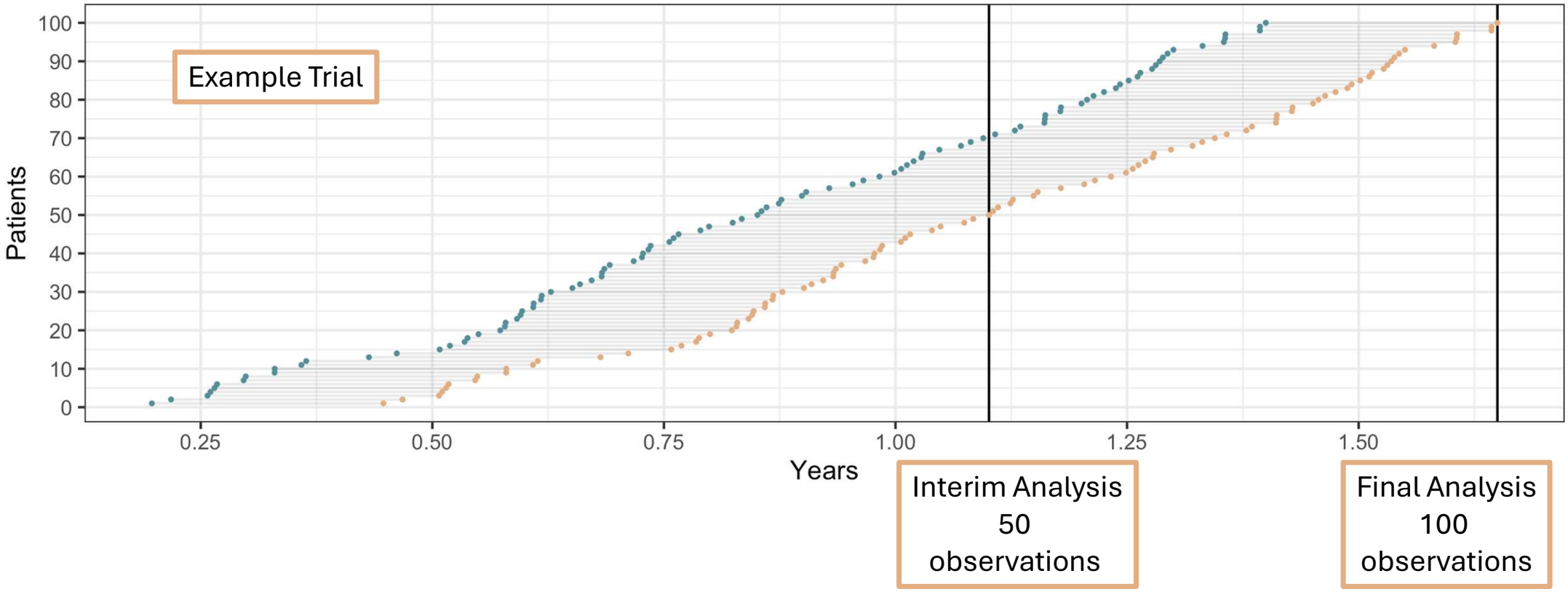
Forecasting with Confidence: Harnessing Predictive Probabilities in Adaptive Clinical Trial Design

Cora Allen-Savietta
Berry Consultants

thanks to my co-authors
Joe Marion, Liz Lorenzi, Kert Viele, & Scott Berry

Berry Consultants

Predicting Future Outcomes from Current Data



Given 50 observed patients, what is the probability of success at 100?

What do current data show?

	posterior probability	conditional power	predictive probability
Assumptions	incorporates prior information	frequentist calculation, no priors	incorporates prior information
Information	currently observed data	currently observed data	currently observed data
Goal	summarizes current information + prior	predicts trial success assuming a **precise** future treatment effect	predicts trial success based on a **distribution of possible future treatment effects**

Berry Consultants

Saville, Detry, & Viele 2023

Given observed interim data, how likely is a win if all future data show an assumed treatment effect?

	posterior probability	conditional power	predictive probability
Assumptions	incorporates prior information	typically frequentist, no priors	incorporates prior information
Information	currently observed data	currently observed data	currently observed data
Goal	summarizes current information + prior	predicts trial success assuming a **single** future treatment effect	predicts trial success based on a **distribution of possible future treatment effects**

Berry Consultants

Saville, Detry, & Viele 2023

Given the observed data and distribution of treatment effects, how likely is a win?

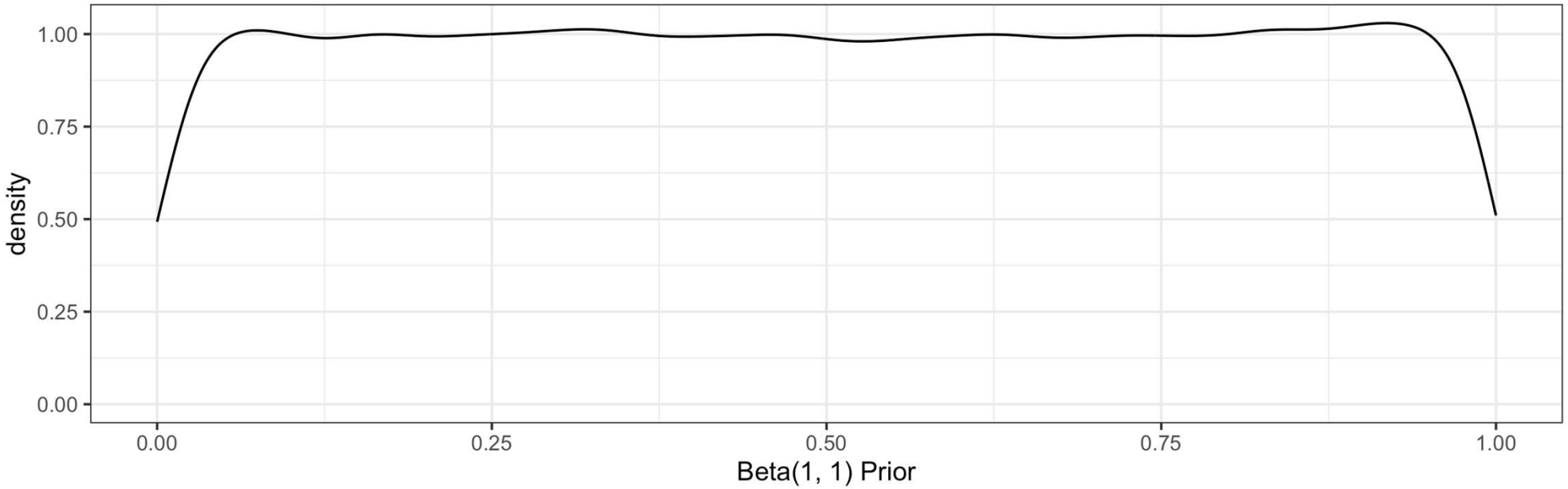
	posterior probability	conditional power	predictive probability
Assumptions	incorporates prior information	typically frequentist, no priors	incorporates prior information
Information	currently observed data	currently observed data	currently observed data
Goal	summarizes current information + prior	predicts trial success assuming a **precise** future treatment effect	predicts trial success based on a **distribution of possible future treatment effects**

Saville, Detry, & Viele 2023

Computing Predictive Probabilities – Closed Form

centered at prior estimate

$$\theta \backsim \text{Beta}(\alpha, \beta)$$



Bayesian Adaptive Methods for Clinical Trials. Berry et al. 2011

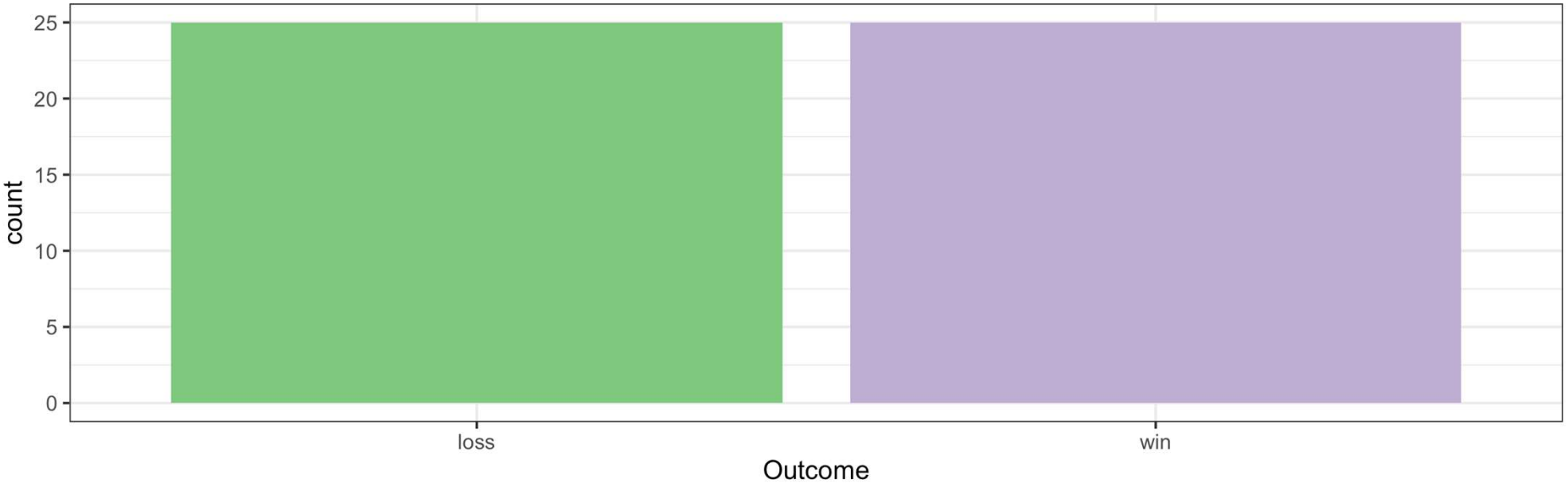
Computing Predictive Probabilities – Closed Form

centered at prior estimate

$$\theta \backsim \text{Beta}(\alpha, \beta)$$

observed data at N = 50
25 wins, 25 failures

$$x_1 \backsim \text{Binomial}(n_1, \theta)$$



Bayesian Adaptive Methods for Clinical Trials. Berry et al. 2011

Computing Predictive Probabilities – Closed Form

centered at prior estimate

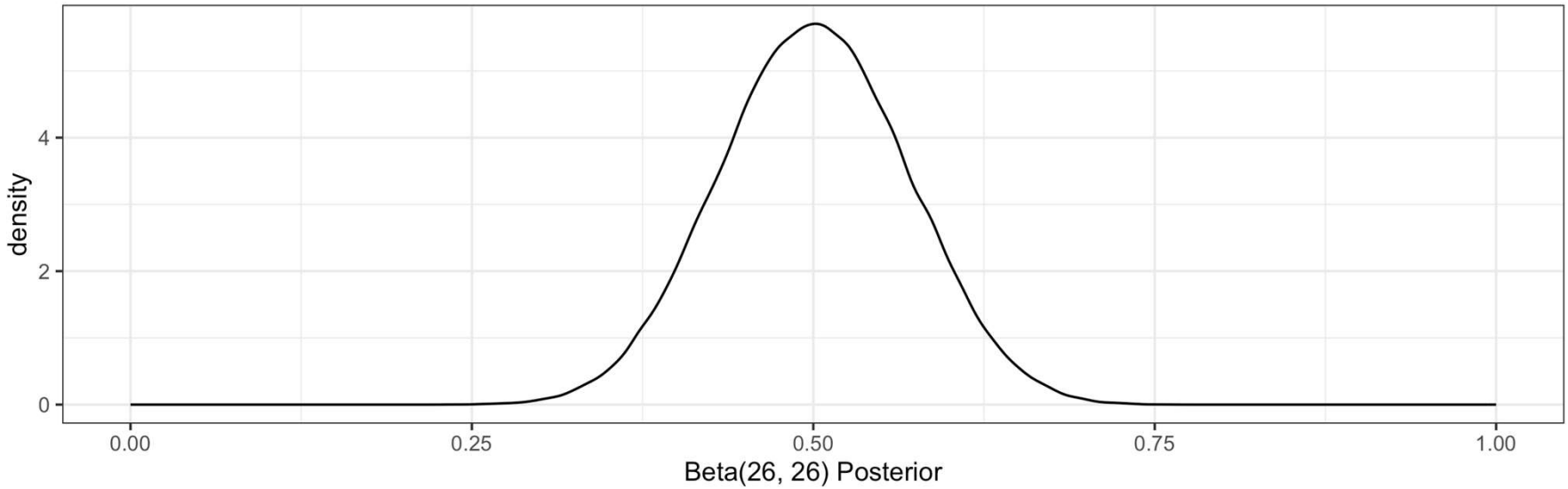
$$\theta \backsim \text{Beta}(\alpha, \beta)$$

