Simple benchmark for planning and evaluating complex dose finding designs

Ken Cheung
Columbia Biostatistics

Cheung (2014). *Biometrics* 70, 389—397
Agenda

• Dose Finding Trials
  – General background
  – Example: A phase 1/2, Eff-Tox design

• Dose Finding benchmark
  – Applications: Design diagnostic
  – (method comparison; sample size calculation)
  – Discussion
Dose Finding Trials

• Phase I and phase I/II
• Not parallel randomized
• Small-group-sequential: Adapt after every small cohort (e.g. 3)
• General design and analysis strategy
  – Observe a few
  – Estimate a “good” dose (model-based, myopic or not)
  – Treat at the good dose, and observe
Dose Finding Trials

Challenge in planning: Complexity

- Assume programming correct without theoretical guidance
- Pathological properties may not be detected by simulation
- Difficult to reproduce by another statistician, and review the plausibility of the simulation results
Some generality and notation

• A pre-specified set of test levels \{1, \ldots, K\}

• Multinomial outcome \(Y\):
  - \(Y_i(k)\) = Outcome for patient \(i\) at dose level \(k\)
  - Take values on \(L+1\) possible values \(\{w_0, w_1, \ldots, w_L\}\)
  - Tail distribution \(\pi_l(k) = Pr\{Y(k) \geq w_l\}\) for \(l = 1, \ldots, L\)

• Objective: Estimate the target dose \(d(\pi)\) in \(\{1, \ldots, K\}\)

• Example 1: Phase I trial with binary toxicity \(Y = 0, 1\)
  - \(\pi_l(k)\) denotes toxicity probability at dose \(k\)
  - \(d(\pi) = \arg \min_k |\pi_l(k) - p|\) for some target \(p\).
Example 2: Thrombolytic agent for acute stroke

- Phase 1/2 study
- Trinary outcome (Efficacy-toxicity):
  - Intracranial hemorrhage (Toxicity; $Y=2$)
  - Reperfusion without hemorrhage (Response; $Y=1$)
  - Neither ($Y=0$)
- Thall and Cook (2004):
  - Define desirability $\delta(\pi_E, \pi_T)$ as a function of response rate $\pi_E$ and toxicity rate $\pi_T$
  - Aim to find a dose that maximizes $\delta(\pi_E, \pi_T)$
  - $d_{TC}(\pi) = \arg \max_k \delta_k$
Example 2: Thrombolytic agent for acute stroke

$K = 5 \text{ levels}$

Thall and Cook (2004)
Example 2: Thrombolytic agent for acute stroke

Thall and Cook (2004):
- Outcome-adaptive
- Bayesian, model-based dose finding method
  - Assign patients at dose with maximum desirability based on interim data, subject to acceptability criteria
  - Consider two dose-response-toxicity models: Proportional odds (PO) and Continuation ratio (CR)

Use simulation at planning: compare models
Simulation results: Which model to use?

**Scenario 3**

<table>
<thead>
<tr>
<th>Model</th>
<th>Dose 1</th>
<th>Dose 2</th>
<th>Dose 3</th>
<th>Dose 4</th>
<th>Dose 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desirability</td>
<td>-0.48</td>
<td>-0.13</td>
<td>0.22</td>
<td><strong>0.32</strong></td>
<td>-0.26</td>
</tr>
<tr>
<td>PO</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>72</td>
<td>7</td>
</tr>
<tr>
<td>CR</td>
<td>0</td>
<td>2</td>
<td>32</td>
<td><strong>49</strong></td>
<td>16</td>
</tr>
</tbody>
</table>

**Scenario 4**

<table>
<thead>
<tr>
<th>Model</th>
<th>Dose 1</th>
<th>Dose 2</th>
<th>Dose 3</th>
<th>Dose 4</th>
<th>Dose 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desirability</td>
<td>0.12</td>
<td>0.29</td>
<td>0.45</td>
<td>0.58</td>
<td><strong>0.69</strong></td>
</tr>
<tr>
<td>PO</td>
<td>0</td>
<td>2</td>
<td>10</td>
<td>34</td>
<td><strong>54</strong></td>
</tr>
<tr>
<td>CR</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td><strong>94</strong></td>
</tr>
</tbody>
</table>
Which model to use?

• Motivation:
  – Numerical performance from simulation can be difficult to interpret without a benchmark

• Proposal:
  – Dose Finding Benchmark Design
Dose Finding Benchmark

• Goal: A theoretical dose finding design that provides an upper limit of accuracy for any dose finding methods for a given design objective under a given scenario.

