Benefit:Risk and Pragmatism:
Use Outcomes to Analyze Patients Rather than Patients to Analyze Outcomes

Scott Evans, PhD, MS, Harvard University

Harvard Catalyst
April 4, 2016

Significant Contributors (p<0.001)

• Chip Chambers, UCSF
• Dean Follmann, NIAID/NIH
• Dan Rubin, CDER/FDA
Outline

- Motivation
  - Pragmatism and Benefit:Risk Analyses
  - Challenges in Clinical Trials
- DOOR
- RADAR
- Partial Credit

The medical community is calling for more:

1. Systematic evaluation of benefits and harms
2. Pragmatism
"Most clinical trials fail to provide the evidence needed to inform medical decision-making. However, the serious implications of this deficit are largely absent from public discourse"

DeMets and Califf, *JAMA*, 2011

With billions of dollars spent each year to test new drugs …, you would think clinical trials would help doctors treat the patient in front of them.

“But you would be wrong.”


---

**On Resistance to Pragmatic Trials**

Scientists might be concerned about a focus on pragmatism and too little on basic biology.

“I intend at FDA to fight that battle.”

Robert Califf, FDA Commissioner in *Science*, 2015
Let's take a simple test to see how pragmatic we are.

Question 1

- Suppose the person that you care about the most, has just been diagnosed with a terrible disease

- 3 treatment options: A, B, and C

- 2 outcomes with similar importance
  - Efficacy (or "benefit"): binary
  - Toxicity: binary
### RCT Comparing A, B, and C

**Analysis of Endpoints**

<table>
<thead>
<tr>
<th></th>
<th>A (N=100)</th>
<th>B (N=100)</th>
<th>C (N=100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benefit</td>
<td>Benefit: 50%</td>
<td>Benefit: 50%</td>
<td>Benefit: 50%</td>
</tr>
</tbody>
</table>
### RCT Comparing A, B, and C

**Analysis of Endpoints**

<table>
<thead>
<tr>
<th></th>
<th>A (N=100)</th>
<th>B (N=100)</th>
<th>C (N=100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benefit</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Toxicity</td>
<td>20%</td>
<td>50%</td>
<td>50%</td>
</tr>
</tbody>
</table>

Which treatment would you choose?
# RCT Comparing A, B, and C

## Analysis of Endpoints

<table>
<thead>
<tr>
<th></th>
<th>A (N=100)</th>
<th>B (N=100)</th>
<th>C (N=100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benefit</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Toxicity</td>
<td>20%</td>
<td>50%</td>
<td>50%</td>
</tr>
</tbody>
</table>

Which treatment would you choose?

Of course any reasonable researcher would tell you is **C**
### Analysis of Patients: 4 Possible Outcomes

<table>
<thead>
<tr>
<th></th>
<th>Benefit: 50%</th>
<th>Toxicity: 20%</th>
<th>Benefit: 50%</th>
<th>Toxicity: 50%</th>
<th>Benefit: 50%</th>
<th>Toxicity: 50%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A (N=100)</strong></td>
<td>+ 10 &lt;br&gt; - 40</td>
<td>+ 10 &lt;br&gt; - 40</td>
<td>+ 50 &lt;br&gt; - 0</td>
<td>+ 0 &lt;br&gt; - 50</td>
<td>+ 0 &lt;br&gt; - 50</td>
<td>+ 50 &lt;br&gt; - 0</td>
</tr>
<tr>
<td><strong>B (N=100)</strong></td>
<td>+ 10 &lt;br&gt; - 40</td>
<td>+ 10 &lt;br&gt; - 40</td>
<td>+ 50 &lt;br&gt; - 0</td>
<td>+ 0 &lt;br&gt; - 50</td>
<td>+ 0 &lt;br&gt; - 50</td>
<td>+ 50 &lt;br&gt; - 0</td>
</tr>
<tr>
<td><strong>C (N=100)</strong></td>
<td>+ 10 &lt;br&gt; - 40</td>
<td>+ 10 &lt;br&gt; - 40</td>
<td>+ 50 &lt;br&gt; - 0</td>
<td>+ 0 &lt;br&gt; - 50</td>
<td>+ 0 &lt;br&gt; - 50</td>
<td>+ 50 &lt;br&gt; - 0</td>
</tr>
</tbody>
</table>

**Rate of saving your loved one**<br>(benefit without toxicity)<br><br>40% 0% 50%
Our culture is to use patients to analyze the endpoints.

