2025 ASA New Jersey and Princeton-Trenton Chapters Spring Symposium

Bayesian Interim Analysis and Sample Size Reestimation

Ming-Dauh Wang, PhD
Bayer Pharmaceuticals

June 13, 2025

Outline of Presentation

1. Prologue
   - // Some reflections on statistical innovation
   - // Examples
2. Bayesian methods for interim analysis and sample size reestimation
   - // Bayesian thinking
   - // An example and points for consideration
3. Epilogue
   - // An insightful prediction
   - // Some parting thoughts

Prologue

01

What is Statistical Innovation?

- Dose it need to be complex?
- Problem solving vs. revolutionary/ground-breaking
- Can statistics save a company?
- How do we think about “advanced analytics”?
- A prudent caution from a former colleague
- A piercing question worth much reflection

A Phase 3 Rare Disease Safety Trial for Label Expansion

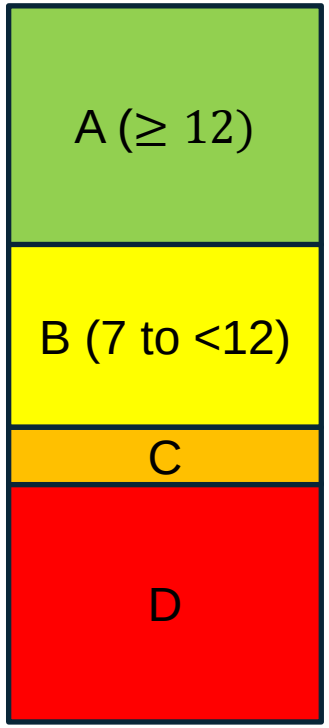
BAYER NEWS

May 19, 2025

U.S. FDA grants approval for Jivi® antihemophilic factor (recombinant), PEGylated-aucl in pediatric patients 7 to under 12 years of age with hemophilia A (congenital Factor VIII deficiency)

Background

- Jivi was first approved by the FDA on August 30, 2018, for the treatment of hemophilia A in previously treated adults and adolescents 12 years of age and older. The pivotal trials (single-arm) were conducted in a larger population (A+B+C+D)
- A very rare adverse event of special interest (AEoSI) was observed, 1 case in A and other cases in D
- B+C has safety profile similar to A
- Sponsor proposed A for initial application by creating B+C as a safety margin
- Sponsor planned to conduct a single-arm trial with the same design and criteria of the pivotal trials in B, still leaving C as a margin, for label expansion



Proposed Trial

- Considered enrollment feasibility, only 30 patients were planned to be enrolled
- Planned to utilize the 25 patients from B of the pivotal studies for analysis
- True incidence of AEoSI was believed to be <5%, which was intended to be demonstrated by the new study
- Statistical inference focused on estimation than hypothesis testing
- Needed a reasonable success criterion and sample size justification

Conventional Analysis

- Analysis of 30 new patients alone
  - If 1 AEoSI is observed, the 90% Clopper-Pearson confidence interval is (0.002, 0.149)
- Analysis of 30 new + 25 old patients
  - If 1 AEoSI is observed, the 90% Clopper-Pearson confidence interval is (0.001, 0.083)
- Neither would be able to reasonably conclude the true incidence is <5%

A Bayesian Analysis Proposal

- Full borrowing from the pivotal studies: pool the 25 old patients with the 30 new patients for analysis
- Bayesian success criterion: >90% posterior probability that the true incidence of AEoSI is <5%
- Justification of the sample size 30:
  - Assume at most 1 patient out of 30 will experience the AEoSI
  - None of the 25 patients in the pivotal study experienced the AEoSI
  - If 1 of 30 patients from the new study experienced the AEoSI, then a Bayesian beta-binomial inference with a prior of Beta(1/4,1/4) for the pooled 55 patients will result in
    - the median of the posterior distribution of the true incidence of AEoSI is 1.7%, and
    - there is 91% posterior probability that the true incidence of AEoSI is <5%

