Likelihood Approaches for Trial Designs in Early Phase Oncology Clinical Trials

Elizabeth Garrett-Mayer, PhD
Cody Chiuzan, PhD

Hollings Cancer Center
Department of Public Health Sciences
Medical University of South Carolina

April 24, 2014
Practicalities of Early Phase Oncology Trials

- Small sample sizes
- Conservation of resources
  - patients
  - time
  - funding
- Ethics
  - early stopping
  - protection of (sick) patients
- Recent Emphasis on “Adaptive” trial designs
  - Bayesian?
  - Controversies over priors
  - CRM: Both Bayesian and Likelihood implementations
Common Problems with Oncology Clinical Trials

1. No early stopping for futility
2. No safety stopping rules in efficacy trials
3. Poor characterization of properties of phase I design
Today’s Talk: Simple Likelihood Approaches for Addressing Common Problems

- The Likelihood Approach
  - principles
  - multiple looks
- The Sequential Probability Ratio Test (SPRT)
- Phase I Designs
- Phase II Designs
The Likelihood or Evidential Paradigm

Based on the Law of Likelihood:

If hypothesis $A$ implies that the probability of observing some data $X$ is $P_A(X)$, and hypothesis $B$ implies that the probability is $P_B(X)$, then the observation $X = x$ is evidence supporting $A$ over $B$ if $P_A(x) > P_B(x)$, and the likelihood ratio, $\frac{P_A(x)}{P_B(x)}$, measures the strength of that evidence. (Hacking (1965); Royall (1997))
**Likelihood Approach**

- Determine “what the data say” about the parameter of interest
- Likelihood function: gives a picture of the data
- Likelihood interval (LI): gives a range of reasonable values for the parameter of interest

![Likelihood Curve](image)
Likelihood Ratios

- Take ratio of heights of $L$ for different values of $\lambda$
- $L(\lambda = 0.030) = 0.78$; $L(\lambda = 0.035) = 0.03$
- $LR = \frac{0.78}{0.03} = 26$
Key difference in likelihood vs. frequentist paradigm

- **Consideration of the alternative hypothesis**
- Frequentist hypothesis testing
  - $H_0$: null hypothesis
  - $H_1$: alternative hypothesis
- Frequentist p-values:
  - Calculated assuming null is true
  - Ignores the alternative hypothesis
- Likelihood Ratio:
  - Compares evidence for the two specified hypotheses
  - Acceptance of rejection of null depends on the alternative
Example

- Assume $H_0 : \lambda = 0.12$ vs. $H_1 : \lambda = 0.08$
- What if true $\lambda = 0.10$?
- Simulated data, $N = 300$
- Frequentist: $p = 0.01$, reject the null
- Likelihood: $LR = \frac{1}{4}$, weak evidence in favor of null
Example

- Why?
- P-value looks for evidence against the null
- LR compares evidence for both hypotheses
- When the “truth” is in the middle, which makes more sense?
Likelihood Inference

- Three possible results: Strong evidence for $H_0$, strong evidence for $H_1$, or weak evidence

  - **Weak evidence**: at the end of the study, there is not sufficiently strong evidence in favor of either hypothesis
    - This can be controlled by choosing a large enough sample size
    - But, if neither hypothesis is correct, can end up with weak evidence even if $N$ is seemingly large

  - **Strong evidence**
    - **Correct evidence**: strong evidence in favor of correct hypothesis
    - **Misleading evidence**: strong evidence in favor of the incorrect hypothesis
      - Analogous to type I and type II errors
      - How can we control these?
Misleading Evidence in Likelihood Paradigm

- Universal bound: Under $H_0$,

$$P \left( \frac{L_1}{L_0} \geq k \right) \leq \frac{1}{k}; \quad \text{(Birnbaum, 1962; Smith, 1953)}$$

- In words, the probability that the likelihood ratio exceeds $k$ in favor of wrong hypothesis can be no larger than $\frac{1}{k}$

- In certain cases, an even lower bound applies Royall, 2000
  - Difference in normal means
  - Large sample size

- Common choices for $k$ are 8 (strong), 10, 32 (very strong).
Implications

- **Important result**: For a sequence of independent observations, the universal bound still holds (Robbins, 1970)