Shouldn’t we use endpoints to analyze the patients?
Scott's father (a math teacher) to his confused son many years ago:

“The order of operations is important”

**Question 2**

- During analyses of a clinical trial, we define analysis populations
- Efficacy analysis: efficacy population (e.g., mITT)
- Safety analysis: safety population (e.g., those > 1 dose)
- Efficacy population ≠ safety population
- We then combine these analyses into a benefit:risk analysis
- To whom does this benefit:risk analysis apply?
### Vision for the Future of Clinical Trials

<table>
<thead>
<tr>
<th>Today</th>
<th>Tomorrow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment Effects</td>
<td></td>
</tr>
<tr>
<td>Endpoints</td>
<td></td>
</tr>
</tbody>
</table>

### Vision for the Future of Clinical Trials

<table>
<thead>
<tr>
<th>Today</th>
<th>Tomorrow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Few (usually 1)</td>
<td>Many (stratified medicine)</td>
</tr>
<tr>
<td>Treatment Effects</td>
<td></td>
</tr>
<tr>
<td>Endpoints</td>
<td></td>
</tr>
</tbody>
</table>
STEPP

- Subpopulation Treatment Effect Pattern Plot
  - Exploratory tool to detect interactions not apparent via modeling
  - A “moving average” approach to examining the pattern of treatment differences across a covariate

Clinical Trial: IBCSG IX

N=1715, Postmenopausal, Node Negative Women

Stratify ER Status RT Planned? Institution TAM to 5 years CMF x 3 TAM to 5 years
### Disease-Free Survival

#### ER-  ER+

<table>
<thead>
<tr>
<th></th>
<th>Percent</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMF-Tam</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Tam</td>
<td>80</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>6</td>
</tr>
</tbody>
</table>

p = 0.003

---

### STEPP: 5-Year DFS by ER

- **TAM ALONE**
- **CMF + TAM**

5 yr DFS

Subpopulations by ER (log scale)

---

All Rights Reserved, Duke Medicine 2007
STEPP

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).

Figure: Kaplan-Meier STEPP (sliding window)

Vision for the Future of Clinical Trials

<table>
<thead>
<tr>
<th>Today</th>
<th>Tomorrow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Few</td>
<td>Many</td>
</tr>
<tr>
<td>(usually 1)</td>
<td>(stratified medicine)</td>
</tr>
<tr>
<td>Treatment Effects</td>
<td>Endpoints</td>
</tr>
<tr>
<td>Many</td>
<td>Few</td>
</tr>
<tr>
<td>(efficacy, toxicity, QOL)</td>
<td>(overall patient outcome)</td>
</tr>
</tbody>
</table>
Good News

- Analyzing patients rather than endpoints has several advantages

- These include:
  - More informative and pragmatic benefit:risk evaluation (of patient outcomes)
  - Alleviation from competing risk challenges
  - More comprehensive picture of intervention effects
  - Perhaps additional efficiencies

Trials Evaluating Strategies of Antibiotic Use

- Considerable interest in comparing new strategies of antibiotic use vs. standard strategies with respect to clinical outcomes and antibiotic use

- Current approaches struggle to address many challenging issues in these trials
Competing Risks

- Common endpoints can be distorted / challenging to interpret
  - Days in the hospital
  - Days in the ICU
  - Days of antibiotic use

- Fewer days is better …or is it?