Bayesian Small-Sample Inference for An Ultra-Rare Disease Trial Intended for Regulatory Submission

***REGENERON* News**

August 18, 2023

Veopoz™ (pozelimab-bbfg) Receives FDA Approval as the First Treatment for Children and Adults with CHAPLE Disease at 1:43 PM EDT

Background

- A single-arm open-label trial of pozelimab for PLE (Protein-Losing Enteropathy) ultra-rare disease (<100 worldwide) just newly described in 2017 was designed and intended for BLA submission
- A minimum of 6 and up to 10 patients were proposed to be enrolled
- Typical statistical approaches would betray inappropriateness for the case
- There was off-label use of an approved drug by an investigator
- The investigator shared patient history and on-treatment data with us
- Met with FDA
  - The agency itself was learning about the new disease even during the pIND meeting
  - Asked for success criteria to guide regulatory review

First Attempt

- The primary endpoint was binary, with response defined by a composite of a primary lab analyte (proposed by sponsor) and clinical symptoms (requested by FDA)
- In the historical control period before initiating the off-label treatment, 0 of 14 patients met the primary endpoint, with a 90% exact CI of (0.00, 0.19)
- If 4 of 6 assessable patients treated with the new drug achieve the primary endpoint, the 90% exact CI for the probability of achieving the primary endpoint is (0.27, 0.94), with the lower limit of 0.27 clearly greater than the upper limit of 0.19 for the historical control period of the 14 patients
- Thus, the proposed study would be regarded as successful if at least 4 of 6 assessable patients with active disease state achieve the primary endpoint
- FDA did not agree with the reasoning

A Bayesian Justification

- By a Bayesian analysis, a comparison between (4 of 6 drug-treated patients achieving the primary endpoint) and (0 of 14 in the historical control data) by a beta-binomial analysis with the Jefferys prior of Beta(0.5, 0.5) would conclude that there is a >99% posterior probability that the new drug has a greater response rate than that of untreated patients
- There would also be a 90% posterior probability that the response rate with the new drug is at least 0.37 higher than that of untreated patients
- Thus, meeting the study success criterion of observing at least 4 of 6 assessable patients with active disease state achieving the primary endpoint indicates strong evidence of effectiveness
- FDA then accepted the Bayesian reasoning and the success criterion

Probability of Study Success

- Needed to justify N=6 in consideration of enrollment feasibility
- All 14 patients treated with off-label use of the approved drug achieved the primary endpoint, with a 90% exact CI: (0.81, 1.00)
- If the true rate of achieving the primary endpoint for patients on the new drug is 0.81 (the lower limit of the CI for the approved drug), then a sample size of 6 gives a probability of 91% for achieving the study success criterion
- Notes:
  - Relying on the success criterion alone is too much simplified
  - Patient efficacy and safety profiles of many labs/clinical endpoints would have been closely examined by both FDA and the sponsor

Bayesian Methods for Interim Analysis and Sample Size Reestimation

02

Bayesian Thinking

Bayesian Inference

- A scientific question at hand is answered by information contained in available data D, e.g. “is the studied drug efficacious?”
- The data D are indexed by an unknown parameter θ (could be a vector) for a representative quantification of the information, expressed by the likelihood $f(D|\theta)$
- Learning about θ given D is derived from $f(D|\theta)$ and a prior knowledge of θ expressed by a prior density $\pi(\theta)$, proper or improper, resulting in a posterior density

$$f(\theta|D) \propto f(D|\theta)\pi(\theta)$$

Bayesian Decision Process

- A decision criterion is defined regarding an answer to the scientific question, e.g. “if the posterior θ is greater than a threshold r, then the drug is efficacious; otherwise, it is not.” That is,

  decision criterion about efficacy: $\theta > r$

- A decision rule (success/failure, go/no-go) is made based on how likely the decision criterion will be achieved based on the posterior inference, e.g. “is the posterior probability $P(\theta > r)$ is greater than a?” That is,

