The Utility of Bayesian Predictive Probabilities for Interim Monitoring of Clinical Trials

Ben Saville, Ph.D.

Berry Consultants

KOL Lecture Series, Nov 2015

How are clinical trials similar to missiles?

How are clinical trials similar to missiles?

- Fixed trial designs are like ballistic missiles:
  - Acquire the best data possible a priori, do the calculations, and fire away
  - They then hope their estimates are correct and the wind doesn't change direction or speed
- Adaptive trials are like guided missiles:
  - Adaptively change course or speed depending on new information acquired
  - More likely to hit the target
  - Less likely to cause collateral damage

Interim analyses in clinical trials

- Interim analyses for stopping/continuing trials are one form of adaptive trials
- Various metrics for decisions of stopping
  - Frequentist: Multi-stage, group sequential designs, conditional power
  - Bayesian: Posterior distributions, predictive power, Bayes factors
- Question: Why and when should I use Bayesian predictive probabilities for interim monitoring?
  - Clinical Trials 2014: Saville, Connor, Ayers, Alvarez

Questions addressed by interim analyses

1. Is there convincing evidence in favor of the null or alternative hypotheses?
    - evidence presently shown by data
2. Is the trial likely to show convincing evidence in favor of the alternative hypothesis if additional data are collected?
    - prediction of what evidence will be available later

- Purpose of Interims
    - ethical imperative to avoid treating patients with ineffective or inferior therapies
    - efficient allocation of resources

Predictive Probability of Success (PPoS)

- Definition: The probability of achieving a successful (significant) result at a future analysis, given the current interim data
- Obtained by integrating the data likelihood over the posterior distribution (i.e. we integrate over future possible responses) and predicting the future outcome of the trial
- Efficacy rules can be based either on Bayesian posterior distributions (fully Bayesian) or frequentist p-values (mixed Bayesian-frequentist)

Calculating predictive probabilities via simulation

1. At an interim analysis, sample the parameter of interest θ from the current posterior given current data $X_{(n)}$.
2. Complete the dataset by sampling future samples $X_{(m)}$, observations not yet observed at the interim analysis, from the predictive distribution
3. Use the complete dataset to calculate success criteria (p-value, posterior probability). If success criteria is met (e.g. p-value < 0.05), the trial is a success
4. Repeat steps 1-3 a total of B times; the predictive probability (PPoS) is the proportion of simulated trials that achieve success

Futility - Possible definitions

1. A trial that is unlikely to achieve its objective (i.e. unlikely to show statistical significance at the final sample size)
2. A trial that is unlikely to demonstrate the effect it was designed to detect (i.e. unlikely that H_a is true)

Illustrative Example: Monitoring for futility

- Consider a single arm Phase II study of 100 patients measuring a binary outcome (favorable response to treatment)
- Goal: compare proportion to a gold standard 50% response rate
- $x \sim \mathrm{Bin}(p, N = 100)$
  $p =$ probability of response in the study population
  $N =$ total number of patients
- Trial will be considered a success if the posterior probability that the proportion exceeds the gold standard is greater than $\eta = 0.95$,

$$\Pr(p > 0.5|x) > \eta$$

Illustrative Example

- Uniform prior $p \sim \mathrm{Beta}(\alpha_0 = 1, \beta_0 = 1)$
- The trial is a “success” if 59 or more of 100 patients respond
- Posterior evidence required for success:
  $\Pr(p > 0.50 | x = 58, n = 100) = 0.944$
  $\Pr(p > 0.50 | x = 59, n = 100) = 0.963$
- Consider 3 interim analyses monitoring for futility at 20, 50, and 75 patients

Notation

- Let $j = 1, ..., J$ index the jth interim analysis
- Let n_j be the number of patients
- x_j = number of observed responses
- m_j = number of future patients
- y_j = number of future responses of patients not yet enrolled
  i.e. $n = n_j + m_j$ and $x = x_j + y_j$

First Interim analysis

- Suppose at the 1st interim analysis we observe 12 responses out of 20 patients (60%, p-value $=$ 0.25)
- $\Pr(p > 0.50 | x_1 = 12, n_1 = 20) = 0.81$, and 47 or more responses are needed in the remaining 80 patients ($\geq$ 59%) in order for the trial to be a success
- $y_1 \sim$Beta-binomial$(m_1 = 80, \alpha = \alpha_0 + 12, \beta = \beta_0 + 8)$
- PPoS $= \Pr(y_1 \geq 47) = 0.54$
- Should we continue?

Second Interim analysis

- 2nd interim analysis: 28 responses out of 50 patients (56%, p-value=0.24)
- Posterior Probability = 0.81
- Predictive Probability of Success = 0.30
- 31 or more responses are needed in the remaining 50 patients ($\geq$ 62%) in order to achieve trial success.
- Should we continue?