observed data at N = 50
25 wins, 25 failures

$$x_1 \backsim \text{Binomial}(n_1, \theta)$$

posterior distribution

$$\theta | x_1, n_1 \backsim \text{Beta}(\alpha + x_1, \beta + n_1 - x_1)$$



Bayesian Adaptive Methods for Clinical Trials. Berry et al. 2011

Computing Predictive Probabilities – Closed Form

centered at prior estimate

$$\theta \backsim \text{Beta}(\alpha, \beta)$$

observed data at N = 50
25 wins, 25 failures

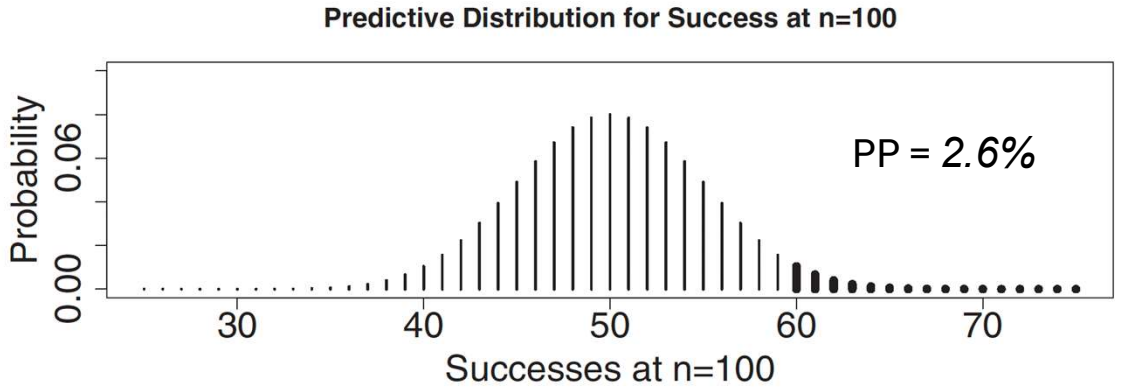
$$x_1 \backsim \text{Binomial}(n_1, \theta)$$

posterior distribution

$$\theta | x_1, n_1 \backsim \text{Beta}(\alpha + x_1, \beta + n_1 - x_1)$$

predictive distribution for next n_2 observations

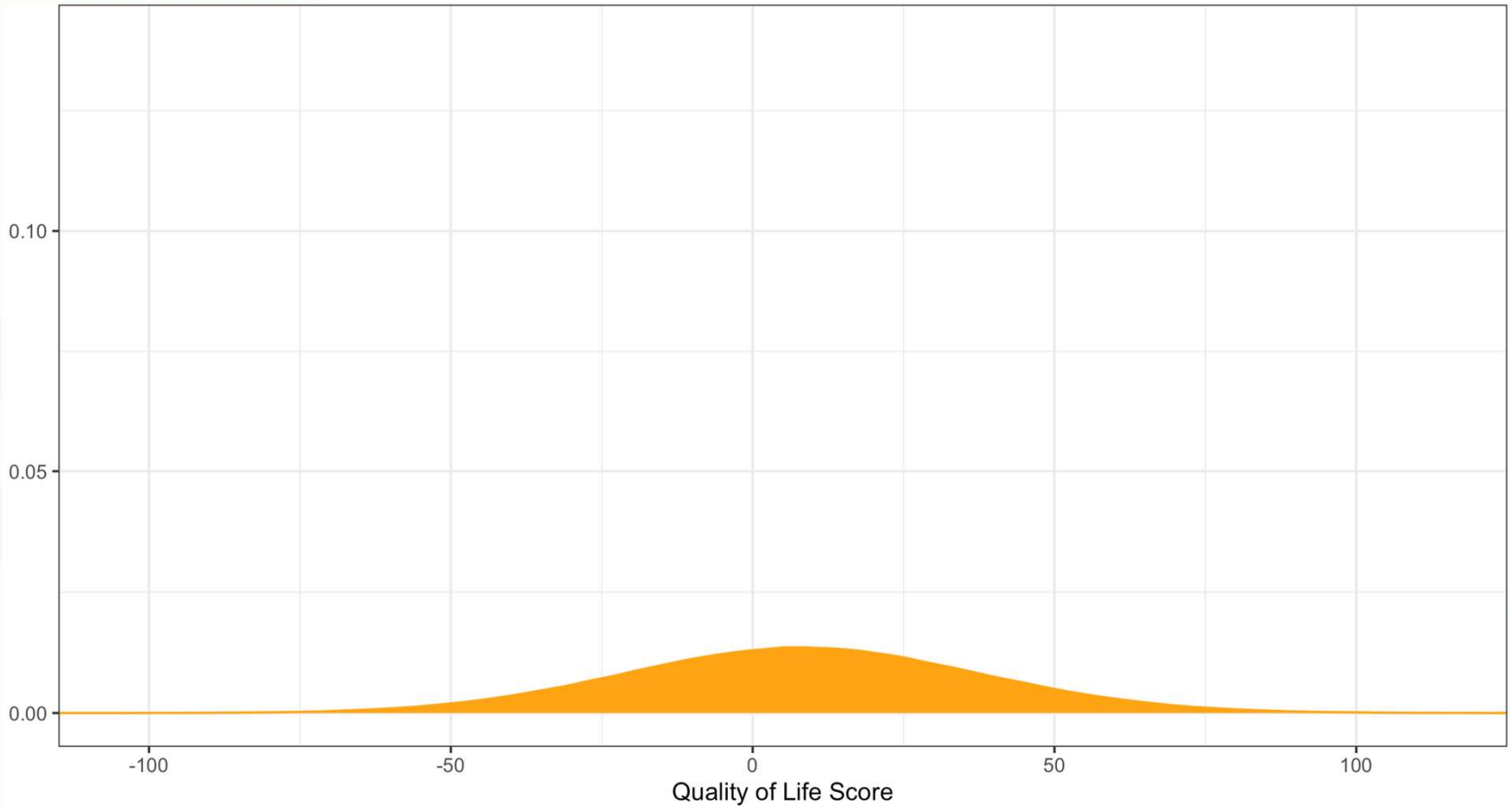
$$\text{x}_2 \mid n_1, \alpha + x, \beta + n_1 - x_1 \backsim \text{Beta-Binomial}(n_2, \alpha + x_1, \beta + n_1 - x_1)$$



