Design Concept for a Confirmatory Basket Trial

Robert A. Beckman, MD¹ and Cong Chen, PhD²

¹Professor of Oncology & of Biostatistics, Bioinformatics, and Biomathematics
Lombardi Comprehensive Cancer Center and Innovation Center for Biomedical Informatics
Georgetown University Medical Center
Founder and Chief Scientific Officer, Oncomind, LLC

²Executive Director, Biostatistics and Research Decision Sciences, Merck & Co, Inc.
Acknowledgements

• Co-authors on the initial design:
  – Cong Chen—led group; co-led concept development; led all statistical and simulation work
  – Zoran Antonijevic, Amgen
  – Rasika Kalamegham, Genentech

• Pathway design subgroup, additional members:
  – Christine Gausse, Merck
  – Sebastian Jobjornsson, Chalmers
  – Lingyun Liu, Cytel
  – Sammy Yuan, Merck
  – Yi (Joey) Zhou, Ultragenyx
  – Advisor: Sue-Jane Wang

• Pathway design subgroup is one of 5 working subgroups of the DIA Small Populations Workstream, a group of 50 statisticians and clinicians from industry, academia, and national health authorities (FDA and EMEA)

• Small populations workstream is part of DIA Adaptive Design Scientific Working Group (ADSWG), a group of > 200 statisticians and clinicians from industry, academia, and national health authorities (FDA and EMEA)
Small Populations Within A Common Disease

- The increasing discovery of molecular subtypes of cancer leads to small subgroups that actually correspond to orphan or “niche” indications, even within larger tumor types.
- Enrolling enough patients for confirmatory trials in these indications may be challenging.
- The shift to a molecular view of cancer requires a corresponding paradigm shift in drug development approaches.
- Exclusive use of “one indication at a time” approaches will not be sustainable.
Basket Trials

- Multiple tumor types with one drug and predictive biomarker
- Evaluation often based on pooled analysis
  - In some designs, pooling can be partial, based on Bayesian hierarchical model. Degree of pooling can be adjusted based on data
  - In some designs, indications are considered individually. Basket is then more of an operational tactic
- Premise is that molecular subtype is more fundamental than histology
- Can be single sponsor or consortium
- Opportunity for multiple indications for the sample size of one
Introduction

General Design Concept for a Confirmatory Basket Trial

Challenges and Recommendations for Overcoming Them

Performance Simulations and Design Considerations

Conclusions
The Original Basket: Imatinib B2225

186 patients with 40 different malignancies with known genomic mechanisms of activation of imatinib target kinases

Kit, PDGFRα or PDGFRβ

Imatinib 400-800 mg BID primary Endpoint ORR

Synovial Sarcoma
Aggressive Fibromatosis
Dermatofibrosarcoma Protuberans
Aggressive Systemic mastocytosis
Hyper-eosinophilic syndrome
Myelo-proliferative disorder

1/16 (6%) 2/20 (10%) 10/12 (83%) 1/5 (20%) 6/14 (43%) 4/7 (58%)

13 centers in consortium: North America, Europe, Australia

Lead to supplemental indications for these 4 subsets after pooling with other trials and case reports

Blumenthal. Innovative trial designs to accelerate the availability of highly effective anti-cancer therapies: an FDA perspective, AACR 2014
Features of These Designs

• A similar design to Imatinib B2225 was endorsed at a Brookings/Friends Conference in 2011

• Common features:
  – Exploratory and opportunistic in nature
  – Single-arm trials with ORR as primary endpoint
  – Intend to use pooled population for primary analysis to gain broader indication across tumor types (individual tumor type is not adequately powered)
  – Involve possibly transformative medicines in patients with great unmet need and seemingly exceptionally strong scientific rationale
Issues

• Clinical data to support pooling may be limited, and treatment effect may differ between tumor types
  – Vemurafenib works in melanoma with BRAF V600E mutation but not colorectal cancer with same mutation
• Not all drugs hoped to be transformational live up to this promise
• Response rate may not predict overall survival
• Single arm trials are subject to patient selection bias
• Predictive effect of a biomarker is confounded with the prognostic value which is often unknown
• Health authorities can be non-committal upfront
DIA Small Population Pathway Subteam