Third Interim analysis

- 3rd interim analysis: 41 responses of 75 patients (55%, p-value = .24)
- Posterior Probability = 0.81
- Predictive Probability of Success = 0.086
- 18 or more responses are needed in the remaining 25 patients ($\geq$ 72%) in order to achieve success
- Should we continue?
- The posterior estimate of 0.80 (and p-value of 0.24) means different things at different points in the study relative to trial "success"

Table

Table: Illustrative example

n_j	x_j	m_j	y_j^*	p-value	$\Pr(p > 0.5)$	PPoS
20	12	80	47	0.25	0.81	0.54
50	28	50	31	0.24	0.80	0.30
75	41	25	18	0.24	0.79	0.086
90	49	10	10	0.23	0.80	0.003

n_j and x_j are the number of patients and successes at interim analysis j
m_j = number of remaining patients at interim analysis j
y_j^* = minimum number of successes required to achieve success
PPoS= Bayesian predictive probability of success

Graphical representation

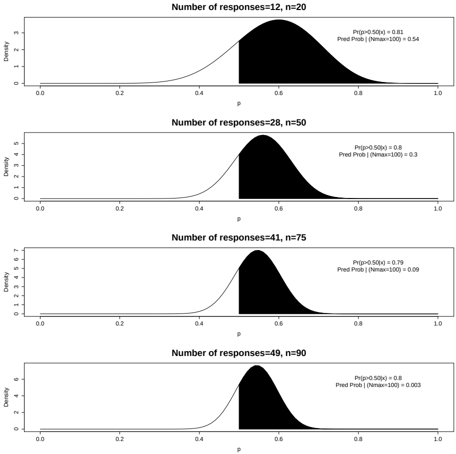

Figure: Posterior distributions for 4 interim analyses

Mapping PPoS to posterior probabilities

- Suppose in our example, the trial is stopped when the PPoS is less than 0.20 at any of the interim analyses
  - Power $= 0.842$
  - Type I error rate $= 0.032$ (based on 10,000 simulations)
- Equivalently, we could choose the following posterior futility cutoffs
  - < 0.577 (12 or less out of 20)
  - < 0.799 (28 or less out of 50)
  - < 0.897 (42 or less out of 75)
- Exactly equivalent to stopping if PPoS < 0.20

Predictive vs. posterior probabilities

- In simple settings where we can exactly map posterior and predictive probabilities: computational advantages of using the posterior probabilities
- In more complicated settings, it can be difficult to align the posterior and predictive probability rules
- It is more straightforward to think about "reasonable" stopping rules with a predictive probability
- Predictive probabilities are a metric that investigators understand ("What's the probability of a return on this investment if we continue?"), so they can help determine appropriate stopping rules

Group sequential bounds

- Group sequential methods use alpha and beta spending functions to preserve the Type I error and optimize power
- Given working example, an Emerson-Fleming lower boundary for futility will stop for futility if less than 5, 25, or 42 successes in 20, 50, 75 patients, respectively.
- Power of design is 0.93, Type I error is 0.05

Emerson-Fleming lower boundary

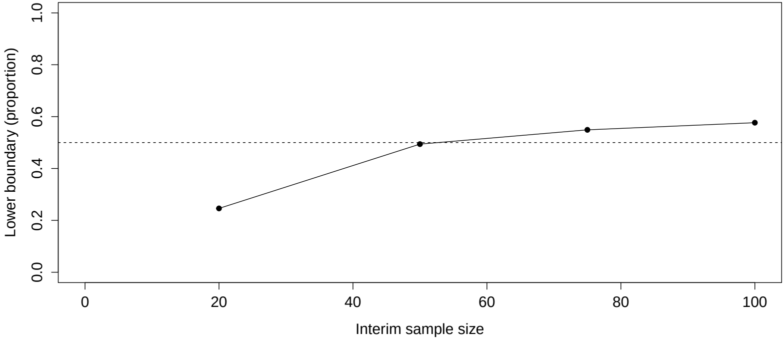

Figure: Emerson-Fleming lower boundary for futility

Emerson-Fleming lower boundary

- The changing critical values are inherently trying to adjust for the amount of information yet to be collected, while controlling Type I and Type II error
- The predictive probabilities of success at 5/20 or 25/50 (which both continue with Emerson-Fleming boundaries) are 0.0004 and 0.041
- Are these reasonable stopping rules?