Bayesian Adaptive Methods for Clinical Trials. Berry et al. 2011

Calculating a Predictive Probability of Success: Monte Carlo Integration

prior information
- clinical expertise
- previous studies
- purposefully diffuse

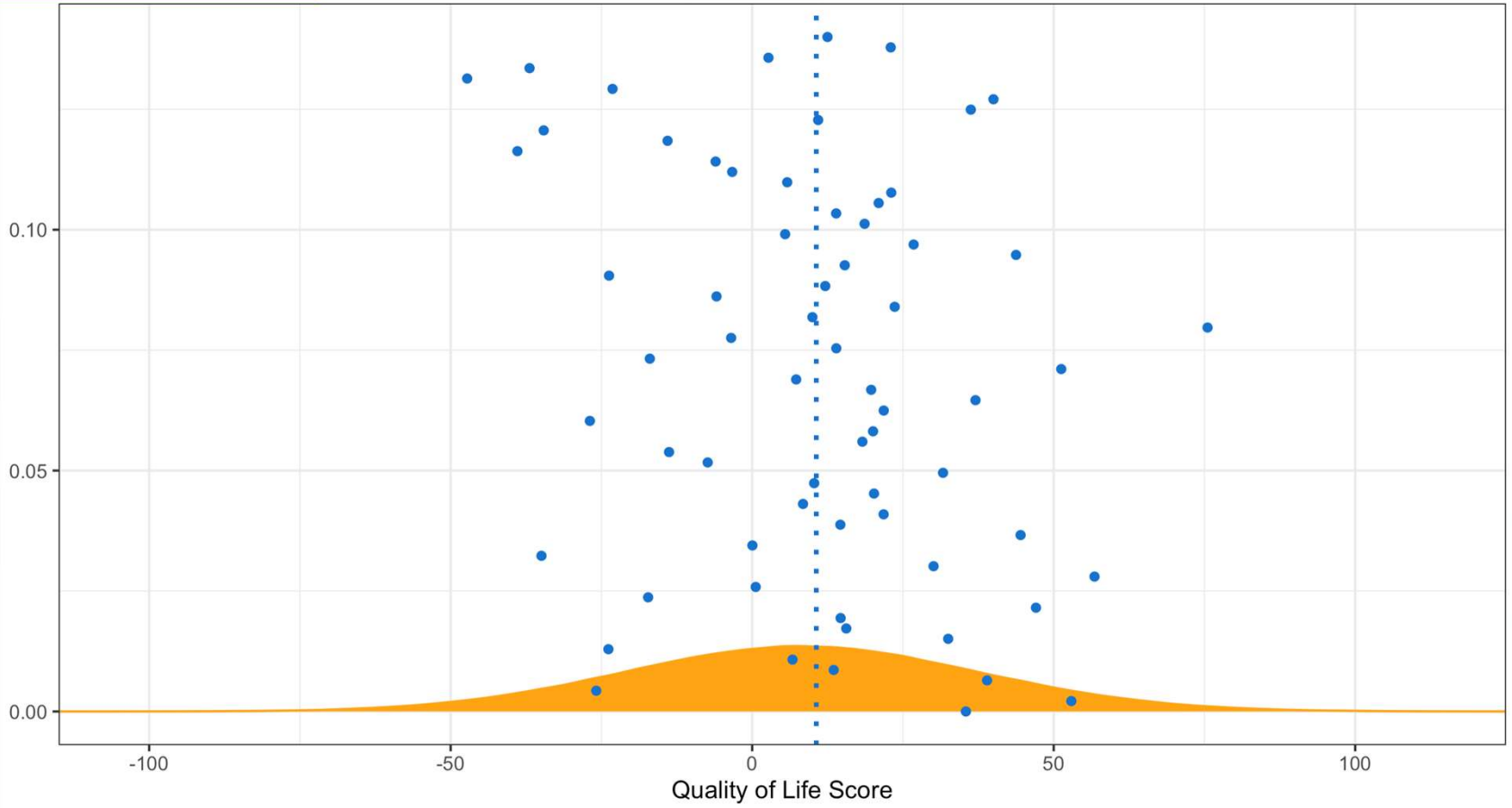


Calculating a Predictive Probability of Success

prior information
- clinical expertise
- previous studies
- purposefully diffuse

+

interim observed data



Calculating a Predictive Probability of Success

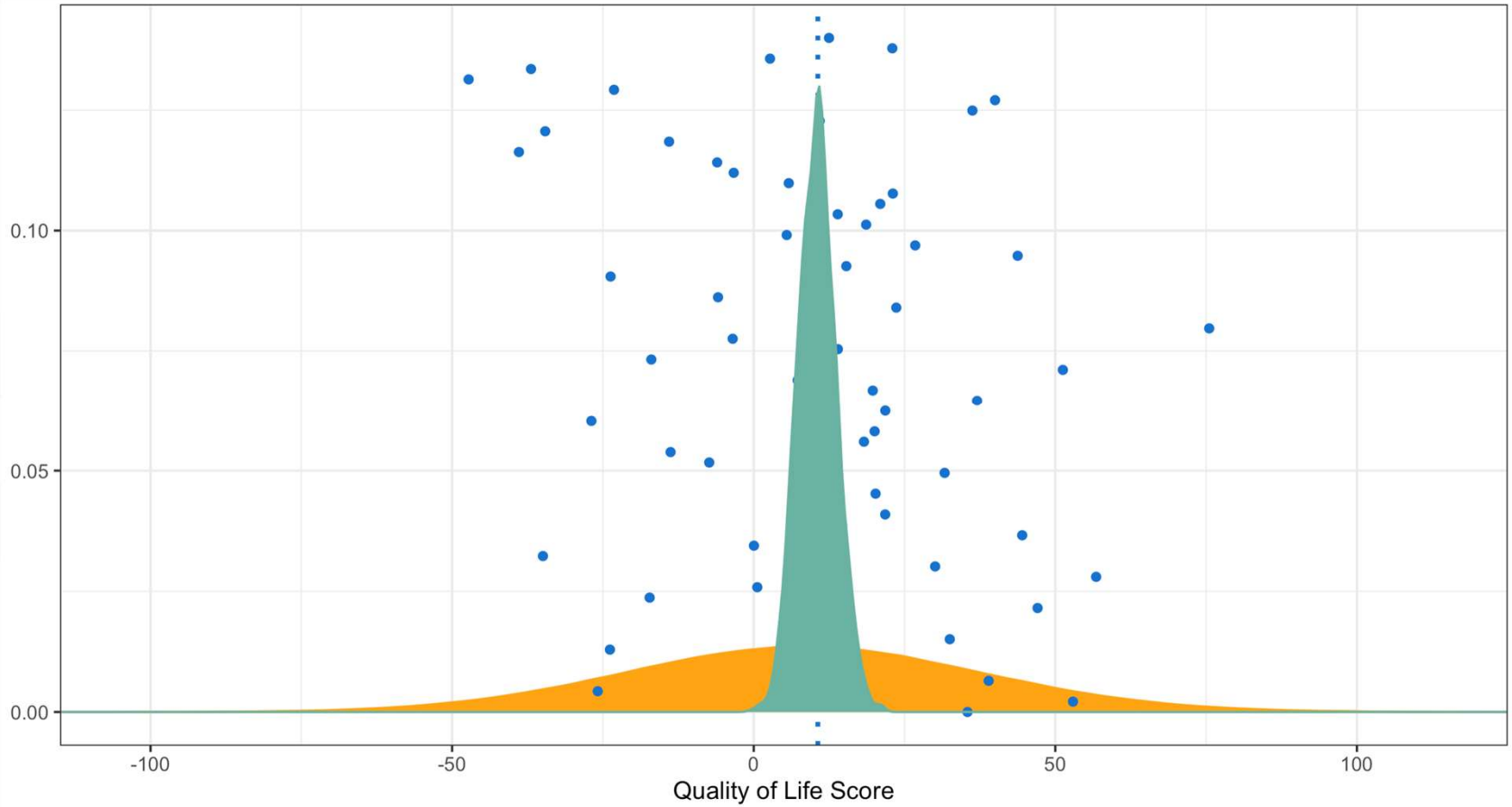
prior information
- clinical expertise
- previous studies
- purposefully diffuse

+

interim observed data

=

posterior distribution of the mean



Calculating a Predictive Probability of Success

prior information
- clinical expertise
- previous studies
- purposefully diffuse

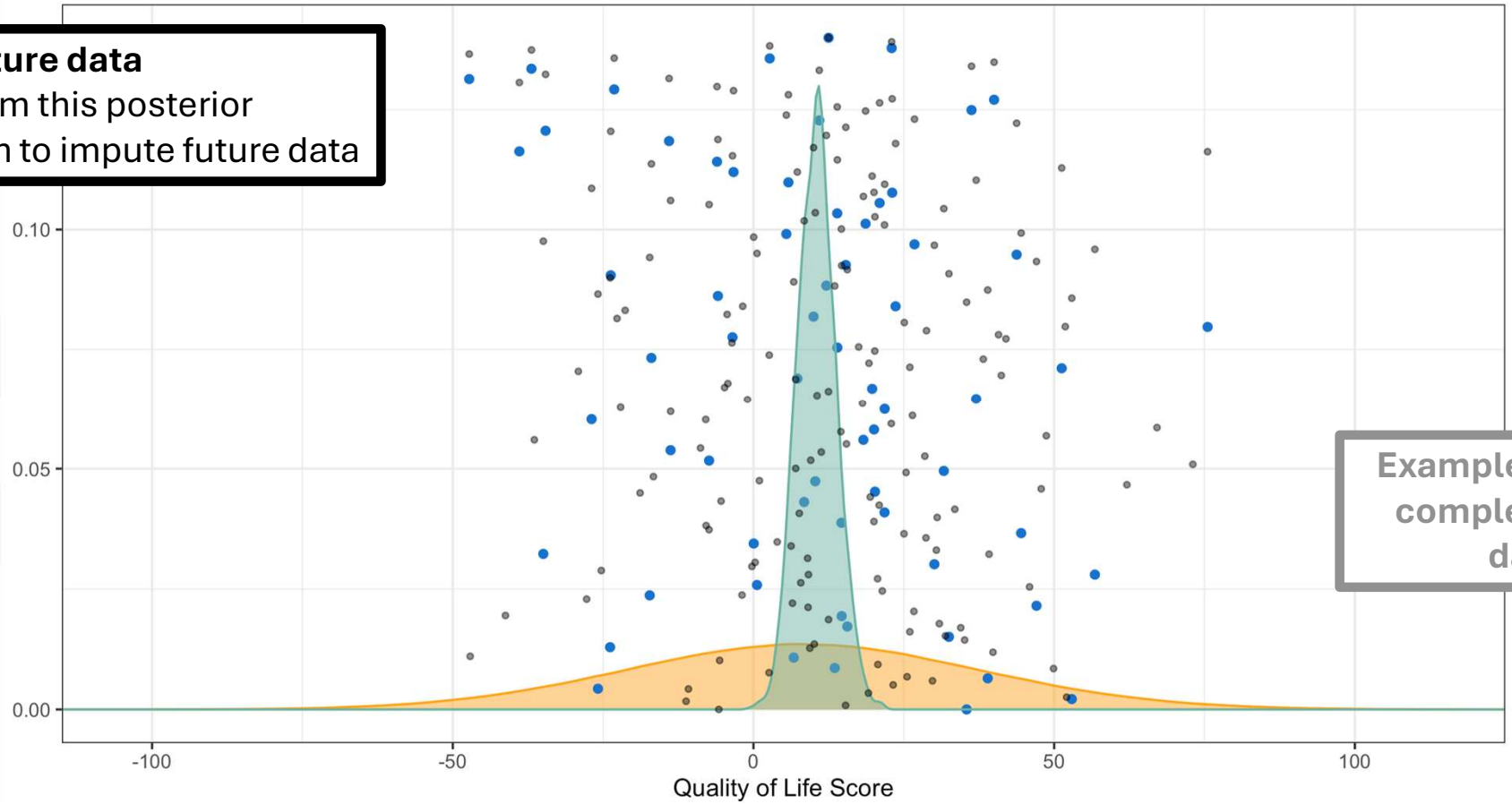
+

interim observed data

=

posterior distribution of the mean

impute future data
sample from this posterior distribution to impute future data



Example imputed completed trial data

Calculating a Predictive Probability of Success

prior information
- clinical expertise
- previous studies
- purposefully diffuse

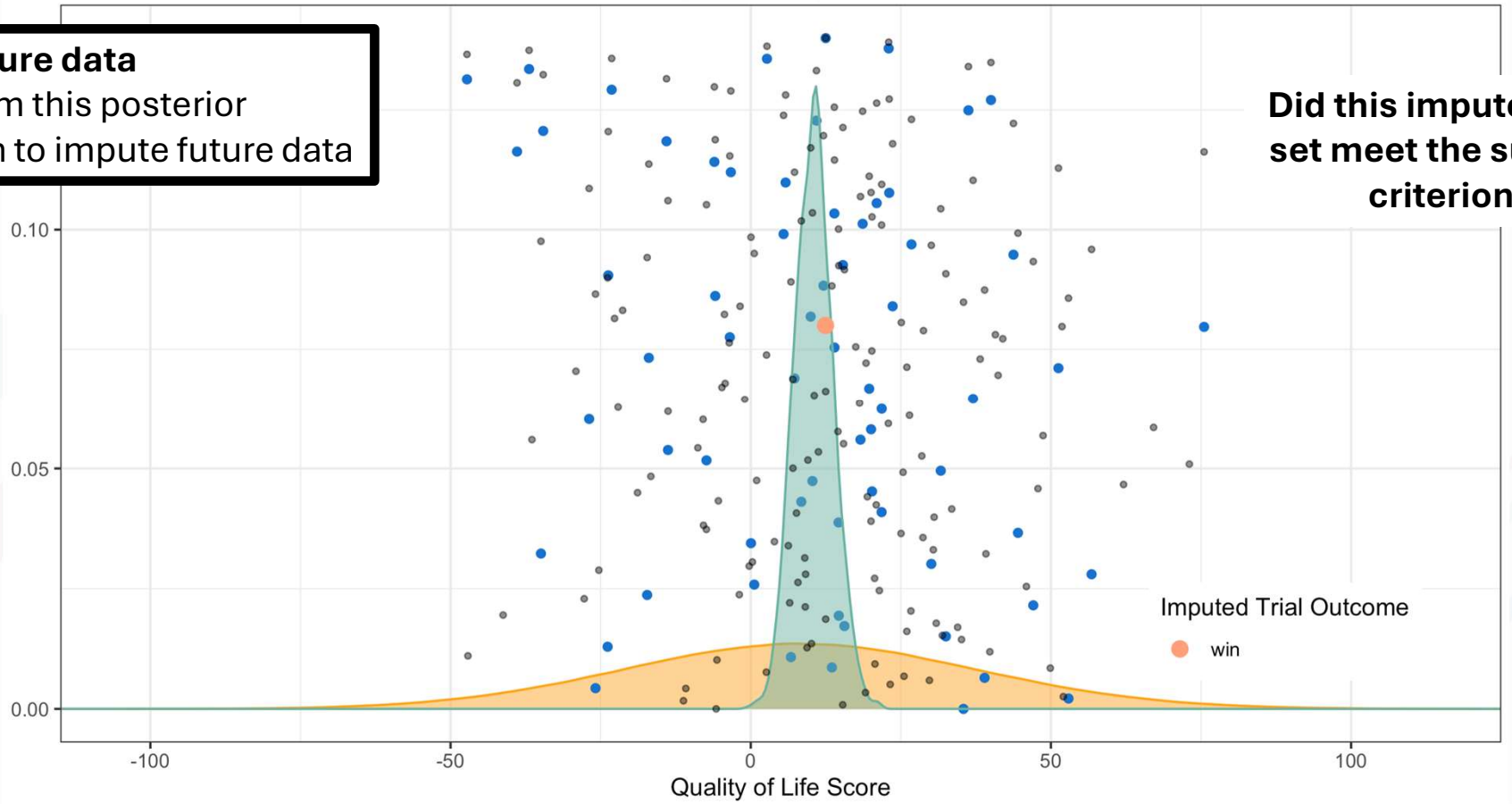
+

interim observed data

=

posterior distribution of the mean

impute future data
sample from this posterior distribution to impute future data



Did this imputed data set meet the success criterion?

Calculating a Predictive Probability of Success

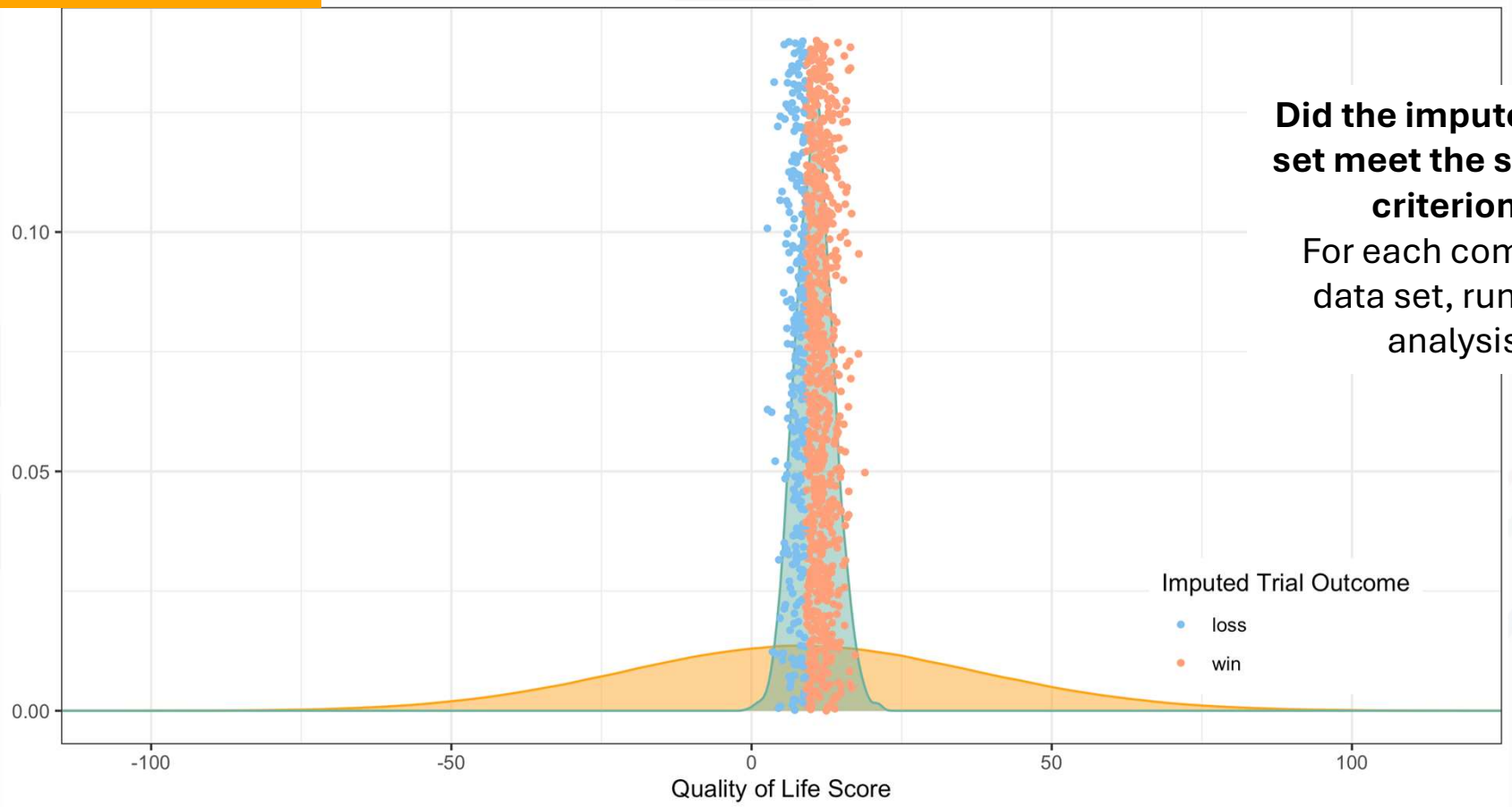
prior information
- clinical expertise
- previous studies
- purposefully diffuse

