Dichotomizing Continuous Biomarkers in the Co-Development of Drug and Companion Diagnostics: Practical Considerations

Liang Fang, Adarsh Joshi, Huan Wang, Yafeng Zhang
Gilead Sciences, Inc.
Co-Development of Drug and Companion Diagnostics

Fridley et al., 2013, Contemporary Clinical Trials
An Ideal Framework

At biomarker exploration stage (Ph1/II):

- Found a GOOD cutoff value such that
  - Treatment benefit in the BM+ group is clinically meaningful and statistically significant
  - while, treatment benefit in the BM- group is not clinically meaningful or statistically significant

At biomarker validation stage (PhIII):

- Pre-specify BM+ group with the cutoff determined at the exploration stage
- Formally test the treatment effect in the BM+ group in the pivotal study
A “GOOD” cutoff value can not be determined in Phase II, because

- Often time, data at the exploration stage (Phase I/II)
  - is limited due to small sample size
  - has quality issue due to analytical performance of the assay (market ready assay is usually not available at this stage)
  - Sometimes, come from a single arm study (especially, in oncology)
- Phase I/II and Phase III studies have somewhat different patient population, may have different control arms
- However, the decision of moving a drug (or biomarker) to late phase depends on: (1) early phase data evidence; (2) biological hypothesis; (3) business considerations (ex, competitor intelligence, company goals, etc)
  - Very typical for target therapy (ex, HER2)
- Development timeline is expedited (“race to market”)
- Biomarker as a ‘rescue plan’ (“in case ITT failed, how about BM subgroup?”)
- It seems we never have enough information to determine a GOOD cutoff in Phase II prior to the initiation of Phase III
Question:

is it even realistic to think that we can find a GOOD cutoff point based on data from Phase I/II?

PERSPECTIVES

Considerations for the successful co-development of targeted cancer therapies and companion diagnostics

Jane Fridlyand, Richard M. Simon, Jessica C. Walrath, Nancy Roach, Richard Buller, David P. Schenkein, Keith T. Flaherty, Jeff D. Allen, Ellen V. Sigal and Howard I. Scher

Although the scientific rationale, pre-clinical data and epidemiological data may be suggestive of a threshold for companion diagnostics that measure continuous biomarkers, the relevant clinical data that are required to establish a reasonable threshold are typically very limited prior to the initiation or analysis of the first pivotal clinical trial. For example, for first-in-class drugs
Some Mathematical Results

- Let $C$ be the continuous biomarker value
  - Let $f(C), g(C)$ be functions associated with the predictive and prognostic effects of the biomarker
  - Let $\beta, \theta, \gamma$ be parameters relating to the treatment indicator, and the predictive and prognostic effects.
  - The hazard:
    $$\lambda(t|X) = \lambda(t)e^{\beta I_{\{\text{Tx}\}} + \theta I_{\{\text{Rx}\}} f(C) + \gamma g(C)}$$
    - The log-hazard ratio between treatment and control: $\log(HR) = \beta + \theta f(C)$
  - Model is simple extension of standard setup involving only Tx effect, and accommodates, through shape of $f(C)$, both a rapidly changing or smoothly changing HR.
  - We seek the variance/covariance $\text{var}(\hat{\beta}, \hat{\theta}, \hat{\gamma})$

- With $C \sim U(0,1)$, $\log(HR) = \beta + \theta \times C$
  - then, approximately, for large $d$,
    $$\text{Var}(\hat{C}_t) = \text{Var}\left\{ \left( \frac{\log(HR_{\text{var}}) - \hat{\beta}}{\hat{\theta}} \right) \right\} = \frac{1}{d} \times \frac{24C_t^2 + 10 - 24C_t}{\hat{\theta}^2}$$
    $$\text{Var}(\hat{C}_t) = \frac{1}{\hat{\theta}^2} \times \frac{4}{d}$$
  - This is minimized for $C_t = .5$, for which $\text{Var}(\hat{C}_t) = \frac{1}{\hat{\theta}^2} \times \frac{4}{d}$
  - Example: let $\beta = 0$, $\theta = -1$ (strong dependence of outcome on C; 66% risk reduction for pats w/ C=1).
    - With $d=400$, 95% conf. interval for $C_t$ is $\hat{C}_t \pm .2$
      - Best case scenario does not sufficiently accurately identify biomarker positive patient subgroup.

