Evaluating the Utility of ROC Analysis in Blinded Evaluation of an Ongoing Clinical Trial

Bonzo, Daniel
XenoPort, Inc., Biometrics and Data Management
3410 Central Expressway, Santa Clara, CA 95051
daniel.bonzo@xenoport.com

Bian, Amy
XenoPort, Inc., Biometrics and Data Management
3410 Central Expressway, Santa Clara, CA 95051
amy.bian@xenoport.com

Kameda, Tosh
XenoPort, Inc., Biometrics and Data Management
3410 Central Expressway, Santa Clara, CA95051
tosh.kameda@xenoport.com

1. Introduction

The importance of statistical monitoring in the ongoing evaluation of a clinical trial's risk-benefit to trial subjects and that of the Sponsor is well recognized. When used appropriately, it can greatly impact the advancement or demise of a clinical development program. We take the view that blinded statistical monitoring is an appropriate tool to use to help reduce development efforts by identifying ineffective drugs when they are still in the exploratory phase of development. Likewise, it is an appropriate tool that can be used to trigger Phase 3 planning and development if it shows that the experimental drug is effective during its exploratory phase of development.

In this paper, we propose the use of ROC analysis in deriving a predictive rule for inferring whether a subject has received active drug or placebo in a placebo-controlled trial. ROC analysis will be performed between a clinical benefit endpoint and a global rating of change (GRC) scale as anchor. Classification based on the derived predictive rule will be used as a surrogate for the actual treatment assignment.

Limited simulation experiment was performed to evaluate the usefulness of the proposal. Comparison of the effect size based on the original treatment assignment and the effect size based on the surrogate treatment assignment was done. The predictive value of the said technique was assessed using measures such as area under the curve (AUC), distance from the perfect marker (DPM) and distance from the non-informative marker (DNM).

The organization of the paper is as follows. Section 1 provides the problem setting background that motivated the proposal. In Section 2 the technical details of the proposal is presented. Section 3 discusses the limited simulation that was done to assess the utility of the proposal. Section 4 provides a discussion of the simulation results. Finally, Section 5 presents a brief summary of the simulation results.

2. Methodology

The methodology proposed in this paper is to utilize ROC analysis in deriving a prediction rule that can be used as a surrogate for treatment assignment in a placebo-controlled clinical trial. If the prediction rule assigns a

subject as a responder then the subject is inferred as having received active treatment. In the case that the prediction rule assigns a subject as a non-responder then the subject is inferred as having received placebo.

Derivation of the prediction rule utilized a subject-reported global rating of change (GRC) as an anchor or predictor. GRCs are designed to quantify a patient's improvement or deterioration over time, usually either to determine the effect of an intervention or to chart the clinical course of a condition. The design of the GRC scale is such that the outcome will assesses current health status versus a recollection of a previous time-point health status. An example of a GRC scale is given by the patient-reported overall treatment effect (Revicki, et. al., 2004) [POTE]. The scale is a 15-point scale and seeks to assess change in a patient's condition relative to some clinical visit in the past (See Table 1 below). The prediction rule will be linked to a clinical endpoint that is seen as the gold standard for determining treatment responders.

Formally, let Y be the patient's clinical endpoint response and X be the patient's GRC measure. Suppose there exists a threshold c_Y such that the subject is classified as a responder if $Y \geq c_Y$ and non-responder, otherwise. Denote by c_X the corresponding threshold using X. Then given c_Y, we wish to determine the value of c_X such that $P(Y \geq c_Y) = P(X \geq c_X)$. If F_Y and F_X are the distribution functions of Y and X, respectively, then $F_Y(c_Y) = F_X(c_X)$. We utilized the method provided in Gonen (2007) to derive the optimal value for c_X. The selection of c_X depended on the profile of the resulting ROC curve. Note that a fixed c_Y will induce an ROC curve in the unit square given by

$$\{(x(t), y(t)) : x(t) = P(X > t \mid Y < c_Y), y(t) = P(X > t \mid Y \geq c_Y), -\infty < t < \infty\}$$

It is clear then that the resulting profile of the ROC curve depended heavily on the dependence structure between Y (clinical endpoint) and X (anchor or GRC)..

Typical measures used to assess the resulting ROC curve were the area under the curve (AUC), distance from the perfect marker (DPM) and distance from the non-informative marker (DNM). The ideal ROC will have an AUC of 1, a DPM of 0, and a DNM close to 1. For further discussion of these measures see Krazanowski and Hand (2009) and Pepe (2003).

Table 1 Patient-Rated Overall Treatment Effect (POTE) Scale

How much better/worse would you say your symptoms have been since your baseline study visit?