+

interim observed data

=

posterior distribution of the mean



Did the imputed data set meet the success criterion?
For each complete data set, run final analysis

Calculating a Predictive Probability of Success

prior information
- clinical expertise
- previous studies
- purposefully diffuse

+

interim observed data

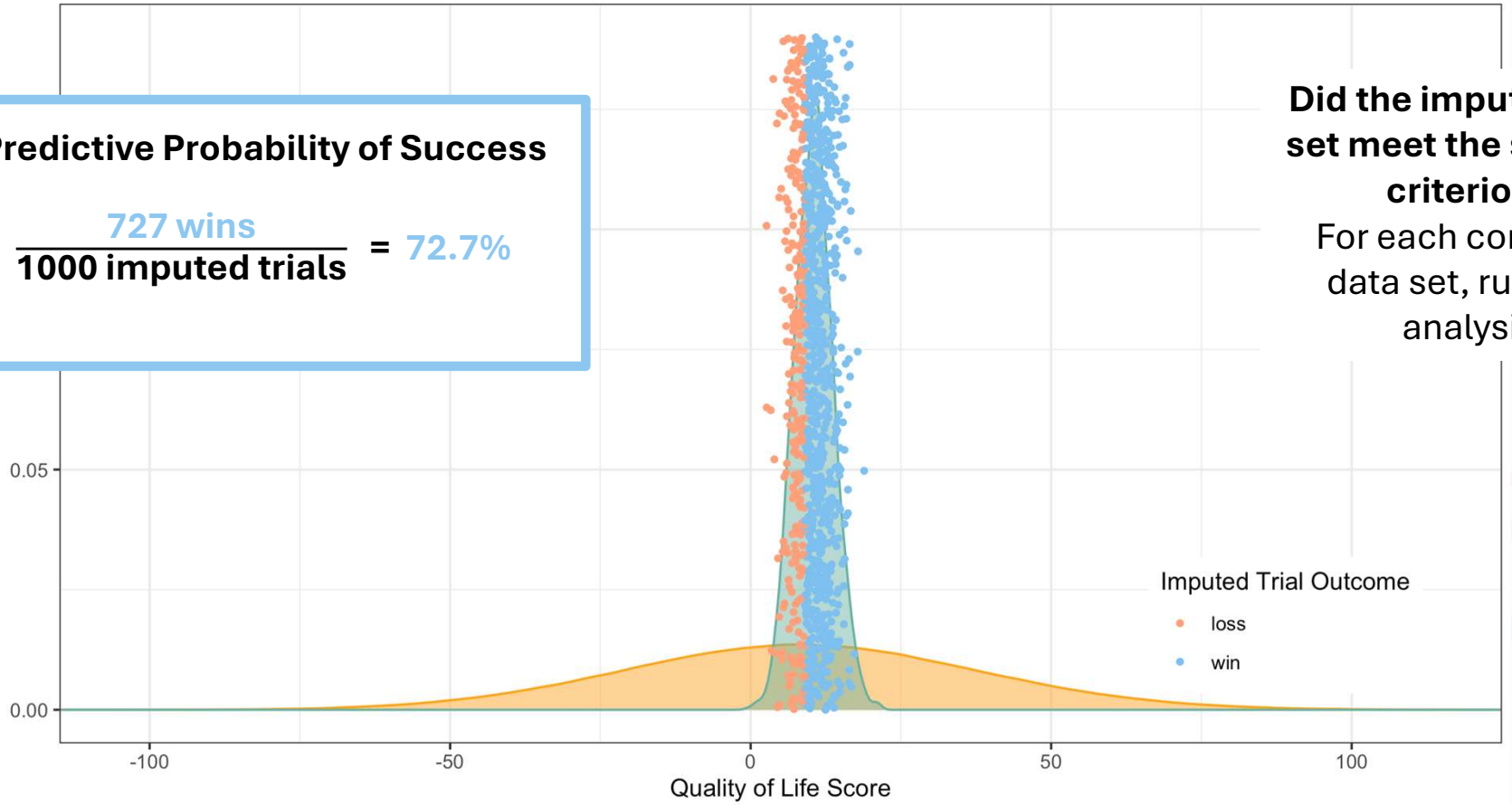
=

posterior distribution of the mean

Predictive Probability of Success

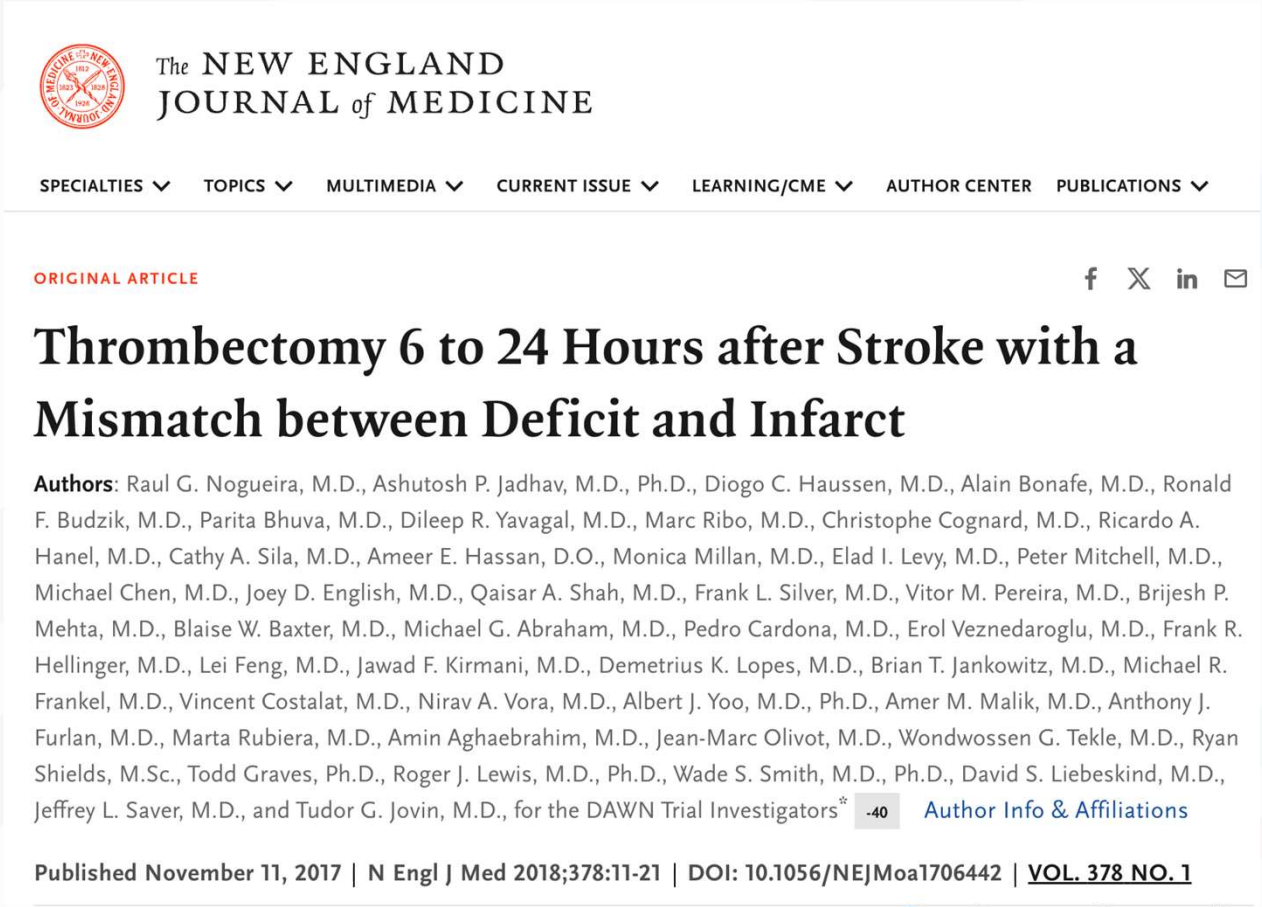
$$\frac{\text{727 wins}}{\textbf{1000 imputed trials}} = 72.7\%$$

Did the imputed data set meet the success criterion?
For each complete data set, run final analysis



When would we need predictive probabilities?

- To **choose a sample size** at a prespecified interim analysis



The NEW ENGLAND JOURNAL of MEDICINE

SPECIALTIES TOPICS MULTIMEDIA CURRENT ISSUE LEARNING/CME AUTHOR CENTER PUBLICATIONS

ORIGINAL ARTICLE

Thrombectomy 6 to 24 Hours after Stroke with a Mismatch between Deficit and Infarct

Authors: Raul G. Nogueira, M.D., Ashutosh P. Jadhav, M.D., Ph.D., Diogo C. Haussen, M.D., Alain Bonafe, M.D., Ronald F. Budzik, M.D., Parita Bhuva, M.D., Dileep R. Yavagal, M.D., Marc Ribo, M.D., Christophe Cognard, M.D., Ricardo A. Hanel, M.D., Cathy A. Sila, M.D., Ameer E. Hassan, D.O., Monica Millan, M.D., Elad I. Levy, M.D., Peter Mitchell, M.D., Michael Chen, M.D., Joey D. English, M.D., Qaisar A. Shah, M.D., Frank L. Silver, M.D., Vitor M. Pereira, M.D., Brijesh P. Mehta, M.D., Blaise W. Baxter, M.D., Michael G. Abraham, M.D., Pedro Cardona, M.D., Erol Veznedaroglu, M.D., Frank R. Hellinger, M.D., Lei Feng, M.D., Jawad F. Kirmani, M.D., Demetrius K. Lopes, M.D., Brian T. Jankowitz, M.D., Michael R. Frankel, M.D., Vincent Costalat, M.D., Nirav A. Vora, M.D., Albert J. Yoo, M.D., Ph.D., Amer M. Malik, M.D., Anthony J. Furlan, M.D., Marta Rubiera, M.D., Amin Aghaebrahim, M.D., Jean-Marc Olivot, M.D., Wondwossen G. Tekle, M.D., Ryan Shields, M.Sc., Todd Graves, Ph.D., Roger J. Lewis, M.D., Ph.D., Wade S. Smith, M.D., Ph.D., David S. Liebeskind, M.D., Jeffrey L. Saver, M.D., and Tudor G. Jovin, M.D., for the DAWN Trial Investigators* -40 Author Info & Affiliations