- Math suggests that accurate estimation of the threshold is not a realistic pursuit in early development
  - unless biomarker effect is very strong at threshold (step function)
- What question(s) can realistically be answered in the early oncology setting?
  - Do patients with biomarker level above a pre-specified value, say 3rd quartile ($C > .75$), benefit clinically?
    $$H_0 : \beta + .75\theta \geq \ln(HR_{\text{target}})$$
    $$H_1 : \beta + .75\theta < \ln(HR_{\text{target}})$$
  - To power, Table 1 Mackey & Bengtsson 2013 gives variance of any contrast.
Can a pivotal study being designed without a pre-specified cutoff?

“To date, widely accepted statistical approaches to design and power a clinical study to estimate a continuous biomarker with an unknown threshold are not available. It should be noted that the FDA are preferential to prospectively designed studies as a regulatory requirement, although retrospective options may not be completely ruled out”

• Fridlyand et al 2013; Contemporary Clinical Trials; “An industry statistician’s perspective of PHC drug development”

“for biomarkers that are not clearly binary: how can Phase III studies be conducted without pre-specifying a threshold cut-off value for biomarker selection? And how can the threshold value be readjusted from a pre-specified value and how may it be pre-planned?”

• Fridlyand et al 2013; Nature Reviews; “Considerations for the successful co-development of targeted cancer therapies and companion diagnostics”
Is Adaptive Design the Solution?

Adaptive Designs – Determine and Confirm subgroup in one trial
  • Adaptive threshold design
  • Adaptive enrichment design

Challenges:
  • no regulatory precedence
  • Methodology is not straightforward to implement or easy for non-statisticians to comprehend
What about Explore-and-confirm approach?

- Explore and Confirm approach

Proposed in the context of multi-marker signatures, but could be considered in this setting
- alpha splitting based on Bonferroni rule in the original paper; but that can be modified to take advantage of the correlation between the all-comers and BM analyses (as described in Spiessens’ (2010))

Operationally complexity: time lost between exploration and confirmation
A Practical 3-Step Approach to Explore, Validate, and Refine the Cutoff Value

**Exploration**
- Explore cutoff in Phase 1/2 with multiple tools
- Nominate 2 (or 3) candidate cutoff points for validation at the next step

**Validation**
- Design a Phase 3 study with multiple primary hypotheses (e.g., ITT, biomarker subgroup 1, and subgroup 2)
- Allocate alpha of the 3 hypotheses according to correlation matrix among the 3 populations and estimated PoS

**Refinement**
- For the biomarker subgroup with positive results at the validation stage, refine the cutoff value around the nominated value
Cutoff Determination at Exploration Stage
Cutoff Determination in Literature

Most published methods focuses on two statistical points:

- Minimize or maximize an objective function
- Multiplicity

Unfortunately, many confusions around these two points occur in practice
Which objective function is the right one?

- Statistical literature often focused on which statistics to minimize/maximize (ex, p-value vs likelihood, etc)
- Which threshold to determine (biology vs clinical) is a more important question, but often ignored
The answer is “it depends”

- Multiplicity adjustment is an important issue for biomarker signal detection, ie, is this biomarker signal real?
  - Ex, bioinformatics, hypothesis free approach
- However, for biomarkers with strong biologic hypothesis and the decision to advance a biomarker to late stage is made primarily based on biological evidence, multiplicity control is less important
  - The decision of moving a drug (or biomarker) to late phase depends on: (1) early phase data evidence; (2) biological hypothesis; (3) business considerations (ex, competitor intelligence, company goals, etc)
  - Very typical for biomarkers associated with targeted pathway (ex, HER2)
- **Thus, our main goal is to estimate a relationship between biomarker and treatment effect (ie, estimation vs hypothesis testing)**
  - This is a well accepted concept for early drug development, however, biomarker developers often found it hard to accept
We will review 5 methods today