• Can we develop a generalizable confirmatory basket design concept with statistical rigor?
  – Applicable not only to exceptional cases, but to all effective medicines in any line of therapy
  – Follow existing accelerated and standard approval pathways to increase potential approvability

• This would have multiple benefits
  – Increase and accelerate access to effective medicines for patients in niche indications
  – Provide sponsors with cost-effective options for development in niche indications
  – Provide health authorities with more robust packages for evaluation of benefit and risk

  – MOST OF DRUG DEVELOPMENT RESOURCES ARE SPENT IN THE CONFIRMATORY PHASE
GENERAL DESIGN CONCEPT
Consistent trend in definitive endpoint

PRUNING (External Data)

PRUNING (Interim endpoints)

FULL APPROVAL (Pooled analysis of definitive endpoint)

September 14, 2018
Tumor histologies are grouped together, each with their own control group (shared control group if common SOC)

Randomized control is preferred
  - Single arm cohorts with registry controls may be permitted in exceptional circumstances as illustrated by imatinib B225 and others

In an example of particular interest, each indication cohort is sized for accelerated approval based on a surrogate endpoint such as progression free survival (PFS)
  - This may typically be 25-30% of the size of a Phase 3 study

In another approach, an interim evaluation of partial information on the definitive endpoint may be used

Initial indications are carefully selected as one bad indication can spoil the entire pooled result
Features of the Design (II)

- Indications are further “pruned” if unlikely to succeed, based on:
  - External data (maturing definitive endpoint from Phase 2; other data from class)
  - Internal data on surrogate endpoint OR partial information on definitive endpoint

- Sample size of remaining indications may be adjusted based on pruning

- Type I error threshold will be adjusted to control type I error (false positive rate) in the face of pruning
  - Pruning based on external data does not incur a statistical penalty
  - Discussed in more detail later in talk

- Study is positive if pooled analysis of remaining indications is positive for the primary definitive endpoint
  - Remaining indications are eligible for full approval in the event of a positive study
  - Full pooling chosen for simplicity
  - Some of the remaining indications may not be approved if they do not show a trend for positive risk benefit as judged by definitive endpoint
Another Possible Source of External Data

- Real World Data (RWD) from Off-Label Use
- Impact of RWD on basket trial performance is currently under study in a project led by postdoctoral fellow Daphne Guinn
CHALLENGES OF BASKET DESIGNS AND RECOMMENDATIONS FOR OVERCOMING THEM
Challenge 1: Risks of Pooling

- One of more bad indications can lead to a failed study for all indications in a basket
- Histology can affect the validity of a molecular predictive hypothesis, in ways which cannot always be predicted in advance
  - Vemurafenib is effective for BRAF 600E mutant melanoma, but not for analogous colorectal cancer (CRC) tumors
  - This was not predicted in advance but subsequently feedback loops leading to resistance were characterized
Basket trials are recommended primarily after there has been a lead indication approved (by optimized conventional methods) which has validated the drug, the predictive biomarker hypothesis, and the companion diagnostic.

- Example, melanoma was lead indication preceding Brookings trial proposal in V600E mutant tumors.

Indications should be carefully selected.

Indications should be pruned in several steps before pooling.
Challenge 2: Clinical validity of the predictive biomarker hypothesis

The clinical validity of the predictive biomarker can only be verified by inclusion of “biomarker negative” patients in the confirmatory study.

Addressing the challenge

- Recommend a smaller pooled, stratified cohort for biomarker negative patients, powered on surrogate endpoint
  - Would need to expand the biomarker negative cohort (to evaluate definitive endpoint) if surrogate endpoint shows possible benefit
- Prior evidence should permit this if:
  - An approved lead indication has already provided clinical evidence for the predictive biomarker hypothesis
  - Prior phase 2 studies support the predictive biomarker hypothesis in other indications
Pruning indications that are doing poorly on surrogate endpoints may be seen as cherry picking
– This can inflate the false positive rate, an effect termed “random high bias”

Addressing the challenge:
– Emphasize use of external data, especially maturing Phase 2 studies, for pruning
  • Pruning with external data does not incur a penalty for random high bias
– Apply statistical penalty for control of type I error when applying pruning using internal data
  • Methods for calculating the penalty are described in stat methods papers (see key references)
  • Rules for applying penalty must be prospective
  • Penalty is not large enough to offset advantages of design
Type I error control under global null hypothesis