About the same ☐4

Almost the same, hardly worse at all ☐1	Almost the same, hardly better at all ☐5
A little worse ☐2	A little better ☐6
Somewhat worse ☐3	Somewhat better ☐7
Moderately worse ☐4	Moderately better ☐8
A good deal worse ☐5	A good deal better ☐9
A great deal worse ☐6	A great deal better ☐10
A very great deal worse ☐7	A very great deal better ☐11

The use of a patient-rated GRC as a surrogate for treatment assignment will induce misclassification. The attractiveness of the resulting ROC analysis can then be evaluated based on the predictive profile of the resulting classification rule based on c_X, i.e, patient is inferred as assigned to active treatment if $X \geq c_X$ and placebo, otherwise. Misclassification rate is the number of patients incorrectly classified by the anchor (GRC) compared to the true responder status as given by the clinical endpoint. Other measures for assessing the predictive profile of the resulting classification rule included the positive predictive value (PPV) and negative predictive value (NPV). PPV is the number of true responders (as given by the clinical endpoint) in the set of responders classified based on the anchor (GRC) divided by the number of patients classified by the anchor (GRC) as responders. NPV is the number of true non-responders (as given by the clinical endpoint) in the set of non-

responders classified based on the anchor (GRC) divided by the number of patients classified by the anchor (GRC) as non-responders. A good predictive profile will have a misclassification rate close to zero and NPV and PPV close to 1.

In addition to the standard measures given above, the predictive profile of the method were further assessed by obtaining the two-sided p-value of a test statistic used in assessing treatment difference between the active and placebo groups. We called this p-value as the surrogate p-value. The calculated effect size based on the surrogate treatment assignment, also known as the surrogate treatment effect, was be used to assess the magnitude of the clinical benefit. A surrogate p-value near 0 and a surrogate treatment effect greater than or equal to 0.8 indicate good separation between the responder and non-responder populations.

3. Simulation Experiment

To demonstrate the utility of the technique, we simulated data based on a clinical trial on patients with Gastroesophageal Reflux Disease (GERD) who were incomplete responders to a Proton Pump Inhibitor (PPI). The trial was designed as a randomized, multi-center, double-blind, placebo-controlled parallel group dose-ranging study. For patients to be randomized into the study, they must have had a minimum of 3 days of heartburn symptoms in the week prior to screening period and must have confirmed to have a minimum of 4 events (episodes) of heartburn symptoms on ≥ 3 days during the baseline week, have documented compliance of PPI dosing and must have no structural abnormalities of the GI tract as verified by an endsocopy. Patients randomized were then asked to take the drug that they were randomized to with food. Efficacy and safety evaluations were performed over 6 weeks of treatment.

An electronic diary was used to capture the frequency and severity of GERD symptoms of heartburn and regurgitation and other associated symptoms. Patients were trained to use the diary during the Screening Period. Patient were then asked to record their symptoms from the start of the Baseline Period through the final visit at Week 6. The primary efficacy endpoint of the study was percent change from baseline in heartburn (a burning discomfort or pain behind the breastbone, denoted by PCHB) events at Week 6.

Simulated data at week 6 were generated for the best active dose and placebo dose groups. Several effect sizes were used to evaluate the utility of the technique when the efficacy signals are weak (0.2), moderate (0.5), and strong (0.8). In order to simplify the data generation, both the primary endpoint and anchor were assumed to be normally distributed. The mean PCHB at Week 6 was set to -20 % for the best active dose and 0 for placebo. The pooled standard deviations were then adjusted in order to obtain the desired effect sizes.

Each simulation consisted of n data points for the primary endpoint for both the best active dose and placebo groups. Corresponding values for POTE were then generated using various correlation values to represent weak (-0.3), moderate (-0.6) and strong (-0.9) association between the primary endpoint and the anchor. Mean POTE at Week 6 was set to 0 (about the same) for placebo and 2 (a little bit better) for the best active dose. Pooled variance was set to 7. Each simulation was then replicated for a total of 1000.

Two sets of n's were utilized: n = 36 and n=72.

4. Analysis

In evaluating the proposed technique, true treatment responders were defined as patients who have achieved a PCHB of less than or equal to -50% at Week 6, i.e. $c_Y = -50\%$. ROC analysis was performed on the true responder data using POTE as the predictor. Patients that had a POTE value of at least equal to the derived threshold (c_X) was identified as a responder. PCHB was then analyzed using a simple analysis of variance (ANOVA) model with fixed effect for responder status based on POTE as the surrogate for treatment group. Table 2 below shows an example of a summary table showing the derived threshold ($c_X = 0$ or no change) and the corresponding measures AUC, DPM and DNM. As expected of a 'no change' threshold, the corresponding values for AUC, DPM and DNM were far from the ideal values of 1, 0 and 1, respectively. Table 3 below shows an example of the analysis results using POTE as the surrogate for treatment group. Results showed that the POTE as the surrogate treatment provided good discrimination of PCHB values.