– Percentile/quartile method
– STEPP
– Janes and Huang
– Spline methods
  • Penalized Splines
  • Constrained GAM (cGAM)
### Percentile/Quartile Method

#### Table

<table>
<thead>
<tr>
<th>Biomarker Risk Factor</th>
<th>Control</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total n</td>
<td># event</td>
</tr>
<tr>
<td><strong>All Patients</strong></td>
<td>991</td>
<td>304</td>
</tr>
<tr>
<td>HER2 qRT-PCR concentration ratio</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min-30th percentile</td>
<td>266</td>
<td>93</td>
</tr>
<tr>
<td>10th-40th percentile</td>
<td>266</td>
<td>87</td>
</tr>
<tr>
<td>20th-50th percentile</td>
<td>266</td>
<td>70</td>
</tr>
<tr>
<td>30th-60th percentile</td>
<td>266</td>
<td>74</td>
</tr>
<tr>
<td>40th-70th percentile</td>
<td>267</td>
<td>77</td>
</tr>
<tr>
<td>50th-80th percentile</td>
<td>267</td>
<td>80</td>
</tr>
<tr>
<td>60th-90th percentile</td>
<td>268</td>
<td>81</td>
</tr>
<tr>
<td>70th percentile-Max</td>
<td>267</td>
<td>84</td>
</tr>
</tbody>
</table>
STEPP Method

- Form subpopulations by selecting subjects whose biomarker values fall into intervals \((l_j, u_j), j=1,\ldots,K\), where \(l_j < l_{j+1} < u_j < u_{j+1}\), so that consecutive subpopulations overlap.
- Intervals \((l_j, u_j)\) are chosen so that proportion \(r_1\) of all subjects fall into each subpopulation, and \(r_2 (< r_1)\) of all subjects overlap between consecutive subpopulations.
- Difference of KM survival probabilities at time \(t\) are calculated between treatment and placebo groups for each subpopulation, \(\theta_j = S\text{A}_j(t) - S\text{B}_j(t)\), \(\theta_1, \ldots, \theta_K\) asymptotically follow a multivariate normal distribution. Simultaneous confidence interval can also be constructed for \(\theta_1, \ldots, \theta_K\).
- A STEPP plot is created by plotting \(\theta_1, \ldots, \theta_K\) and simultaneous CI against the median biomarker values of each subpopulation.

Reference: 17
Fig. 3. STEPP plot for IBCSG Trial IX data—ER subgroups: (a) 5-year disease-free survival (DFS) percentages for CMF followed by tamoxifen and for tamoxifen alone; (b) 5-year DFS difference (CMF followed by tamoxifen minus tamoxifen alone).
Methods for Estimation and Inference

Marker-by-treatment predictiveness curves can be estimated by modeling the response rate as a function of treatment and marker value

\[ P(S = 1|T,Y) \]

for example, by using logistic regression (7, 8). The marker distribution can be estimated empirically in the entire trial population and used to calculate the marker value that corresponds to each fixed percentile. At a specific marker positivity threshold \( y \), which corresponds to the percentile

\[ \nu = F(y) \]

the proportion of marker-negative individuals is estimated by

\[ \hat{F}(y) \]

The response rate under the corresponding marker-based treatment policy

\[ P(S = 1|T = 0, F(Y) < \nu)\nu + P(S = 1|T = 1, F(Y) > \nu)(1 - \nu) \]

can be estimated by (23):

\[
\int_0^\nu \hat{P}(S = 1|T = 0, Y = F^{-1}(w))dw + \int_\nu^1 \hat{P}(S = 1|T = 1, Y = F^{-1}(w))dw
\]