  decision rule based on θ: $P(\theta > r) > a$

Predictive Inference and Decision

- Inference often dose not stop at posterior summary
  - Before the study: “how likely is the new study to succeed?”
  - During the study: “What will the rest of patients planned to enroll fare?”
  - After the study: “how about the subsequent studies?”
- The continuing task is “prediction”, which involves additional uncertainty from the yet unobserved future
- Based on the posterior inference, the predictive density is

$$f(\tilde{D}|D) = \int f(\tilde{D}|\theta) f(\theta|D) d\theta$$

- Predictive inference is then based on a summary of $\tilde{D}$

Predictive Inference of Seed Quality

Background

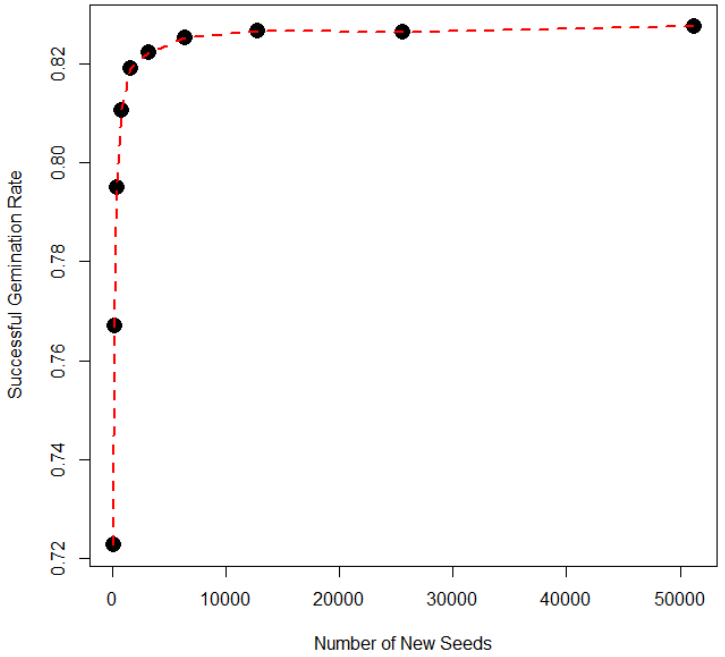
- A Bayer Crop Science colleague presented a Bayesian approach to assessment of seed batch quality
- For each batch, seed sample(s) are tested by warm gemination test or radical emergence test (RET)
- Bayesian binomial-beta conjugate analysis was performed for batches, and posterior studentized residuals or means are ranked for decision of retention or discarding

A Predictive Approach

- Instead of posterior inference, can predict what would be the successful gemination rate for each batch given warm germination or RET test data
- Then set a rule based on the predictive distribution to make sure the future probability of successful gemination rate of >92% of a batch is high enough, such as 80%
- Sell only batches that pass the rule
- If a test of a sample of 100 seeds from a batch resulted in 95 successes and 5 failures
  - Given a prior Beta(1,1) for the successful gemination rate, the posterior would be Beta(96,6)
  - Then to observe >92% successful gemination rate of a new sample of 100 seeds, the predictive probability is 72%
  - Given a threshold probability a decision can be made on sale of the batch

A Predictive Approach (Cont'd)

- An immediate question would be "Isn't that every seed of the millions in the batch should count?" Then what would be the correct predictive probability?
- If we look at the predictive probability by number of new seeds, it increases and levels at about 0.827, which can be used for decision for the batch
- Might also consider collective inference across batches by adding a parameter to account for heterogeneity among batches and use each batch's estimated posterior successful gemination rate to predict the batch's overall predictive performance



Bayesian Interim Analysis

Dimitris A and Wang M-D. (2006). *Bayesian predictive approach to interim monitoring in clinical trials.* Statistics in Medicine