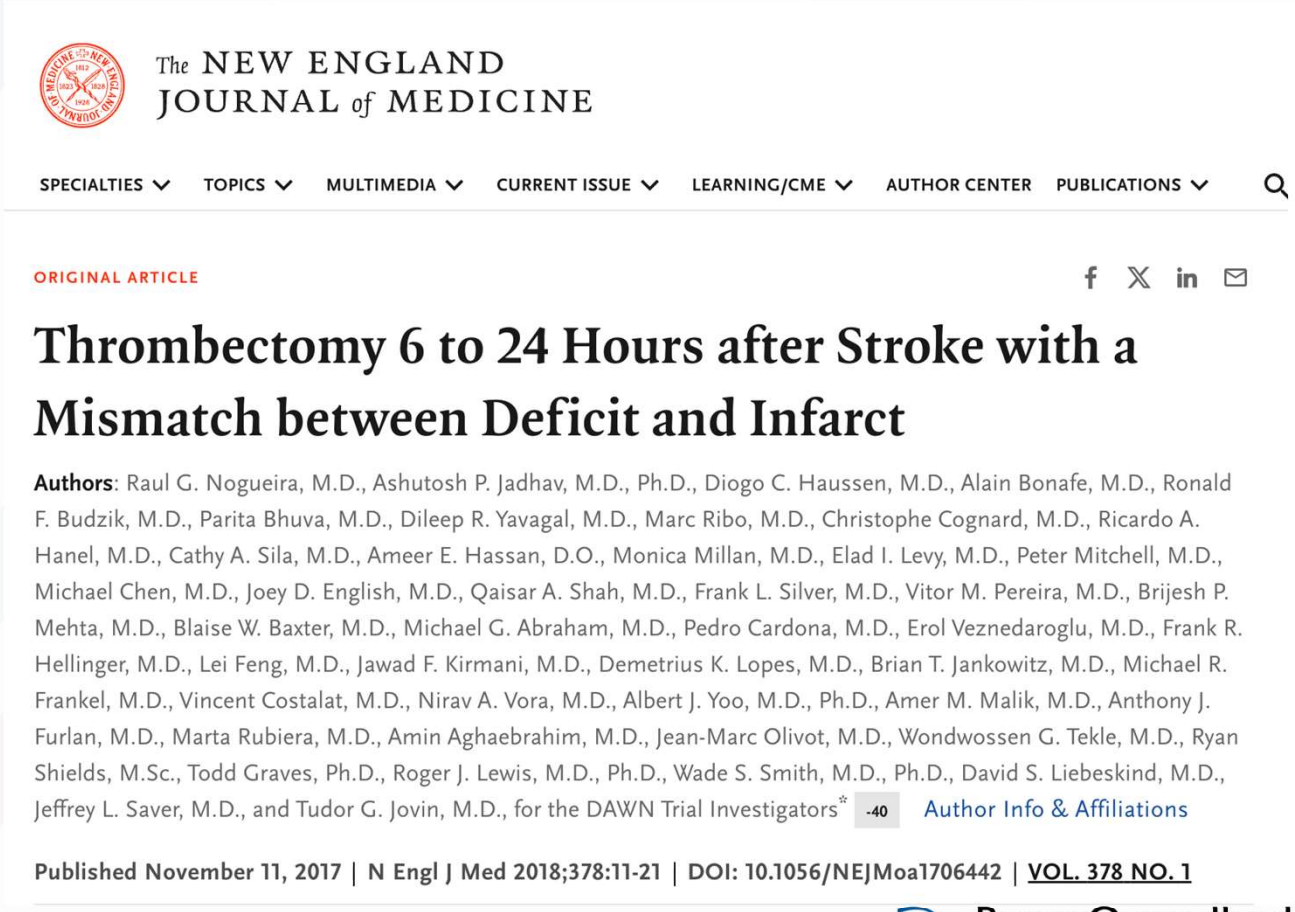
Published November 11, 2017 | N Engl J Med 2018;378:11-21 | DOI: 10.1056/NEJMoa1706442 | VOL. 378 NO. 1

STATISTICAL ANALYSIS

The adaptive trial design allowed for a sample size ranging from 150 to 500 patients. During interim analyses, the decision to stop or continue enrollment was based on a prespecified calculation of the probability that thrombectomy plus standard care would be superior to standard care alone with respect to the first primary end point. The enrichment trial design gave us the flexibility to identify whether the benefit of the trial intervention was restricted to a subgroup of patients with relatively small infarct volumes at baseline. The interim analyses, which included patients with available follow-up data at the time of the analysis, were prespecified to test for the futility, enrichment, and success of the trial.

When would we need predictive probabilities?

- To **identify subgroups** benefiting most from a treatment



The NEW ENGLAND JOURNAL of MEDICINE

SPECIALTIES TOPICS MULTIMEDIA CURRENT ISSUE LEARNING/CME AUTHOR CENTER PUBLICATIONS

ORIGINAL ARTICLE

Thrombectomy 6 to 24 Hours after Stroke with a Mismatch between Deficit and Infarct

Authors: Raul G. Nogueira, M.D., Ashutosh P. Jadhav, M.D., Ph.D., Diogo C. Haussen, M.D., Alain Bonafe, M.D., Ronald F. Budzik, M.D., Parita Bhuva, M.D., Dileep R. Yavagal, M.D., Marc Ribo, M.D., Christophe Cognard, M.D., Ricardo A. Hanel, M.D., Cathy A. Sila, M.D., Ameer E. Hassan, D.O., Monica Millan, M.D., Elad I. Levy, M.D., Peter Mitchell, M.D., Michael Chen, M.D., Joey D. English, M.D., Qaisar A. Shah, M.D., Frank L. Silver, M.D., Vitor M. Pereira, M.D., Brijesh P. Mehta, M.D., Blaise W. Baxter, M.D., Michael G. Abraham, M.D., Pedro Cardona, M.D., Erol Veznedaroglu, M.D., Frank R. Hellinger, M.D., Lei Feng, M.D., Jawad F. Kirmani, M.D., Demetrius K. Lopes, M.D., Brian T. Jankowitz, M.D., Michael R. Frankel, M.D., Vincent Costalat, M.D., Nirav A. Vora, M.D., Albert J. Yoo, M.D., Ph.D., Amer M. Malik, M.D., Anthony J. Furlan, M.D., Marta Rubiera, M.D., Amin Aghaebrahim, M.D., Jean-Marc Olivot, M.D., Wondwossen G. Tekle, M.D., Ryan Shields, M.Sc., Todd Graves, Ph.D., Roger J. Lewis, M.D., Ph.D., Wade S. Smith, M.D., Ph.D., David S. Liebeskind, M.D., Jeffrey L. Saver, M.D., and Tudor G. Jovin, M.D., for the DAWN Trial Investigators* -40 Author Info & Affiliations

Published November 11, 2017 | N Engl J Med 2018;378:11-21 | DOI: 10.1056/NEJMoa1706442 | VOL. 378 NO. 1

STATISTICAL ANALYSIS

The adaptive trial design allowed for a sample size ranging from 150 to 500 patients. During interim analyses, the decision to stop or continue enrollment was based on a prespecified calculation of the probability that thrombectomy plus standard care would be superior to standard care alone with respect to the first primary end point. The enrichment trial design gave us the flexibility to identify whether the benefit of the trial intervention was restricted to a subgroup of patients with relatively small infarct volumes at baseline. The interim analyses, which included patients with available follow-up data at the time of the analysis, were prespecified to test for the futility, enrichment, and success of the trial.

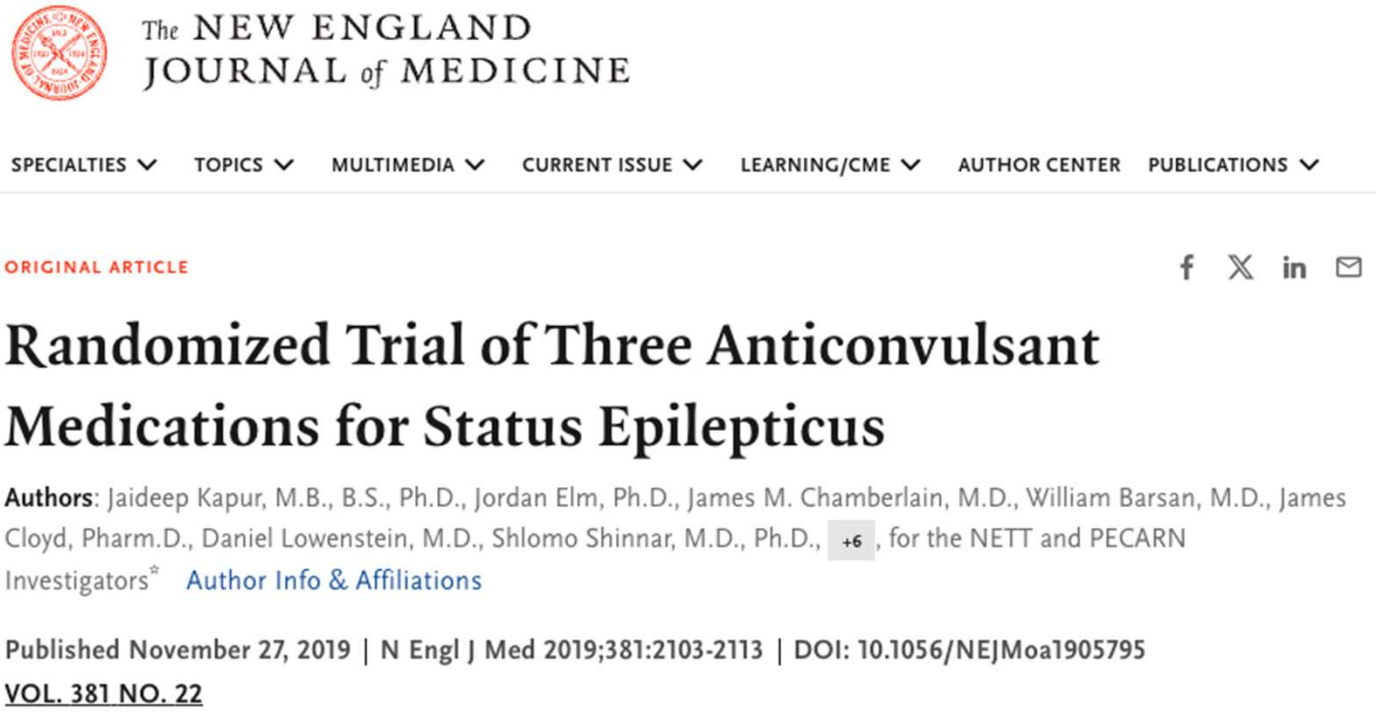
When would we need predictive probabilities?

- To **identify subgroups** benefiting most from a treatment

https://www.fda.gov/news-events/press-announcements/fda-expands-treatment-window-use-clot-retrieval-devices-certain-stroke-patients

When would we want to use a predictive probability?

- To determine if additional data are likely to provide convincing evidence of a treatment effect. In other words, **should the trial stop for futility**?



The NEW ENGLAND JOURNAL of MEDICINE

SPECIALTIES TOPICS MULTIMEDIA CURRENT ISSUE LEARNING/CME AUTHOR CENTER PUBLICATIONS

ORIGINAL ARTICLE

Randomized Trial of Three Anticonvulsant Medications for Status Epilepticus

Authors: Jaideep Kapur, M.B., B.S., Ph.D., Jordan Elm, Ph.D., James M. Chamberlain, M.D., William Barsan, M.D., James Cloyd, Pharm.D., Daniel Lowenstein, M.D., Shlomo Shinnar, M.D., Ph.D., +6, for the NETT and PECARN Investigators* Author Info & Affiliations

Published November 27, 2019 | N Engl J Med 2019;381:2103-2113 | DOI: 10.1056/NEJMoa1905795

VOL. 381 NO. 22

When would we want to use a predictive probability?

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Randomized Trial of Three … Medications for Status Epile…

Authors: Jaideep Kapur, M.B., B.S., Ph.D., Jordan Elm, Ph.D., Ja… Cloyd, Pharm.D., Daniel Lowenstein, M.D., Shlomo Shinnar, M.… Investigators* Author Info & Affiliations

Published November 27, 2019 | N Engl J Med 2019;381:2103-

VOL. 381 NO. 22

Table S6. Computations of the futility analysis

Look	Predictive probability that an arm is identified as best / worst at maximum sample size*			Predictive probability that any arm Wins**
	Levetiracetam	Fosphenytoin	Valproate	
Analysis after 400th Enrollment (N=384 unique subjects)				

* Maximum sample size was assumed to be 720 unique subjects for calculation of the predictive probabilities.

** This represents the sum of the predictive probabilities arm is best/worst at the maximum sample size for each of the 3 groups. If this sum is < 5%, the trial stops for futility.

When would we want to use a predictive probability?

The NEW ENGLAND JOURNAL of MEDICINE

SPECIALTIES TOPICS MULTIMEDIA CURRENT ISSUE

ORIGINAL ARTICLE

Randomized Trial of Three … Medications for Status Epile…

Authors: Jaideep Kapur, M.B., B.S., Ph.D., Jordan Elm, Ph.D., Ja… Cloyd, Pharm.D., Daniel Lowenstein, M.D., Shlomo Shinnar, M.… Investigators* Author Info & Affiliations

Published November 27, 2019 | N Engl J Med 2019;381:2103-

VOL. 381 NO. 22

Table S6. Computations of the futility analysis

Look	Predictive probability that an arm is identified as best / worst at maximum sample size*			Predictive probability that any arm Wins**
	Levetiracetam	Fosphenytoin	Valproate	
Analysis after 400th Enrollment (N=384 unique subjects)	.0013 / .0008	.002 / .0027	.0022 / .0013	0.01

* Maximum sample size was assumed to be 720 unique subjects for calculation of the predictive probabilities.