We recommend bootstrapping for inference, to account for uncertainty in both the response rate model and the estimated marker distribution (26).
Figure 2: Distribution of the treatment effect, as measured by the difference in disease rate without vs. with treatment, $\Delta(Y) = P(D = 1|T = 0, Y) - P(D = 1|T = 1, Y)$, for the Oncotype-DX-like marker ($Y_1$) and the strong marker ($Y_2$). Horizontal pointwise 95% confidence intervals are shown.
Monotonously increasing p-spline

- likelihood:
  \[ S_i \sim \text{Weibull}(\nu, \lambda_i) \]  
  # Survival time of patient \( i \)
  \[ \log(\lambda_i) = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \ldots \]  
  # hazard of patient \( i \)
  # \( x \)'s are the elements of the design matrix constructed
  # with a truncated spline basis generating function

- \( \beta_3, \beta_4, \beta_5, \ldots \) are changes in the slope of the of the spline function at the knot locations
- Can ensure monotonicity with a reparametrization
  \[ \beta_1 = \eta_1; \beta_2 = \eta_2 \]
  \[ \beta_3 = \eta_3 - \eta_2 \]
  \[ \beta_4 = \eta_4 - \eta_3 \]
  \[ \beta_5 = \eta_5 - \eta_4 \]

- Right kind of priors on the \( \eta \)'s ensures that the function above is monotonically increasing
  \[ \eta_3, \eta_4, \eta_5, \eta_6, \ldots \sim \log-N(0, \sqrt{1/ \tau_\beta}) \]  
  # these are the terms that are penalized

- Priors to complete a Bayesian Model
  \[ \eta_1 \sim N(0, \sqrt{1/ \tau_\theta}); \eta_2 \sim \log-N(0, \sqrt{1/0.25}) \]
  \[ \tau_\beta \sim \text{Ga}(0.001, 0.001); \nu \sim \text{Ga}(a, b) \]

*P-spline method for cutoff determination was developed by Adarsh Joshi*
Modeling Strategy

- **Within each treatment arm**, do the following
  - Fit a monotonously increasing spline to model log-hazard
  - Fit a monotonously decreasing spline to model log-hazard
  - Fit a constant function to model log-hazard
  - Choose one of the above based on a model selection criterion (DIC in WinBUGS)

- Treatment differential is the difference of the two log-hazard functions
  - Once we get posterior samples of log-hazard function in individual treatment arms, the posterior samples of the differential are also available

- Cutpoint is the point where the hazard ratio crosses a pre-determined clinical utility threshold (0.7 in my simulations)
  - Subpopulation is declared to exist if such a cutpoint exists
A simulated dataset
Constrained Generalized Additive Model (cGAM)*

Assume that the log-odds of success is linear in the predictors:

\[ g(E(Y)) = \beta_0 + \beta_1 T + \beta_2 BM + \beta_3 T \times BM \]

where, \( g() \) is link function, \( \beta = (\beta_0, \beta_1, \beta_2, \beta_3) \) are model coefficients.

Re-parameterization:
Let \( X_1 = BM \times T \), and \( X_2 = BM \times (1 - T) \), then let

\[ g(E(Y)) = \eta_0 + \eta_1 T + \eta_2 X_1 + \eta_3 X_2 \]

- \( \eta_1 = \beta_1 \): treatment effect
- \( \eta_2 = \beta_2 + \beta_3 \): prognostic effect and predictive effect (treatment group)
- \( \eta_3 = \beta_2 \): prognostic effect (placebo group)

Generalized additive model does not assume any parametric format of the relationship between log-odds of success and predictors:

\[ g(E(Y)) = \eta_0 + \eta_1 T + f_1(X_1) + f_2(X_2) \]

where, \( f1 \) and \( f2 \) are nonparametric smooth functions, estimated by I-Spine (ie, integration of M-spline).