- Notation
  - x_1: data up to IA, x_2: data after IA, $x = (x_1, x_2)$
  - T: a statistical test, H_a: alternative hypothesis, t_α: significance threshold for level α
  - η: a Bayesian success threshold
  - $f(x_2|x_1) = \int f(x_2|\theta) f(\theta|x_1)\, \pi(\theta) d\theta$: predictive density of x_2 given x_1
- Predictive power approach

$$\text{Power (PoS)} = \int I(T(x) > t_\alpha) f(x_2|x_1) d\, x_2$$

- Bayesian predictive approach

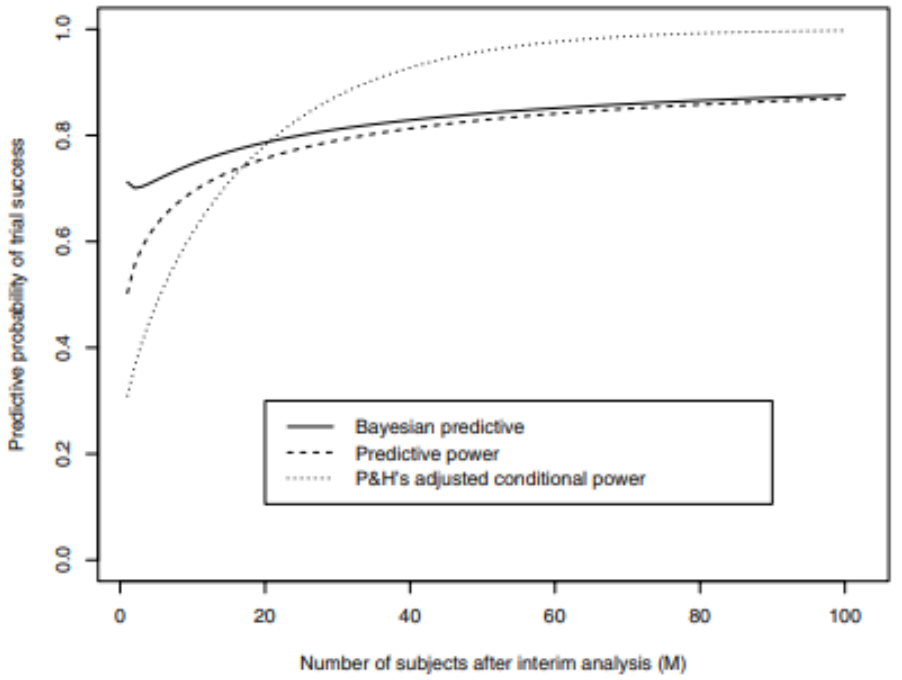
$$\text{Power (PoS)} = \int I(\Pr(\theta \in H_a|x) > \eta) f(x_2|x_1) d\, x_2$$

Bayesian Predictive Sample Size Reestimation

Wang M-D. (2007). *Sample size reestimation by Bayesian prediction.* Bimetrical Journal

- Can not always get whatever wanted by increasing sample size
- As the sample size increases, the power (PoS) will converge to the interim learning of the parameter
- That is, the power (PoS) will be capped by the interim posterior probability of success

- An example from Wang (2007) comparing the conditional power, predictive power, and Bayesian predictive approaches



A Phase 2 Trial with a Bayesian Goldilocks Design

Background

- A Phase 2 trial was planned to enroll 90 to 150 patients randomized 1:1:1 to two active doses and control
- The primary endpoint was a hierarchical composite endpoint
- A Goldilocks adaptive design was adopted and assessed by “intensive” simulation
  - Early stopping for efficacy
  - Non-binding early stopping for futility
- Assumptions for design were highly variable

Interim and Final Analyses

- Active groups combined was compared with control group
- The non-parametric test of Finkelstein and Schoenfeld was applied for the primary hierarchical composite endpoint
- A flexible number of interim analyses determined by enrollment rate
- Control of type 1 error

Bayesian Modeling and Prediction

- Modeling: components in the composite were independently modeled
- Bayesian analysis: only complete cases were considered for interim analysis, i.e. to derive posterior distributions for the components
- Prediction: posterior distributions of the components were used to predict outcomes for incomplete cases and those yet to be enrolled
- Thresholds for early stoppings were set
  - Fixed for futility
  - Variable for efficacy due to control of type 1 error, depending on the actual number of IAs