- The faster a patient dies, the fewer the days…

- Interpretation of these endpoints needs clinical context of other outcomes (e.g., survival) for the same patient

Noninferiority (NI) Complexities

- Validity relies upon several foundational requirements

- Constancy assumption threats
  - Development of antibiotic resistance
  - Improvements in supportive care

- Measures take to assure assay sensitivity often limit pragmatism
  - E.g., Strict entry criteria, e.g., in antibiotic studies we often exclude patients with prior therapy (despite it being common)
  - Analysis populations limited to per protocol (PP) or modified ITT (mITT) but patient status in these subgroups is unknown until after treatment initiation
Question the Question

“The formulation of a problem is far more essential than its solution ...”
Albert Einstein

- Many NI trials are conducted based on the premise that the new intervention has (non-efficacy) advantages (e.g., less toxic; better quality of life) over the standard of care...the rationale being if NI on a primary efficacy outcome is demonstrated then the new intervention may be preferred to the standard of care
- But in this scenario, the broader perspective reveals that the research question is not one of NI but one of superiority, i.e., whether the new intervention is (globally) better than the standard of care when all of the important outcomes are appropriately considered.
  - Stewardship is not a “me too” scenario
  - If a new antibiotic-use strategy is not better than the status quo (overall), then what is it’s value?
- Creative endpoint and analyses construction may allow for superiority designs

We cannot solve problems using the same thinking that we used to create them.
Albert Einstein
Desirability Of Outcome Ranking (DOOR)

- Vision
  - Rank the global patient outcomes based on a synthesis of benefits and harms
    - Often begins by constructing an ordinal clinical outcome
  - Compare DOOR distributions between intervention strategies
    - Estimate (using CIs) the probability that a randomly selected patient will have a better DOOR if assigned to a new strategy vs. a control
    - Evaluate the superiority of DOOR by testing if the probability is greater than something meaningful
      - >50% implies superiority
Ordinal Clinical Outcome

Before we analyze several hundred patients, we must understand how to analyze one.

- Classify each patient according to a *ordinal clinical outcome*
  - Synthesized benefits, harms, QOL using ordinal categories
  - Longitudinal snapshot (in contrast to TOC timepoint)
    - “Exit Examination” or “Discharge Review”

Ordinal Clinical Outcome

- The number and definition of levels is tailored
  - The top and bottom categories are often obvious
  - Layers between

Generic Example

- Benefit w/o toxicity
- Benefit w/ toxicity
- Survive, no benefit w/o toxicity
- Survive, no benefit w toxicity
- Death
Response-Adjusted for Duration of Antibiotic Risk (RADAR)

- DOOR is constructed using 2 rules:
  1. When comparing 2 patients with different clinical outcomes
     - The patient with the better clinical outcome receives a higher rank
  2. When comparing 2 patients with the same clinical outcome
     - The patient with a shorter duration of Ab use receives a higher rank

- Consistent with "reduce Ab use w/o clinical compromise"

Handling Adherence

- Evaluation of strategy is the most clinically relevant question

- Patients may change therapy because they fail but they do not fail because they change therapy
  - Adjustments to therapy are part of practice and the strategy

- RADAR is pragmatic, evaluating strategy
  - Patients are evaluated as part of their randomly assigned strategy consistent with the ITT principle
  - Observed Ab use is utilized to determine the DOOR
Example: ARLG SCOUT-CAP Trial
Short vs. Standard Course Outpatient Antibiotic Therapy of Community Acquired Pneumonia in Children

- Phase IV, DB, PC, RCT comparing DOOR among children ≥6 months to 5 years of age with CAP assigned to a strategy of short course (5 days) vs. standard course (10 days) outpatient beta-lactam therapy

- Original design
  - Debate over appropriate NI margin
  - Questionable feasibility w/ N=800 required for 90% power

- RADAR design
  - Superiority trial (avoiding NI)
  - N=360 (>50% reduction in the required N)

Ordinal Clinical Outcome

<table>
<thead>
<tr>
<th>Adequate clinical response* (Assessed at end of Rx &amp; 1 month)</th>
<th>Solicited events* (Assessed at end of Rx &amp; 1 month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Yes</td>
<td>None</td>
</tr>
<tr>
<td>2 Yes</td>
<td>Mild (Grade 1)</td>
</tr>
<tr>
<td>3 Yes</td>
<td>Moderate (Grade 2)</td>
</tr>
<tr>
<td>4 Yes</td>
<td>Severe (Grade 3)</td>
</tr>
<tr>
<td>5 No - no ED or outpatient clinic visit or hospitalization</td>
<td>None or any grade</td>
</tr>
<tr>
<td>6 No - with ED or outpatient clinic visit but no hospitalization</td>
<td>None or any grade</td>
</tr>
<tr>
<td>7 No - with hospitalization</td>
<td>None or any grade</td>
</tr>
<tr>
<td>8 Death from any cause</td>
<td>None or any grade</td>
</tr>
</tbody>
</table>
Challenges in Developing an Ordinal Outcome