*cGAM method for cutoff determination was developed by Huan Wang
Application on Biomarker Cutoff Point Selection

Baseline Biomarker Value

Baseline Biomarker Value

Baseline Biomarker Value

Baseline Biomarker Value

Baseline Biomarker Value

Baseline Biomarker Value

Baseline Biomarker Value

Baseline Biomarker Value

Baseline Biomarker Value
All above methods require data from RCTs, what if there is no RCT Phase II data?

- Explore and Confirm approach
Cutoff Determination at Validation Stage
A Typical Dilemma

When design a pivotal Phase III study for a target therapy:

- “Believer”:
  - “it is a target therapy, patients in the study have diverse baseline biomarker values, so, it is reasonable to believe that patients with biomarker high values will have higher inhibition of the pathway and therefore benefit more from the drug, can we have it as a co-primary endpoint or at least a pre-specified secondary test on the biomarker positive population?”

- “Criticizer”:
  - “But we don’t have enough data evidence supporting this biomarker is of any predictive value; also, without enough prior data, how do we determine cutoff value to define a biomarker positive subgroup?”

- Decision-maker:
  - “Both parties have their points, what do we do now?”
Sequential Testing Method (Holmgren)

Similar idea as group sequential design
- Group sequential design: to reconstruct a group sequential analysis after a study has been completed one would order subjects by the calendar time they entered the study.
- Sequential testing of biomarker subgroups: Instead of ordering subjects by the time they enter the study, we can order them by their marker expression
- Test the following 4 hypotheses sequentially

<table>
<thead>
<tr>
<th>Marker Expression</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;75’th Percentile</td>
<td>$H_1: \lambda_1 = 1$</td>
</tr>
<tr>
<td>&gt;50’th Percentile</td>
<td>$H_2: \lambda_1 = 1$ and $\lambda_2 = 1$</td>
</tr>
<tr>
<td>&gt;25’th Percentile</td>
<td>$H_3: \lambda_1 = 1$ and $\lambda_2 = 1$ and $\lambda_3 = 1$</td>
</tr>
<tr>
<td>All Subjects</td>
<td>$H_4: \lambda_1 = 1$ and $\lambda_2 = 1$ and $\lambda_3 = 1$ and $\lambda_4 = 1$</td>
</tr>
</tbody>
</table>
An Example

Sequential testing procedure can also be easily modified

- Testing strategy, 3 tests at 1 interim and 1 final analysis:
  - 1st test: ITT
  - 2nd test: BM >= C1
  - 3rd test: BM >= C2
  - Alpha spending plan:

<table>
<thead>
<tr>
<th></th>
<th>Interim</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st test: ITT</td>
<td>a1</td>
<td>b1</td>
</tr>
<tr>
<td>2nd test: BM &gt;= C1</td>
<td>a2</td>
<td>b2</td>
</tr>
<tr>
<td>3rd test: BM &gt;= C2</td>
<td>a3</td>
<td>b2</td>
</tr>
</tbody>
</table>

- How to determine C1 and C2?
  - If there is Ph II RCT, use the methods (on slide 14-27), determine two best cutoff values (with the consideration of PoS)
  - If there is no Ph II RCT, nominate based on biologic hypothesis (ex, distribution of the biomarker in healthy subjects vs patients) or “explore cohort” in PhIII

- Alpha spending a1 – 3, b1 – b3, can be determined based on correlation matrix (Spiessens 2010, Contemporary Clinical Trials)
  - Simulation study can help determine the optimal allocation of alpha to maximize the PoS

- An analog: take two dose levels into Phase III (when you are not 100% sure which dose level is the right dose)
## Comparisons of 3 Approaches: Practical Considerations