- k tumor indications each with sample size of $N$ and all with 1:1 randomization
- An interim analysis is conducted at information fraction $t$ for each tumor indication and a tumor will not be included in the pooled analysis if p-value $> \alpha_t$
- The pooled analysis will be conducted at $\alpha^*$ so that the overall Type I error is controlled at $\alpha$ when there is no treatment effect for any tumor (H0)
- What is $\alpha^*$?
Solving for adjusted alpha ($\alpha^*$)

- Let $Y_{i1}$ be the test statistics based on information fraction $t$, and $Y_{i2}$ be the test statistics based on the final analysis of data in the $i$-th cohort ($i=1, 2, ..., k$).

- Suppose that $m$ cohorts are included in the final analysis ($m \geq 1$), and let $V_m$ be the corresponding test statistics. The probability of a positive outcome in pooled analysis is

$$Q_0(\alpha^*|\alpha_t, m) = \Pr_{H_0}(\cap \{Y_{i1} > Z_{1-\alpha_i} \text{ for } i=1, ..., m\}, \cap \{Y_{j1} < Z_{1-\alpha_j} \text{ for } j=m+1, ..., k\}, V_m > Z_{1-\alpha^*})$$

or

$$Q_0(\alpha^*|\alpha_t, m) = \Pr_{H_0}(\cap \{Y_{i1} > Z_{1-\alpha_i} \text{ for } i=1, ..., m\}, V_m > Z_{1-\alpha^*})(1-\alpha_t)^{(k-m)}$$

- $\alpha^*$ is solved from below where $c(k, m) = k!/(k-m)!m!$)

$$\sum_{m=1}^{k} c(k, m)Q_0(\alpha^*|\alpha_t, m) = \alpha$$
Challenge #4: Strong Control of FWER

- This problem is still open

- Challenge:
  - One or more strongly positive indications can drive an overall pooled positive result and negative indications are carried along
  - Simulation involves a large number of cases and the degree to which active indications are active affects the results

- A recent MSKCC study* simulated a popular Bayesian basket trial design and found FWER of up to 57%.
  - Authors advocate characterization of FWER by simulation

Should Basket Trials Control FWER by Indication?

**Conventional**

- Inconsistency in treatment effect (e.g., gender, age) is not suspected.
- Primary analysis in pooled population, and, if positive, consistency will be assessed ad hoc.

**Basket (two-stage)**

- Consistency in treatment effect across tumor indication is less certain.
- Inactive ones are pruned at an interim analysis.
- Primary analysis in pooled population of remaining ones, and, if positive, consistency will be assessed ad hoc.
Other FWER Considerations

• A basket trial with k indications replaces k independent trials that collectively would have a family-wise error rate of approximately \( k \times 0.025 \)

• Should we therefore allow approximately \( k \times 0.025 \) for FWER of a basket trial?

• Under would conditions would FDR be a better measure than FWER?
PERFORMANCE SIMULATIONS AND DESIGN CONSIDERATIONS
Comparison of operating characteristics

• k=6 tumor indications with total planned event size \((kN)\) ranging from 150-350
  – The true treatment effect is \(-\log(0.6)\), or hazard ratio of 0.6 in a time-to-event trial

• Pruning occurs at when half of the events have occurred

• Number of active indications \((g)\) with target effect size ranges from 3 to 6, with remaining ones inactive
Study power and sample sizes under different pruning and pooling strategies

<table>
<thead>
<tr>
<th>Planned events</th>
<th>Number of active tumors</th>
<th>Power (%) for a positive study</th>
<th>Exp. number of events for pooled population</th>
<th>Exp. number of events for overall study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>D0</td>
<td>D1</td>
<td>D2</td>
</tr>
<tr>
<td>200</td>
<td>6</td>
<td>95</td>
<td>85</td>
<td><strong>95</strong></td>
</tr>
<tr>
<td>200</td>
<td>5</td>
<td>85</td>
<td>75</td>
<td><strong>91</strong></td>
</tr>
<tr>
<td>200</td>
<td>4</td>
<td>67</td>
<td>62</td>
<td><strong>82</strong></td>
</tr>
<tr>
<td>200</td>
<td>3</td>
<td>44</td>
<td>45</td>
<td><strong>68</strong></td>
</tr>
<tr>
<td>300</td>
<td>6</td>
<td>99</td>
<td>96</td>
<td><strong>99</strong></td>
</tr>
<tr>
<td>300</td>
<td>5</td>
<td>96</td>
<td>81</td>
<td><strong>98</strong></td>
</tr>
<tr>
<td>300</td>
<td>4</td>
<td>84</td>
<td>81</td>
<td><strong>94</strong></td>
</tr>
<tr>
<td>300</td>
<td>3</td>
<td>60</td>
<td>64</td>
<td><strong>84</strong></td>
</tr>
</tbody>
</table>
An Application of Special Interest