** This represents the sum of the predictive probabilities arm is best/worst at the maximum sample size for each of the 3 groups. If this sum is < 5%, the trial stops for futility.

was not included in the intention-to-treat analysis. In November 2017, enrollment was discontinued at the recommendation of the data and safety monitoring board after the trial met the predefined futility criterion in a planned interim analysis, since there was a 1% chance of showing a most effective or least effective treatment if the trial were to continue to the maximum sample size. Computations for the futility analysis are given in Table S6. A

Computing Bayesian Predictive Probabilities

Computing Bayesian Predictive Probabilities

Analytical Calculation

- Mathematical formula to directly calculate predictive probability
- Feasible when integral has closed form
- Very fast

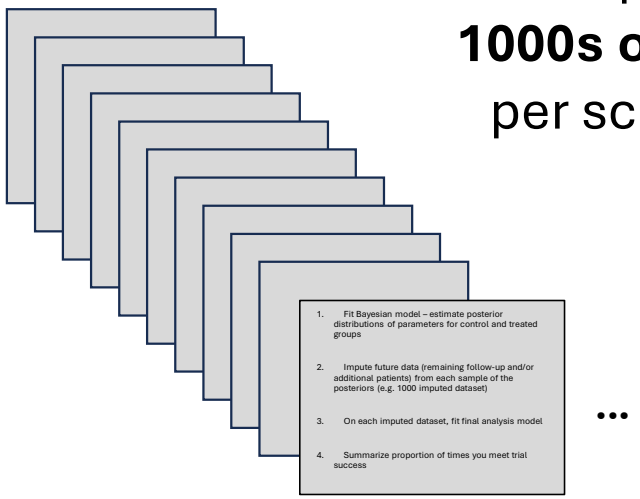
Monte Carlo integration

- if no closed form, requires Monte Carlo integration
- **Can become computationally restrictive**

Monte Carlo Integration for Bayesian Predictive Probabilities

1. Fit Bayesian model
2. Impute future data & fit final analysis model
3. Summarize % meeting trial success

Clinical trial simulations for operating characteristics…



Repeat **1000s of times** per scenario

…

Computing Bayesian Predictive Probabilities

Analytical Calculation

- Mathematical formula to directly calculate predictive probability
- Feasible when integral has closed form
- Very fast

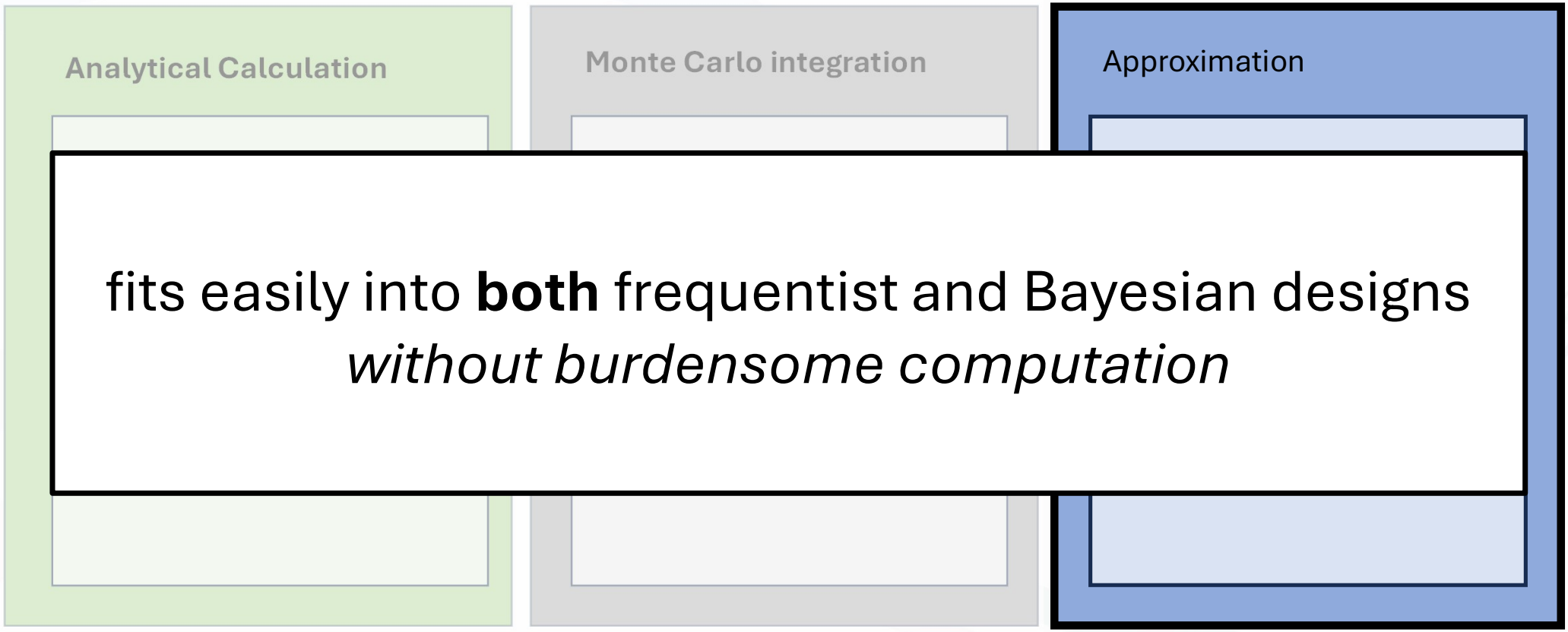
Monte Carlo integration

- if no closed form, requires Monte Carlo integration
- **Can become computationally restrictive**

Approximation

- accurate
- simple and fast

Computing Bayesian Predictive Probabilities



fits easily into **both** frequentist and Bayesian designs
without burdensome computation

Predictive Probability Approximation

Interim analysis

Z_n

- n patients enrolled
- I_n Information
- p_n associated *p*-value

Predictive probability PP_N is probability null hypothesis is rejected if analysis performed at N patients

Rewrite final test statistic as weighted sum of Z_n and Z_{N-n}

$$Z_N\sqrt{I_N} = Z_n\sqrt{I_n} + Z_{N-n}\sqrt{I_N - I_n}$$

Assume uninformative prior distribution $\theta \propto 1$ which yields posterior:

$$\theta|(Z_n = z_n) \sim N\left(z_n/\sqrt{I_n}, 1/I_n\right)$$

Results in predictive distribution for Z_{N-n}

$$Z_{N-n}\,|(Z_n = z_n) \sim N\left(z_n\sqrt{\frac{I_N - I_n}{I_n}}, \frac{I_N}{I_n}\right)$$

Final analysis

Z_N

- N patients enrolled
- I_N Information level
- p_N associated *p*-value

$$PP(p_n, r, \alpha) = \Phi\left(\frac{\Phi^{-1}(1-p_n) - \Phi^{-1}(1-\alpha)\sqrt{r}}{\sqrt{1-r}}\right)$$

Predictive Probability Approximation

$$PP(p_n, r, \alpha) = \Phi\left(\frac{\Phi^{-1}(1-p_n) - \Phi^{-1}(1-\alpha)\sqrt{r}}{\sqrt{1-r}}\right)$$

Requires only:

- p: interim p-value
- n: information at interim
- N: expected information at trial end

Easy-to-use R functions at
github.com/BerryConsultants/approximatePredictiveProbability

Applying the Approximate Predictive Probability

Endpoint	Example analysis	I_n	I_N
Continuous	T-tests ANOVA/ANCOVA	Interim sample size	Final sample size
Binary	z-tests Chi-squared tests	Interim sample size	Final sample size
Time-to-event	Log-rank test Proportional hazards models	Events at interim	Events at final
Ordinal/ Non-parametric	Ordinal regression Wilcoxon rank-sum	Interim sample size	Final sample size
Count data	Generalized linear regressions (e.g. Poisson regression)	Interim exposure	Final exposure

Key Assumptions:

- primary analysis test statistic ~ Normal
- $r = I_n/I_N$ known

Example: Frequentist Binary Endpoint

- Primary Endpoint: Did a participant die by 90 days?
  - Chi-square analysis

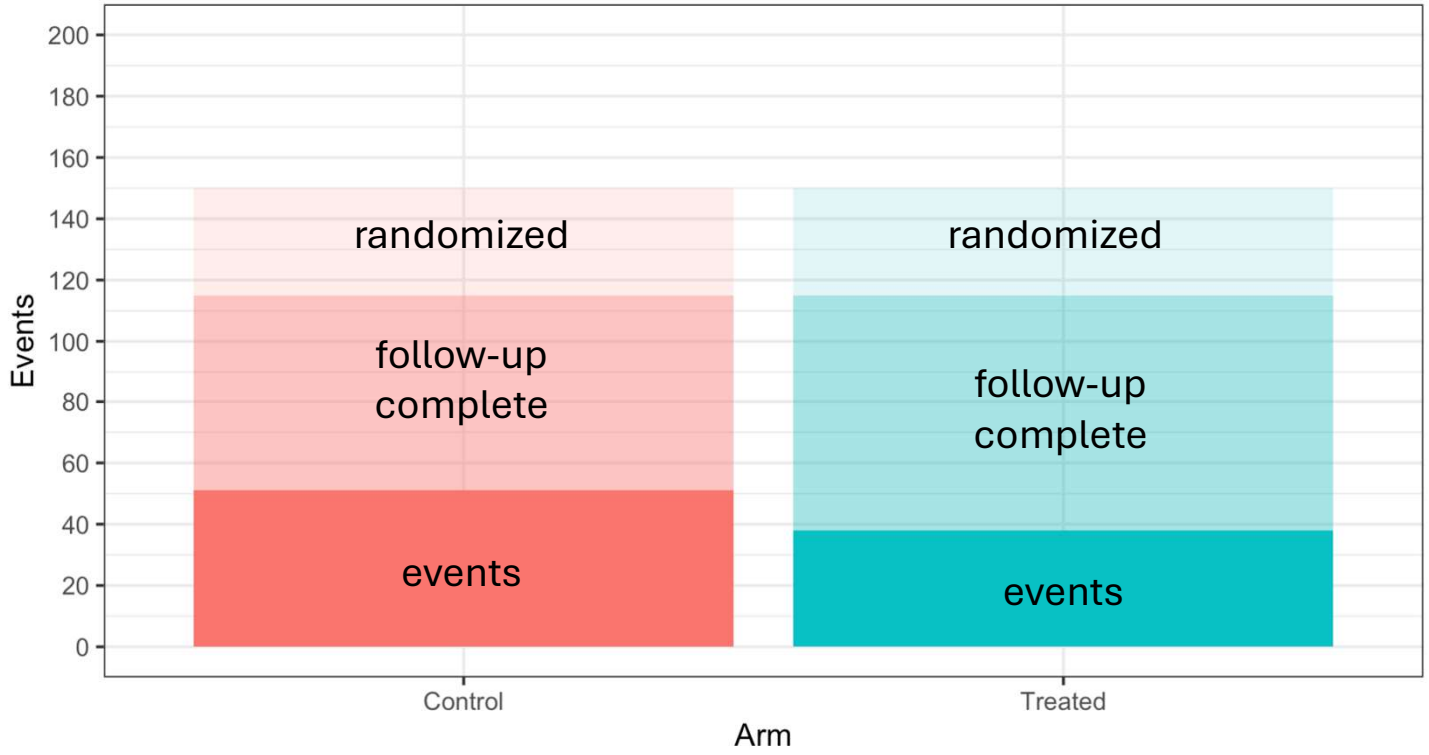
- Maximum Sample Size: 500
  - Interim Goldilocks-style* sample size re-estimations:
    - n = 300 randomized
    - n = 400 randomized

- At each interim, the algorithm can:
  - Stop trial enrollment for expected success at this sample size if $\boldsymbol{PP_n > 90\%}$
  - Stop trial enrollment for futility if $\boldsymbol{PP_{500} < 5\%}$ or
  - Continue trial enrollment

*Broglio et al. 2014 *Not Too Big, Not Too Small: A Goldilocks Approach To Sample Size Selection*

Interim 1: 300 Randomized

Randomized			Follow-up Complete			Events by 90 Days (%)		Predictive Probability of Success at Current N		Predictive Probability of Success at Max N	
Total	Control	Treated	Total	Control	Treated	Control	Treated	PPn	aPPn	PPmaxN	aPPmaxN
300	150	150	230	115	115	0.4435	0.3304	0.5031	0.4940	0.7112	0.7023