Simulations

- Thresholds for early stoppings for efficacy were identified through simulation

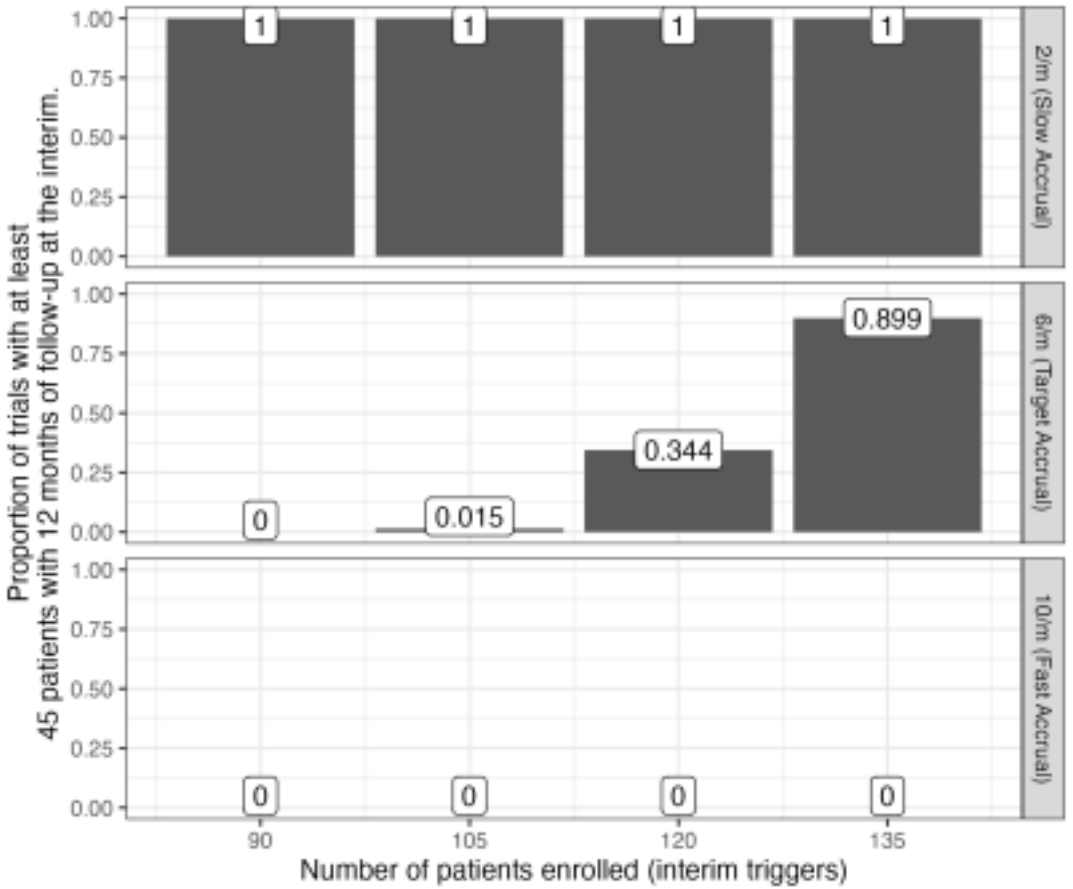
First interim occurrence	S_{90}	S_{105}	S_{120}	S_{135}	Threshold for FS test p-value
$n = 90$	0.980	0.980	0.980	0.980	0.0235
$n = 105$	–	0.975	0.975	0.975	0.0235
$n = 120$	–	–	0.970	0.970	0.0235
$n = 135$	–	–	–	0.965	0.0235
None	–	–	–	–	0.0250

- Simulations were conducted for various scenarios of enrollment rate and effect assumptions for the assessment of
  - Distribution of the number of interim analyses
  - Probability of early futility
  - probability of early efficacy
  - Overall sample size
  - Power