- **Implication**: We can look at the data as often as desired and our probability of misleading evidence is bounded
  - That is, if \( k = 10 \), the probability of misleading strong evidence is \( \leq 10\% \)
  - \( k = 10 \) seems reasonable considering \( \beta = 10 - 20\% \) and \( \alpha = 5 - 10\% \) in most studies
Sequential Probability Ratio Test

- This is not a new idea
  - Abraham Wald (1945) published “Sequential Tests of Statistical Hypotheses” in the Annals of Mathematical Statistics (it’s 70 pages long)
  - Commonly seen in DSMB plans
  - SPRT used to monitor secondary safety outcomes
- Classical SPRT assumes two boundaries:
  - Stop if there is strong evidence for $H_0$
  - Stop if there is strong evidence for $H_1$
  - Generally setup based on $A$ and $B$ which are based on selected $\alpha$ and $\beta$ levels
Wald’s Figure 2

- The context was not clinical trials
- The initial result was useful for the “war effort” in 1943 and essentially classified until 1945
- Popular in manufacturing, item response theory, etc.
- Max SPRT (Kulldorf et al., 2011) with applications to medical research, specifically rare adverse events
Most immunotherapy trials in cancer use a two-step approach to dose finding:

1. Perform an algorithmic design to identify safe doses
2. Collect immunologic data and “explore” it to see if there appears to be an optimal dose

Optimal dose?

- we imagine there will be a clear plateau in the association between dose and immunoresponse
- unrealistic and simplistic due to small sample sizes at each dose
- unrealistic and simplistic due to heterogeneity across patients
**Goal:** Develop an adaptive early phase design for assessing toxicity and efficacy outcomes in cancer immunotherapy trials.

- Identify the optimal dose to maximize efficacy while maintaining safety.
- **Two-stage design:**
  - Stage 1: Explore doses for safety and obtain information on immunotherapy outcomes
  - Stage 2: Allocate patients to allowable doses with emphasis towards doses with higher efficacy
- Uses both:
  - continuous (immunologic) outcomes
  - binary toxicity information
- Optimize efficacy while setting a threshold on acceptable toxicity
Stage 1: Confirm Safety

- Define $p_1$ and $p_2$ as unacceptable and acceptable DLT rates.
- Use cohorts of size $m$ to explore selected dose levels.
- Likelihood inference used to declare dose levels “allowable” based on $p_1$ and $p_2$ and observed data.
- Define $k$ as the threshold of evidence required for declaring a dose to be toxic.
- At end of Stage 1, there will be a set of doses for Stage 2.
- Continue to Stage 2 if two or more allowable doses.
- Details in Chiuzan et al. 2014
Likelihood method with cohorts of size 3

Example:

- \( p_1 = 0.40; \ p_2 = 0.15; \ m = 3 \)
- Require likelihood ratio \( \geq 2 \) (in favor of \( p_1 \)) to declare toxic.
- 0 or 1 DLT in 3 pts: allowable dose
- 2 or 3 DLTs in 3 pts: unacceptable dose (and all higher doses unacceptable)

<table>
<thead>
<tr>
<th>DLT</th>
<th>'3+3' rule</th>
<th>( L(p_1)/L(p_2) )</th>
<th>( k = 1 )</th>
<th>( k = 2 )</th>
<th>( k = 8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>acceptable</td>
<td>( LR = 0.35 )</td>
<td>acceptable</td>
<td>acceptable</td>
<td>weak</td>
</tr>
<tr>
<td>1</td>
<td>expand to 6</td>
<td>( LR = 1.33 )</td>
<td>toxic</td>
<td>weak</td>
<td>weak</td>
</tr>
<tr>
<td>2</td>
<td>toxic</td>
<td>( LR = 5.0 )</td>
<td>toxic</td>
<td>toxic</td>
<td>weak</td>
</tr>
<tr>
<td>3</td>
<td>toxic</td>
<td>( LR = 19.0 )</td>
<td>toxic</td>
<td>toxic</td>
<td>toxic</td>
</tr>
</tbody>
</table>

- toxic: \( LR \geq k \); acceptable: \( LR \leq \frac{1}{k} \); weak: \( \frac{1}{k} < LR < k \)

E Garrett-Mayer, C Chiuzan

Likelihood Approaches for Trial Designs in Early Phase Oncology
Stage 2: Adaptive randomization

Using data from Stage 1, estimate immunologic parameter estimate at each of $J$ allowable doses.

