been used, and there is no consensus as to which is the most appropriate one [3]. If a cutoff is chosen post hoc, this is potentially an inappropriate manipulation of the data.

The issues and challenges inherent in the responder approach deserve particular attention in the development and licensing of new therapeutics. The present paper addresses the two issues mentioned above, offering a novel approach for responder analysis that could both improve the efficiency and power of responder analysis and explore different responder cutoffs if an agreed-upon common cutoff is not present.

Pre-specification of the responder cutoff and a properly planned statistical analysis are essential to avoid multiple comparisons and inflated type I error rates. But how can we pre-specify when we are not certain which responder cutoff is the optimal one? Ganju et al. recently proposed to analyze clinical trial data by pre-specifying multiple test statistics and using a combined statistic – the minimum p-value – for inference when there is uncertainty about what candidate primary endpoint, hypothesis, or statistical test to use in planning a clinical trial [4–7]. The critical value for hypothesis testing comes from permutation which consists of re-randomizing the treatment assignments and calculating the combined statistic. For instance, for a trial with a time-to-event endpoint, it might be unclear at the planning stage of the trial whether a log-rank test or a stratified log-rank test would be more appropriate for the analysis. Using the proposed method, the trialists can pre-specify both tests and use the minimum of the p-values as the new test statistic. It has been shown that the method, referred to as MinP, is robust, controls the type I error rate, and provide statistical power that is closest to the best-performing statistic.

In this paper, we borrow the idea from Ganju et al. and extend the use of MinP to clinical trials analyzed by the responder approach. We propose a statistically rigorous clinical trial design that pre-specifies multiple tests of responder rates between treatment groups based on a range of pre-specified responder cutoffs, and uses the minimum of the p-values for formal inference. The null hypothesis associated with the multiple tests is that there is no treatment effect however the "responder" is defined. The alternative hypothesis is that there is a significantly greater proportion of responders in the new treatment group, with the criterion for "responding" being one of the pre-specified cutoffs. The proposed method therefore provides not only a formal test for the treatment effect, but also an estimate of the optimal responder cutoff, which could be carried forward into future trials. More importantly, we show that the proposed method, which we will refer to as MinP responder analysis in the rest of this paper, substantially improves the power of responder analysis. In many cases, the MinP responder analysis yields power that is approaching the analysis using the original continuous (or ordinal) measure.

The rest of the paper is structured as follows. In Section 2, we describe the proposed method. The method is then illustrated on a real data example in Section 3, and simulation studies evaluating the performance of the MinP responder analysis are presented in Section 4. Discussions and conclusion are given in Section 5.

2. Method

2.1. Design considerations

In general, suppose that the clinical endpoint is a continuous variable, Y, such that larger values represent better efficacy. Note that Y could represent a measurement taken at the conclusion of the trial or a change in that measurement from its baseline value. Assume, without loss of generality, a two-treatment trial with N_A subjects randomized to treatment A (e.g. experimental treatment) and N_B to treatment B (e.g. control). There is interest in the mean difference in this endpoint, μ, between the experimental treatment and the control.

The difference in treatment effects can be determined using the original continuous scale. In this case, the typical null hypothesis (assuming one-sided testing) is that of no difference, or $\mu \leq 0$, versus the alternative hypothesis $\mu > 0$.

Alternatively, with responder analysis, a threshold value is defined above which a subject is considered to be a "responder", and below which a subject is considered to be a "non-responder". If we let y_0 represent the threshold value, then

$$W = \begin{cases} 1 & \text{if } Y \geq y_0 \\ 0 & \text{if } Y < y_0 \end{cases}$$

is a binary variable indicating whether or not the subject is a responder. Now let p_A and p_B be the response rates in the experimental group and the control group, respectively. Therefore the null hypothesis for the responder analysis is $p_A \leq p_B$, and the alternative is $p_A > p_B$. If the responder null hypothesis is rejected then both statistical significance and clinical relevance are concluded. When the responder cutoff value y_0 is not well established and properly validated before the study, however, the results from such responder analysis could be inadequate or irrelevant. Moreover, as pointed out before, this approach substantially reduce the power of the study as information is lost through dichotomization the continuous endpoint.