- Choosing between multiple outcomes: clinical response, adverse events, antibiotic exposure, QoL, cost
- Identifying factors that distinguish outcomes: avoidance of clustering of patients in few strata
- Relies on judgement of expert clinicians. But what about other perspectives (subjects/families, clinicians, payers)?
- Involves value judgments in assigning ranks
  - Which is more important, clinical response or AEs?

DOOR Challenges

- Cultural change
- Construction of ordinal outcome is novel and challenging
- Careful deliberation is essential to synthesize the outcomes
- An example strategy …
The BAC DOOR

- ARLG is conducting a pre-trial sub-study to develop / validate a DOOR in *Staphylococcus aureus* bacteremia

- 20 representative patient profiles (including benefits, harms, and QoL) constructed based on experiences observed in prior trials

- Profiles sent to 43 expert clinicians. They were asked to rank the patient profiles by desirability of outcome.

- 42 clinicians responded

---

The BAC DOOR

- Clinician ranking analyses
  - Consensus among the ranks being examined
  - Variation of patient ranks examined
    - Evaluate reasons for small or large variation, e.g., LFU
  - Characteristics that guide the ranks evaluated
  - Fun with outlier clinicians

- Comparison of consensus clinician rank with team-constructed strategies
  - Correlations
  - Agreement on 190 pairwise comparisons

- Validation of team-constructed strategy for use in future trials
Boxplots of patient rankings

SD's of Patient Rankings

Heat-map for between-patient pairwise p-values
Weighting outcomes / scoring categories was avoided
  - Ranking equates to weighting indirectly

Concern that drop in (more important) clinical outcome would be trumped by improvement in (less important) Ab use
  - When sizing trials, consider sizing for most important component outcomes as well (Molina and Cisneros, 2015)

Potential solutions
  - Composite endpoint fundamentals
  - Sensitivity analyses
  - Partial credit
Sensitivity Analyses:
Require a larger difference in Ab use

- Current version of RADAR breaks ties based on 1-day difference

- If one day is not considered an important enough difference then consider breaking ties only if difference is larger, e.g., 3 or 5 days. Otherwise it is considered a tie.

- Plot the (probability of a higher DOOR for a randomly selected from A vs. B) vs. required difference (1, 2, 3, 4, 5 days). Use pointwise confidence bands.
Example: Colistin vs. New Drug


- Colistin
  - Used against Gram-negatives that cause life-threatening infections and are resistant to other antibiotics
  - Questionable efficacy
  - Causes nephrotoxicity and neurotoxicity

- A new drug could provide a superior alternative to colistin if reduced mortality, or has similar mortality but reduces clinically meaningful adverse effects

Example: Colistin vs. New Drug

- In an RCT comparing colistin to a new therapy, DOOR could be defined based on an ordinal clinical outcome:
  - Survives without a major adverse event (AE)
  - Survives with a major AE
  - Death

- Important to utilize major AEs of unquestioned importance to the patient (e.g., irreversible renal failure; need for hemodialysis)

- If the new drug reduces major AEs, then using DOOR may have greater power than a mortality trial to detect a benefit over colistin
Partial Credit

- A potential concern is that a major AE may be given too much (or to little) influence relative to survival

- The influence of a major AE is not directly assigned and is unknown until after the trial, as it depends on the resulting distribution of the ordinal outcome

- Consider assigning the influence directly

Partial Credit: Academic Test Scoring

- Patient survives without a major AE
  - Score = 100%

- Patient dies
  - Score = 0

- Patient survives with a major AE
  - Partial credit is given

- How much partial credit should be provided?
Engaging Patients is Increasingly Important

“I plan to give special emphasis to further developing the critical role that patients play in our work. When it comes to finding solutions for challenges facing the FDA, there is no greater resource than the one presented through engagement and outreach with patients. Including their perspectives and voices in our work along the entire medical product continuum, from development to review and evaluation to post-market surveillance, offers opportunities to enhance our knowledge of the benefits and risks of medicines.”