• Definition:
  – Recall $d(\pi)$ is the target dose (estimand)
  – Benchmark design: $d(\pi^*)$ where $\pi^*$ is a nonparametric optimal estimate of $\pi$ based on complete outcome profile
Complete outcome profile: Example 1

• In an actual trial, we observe a partial outcome profile, e.g., a patient at dose 3 with toxicity

<table>
<thead>
<tr>
<th>Dose 1</th>
<th>Dose 2</th>
<th>Dose 3</th>
<th>Dose 4</th>
<th>Dose 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>?</td>
<td>Toxicity</td>
<td>Toxicity</td>
<td>Toxicity</td>
</tr>
</tbody>
</table>

• In computer simulation, we can observe a complete profile by generating a uniform tolerance

<table>
<thead>
<tr>
<th>Dose 1</th>
<th>Dose 2</th>
<th>Dose 3</th>
<th>Dose 4</th>
<th>Dose 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>No toxicity</td>
<td>Toxicity</td>
<td>Toxicity</td>
<td>Toxicity</td>
<td>Toxicity</td>
</tr>
</tbody>
</table>

• The nonparametric optimal estimate $\pi^*(k)$ is evaluated by the proportion of toxicity at dose $k$ in a simulated trial
Complete outcome profile: General (inc. Example 2)

- **Ordinal outcome** $Y$: Takes values on $L+1$ possible values \{w_0, w_1, ..., w_L\} with tail distribution $\pi(k)$ at dose $k$
- $Y_i(k) =$ Outcome for patient $i$ at dose level $k$
- In simulation, randomly draw a tolerance profile: $U_{i1}, U_{i2}, \ldots, U_{iL}$ iid Uniform(0,1)
- Generate complete outcome profile $Y_i(k)$ for patient $i$ at dose level $k$ as follows:
  - $Y_i(k) = w_l$ if $U_{i,l+1} > r_{l+1}(k)$ and $U_{ij} \leq r_j$ for all $j=1,\ldots,l$
  - $r_j(k) = \pi_j(k) / \pi_{j-1}(k)$
- Nonparametric optimal $\pi^*(k) =$ average of $I\{Y_i(k) \geq w_l\}$
Thall and Cook (2004), revisit

Scenario 3

<table>
<thead>
<tr>
<th>Model</th>
<th>Dose 1</th>
<th>Dose 2</th>
<th>Dose 3</th>
<th>Dose 4</th>
<th>Dose 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desirability</td>
<td>-0.48</td>
<td>-0.13</td>
<td>0.22</td>
<td>0.32</td>
<td>-0.26</td>
</tr>
<tr>
<td>PO✔</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>72</td>
<td>7</td>
</tr>
<tr>
<td>CR</td>
<td>0</td>
<td>2</td>
<td>32</td>
<td>49</td>
<td>16</td>
</tr>
<tr>
<td>d(\pi^*)</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>85</td>
<td>1</td>
</tr>
</tbody>
</table>

Scenario 4

<table>
<thead>
<tr>
<th>Model</th>
<th>Dose 1</th>
<th>Dose 2</th>
<th>Dose 3</th>
<th>Dose 4</th>
<th>Dose 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desirability</td>
<td>0.12</td>
<td>0.29</td>
<td>0.45</td>
<td>0.58</td>
<td>0.69</td>
</tr>
<tr>
<td>PO</td>
<td>0</td>
<td>2</td>
<td>10</td>
<td>34</td>
<td>54</td>
</tr>
<tr>
<td>CR✔</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>94</td>
</tr>
<tr>
<td>d(\pi^*)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>95</td>
</tr>
</tbody>
</table>
Thall and Cook (2004), revisit

Benchmark as “effect size”

- Benchmark $d(\pi^*)$ performs better in S4 than in S3 suggesting S4 is an “easier” scenario than S3; analogous to large effect size in hypothesis test
- Eff-tox using proportional odds model is idiosyncratic in that it does comparatively poorly in an easy scenario (S4).
- Continuation ratio model wins in this example
Benchmark for Method Comparison


Least squares fit (logit scale)

Prop Odds
Cont. Ratio

Multiple toxicity

Least squares fit (logit scale)

Method A
Method B

Benchmark accuracy (logit scale)
Benchmark for “Power” Calculation

Cheung (2013): Sample size formulae for CRM
Summary & Discussion

• The proposed benchmark is applicable to general early phase dose finding settings:
  – Discrete test levels, including combination therapy
  – Multinomial outcome (multiple tox; bivariate; etc)

• Applications: effect size; method comparison; power calculation

• Features of a good benchmark:
  – Easy and quick to compute (not error prone)
  – Nonparametric: not favoring one model over another
  – Upper bound of accuracy for parametric methods