- Example: T cell persistence at 14 days.
- $y_i = \%$ CD3 cells of patient $i$ at 14 days compared to baseline
- $d_i =$ dose level for patient $i$

Estimation is based on a standard linear regression model using a log transformation of $y_i$:

$$\log(y_i) = \beta_0 + \sum_{j=1}^{J} \beta_j I(d_i = j) + e_i$$

$$e_i \sim N(0, \sigma^2); \quad \sum_j \beta_j = 0$$
Stage 2: Adaptive randomization

Define $p_j$ as the estimated persistence ($\%$) at dose $j$:

$$\hat{p}_j = e^{\hat{\beta}_0 + \hat{\beta}_j}$$

Calculate the randomization probabilities $\pi_j$ for doses $j = 1, \ldots, J$:

$$\pi_j = \frac{\hat{p}_j}{\sum_r \hat{p}_r}$$

or

$$\pi_j = \frac{\sqrt{\hat{p}_j}}{\sum_r \sqrt{\hat{p}_r}}$$
Example 1: shallow slope
Example 1: shallow slope
Example 1: shallow slope

\[
\begin{align*}
\pi_1 &= 0.14 \\
\pi_2 &= 0.2 \\
\pi_3 &= 0.33 \\
\pi_4 &= 0.33
\end{align*}
\]
Example 2: steep slope
Example 2: steep slope

![Graph showing Tcell persistence (% of baseline) vs Dose Level](image-url)
Example 2: steep slope
Stage 2

- For the first patient in Stage 2, randomize to allowable doses $j = 1, \ldots, J$ based on $\pi_j$.
- As data becomes available, update randomization probabilities for accruing patients.
- Repeat until total sample size is achieved, or some other stopping criteria is met.
- When DLTs are observed, utilize likelihood inference to determine if dose is “toxic”.

E Garrett-Mayer, C Chiuzan

Likelihood Approaches for Trial Designs in Early Phase Oncology
Motivation

Likelihood Paradigm

Phase I Dose Finding: Immunotherapies

Phase II: Sequential Trial Design for Single Arm Study

Discussion

Thanks

Example: cohort of size 5

**Example:**

- \( p_1 = 0.40; p_2 = 0.15 \)
- Require likelihood ratio \( \geq 2 \) (in favor of \( p_1 \)) to declare toxic.
- 0 or 1 DLT in 5 pts: allowable dose
- \( \geq 2 \) DLTs in 5 pts: unacceptable dose (and all higher doses unacceptable)

<table>
<thead>
<tr>
<th>DLT</th>
<th>( L(p_1)/L(p_2) )</th>
<th>( k = 1 )</th>
<th>( k = 2 )</th>
<th>( k = 8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( LR = 0.18 )</td>
<td>acceptable</td>
<td>acceptable</td>
<td>weak</td>
</tr>
<tr>
<td>1</td>
<td>( LR = 0.66 )</td>
<td>acceptable</td>
<td>weak</td>
<td>weak</td>
</tr>
<tr>
<td>2</td>
<td>( LR = 5.0 )</td>
<td>toxic</td>
<td>toxic</td>
<td>weak</td>
</tr>
<tr>
<td>3</td>
<td>( LR = 19.0 )</td>
<td>toxic</td>
<td>toxic</td>
<td>toxic</td>
</tr>
</tbody>
</table>

- toxic: \( LR \geq k \); acceptable: \( LR \leq \frac{1}{k} \); weak: \( \frac{1}{k} < LR < k \)
Aside: The ‘3+3’ provides weak evidence for DLT rate

\[ H_0 : p_0 = 0.40 \quad H_1 : p_1 = 0.15 \]