Futility: Repeated testing of alternative hypothesis

- Assess current evidence against targeted effect (H_a) using p-values
- At each interim look, test the alternative hypothesis at alpha $= 0.005$ level
- Requires specification of H_a, e.g. $H_a : p_1 = 0.65$
- Example: Stop for futility if less than 8, 24, 38, or 47 responses at 20, 50, 75, or 90 patients
  - Predictive Probabilities are 0.031, 0.016, 0.002, and 0.0, where above rules allow continuation

Conditional Power: Example

- Definition: The probability of a successful trial at the final sample size, given observed data and an assumed effect size
- Commonly used effect sizes: original H_a (CP_{H_a}), current MLE (CP_{MLE}), and null hypothesis H_0 (CP_{H_0})
- Even when the likelihood that 0.65 is the true response rate becomes less and less likely during the course of the trial, CP_{H_a} continues to use 0.65
- CP_{MLE} uses the MLE at each analysis but fails to incorporate the variability of that estimate
- CP_{H_0} only gives the probability assuming that the treatment doesn't work (given observed data)

Table

Table: Illustrative example

n_j	x_j	m_j	y_j^*	p-value	$\Pr(p > 0.5)$	CP_{H_a}	CP_{MLE}	PPoS
20	12	80	47	0.25	0.81	0.90	0.64	0.54
50	28	50	31	0.24	0.80	0.73	0.24	0.30
75	41	25	18	0.24	0.79	0.31	0.060	0.086
90	49	10	10	0.23	0.80	0.013	0.002	0.003

n_j and x_j are the number of patients and successes at interim analysis j

m_j = number of remaining patients at interim analysis j

y_j^* = minimum number of successes required to achieve success

CP_{H_a} and CP_{MLE}: Conditional power based on original H_a or MLE

PPoS= Bayesian predictive probability of success

Conditional Power

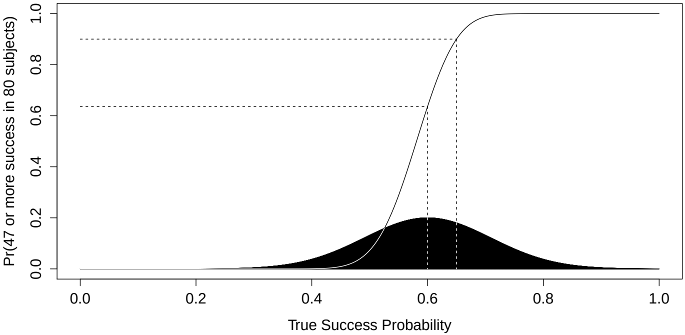

Figure: Conditional Power given 12 success in 20 patients

Predictive probabilities

- Predictive probabilities are weighted averages of the the conditional powers across the current probability that each success rate is the true success rate (i.e. weighted by the posterior)
- Hence, predictive probabilities are a much more realistic value of predictive trial success than any single estimate of conditional power

Efficacy

- Success: There is convincing evidence that the treatment is effective
  - Question naturally corresponds to evidence currently available
  - If outcomes of accrued patients are all observed, prediction methods are not needed
- If we use PPoS to monitor for early success, one typically needs to already meet the posterior success criteria
  - e.g., if PPoS > 0.95 at interim look, typically implies $\Pr(p > p_0 | x_j) > 0.95$, which implies trial success

Efficacy: Delayed outcomes

- Using PPoS for stopping for efficacy is primarily useful for delayed outcomes, e.g. time to event
  - With incomplete data, question of success becomes a prediction problem
  - At an interim analysis, PPoS with the current patients (some of which have yet to observe their complete follow-up time)
  - Trial stopped for expected efficacy, current patients followed until outcomes are observed, final analysis completed

Efficacy: Delayed outcomes

- Traditional group sequential methods
  - If trial is stopped due to an efficacy boundary being met, typically a final analysis is done after all lagged outcomes are observed on the current set of patients
  - Efficacy is determined by interim, not final analysis
  - Hence, DMC's may be unlikely to stop trials for efficacy unless the data are convincing and p-value would not lose significance if a few negative outcomes occurred in the follow-up period
- Predictive probabilities formalize this decision making process, i.e. stop trials for efficacy if they currently show superiority and are likely to maintain superiority after remaining data are collected

Efficacy: Time-lag with auxiliary variables

- PPoS can be used to model a final primary outcome using earlier information that is informative about the final outcome
- For example, if the primary outcome is success at 24 months, many of the accrued patients at a given interim analysis will not have 24 months of observation time
- However, there exists information on the success at 3, 6, and 12 months that is correlated with the outcome at 24 months
- These earlier measures are auxiliary variables, and can be used to model various types of primary outcomes, including binary, continuous, time-to-event, and count data

Efficacy: Time-lag with auxiliary outcomes

- These auxiliary variables may not be valid endpoints from a regulatory perspective
- Incorporates partial information into the predictive distribution of the final outcome to provide a more informative predictive probability of trial success
- If the predictive probability at final N is sufficiently small, the trial can be stopped for futility immediately
- If the predictive probability with current n and more follow-up is sufficiently large, one can stop accrual and wait until the primary outcome is observed for all currently enrolled patients, at which point trial success is evaluated
- Note the auxiliary variables do not contribute to the final analysis