2.2. MinP responder analysis procedure

Consider a setting for which there is a lack of consensus on the proper responder cutoff to use. Without loss of generality, assume that the continuous endpoint (and hence the responder cutoff y_0) take values in the interval (0, 100). The objective of the proposed design is to

1. Formally test for any treatment effect, i.e. determining whether a significantly greater proportion of subjects in the experimental arm "respond" to the treatment compared to the control arm based on a certain responder cutoff; and
2. Identify optimal responder cutoff which could be carried forward into future trials.

As the responder cutoff point is not well-established, we design the trial by pre-specifying multiple tests of responder rates between treatment groups based on a range of pre-specified responder cutoffs. Based on prior medical knowledge and discussion with the clinical team, a series of plausible candidate responder cutoffs $\{y_{0,k}: k = 1, 2, \ldots, K\}$ in the interval (0, 100) can be pre-specified. For instance, $\{y_{0,k}\} = \{10, 20, 30, \ldots, 90\}$. For each candidate cutoff $y_{0,k}$, a proportion test T_k will be performed to test the null hypothesis that $p_A \leq p_B$, resulting in a series of p-values $\{p_k\}$. A natural approach to converting a series of p-values that are calculated over the range of possible cutoff values into a single statistic is then to take the minimum:

$$minP = \min(p_1, \ldots, p_K)$$

Because of the well-known multiple testing problem, the standard asymptotic theory does not apply to the new statistic, *minP*. To provide a statistically valid p-value for *minP*, we propose to use the permutation distribution of *minP*, in which the treatment group

labels are permuted. The permutation procedure as described in Ganju et al. is summarized below [4]:

i. From the K pre-specified proportion tests fit to the data, obtain $minP_{obs} = \min(p_1, \ldots, p_K)$;
ii. Re-randomize treatments A or B to the $N_A + N_B$ subjects maintaining the same N_A/N_B ratio as that observed. This can be more easily understood by envisioning the data as an $(N_A + N_B)$ by 2 matrix, where one column is the response y and the other column contains the actual treatment assignments. Permutation or re-randomization is achieved by randomly re-ordering the labels 'A' or 'B' of the treatment vector. If the trial design is stratified such that randomization occurred within stratification levels, then permutations should be carried out within strata to reflect the design.
iii. Repeat step ii L times, and denote the minimum p-value for the lth permutation $minP_l^*$ ($l = 1, 2, \ldots, L$). A large enough subset of all possible permutations, rather than the complete set, which is often infeasible to generate, suffices. The smaller the α, the larger the number of permutations might be needed for more accurate estimation. The distribution of $minP^*$ denotes the permutation null distribution.
iv. The p-value of $minP_{obs}$ is $\frac{1}{L}\sum_{l=1}^{L} I(minP_l^* < minP_{obs})$, where $I()$ denotes the indicator function that takes the value 1 if the condition in parentheses is true and 0 otherwise.
v. The null hypothesis is rejected at level α if the p-value of $minP_{obs} < \alpha$.

If the permutation procedure rejects the null hypothesis of no treatment effect, the next step is to identify the responder cutoff above which a subject will be considered a "responder". Naturally the responder cutoff that produces the minimum p-value can be selected as the optimal cutoff point to be carried forward to future trials. In some cases, a few neighboring cutoff points might produce similarly small p-values. In this case, we recommend that other practical considerations, such as clinical interpretation, sample size of the potential "responder" group, etc., should be taken into account when selecting the responder cutoff for future studies.

3. Simulation studies

In this section, we evaluate the performance of MinP procedure for responder analysis in the context of trials with two treatment groups, active and control. The clinical endpoint of the trial is assumed to be continuous. Different analysis strategies are considered, including 1) comparison of the means of the two groups; 2) responder analysis, i.e. comparison of responder rates with a pre-specified responder cutoff; and 3) the MinP procedure for responder analysis. Two main scenarios are considered, and in each scenario we investigate the behavior of the type I error and the power under each analysis strategy. For both simulation studies we consider clinical trials with a sample size of 200 per group, which represents a typical Phase II clinical trial. The treatment difference is considered significant at the 0.05 level if the p-value is less than 0.05.