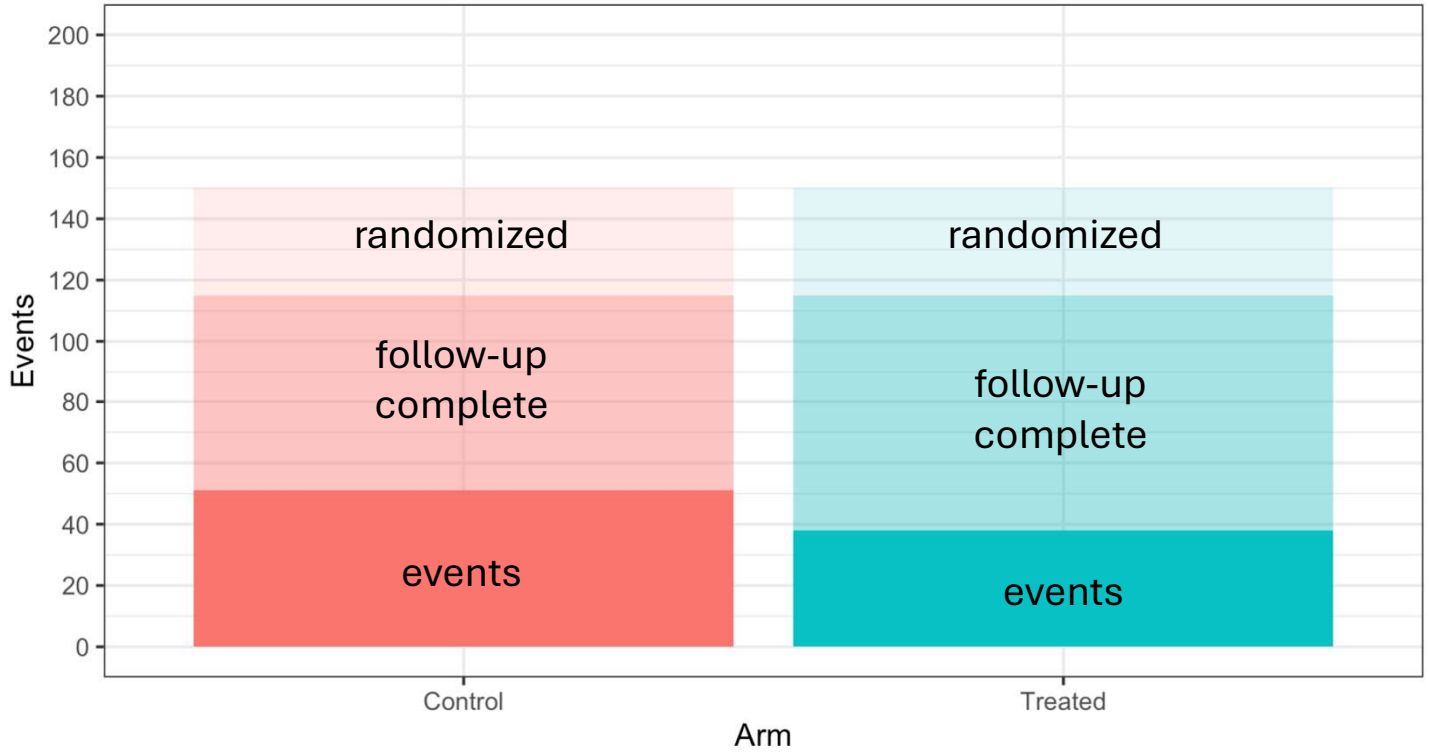


Interim 1: 300 Randomized

Randomized			Follow-up Complete			Events by 90 Days (%)		Predictive Probability of Success at Current N		Predictive Probability of Success at Max N	
Total	Control	Treated	Total	Control	Treated	Control	Treated	PPn	aPPn	PPmaxN	aPPmaxN
300	150	150	230	115	115	0.4435	0.3304	0.5031	0.4940	0.7112	0.7023

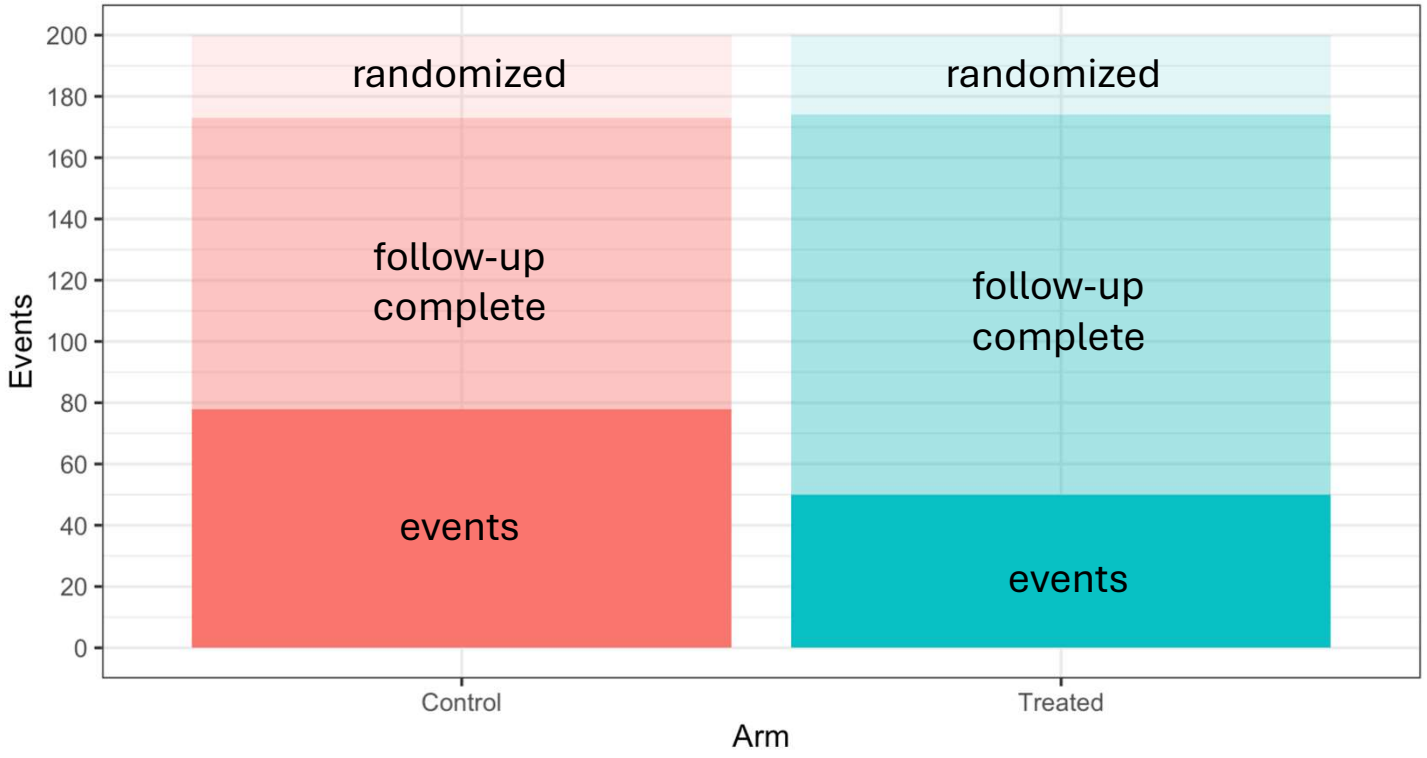
met stopping criteria for expected success? **no**

met stopping criteria for futility? **no**



Interim 2: 400 Randomized

Randomized			Follow-up Complete			Events by 90 Days (%)		Predictive Probability of Success at Current N		Predictive Probability of Success at Max N	
Total	Control	Treated	Total	Control	Treated	Control	Treated	PPn	aPPn	PPmaxN	aPPmaxN
300	150	150	230	115	115	0.4435	0.3304	0.5031	0.4940	0.7112	0.7023
400	200	200	347	173	174	0.4509	0.2874	0.9995	0.9998	0.9973	0.9965

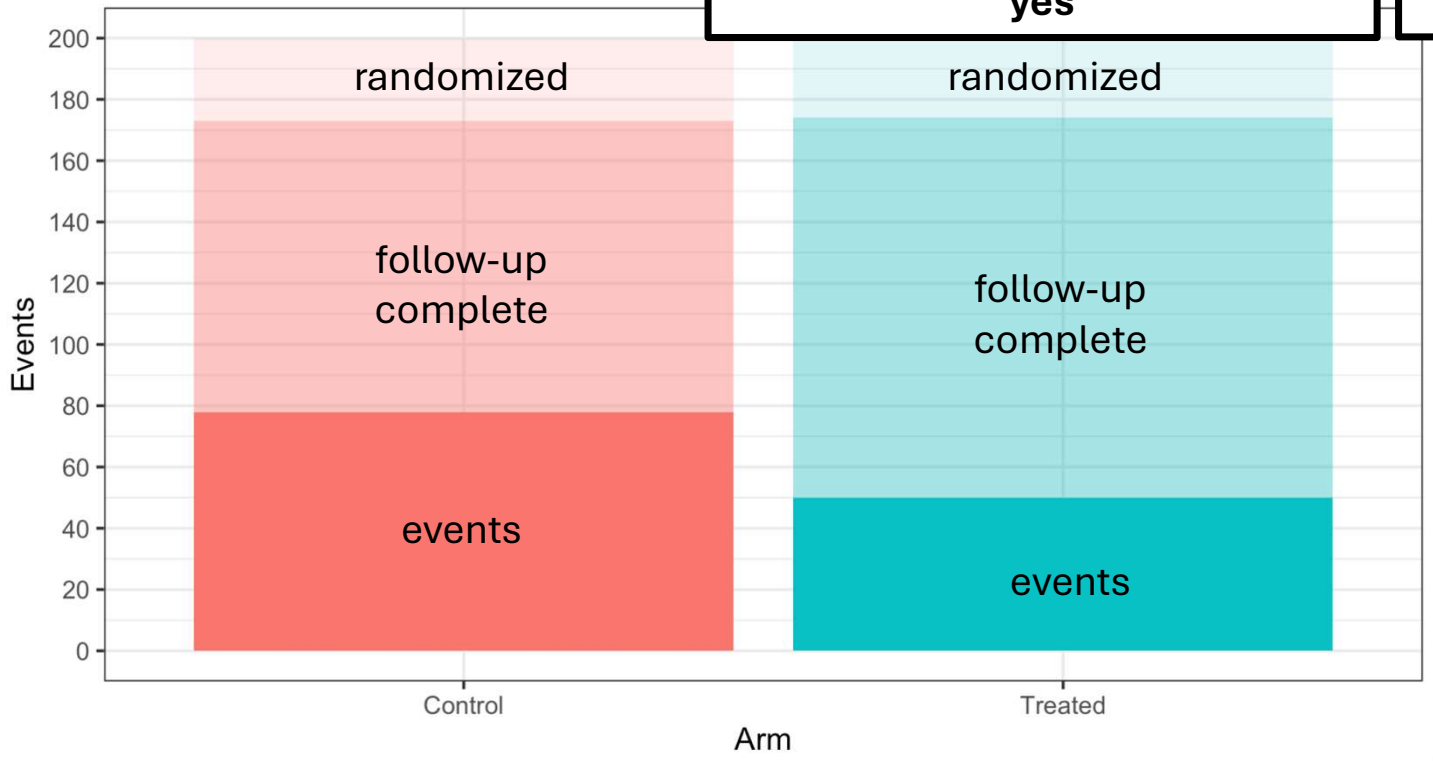


Interim 2: 400 Randomized

Randomized			Follow-up Complete			Events by 90 Days (%)		Predictive Probability of Success at Current N		Predictive Probability of Success at Max N	
Total	Control	Treated	Total	Control	Treated	Control	Treated	PPn	aPPn	PPmaxN	aPPmaxN
300	150	150	230	115	115	0.4435	0.3304	0.5031	0.4940	0.7112	0.7023
400	200	200	347	173	174	0.4509	0.2874	0.9995	0.9998	0.9973	0.9965