<table>
<thead>
<tr>
<th>DLT(n)*</th>
<th>‘3+3’ rule</th>
<th>( L(p_1)/L(p_0) )</th>
<th>( k = 1 )</th>
<th>( k = 2 )</th>
<th>( k = 8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0(3)</td>
<td>acceptable</td>
<td>LR = 2.84</td>
<td>acceptable</td>
<td>acceptable</td>
<td>weak</td>
</tr>
<tr>
<td>1(3) + 0(3)</td>
<td>acceptable</td>
<td>LR = 2.14</td>
<td>acceptable</td>
<td>acceptable</td>
<td>weak</td>
</tr>
<tr>
<td>2(3)</td>
<td>toxic</td>
<td>LR = 0.20</td>
<td>toxic</td>
<td>toxic</td>
<td>weak</td>
</tr>
<tr>
<td>3(3)</td>
<td>toxic</td>
<td>LR = 0.05</td>
<td>toxic</td>
<td>toxic</td>
<td>toxic</td>
</tr>
<tr>
<td>1(3) + 1(3)</td>
<td>toxic</td>
<td>LR = 0.57</td>
<td>toxic</td>
<td>weak</td>
<td>weak</td>
</tr>
<tr>
<td>1(3) + 2(3)</td>
<td>toxic</td>
<td>LR = 0.15</td>
<td>toxic</td>
<td>toxic</td>
<td>weak</td>
</tr>
<tr>
<td>1(3) + 3(3)</td>
<td>toxic</td>
<td>LR = 0.04</td>
<td>toxic</td>
<td>toxic</td>
<td>toxic</td>
</tr>
</tbody>
</table>

Dose limiting toxicities (cohort size)*
Acceptable: \( LR \geq k \); toxic: \( LR \leq 1/k \); weak: \( 1/k < LR < k \)
Constrained Adaptive Randomization

Simulations:

- Six toxicity and five efficacy models; two variance structures; two sample sizes
- Safety constraints implemented using $H_1 : p_1 = 0.40$; $H_2 : p_2 = 0.15$; $k = 2$
- Subset presented: six toxicity scenarios
  - “Plateau” efficacy model
  - Large variance across patients
  - Smaller sample size ($N = 25$)
Results: Stage 1 & 2

DLT rates: 0.05, 0.07, 0.10, 0.12, 0.15
Results: Stage 1 & 2

DLT rates: 0.25, 0.25, 0.25, 0.25, 0.25
Results: Stage 1 & 2

DLT rates: 0.05, 0.18, 0.29, 0.40, 0.51
Results: Stage 1 & 2

DLT rates: 0.05, 0.42, 0.48, 0.50, 0.52
Results: Stage 1 & 2

DLT rates: 0.02, 0.02, 0.02, 0.02, 0.02
Results: Stage 1 & 2

DLT rates: 0.02, 0.05, 0.28, 0.40, 0.50
Phase II: Sequential Trial Design Motivating Example

- Single arm cancer clinical trial
- Outcome = clinical response
- Early stopping for futility
- Standard frequentist approach
  - Simon two-stage design
  - Only one look at the data
  - Determine optimality criterion
    - minimax
    - minimum E(N) under H0 (Simon’s optimal)
- Likelihood approach
  - Use binomial likelihood
  - Can look at the data after each observation
Phase II: Sequential Trial Design Motivating Example

- New cancer treatment agent
- Anticipated response rate is 40% \((p_2 = 0.40)\)
- Null response rate is 20% \((p_1 = 0.20)\)
  - the standard of care yields 20%
  - not worth pursuing new treatment with same response rate as current treatment
- Using frequentist approach:
  - Simon two-stage with \(\alpha = \beta = 0.10\)
  - Optimum criterion: smallest \(E(N)\)
  - First stage: enroll 17. If 4 or more respond, continue
  - Second stage: enroll 20. If 11 or more respond, conclude success.
  - Maximum sample size = 37
Implementation

- Look at the data after each patient
- Estimate the difference in $\log(L_1)$ and $\log(L_2)$
- Rules:
  - If $\log(L_1) - \log(L_2) \geq \log(k)$: stop for futility
  - If $\log(L_1) - \log(L_2) < \log(k)$: continue enrolling patients
Stopping Rules

- Given discrete nature, only certain looks provide an opportunity to stop.