Operating Characteristics

– Impact of enrollment rate on the number of interim analyses

- The number of interim analyses depends on the enrollment rate
- Not necessary to consider fixed event rates for various enrollment scenarios
- Instead
  - can apply an enrollment model with certain priors
  - the model can be updated based on interim data



Operating Characteristics

Impact of enrollment rate and effect assumptions on power, prob of early efficacy, prob of early futility

- Simulations conducted for various scenarios of effect assumptions
- Not necessary to assume fixed scenarios
- Instead
  - Can assume prior distributions for all components, i.e. “design priors”
  - While interim analysis is still conducted using non-informative “analysis priors”
- Need to consider how much gain and loss
  - Operational complexities
  - Time and money savings
  - Flexibility for modifications

Accrual	Scenario	Avg n	Prob ES	Prob EF	Power
2/m	Null	113	0.004	0.863	0.0211
	All Moderate	126	0.520	0.058	0.860
	Variation 1	133	0.208	0.238	0.516
	Variation 2	104	0.906	0.003	0.992
	Variation 3	132	0.158	0.302	0.428
	Variation 4	120	0.017	0.711	0.083
	Variation 5	101	0.944	0.001	0.998
6/m	Null	143	0.001	0.343	0.0228
	All Moderate	146	0.210	0.007	0.875
	Variation 1	148	0.067	0.035	0.536
	Variation 2	139	0.537	0.000	0.995
	Variation 3	148	0.055	0.048	0.443
	Variation 4	145	0.005	0.232	0.089
	Variation 5	138	0.613	0.000	0.997
10/m	Null	150	0.000	0.000	0.0248
	All Moderate	150	0.000	0.000	0.884
	Variation 1	150	0.000	0.000	0.550
	Variation 2	150	0.000	0.000	0.995
	Variation 3	150	0.000	0.000	0.456
	Variation 4	150	0.000	0.000	0.094
	Variation 5	150	0.000	0.000	0.998

Epilogue

03

A Prediction by Seymour Geisser (1929-2004)

"While Jennison and Turnbull, in my view, correctly anticipate the increased use of Bayesian methods for in-house studies in drug development programs, they are, I believe, mistaken in their view that frequentist requirements will remain the fundamental basis for studies to demonstrate efficacy and safety to the public. I believe the Bayesian approach has already supplemented the frequentist approach and sooner or later will entirely supplant it."

In: Geisser S. (1992). *On the curtailment of sampling. Canadian* Journal of Statistics.

Some Parting Thoughts

- Geisser's prediction has been considerably fulfilled
- One area to find greater impact of Bayesian Statistics: Quantitative Decision Making
  - Uses communicative and comprehensible concepts as Bayesian prediction, PoS, PoTS, utility index
  - Covers most (if not all) decision problems in drug development, including
    - Optimization of design, analysis, and interpretation of clinical trials
    - Interim trial decisions
    - Development phase transitions
    - Selection of the most promising drug candidates for advancement
    - Effectiveness comparisons with external competitors

References

Selected publications

- Geisser S. (1992). *On the curtailment of sampling. Canadian* Journal of Statistics.
- Dimitris A and Wang M-D. (2006). *Bayesian predictive approach to interim monitoring in clinical trials.* Statistics in Medicine.
- Wang M-D. (2007). *Sample size reestimation by Bayesian prediction.* Bimetrical Journal.
- Kristine R et al. (2014). *Not too big, not too small: a Goldilocks approach to sample size selection.* Journal of Biopharmaceutical Statistics.
- Ozen A et al. (2024). *Evaluating the efficacy and safety of pozelimab in patients with CD55 deficiency with hyperactivation of complement, angiopathic thrombosis, and protein-losing enteropathy disease: an open-label phase 2 and 3 study.* The Lancet.
- U.S. FDA grants approval for Jivi® antihemophilic factor (recombinant), PEGylated-aucl in pediatric patients 7 to under 12 years of age with hemophilia A (congenital Factor VIII deficiency)

Health for all, Hunger for none

Thank you!

bayer.com