The predictive profile of the method was assessed by calculating misclassification rates, PPV and NPV. In addition, the two-sided p-value of the of the t-test used in assessing treatment difference was obtained utilizing an analysis of variance (ANOVA) model with fixed effect for the surrogate treatment group.

The predictive statistics were obtained for each set of simulation parameters for a total of 1000 replications. Mean and standard deviation of the predictive statistics were calculated for each simulation parameter. Summary results for n = 36 are provided in Table 4 and results for n = 72 are provided in Table 5..

Results showed that the predictive profile was invariant with respect to the original effect sizes. Changes in the predictive profile were evident as correlation increased in negative value. Both the two-sided surrogate p-value and the misclassification rate decreased as correlation increased in negative value. Both NPV and PPV increased as correlation increased in negative value.

Interestingly, the surrogate effect sizes were inflated compared to the original effect sizes. Inflation of the effect sizes increased as the correlation increased in negative value. This showed that the surrogate treatment effect is somehow problematic when used to predict the original effect sizes. However, results showed that prediction of positive outcome for the trial (as given by the surrogate p-value) increased as correlation increased in negative value. These results were true for both sample size cases with greater reliability for n = 72 as compared to n = 36.

5. Concluding Remarks

The proposed approach for blind monitoring utilizes an ROC analysis of a clinical endpoint with a GRC scale as an anchor to derive the optimal threshold for determining treatment responders. Subjects classified as responders by the threshold are inferred as having received active treatment and those classified as non-responders are inferred as having received placebo. The surrogate treatment assignment can then be used to evaluate if treatment difference can be deduced for the clinical endpoint by specifying a suitable analysis model relating the clinical endpoint with the surrogate treatment group as one of the fixed effects. Limited simulation performed to assess the performance of this method showed that as correlation between the clinical endpoint and the GRC scale increases, the ability to predict a positive trial increases. In particular, the method starts to have predictive value when the correlation reaches the moderate range. However, the resulting surrogate treatment effect was not as useful as they tended to be inflated with inflation increasing as the correlation increased.

Utility of this approach in the context of blind monitoring can easily be gauged as the correlation can be computed based on available blinded data for the clinical endpoint and the GRC scale. We take the view that this method can be used in blinded statistical monitoring when there is an overriding concern to screen ineffective drug candidates or as a tool to help trigger Phase 3 planning and development of a candidate drug during its

exploratory phase of development.

REFERENCE

Gonen M. Analyzing receiver operating characteristic curves with SAS, SAS Institute, Inc.: Cary (2007).

Krzanowski WJ and Hand DJ. ROC curves for continuous data, CRC Press: New York (2009).

Pepe MS. The statistical evaluation of medical tests for classification and prediction, Oxford University Press: Oxford (2003).

Revicki, D.A., Rentz, A.M., Tack, J., Stanghellini, V., Talley, N.J., Kahrilas, P., de la Loge, C., Trudeau, E., Dubois, D. Responsiveness and interpretation of a symptom severity index specific toupper gastrointestinal disorders. Clinical Gastroenterology and Hepatology, 2004 2:769-777.

Table 2 Summary of ROC Analysis Results for POTE using Percent Change in Heartburn as True Response Indicator

Percent Change in NHB Threshold	Visit	NHB Responders n (missing)	NHB Responders n (%)	AUC (95% CI)	DPM[1]	DNM[1]	POTE Threshold
<= -50%	Week 6	144 (0)	23 (16%)	0.634 (0.526, 0.740)	0.191	0.572	0

[1] DPM = Distance from the Perfect Marker. DNM = Distance from the Non-Informative Marker.

Table 3 Summary of Responder Analysis Results Based on POTE Responder Definition

Characteristic[1] Visit	Statistics	Responder Status[2] (N=144) Responder (N=80)	Non-Responder (N=64)
Percent change from Baseline in: HEARTBURN EVENTS PER WEEK			
Week 6	n (missing)	80 (0)	64 (0)
	Mean (SD)	-15.9 (44.04)	-0.6 (33.93)
	Median	-19.51	-2.72
	Min, Max	-100.0, 92.2	-60.5, 77.8
	LS Mean (SE)	-15.9 (4.5)	-0.6 (5.0)
	LS Mean Difference (SE)	15.2 (6.7)	
	p-value	0.024	

[1] Endpoint was analyzed using a mixed model with fixed effects for responder status.
[2] Responders are defined as those subjects that showed a score of >=0 in POTE at Week 6 of treatment.