- Current example: stop the study if
  - 0 responses in 9 patients
  - 1 response in 12 patients
  - 2 responses in 15 patients
  - 3 responses in 19 patients
  - 4 responses in 22 patients
  - 5 responses in 26 patients
  - 6 responses in 29 patients
  - 7 responses in 32 patients

- Although total N can be as large as 37, there are only 8 thresholds for futility early stopping assessment.
How does the proposed approach compare to the optimal Simon two-stage design?

What are performance characteristics we would be interested in?

- small $E(N)$ under the null hypothesis
- frequent stopping under null
- infrequent stopping under alternative
- acceptance of $H_1$ under $H_1$
- acceptance of $H_2$ under $H_2$
Simon Optimal vs. Likelihood

![Graph showing comparison between Simon Optimal and Likelihood approaches for trial designs in early phase oncology clinical trials. The graph illustrates the probability of accepting the null hypothesis (H0) or the alternative hypothesis (HA) under different conditions. The y-axis represents the probability, and the x-axis represents different values related to the trial outcomes. The graph indicates that Simon Optimal approaches have a different distribution compared to the Likelihood approaches, with Simon Optimal showing a higher probability of accepting the null hypothesis for certain outcomes.](image-url)
Motivation
Likelihood Paradigm
Phase I Dose Finding: Immunotherapies
Phase II: Sequential Trial Design for Single Arm Study
Discussion
Thanks

Probability of Early Stopping

E Garrett-Mayer, C Chiuzan
Likelihood Approaches for Trial Designs in Early Phase Oncology Clinical Trials
Expected Sample Size

- **Likelihood (optimal N)**
- **Likelihood (minmax N)**
- **Simon Optimal**
- **Simon MinMax**

E Garrett-Mayer, C Chiuzan

Likelihood Approaches for Trial Designs in Early Phase Oncology Clinical Trials
Bayesian Competitor: Predictive Probability

- Bayesian but without loss function (no Bayes risk)
- Searches for design parameters to ensure size and power
- Prior is chosen so that mean of prior is the null hypothesis, but it’s a weak prior.
- Predictive probability \( PP = \) probability that the end result of the trial is positive given current data and data to be observed
  - Based on the probability that the true response rate is greater than the null response rate.
  - Ignores the alternative
Bayesian Competitor Predictive Probability

- **Stopping**: Based on pre-defined thresholds, $\theta_L$ and $\theta_U$
  - If $PP < \theta_L$: stop the trial and reject alternative
  - If $PP > \theta_U$: stop the trial and reject the null (but often $\theta_U = 1$)
- At pre-specified times, the predictive probability is calculated
  - Lee and Liu explore difference frequency of looks
  - Comparisons here are for looking at every patient
- $\theta_T$ is defined as the threshold for determining efficacy at the trial’s end
  - $\theta_T$ and $\theta_U$ do not have the same stringency
## Comparison of Designs

<table>
<thead>
<tr>
<th></th>
<th>$PET_1$</th>
<th>$E(N_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simon Two-Stage</td>
<td>0.55</td>
<td>26.0</td>
</tr>
<tr>
<td>Predictive Probability</td>
<td>0.85</td>
<td>25.1</td>
</tr>
<tr>
<td>Likelihood</td>
<td>0.80</td>
<td>22.1</td>
</tr>
</tbody>
</table>
Monitoring Safety

- Can be used to monitor safety, e.g. transplant-related mortality
- It is set-up so that too many events trigger stopping (instead of too few)
- More appropriate than:
  - Repeated confidence intervals
  - repeated p-values (controlling alpha?)
- Like a one-sided SPRT
**Discussion**

- Evidential methods have been around a long time
- They allow us to:
  - Simply quantify evidence for competing hypotheses
  - Evaluate the data frequently with bounded probability of misleading evidence
- Likelihood methods are simple to implement and require no prior
- Sufficient work has been done to justify $K$ thresholds for Phase III trials
- More work should be done to provide reasonable guidance on appropriate $K$ in early phase trials
Thank you!

Email: garrettm@musc.edu; chiuzan@musc.edu

Work supported by the NIH grant, P01 CA 154778-01 and by the Biostatistics Shared Resource, Hollings Cancer Center, MUSC (P30 CA138313)