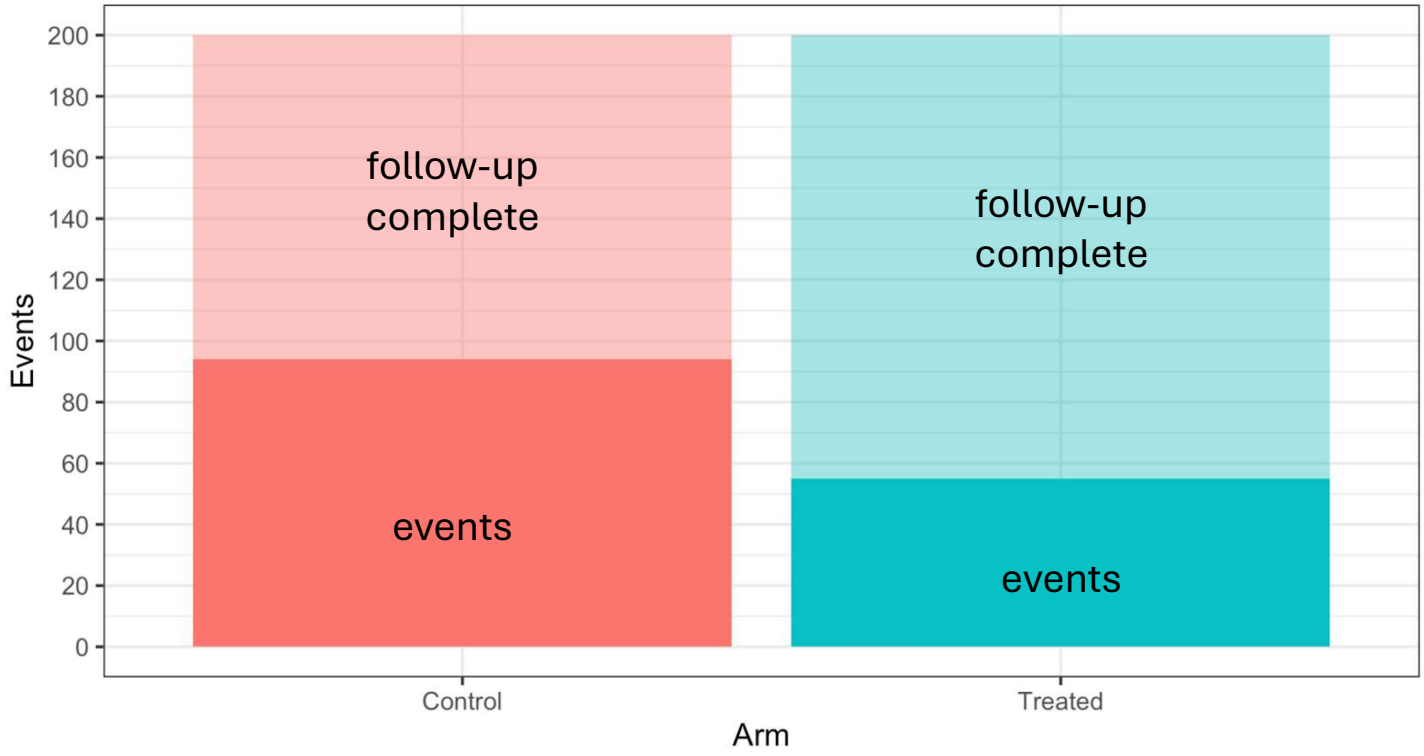
met criteria for expected success?
yes

met criteria for futility?
no



Final Analysis

Randomized			Follow-up Complete			Events by 90 Days (%)		Predictive Probability of Success at Current N		Predictive Probability of Success at Max N		Final Analysis
Total	Control	Treated	Total	Control	Treated	Control	Treated	PPn	aPPn	PPmaxN	aPPmaxN	p
300	150	150	230	115	115	0.4435	0.3304	0.5031	0.4940	0.7112	0.7023	
400	200	200	347	173	174	0.4509	0.2874	0.9995	0.9998	0.9973	0.9965	
400	200	200	400	200	200	0.4700	0.2750					0.0000



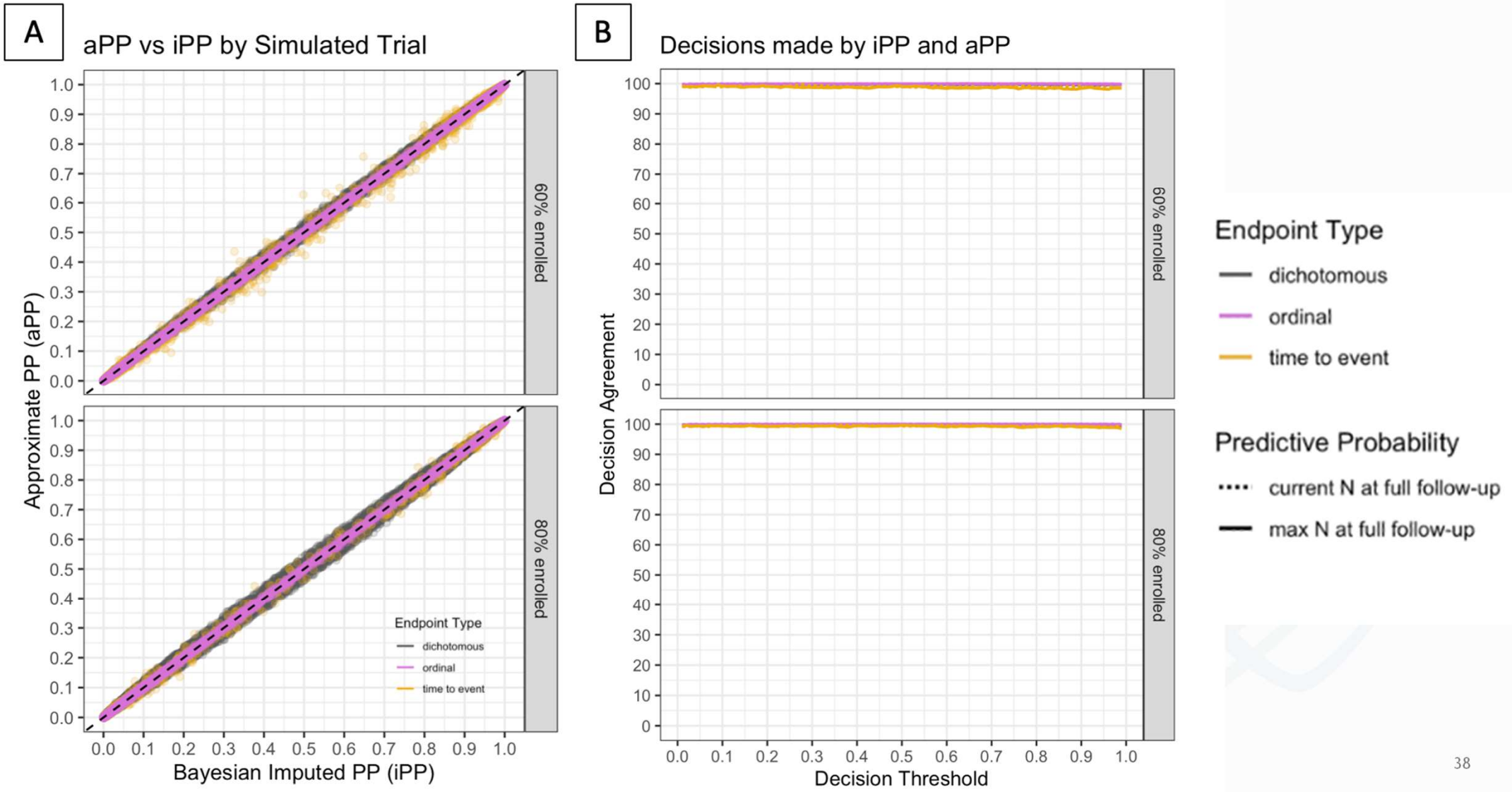
success

Simulation Studies

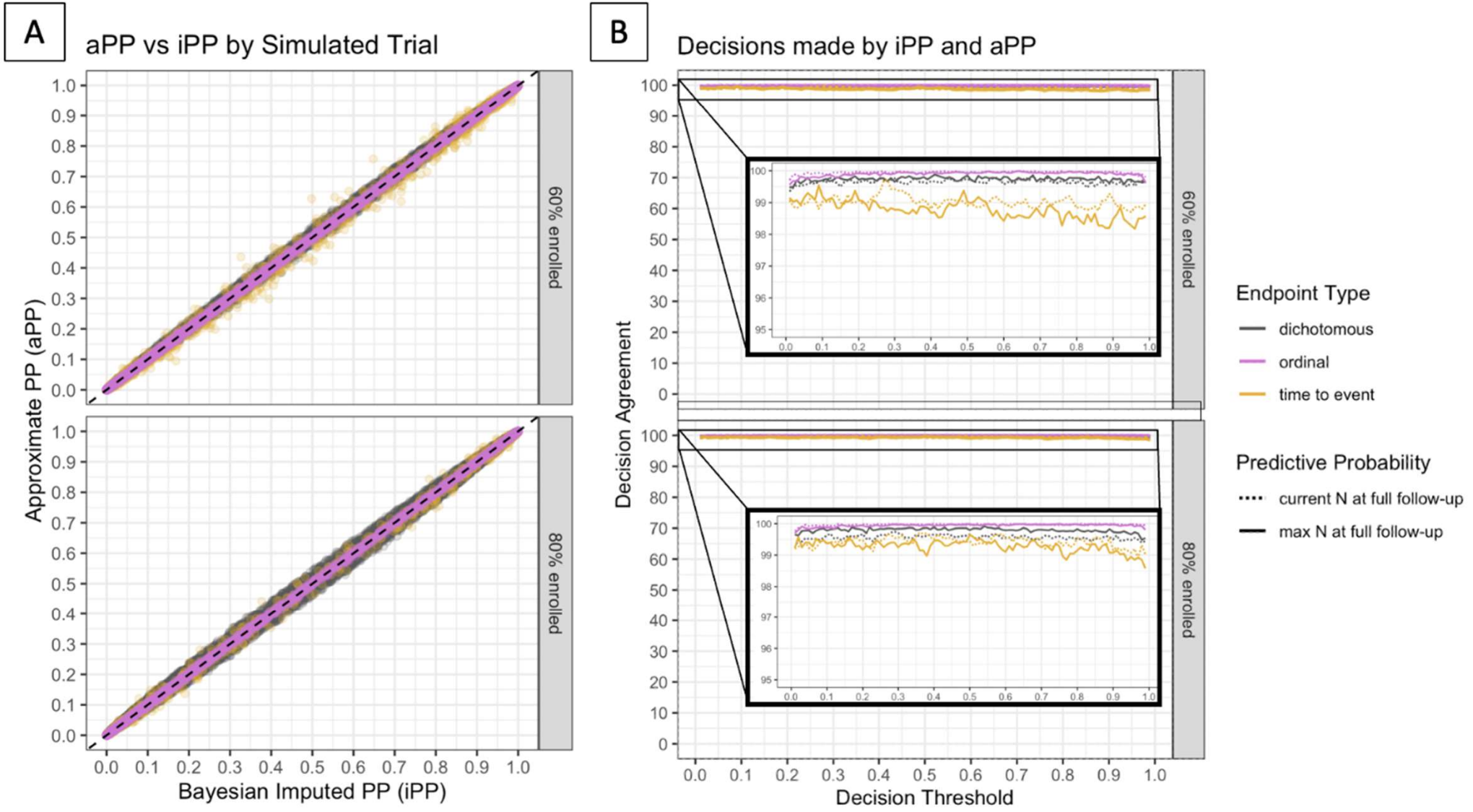
- Interims at 60%, 80% information
- imputed predictive probabilities (iPP) vs approximate predictive probabilities (aPP)
- **Do the iPP and aPP look similar? Make the same decisions?**

Endpoint	Example analysis	I_n	I_N
Continuous	T-tests ANOVA/ANCOVA	Interim sample size	Final sample size
Binary	Z-tests Chi-squared tests	Interim sample size	Final sample size
Time-to-event	Log-rank test Proportional hazards models	Events at interim	Events at final
Ordinal/Non-parametric	Ordinal regression Wilcoxon rank-sum	Interim sample size	Final sample size
Count data	Generalized linear regressions (e.g. Poisson regression)	Interim exposure	Final exposure

iPP vs aPP across endpoint types



iPP vs aPP across endpoint types



Summary

- **Approximate predictive probability from interim z-scores**
- **Fits easily into both frequentist and Bayesian designs**
- High similarity between imputed PP and approximate PP
  - Though there are cases where they disagree (win ratio, analyses with hard-to-compute information)
- Fast: reduces computational burden
  - especially during clinical trial simulations

Marion*, Lorenzi*, Allen-Savietta*, Viele, & Berry. **Predictive Probabilities Made Simple: A Fast and Accurate Method for Clinical Trial Decision Making**

under review, available on arXiv

GitHub.com/BerryConsultants/approximatePredictiveProbability

easy-to-use R functions