<table>
<thead>
<tr>
<th>Method</th>
<th>Practical considerations</th>
</tr>
</thead>
</table>
| (Modified) Sequential testing | • Simplest to implement;  
• low regulatory risk;  
• No Phase III data involved in the determination of the cutoff values                             |
| Explore-n-Confirm          | • Simple to implement;  
• Low/Moderate regulatory risk;  
• Phase III data involved in the determination of the cutoff values                                    |
| Adaptive Designs           | • Complicated to implement;  
• Moderate/high regulatory risk;  
• Phase III data involved in the determination of the cutoff values                                   |
High-level Summary of Simulation Results (1)

- Sequential testing method provided either highest power or close to highest power.
- Nomination method:
  - provided high power when the nominated value was close to true threshold value;
  - the loss of power could be big, when nominated value was far from true value.
- Explore-n-confirm method did not outperform sequential testing method, but outperformed nomination method when nominated value was ‘way off’.
High-level Summary of Simulation Results (2)

True threshold = median; HR- = 1

- Sequential testing method provided either highest power or close to highest power unless the true threshold is at extreme value (90th percentile)
  - because the 2nd biomarker test was done at 75th percentile, which was still quite far away from the true value
  - As expected, if nomination is right on target (ie, 90th percentile), the nomination approach had the highest power
- This finding highlights the importance of a good “starting point” (ie, the 1st threshold value) of the sequential testing procedure
Cutoff Determination at Refinement Stage
Can we ‘refine’ the cutoff at the end of Phase III?

More of a regulatory question, remain to be answered
- However, even without biomarker involved, patient population in Phase III ≈ population in Sponsor’s proposed label ≈ population in FDA approved label
- If refining population in drug label is common, refinement of biomarker cutoff value doesn’t sound that crazy!

Table 4 | Factors involved in diagnostic threshold readjustment*

<table>
<thead>
<tr>
<th>Reason</th>
<th>Readjustment proposal</th>
<th>Conditions required for proposal</th>
<th>Effect as compared with using the original threshold per statistical plan</th>
</tr>
</thead>
<tbody>
<tr>
<td>New treatment options have emerged while the trial was underway that result in a change in how the disease is molecularly classified and treated; as a result, higher efficacy may be needed for approval</td>
<td>Readjust the threshold to ensure that the diagnostic-selected population has a clinically meaningful benefit based on currently available knowledge</td>
<td>Meeting the primary end point in the pre-specified diagnostic subset (using original threshold)</td>
<td>Higher estimated expected benefit in indicated diagnostic-positive patients; lower prevalence of the selected population</td>
</tr>
<tr>
<td>There is a statistically significant benefit in the patient population selected with the pre-defined threshold, but the estimate of the benefit is lower than was expected</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>There is an observed benefit in the patient population selected with the predefined threshold as well as in the subset of patients with diagnostic values that were not captured by the predefined threshold</td>
<td>Adjust the threshold to be more inclusive of biomarker values to include all patients with relevant clinical benefit</td>
<td>Meeting the primary end point in the pre-specified diagnostic subset (using original threshold), as well as in the diagnostic-negative patients or all patients, as pre-specified in the analysis plan</td>
<td>Lower estimated expected benefit in indicated diagnostic-positive patients; higher prevalence of the selected population</td>
</tr>
</tbody>
</table>

*Examples are given here of the reasons why the diagnostic threshold might need readjustment, the resulting proposal, required conditions for proposal implementation and the effect on the label.

Reference: Fridley et al 2013 Nature
More Challenges

All the discussions today are around single marker and single assay. What about cutoff determination for multiple biomarkers and/or multiple assays?

- FDA public meeting on “Complexities in Personalized Medicine: Harmonizing Companion Diagnostics”, March 24 2015


MARCO BONETTI, RICHARD D. GELBER; Patterns of treatment effects in subsets of patients in clinical trials; Biostatistics (2004), 5, 3, pp. 465–481

Howard M. Mackey, Thomas Bengtsson; Sample size and threshold estimation for clinical trials with predictive biomarkers; Contemporary Clinical Trials 36 (2013) 664–672