In **Scenario 1**, consider the measurement, Y, being a normally distributed variable with a known variance. Without loss of generality, assume a true mean value in the control group of $-\mu/2$, a true mean value in the experimental group of $\mu/2$, and an equal standard deviation of 24 such that most of the observations fall within [-100, 100]. Suppose a measurement value greater than 0 indicates improvement. For the MinP procedure, we use a grid of candidate responder cutoff values of 10, 20, 30, …, 90, i.e. $\{y_{0,k}\} = \{10, 20, 30, \ldots, 90\}$. In real clinical trial scenarios, a narrower range of candidate cutoff values might be used given prior medical knowledge and clinical input.

We performed simulations with $\mu = 0, 2, 3$, and 4. Table 1 shows the type I error and the power of detecting a treatment effect with different analysis strategies. As expected, the type I error is controlled under all analysis strategies. Under the alternatives, testing the difference in means (a t-test in this case) always gives a greater power compared to a responder analysis, which is not surprising as the t-test utilizes more information than the proportion test. Compared to the single proportion tests, the MinP procedure improves the power of detecting a treatment effect substantially, by around 4%, 7%, and 5% respectively, under each configuration. In other words, MinP is a combination of all single responder tests specified, yet its power is greater than the best of these single tests. This would translate to a much smaller sample size needed if the responder analysis is the desirable analysis method for the trial.

Scenario 2 assumes the measurement Y follows a beta distribution with shape parameters α and β. Without loss of generality, assume true $\alpha = 2$ for both treatment groups, a true β parameter in the control group of β_A, and a true β parameter in the experimental group of β_B. All observations fall within [0, 1] and a grid of candidate responder cutoff values of 0.1, 0.2, 0.3, …, 0.9, i.e. $\{y_{0,k}\} = \{0.1, 0.2, 0.3, \ldots, 0.9\}$, is pre-specified for the MinP procedure. The type I error and power results under 4 configurations are presented in Table 2. Again, the type I error is controlled under all analysis strategies. Under the alternatives, we observe the power may suffer dramatically if an inappropriate responder cutoff is chosen. For instance, under configurations 3 ($\beta_A = 2.4$) and 4 ($\beta_A = 2.6$), the lost in power can be 25% or more if a cutoff value of less than 0.4 (including 0.4) is chosen. Overall, compared to the single proportion tests, the MinP procedure improves the power of detecting a treatment effect by 2%–18% under configuration (2), 8%–51% under configuration (3), and 7%–74% under configuration (4). The optimal power is again achieved by testing the mean difference of the continuous variable, but the advantage over MinP is minimum (5%–6% in all alternative configurations).

4. Example

Finally we illustrate the performance of the proposed MinP method using data from published clinical trials of heart failure therapies. For clinical studies of cardiac resynchronization therapies (CRT), quality of life (QoL) measures and exercise tolerance (ET) measures have emerged as clinically relevant primary endpoints. One objective measure of ET that has been used in several studies is peak oxygen consumption (VO_2). However, there is currently no objectively justifiable precedence for how to select a universal signal value to define "responder" or "success" for the peak VO_2 measure used in clinical trials. In this section, we use data from two published CRT studies to illustrate how the MinP responder analysis method can be applied to such studies.

The Multicenter InSync Randomized Clinical Evaluation (MIRACLE) trial was a double-blind study of cardiac resynchronization in patients with moderate-to-severe heart failure and a prolonged QRS interval [8]. Approximately 400 patients were randomized and implanted with a CRT device which was turned on in half of the patients and left off in the other half for 6 months. A traditional analysis evaluating between-group differences in mean changes of the ET measurements was performed. The mean (±SD) change in peak VO_2 between baseline and 6 months was 0.2 ± 3.8 mL/kg/min in the control group and 1.1 ± 3.5 mL/kg/min in the treatment group. Significant treatment effect was established based on this analysis of peak VO_2 in continuous scale. Assuming that the data are normally distributed, these numbers can be used to construct estimated distributions of the data.