Table 4 Predictive Profile of POTE when Used as a Surrogate for Treatment (N=36 per dose group)

Correlation (PCHB vs POTE)	PCHB Effect Size	Surrogate Effect Size	Surrogate p-value	Overall Misclassification Rate	PPV	NPV
-0.3	0.2	0.49 (0.20)	0.11 (0.18)	0.48 (0.06)	0.22 (0.06)	0.91 (0.05)
	0.5	0.60 (0.19)	0.05 (0.12)	0.48 (0.05)	0.23 (0.06)	0.91 (0.05)
	0.8	0.65 (0.18)	0.03 (0.07)	0.47 (0.06)	0.23 (0.07)	0.92 (0.05)
-0.6	0.2	1.06 (0.15)	<0.01 (0.00)	0.31 (0.05)	0.33 (0.08)	0.95 (0.03)
	0.5	1.08 (0.14)	<0.01 (0.00)	0.33 (0.05)	0.31 (0.08)	0.96 (0.03)
	0.8	0.99 (0.16)	<0.01 (0.00)	0.32 (0.05)	0.32 (0.08)	0.95 (0.03)
-0.9	0.2	1.41 (0.10)	<0.01 (0.00)	0.26 (0.05)	0.39 (0.09)	0.99 (0.01)
	0.5	1.45 (0.09)	<0.01 (0.00)	0.27 (0.05)	0.38 (0.09)	1.00 (0.01)
	0.8	1.45 (0.09)	<0.01 (0.00)	0.28 (0.05)	0.37 (0.09)	1.00 (0.01)

Table 5 Predictive Profile of POTE when Used as a Surrogate for Treatment (N=72 per dose group)

Correlation (PCHB vs POTE)	PCHB Effect Size	Surrogate Effect Size	Surrogate p-value	Overall Misclassification Rate	PPV	NPV
-0.3	0.2	0.59 (0.14)	0.01 (0.04)	0.48 (0.04)	0.23 (0.05)	0.92 (0.03)
	0.5	0.51 (0.14)	0.02 (0.04)	0.48 (0.04)	0.22 (0.05)	0.91 (0.04)
	0.8	0.50 (0.14)	0.03 (0.08)	0.49 (0.04)	0.22 (0.04)	0.90 (0.04)
-0.6	0.2	1.06 (0.10)	< 0.01 (0.00)	0.32 (0.04)	0.32 (0.06)	0.96 (0.02)
	0.5	1.06 (0.11)	< 0.01 (0.00)	0.31 (0.04)	0.33 (0.06)	0.96 (.0.02)
	0.8	1.06 (0.11)	< 0.01 (0.00)	0.31 (0.04)	0.33 (0.06)	0.96 (0.02)
-0.9	0.2	1.45 (0.06)	< 0.01 (0.00)	0.26 (0.04)	0.38 (0.06)	1.00 (0.01)
	0.5	1.46 (0.06)	< 0.01 (0.00)	0.27 (0.04)	0.38 (0.06)	1.00 (0.01)
	0.8	1.46 (0.06)	< 0.01 (0.00)	0.27 (0.04)	0.38 (0.06)	1.00 (0.01)

RÉSUMÉ

The importance of statistical monitoring in the ongoing evaluation of a clinical trial's risk-benefit to trial subjects and that of the Sponsor is well-recognized. When used appropriately, it can greatly impact the advancement or demise of a clinical development program. We take the view that blinded statistical monitoring is an appropriate tool to use to help reduce development efforts by identifying ineffective drugs when they are still in the exploratory phase of development. Likewise, it is an appropriate tool that can be used to trigger Phase 3 planning and development if it shows that the experimental drug is effective during its exploratory phase of development.

In this paper we propose an approach for use during a blinded evaluation of an experimental drug's risk-benefit which hinges on the identification of a useful surrogate for the actual treatment assignment. The method uses a treatment response indicator whose value can be determined by utilizing Receiver Operating Characteristic (ROC) analysis on the clinical trial's primary efficacy endpoint and using an appropriate global rating of change (GRC) scale as treatment response anchor.

The utility of the proposed approach was evaluated through limited simulation using various effect sizes of the primary efficacy endpoint and correlations between the GRC scale and the primary efficacy endpoint. Limited simulation performed to assess the performance of this method showed that as correlation increased the ability to predict a positive trial outcome also increased. In particular, the method started to have predictive value when the correlation reached the moderate range. However, the resulting surrogate treatment effect estimates were not as useful as they tended to be inflated with increasing inflation as correlation increased.

Mots clés: receiver operatic characteristic (ROC) curve, global rating of change (GRC) scale, statistical monitoring, predictive profile, patient-rated overall treatment effect (POTE)