• A randomized controlled basket trial with 1:1 randomization in 6 tumor indications, each targeting a hazard ratio of 0.5 in PFS with 90% power at 2.5% alpha for global null hypothesis
  – 88 PFS events and 110 patients planned for each indication
  – PFS analysis is conducted when all are enrolled
• D2 is applied to keep total sample size at 660 in pooled population targeting 430 death events
  – The study has ~90% power to detect a hazard ratio of 0.7 in OS at 0.8% alpha (after taking the penalty) assuming $\rho=0.5$
  – Observed hazard ratio ~0.79 or lower for a positive trial in pooled population (vs ~0.84 under D0) for alpha control under global null
• Potential to gain approvals in 6 indications based on comparable sample size to a conventional Phase 3 trial
Characterization of Performance Constrained by FWER (ongoing)

- Team includes Yuru Ren, Valeriy Korostyshevskiy, and Sammy Yuan
- Currently studying single TTE endpoint with normally distributed hazard ratios, mean of 1.0 for inactive, 0.7 for active
- Simulate different scenarios of how many indications in basket are inactive. Maximum Type I error (worst case scenario) is FWER
- What power is achievable when FWER must be $\leq k \times 0.025$?
Current Approaches

• In order to control FWER, we must add an additional post-correction step

• Each indication is tested up to twice individually*
  – at interim information time $t \in [0,1]$ at significance level $\alpha-t$, AND
  – if part of a successful pool, in a post check at significance level $\alpha$-post

Preliminary Results

• $k = 6$; $HR = 0.7$, nominal power of pool = 95%; $t = 0.5$, $\alpha_t = 0.4$, $\alpha_{post} = 0.1$:

<table>
<thead>
<tr>
<th>$k$</th>
<th>$0.5276$</th>
<th>$0.9467$</th>
<th>$0.0000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.5262</td>
<td>0.8954</td>
<td>0.0631</td>
</tr>
<tr>
<td>5</td>
<td>0.5079</td>
<td>0.8032</td>
<td>0.1111</td>
</tr>
<tr>
<td>4</td>
<td>0.4654</td>
<td>0.6675</td>
<td>0.1432</td>
</tr>
<tr>
<td>3</td>
<td>0.3780</td>
<td>0.4739</td>
<td>0.1451</td>
</tr>
<tr>
<td>2</td>
<td>0.2316</td>
<td>0.2402</td>
<td>0.0986</td>
</tr>
</tbody>
</table>

• $k = 3$; $HR = 0.7$, nominal power of pool = 95%; $t = 0.5$, $\alpha_t = 0.3$, $\alpha_{post} = 0.1$:

<table>
<thead>
<tr>
<th>$k$</th>
<th>$0.6940$</th>
<th>$0.9534$</th>
<th>$0.0000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.6815</td>
<td>0.8435</td>
<td>0.0493</td>
</tr>
<tr>
<td>2</td>
<td>0.5813</td>
<td>0.5871</td>
<td>0.0731</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

September 14, 2018
Future Plans

• Further parameter optimization
• Application of heterogeneity detection methods (Simon)
• Study of application with using surrogate interim endpoint
• Application of RWD to study design
Conclusions

- It is feasible to create a general design concept for a basket study that is suitable for many agents.
- Multiple challenges can be addressed with careful planning.
- Benefits include:
  - Increased and earlier patient access to targeted therapies for small subgroups.
  - Cost-effective methods for sponsors to develop targeted agents in small subgroups.
  - More robust datasets for health authorities to assess benefit-risk in these small patient groups.
Key References


September 14, 2018