We intend to determine whether a clinically significant

Table 1
Type I error and power for detecting treatment difference (A) in means; (B) in responder rates with a pre-specified responder cutoff (10, 20, 30, …, 90); and (C) with MinP procedure. The sample size per group is 200.

Configuration	(A) difference in means	(B) difference in responder rates with pre-specified cutoff values:									(C) MinP
		10	20	30	40	50	60	70	80	90	
(1) $\mu = 0$	0.050	0.038	0.039	0.040	0.038	0.039	0.036	0.043	0.040	0.037	0.049
(2) $\mu = 2$	0.483	0.324	0.325	0.333	0.324	0.333	0.320	0.311	0.324	0.326	0.363
(3) $\mu = 3$	0.826	0.622	0.622	0.632	0.647	0.635	0.632	0.639	0.642	0.640	0.704
(4) $\mu = 4$	0.964	0.834	0.831	0.827	0.830	0.831	0.825	0.824	0.823	0.834	0.877

Results are based on 5000 simulations and 2000 permutations.

Table 2
Type I error and power for detecting treatment difference (A) in means; (B) in responder rates with a pre-specified responder cutoff (0.1, 0.2, 0.3, …, 0.9); and (C) with MinP procedure. The sample size per group is 200.

Configuration	(A) difference in means	(B) difference in responder rates with pre-specified cutoff values:									(C) MinP
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
(1) $\beta_A = 2.0$, $\beta_B = 2.0$	0.050	0.025	0.036	0.038	0.042	0.036	0.036	0.039	0.038	0.024	0.046
(2) $\beta_A = 2.2$, $\beta_B = 2.0$	0.274	0.050	0.092	0.120	0.161	0.182	0.210	0.198	0.154	0.072	0.225
(3) $\beta_A = 2.4$, $\beta_B = 2.0$	0.673	0.096	0.202	0.325	0.410	0.504	0.531	0.509	0.418	0.182	0.607
(4) $\beta_A = 2.6$, $\beta_B = 2.0$	0.917	0.127	0.334	0.534	0.661	0.763	0.789	0.762	0.670	0.258	0.867

Results are based on 5000 simulations and 2000 permutations.

Table 3
Power for detecting treatment difference (A) in means; (B) in responder rates with a pre-specified responder cutoff (0.5, 0.6, 0.7, 0.8, 0.9, and 1.0 mL/kg/min); and (C) with the MinP procedure. The sample size per group is 200 for both the MIRACLE trial and the COMPANION trial.

Scenario	(A) difference in means	(B) difference in responder rates with pre-specified cutoff values (mL/kg/min):						(C) MinP
		0.5	0.6	0.7	0.8	0.9	1.0	
(1) MIRACLE trial	0.743	0.582	0.594	0.592	0.582	0.588	0.572	0.631
(2) COMPANION trial	0.667	0.399	0.427	0.435	0.443	0.445	0.460	0.510

Results are based on 5000 simulations and 2000 permutations.

difference would have been demonstrated in terms of difference in responder rates had a responder analysis, either the traditional responder analysis with a single pre-specified cutoff point or the MinP method, been planned as the primary analysis. A total of 5000 simulated trials were conducted. Seven (7) plausible cutoffs for peak VO_2 measure are used to defined response: at least 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0 mL/kg/min increase from baseline 6 months. Three analysis strategies are compared: 1) comparison of the means of the two groups; 2) responder analysis, i.e. comparison of responder rates with the pre-specified responder cutoff; and 3) the MinP procedure for responder analysis. The power of each test to detect significant differences is reported in Table 3. As can be seen, the continuous test still provides the greatest power (74.3%), but MinP is able to improve power by 4%–6% for responder analysis.

This evaluation can be repeated with data from the Comparison of Medical, Pacing, and Defibrillation Therapies in Heart Failure (COMPANION) clinical trial, where subjects were randomized in a 1:4 ratio to optimal medical therapy (OPT) or to OPT plus CRT [9]. In the COMPANION Sub-study, cardiopulmonary exercise testing (peak VO_2) was assessed as the primary endpoint, and success was defined as occurring if peak VO_2 improved $\geq$ 0.7 mL/kg/min at 6 months of assigned therapy. The mean (±SD) change in peak VO_2 between baseline and 6 months was 0.6 ± 2.7 mL/kg/min in the control OPT group and 1.2 ± 3.0 mL/kg/min in the CRT group. The primary endpoint was not met in this study. We are interested to see if the study power could have been improved if a MinP responder analysis approach was used. Similarly using simulated data from the trial, we report in Table 3 the power of each analysis strategy to detect significant differences. In this case, the MinP method is able to improve the power of responder analysis by 5%–11%.