Efficacy

- Time-lags are extremely common in clinical trials; very rare to observe an outcome immediately upon enrollment
- Other competing methods (group sequential, conditional power, posterior probabilities, etc.) are not easily adapted to account for time-lags or auxiliary variables
- Predictive probabilities are also extremely useful for calculating predicted success of future phase III study while in a phase II study

Relationship between predictive probability and posterior

- When an infinite amount of data remains to be collected, PPoS equals the current posterior estimate of efficacy, $\Pr(p > p_0 | x_j, n_j)$
- For example, suppose an interim analysis yields 25 responses from 50 patients. The current estimate of $\Pr(p > 0.50 | x = 25, n = 50)$ equals 0.50
- If the trial claims efficacy for a posterior cutoff of 0.95, i.e. $\Pr(p > 0.50 | N) \geq 0.95$, then for a maximum sample size $N = 100$ patients, PPoS equals 0.04
- Given the same interim data, PPoS for maximum sample sizes of 200, 500, 1000, and 10000 patients are 0.17, 0.29, 0.35, and 0.45 (converging to 0.50 as N approaches infinity)

Predictive Probability vs. Posterior

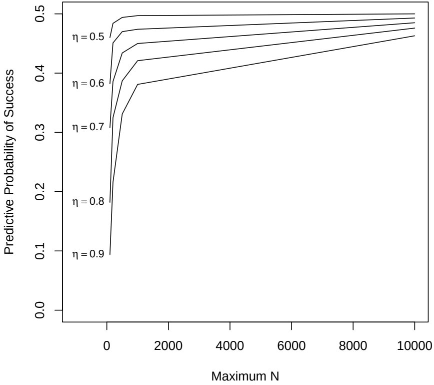

Figure: Predictive probabilities vs. maximum sample size N by posterior threshold η, with interim $n = 50$ and observed $x = 25$

Predictive Probability vs. Posterior

- For a fixed maximum sample size (e.g. $N = 100$) and a fixed posterior probability, PPoS converges to either 0 or 1 as the interim sample size increases
- Logical because the trial success or failure becomes more certain as trial nears its end

Predictive Probability vs. Posterior

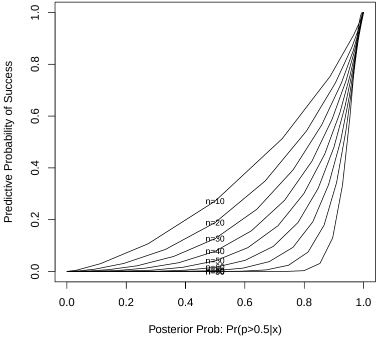

Figure: PPoS vs. posterior estimate $\Pr(p > 0.50|x)$ by interim sample size n, with maximum sample size $N = 100$ and posterior threshold $\eta = 0.95$

Computational challenges

- Simulations are typically used to calculate predictive probabilities; can be problematic for calculating operating characteristics
- Let K trials be needed to assess operating characteristics, J the number of interim analyses, and B the number of simulations required to calculate a single predictive probability
- Trial requires $J \times B \times K$ imputations for a single setting of parameters (e.g. under H_0)
- For example, a trial with 3 interim analyses and $B =$1000, the trial would require a total of $3 \times 1000 \times 1000 =$3,000,000 simulated complete data sets
- Further complicated if Bayesian posterior distributions are not available in closed form (MCMC)

Prior distributions

- Large literature exists on selection of prior distributions for Bayesian analyses of clinical trials
- Common choices: "non-informative" prior, skeptical prior, enthusiastic prior, and historical prior
- Clinical trial designs using predictive probabilities for interim monitoring do not claim efficacy using predictive probabilities; the claim of efficacy is based on either Bayesian posterior probabilities or frequentist criteria (p-values)
- Same discussions of prior distributions in the literature are applicable to Bayesian designs with interim monitoring via predictive probabilities

Prior distributions

- One can calculate the predictive probability of trial success at interim looks using historical prior information, even though the final analysis may use the flat or skeptical prior
- For example, simulating complete data sets under the historical prior, but using the flat or skeptical prior to determine whether each simulated trial is a success
- Uses all available information to more accurately predict whether the trial will be a success, but maintain objectivity or skepticism in the prior for the final analysis
- Hence a historical (i.e. "honest") prior can be more efficient in making decisions about the conduct of a trial

Conclusion

- Predictive probabilities
    - Closely align with the clinical decision making process, particularly with prediction problems such as futility, efficacy monitoring with lagged outcomes, and predicting success in future trials
    - Thresholds can be easier for decision makers to interpret compared to those based on posterior probabilities or p-values
    - Avoids illogical stopping rules
    - In many settings, the benefits are worth the computational burden in designing clinical trials