Rob Califf, FDA Commissioner
(February 26, 2016)

Patient Outreach to Calibrate Partial Credit

- A QoL instrument is given during or at the end of the trial to the survivors
  - Higher QoL scores indicate a better QoL
  - Death = 0

- Survivors without a major AE: mean QoL score = A

- Survivors with a major AE: mean QoL score = B
**Patient Outreach to Calibrate Partial Credit**

- Partial credit for survivors with a major AE = B/A
  - If the AE was as bad as death then the partial credit would be near 0
  - If the AE was irrelevant, then we would have A=B and the credit would be near 1

- One attractive analysis presentation
  - Display the estimated magnitude of effect and associated precision for a range (e.g., 0-100%) of partial credits

---

**Attractive Feature:** Allows for patients / clinicians to select therapy based upon their own preferences rather than have a partial credit selected for them
Analysis of Survival (binary)

Difference in mean scores

Confidence interval for difference in survival

0

Favors Colistin

Favors New Drug

Full Credit
(only survival matters; major AE is irrelevant)

0 0.25 0.5 0.75 1
Credit for Survival w/ Major AE

Analysis of Survival w/o Major AE (binary)

No Credit (only surviving w/o major AE is important)

0

Favors Colistin

Favors New Drug

0 0.25 0.5 0.75 1
Credit for Survival w/ Major AE
Other Features of the Curve

- The figure displays the difference in mean scores as a function of the partial credit (denoted a) for survival with a major AE. This difference can be described as:
  \[(P_1 - Q_1) + a(P_A - Q_A)\]
- where \(P_1\) and \(P_A\) are the rates of survival without AE and survival with a major AE rate in the new drug arm respectively, and \(Q_1\) and \(Q_A\) are analogous for the colistin arm.
- The curve is flat when \((P_A - Q_A) = 0\), i.e. if there is no difference in the rates of survival with a major AE.
- If \((P_A - Q_A)\) is relatively small compared to \((P_1 - Q_1)\), then the selection of partial credit (a) has little impact.

Partial Credit Derived from Patients

![Graph showing the relationship between partial credit and weighted survival](image-url)
**Sizing the Trial**

- Partial credit determined using patient QoL will not be available until after the trial is complete and such data are available.

- For sizing the trial, clinician surveys regarding their perspectives regarding of appropriate partial credit could be conducted before the trial.
Sample Size

- Suppose that a partial credit of >0.5 is expected since survival with a major AE may be viewed as being closer to survival without a major AE than death
- 80% power and alpha=0.05
- Desire to detect a shift from probabilities of 0.25, 0.25 and 0.50 for the 3 outcomes of death, survival with major AE, and survival without major AE for colistin to 0.10, 0.12, and 0.78 for the new drug respectively
- Sample size can be derived assuming use of a t-test

<table>
<thead>
<tr>
<th>Partial Credit</th>
<th>Difference in Means</th>
<th>SD (colistin)</th>
<th>SD (new drug)</th>
<th>Required N (Total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.202</td>
<td>0.409</td>
<td>0.312</td>
<td>104</td>
</tr>
<tr>
<td>0.8</td>
<td>0.176</td>
<td>0.412</td>
<td>0.299</td>
<td>134</td>
</tr>
<tr>
<td>1.0</td>
<td>0.15</td>
<td>0.433</td>
<td>0.300</td>
<td>196</td>
</tr>
</tbody>
</table>

- Although the partial credit strategy reduces the required sample size relative to a binary survival (partial credit = 1) endpoint, the goal is to accurately reflect the impact of the major AE on the statistical method of evaluation rather than to optimize power.
## DOOR / RADAR Advantages

- More pragmatic benefit:risk analyses (of patients)
- More comprehensive picture of intervention effects compared to traditional analyses
  - Competing risks are part of the outcome
  - E.g., a standard primary analysis in cardiovascular disease evaluates the time to the FIRST event (e.g., stroke, myocardial infarction, or death). But patients may experience more than one event. An ordinal composite outcome could incorporate these multiple events
- Superiority design; avoids NI complexities
- Reduction of sample size in some cases