5. Discussion and conclusions

A responder analysis is one in which each subject is classified as either a “responder” or a “non-responder”, and the proportions of patients who benefit are quantified and compared between treatment groups. The use of responder analysis is often recommended by regulatory guidance, especially in trials where “soft” clinical endpoints such as rating scales are used to evaluate treatments. Although the responder analysis is in common use, it has substantial disadvantages. It has been widely acknowledged that the main concern about the responder analysis is the arbitrary nature of the definition of a response. A second problem with the analysis is the dramatic reduction in statistical power by dichotomizing continuous endpoints. In this paper, we address these two issues together by proposing a novel approach for responder analysis that could both improve the power of responder analysis and explore different responder cutoffs if an agreed-upon common cutoff is not present. Specifically, we propose a statistically rigorous clinical trial design that pre-specifies multiple tests of responder rates between treatment groups based on a range of pre-specified responder cutoffs, and uses the minimum of the p-values for formal inference.

The MinP procedure enables us to both establish the treatment effect and find the responder cutoff at the same time. Hence we recommend prospectively incorporating the multiple testing and permutation procedures into the study design and describe them in the study protocol. his method can be most useful in Phase II studies where such exploration and selection of responder cutoff points are appropriate. The information learned and the optimal cutoff selected from this study can then be carried forward to future, Phase III trials which perhaps are less likely to implement such flexible design.

Simulations can be used in determining the sample size required for a study designed using the proposed MinP method. The clinical trialists should be considerate in the selection of candidate cutoff points; only cutoff values that represent a clinically meaningful treatment effect should be included in the design and analysis plan. Statistical judgment should also be in place to include a reasonable number of candidate cutoff values to avoid adversely affecting the sample size required.

References

[1] L. Burke, T. Stifano, S. Dawish, Guidance for industry: patient-reported outcome measures: use in medical product development to support labelling claims: draft guidance, Health Qual. Life Outcomes 4 (2006) 79.

[2] S.M. Snapinn, Q. Jiang, Responder analyses and the assessment of a clinically relevant treatment effect, Trials 8 (2007) 31.

[3] S. Leucht, J.M. Davis, R.R. Engel, J.M. Kane, S. Wagenpfeil, Defining "response" in antipsychotic drug trials: recommendations for the use of scale-derived cutoffs, Neuropsychopharmacology 32 (2007) 1903–1910.

[4] J. Ganju, X. Yu, J. Ma, Robust inference from multiple statistics via permutations: a better alternative to the single statistic approach, Pharm. Stat. 12 (2013) 282–290.

[5] J. Ganju, G. Ma, The potential for increased power from combining p-values testing the same hypothesis, Stat. Methods Med. Res. (2014), http://dx.doi.org/10.1177/0962280214538016.

[6] Y. Lin, K. Zhou, J. Ganju, A single test for rejecting the null hypothesis in subgroups and in the overall sample, J. Biopharm. Stat. http://dx.doi.org/10.1080/10543406.2016.1148718

[7] J. Ganju, Y. Lin, K. Zhou, Robust inference for group sequential trial, *Pharm. Stat.* Under Review.

[8] W.T. Abraham, W.G. Fisher, A.L. Smith, et al., Cardiac resynchronization in chronic heart failure, N. Engl. J. Med. 346 (24) (2002) 1845–1853.

[9] T. De Marco, E. Wolfel, A.M. Feldman, B. Lowes, M.B. Higginbotham, J.K. Ghali, et al., Impact of cardiac resynchronization therapy on exercise performance, functional capacity, and quality of life in systolic heart failure with QRS prolongation: COMPANION trial sub-study, J. Card. Fail. 14 (1) (2008) 9–18.