Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparison of Predictive Probability Calculation Methods for Bayesian Interim Monitoring with Delayed Binary Responses: A Simulation Study of the Handling of Pipeline Subjects

Mika Tominaga
Saga University
Yuko Ishimaru
Saga University
Ayako Takamori
Hiroshima University
Atsushi Kawaguchi
akawa@cc.saga-u.ac.jp

Saga University

Research Article

Keywords: Bayesian adaptive design, predictive probability, delayed response, pipeline subjects, interim analysis, phase II trial, simulation study

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11276556/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

In Bayesian interim monitoring with delayed responses, the handling of pipeline subjects whose outcomes remain pending at the time of interim analysis may affect predictive probability (PP) estimates and interim decision-making. To investigate this, we considered a single-arm Bayesian phase II trial with delayed binary responses and compared five approaches for handling pipeline subjects.

Methods

Available-case (AC), all-nonresponse (ANR), all-response (AR), fixed-rate plug-in (FRP), and observed-rate plug-in (ORP) methods were compared. Maximum sample size was 40, with 20 subjects whose outcomes had been ascertained at the recruitment decision point, at which a PP-based interim early-efficacy decision was made. We considered 36 response-rate scenarios, resulting in 72 analysis conditions with five or 10 pipeline subjects. Binary outcomes for subjects with ascertained outcomes and pipeline subjects were generated from Bernoulli distributions with a common true-response probability. Reference PP was used as a retrospective oracle benchmark. Performance was evaluated using the Pearson correlation coefficient between the PP obtained using each method and the reference PP, bias, mean absolute error (MAE), root mean square error (RMSE), and agreement with the reference PP–based interim decision.

Results

AC and ORP performed favorably across several metrics. Median coefficients for five pipeline subjects were 0.815 for AC and 0.807 for ORP; for 10, they were 0.647 and 0.661. Across the 72 analysis conditions, mean bias was − 0.018 for AC and − 0.008 for ORP. MAEs were 0.059 and 0.055; RMSEs were 0.098 and 0.103. Mean decision agreement for five pipeline subjects was 0.914 for AC and 0.938 for ORP; for 10, they were 0.872 and 0.919. For all methods, increasing pipeline subjects from five to 10 resulted in lower median Pearson correlation coefficients and lower mean decision agreement.

Conclusions

ORP had the mean bias closest to zero, lowest mean MAE, and highest mean decision agreement; AC had the lowest mean RMSE. AC and ORP showed relatively high Pearson correlation coefficients with the reference PP. These findings indicate that the performance of approaches to handling pipeline subjects varies and should be assessed using multiple measures, including association with the reference PP, estimation error, and interim decision-making agreement.

Background

Bayesian interim monitoring

In phase II clinical trials, interim monitoring may be conducted using data accumulated during the trial to evaluate treatment efficacy or futility and determine whether the trial should continue or stop early. Bayesian interim monitoring is one decision-making framework that has been used in adaptive clinical trials [1, 2].

In Bayesian interim monitoring, the predictive probability (PP) of meeting a prespecified success criterion at the final analysis can be calculated based on the data available at the time of the interim analysis. PP integrates interim analysis information with uncertainty regarding future outcomes, thereby quantifying the probability of meeting the prespecified success criterion at the final analysis [3, 4]. Thus, it can be used to inform decisions regarding trial continuation or early termination.

Challenges posed by delayed responses

In trials that require a certain observation period before the primary endpoint can be assessed, some subjects may have already been enrolled but might not have an ascertained outcome for the primary endpoint at the time of the interim analysis. In this study, we refer to such subjects as pipeline subjects [5, 6].

When pipeline subjects are present, it is necessary to specify how uncertainty regarding their pending outcomes is handled in PP calculation. One approach is to not directly incorporate pending outcomes into the current observed data at the time of interim analysis. By contrast, approaches that incorporate pipeline subjects into PP calculation by assuming a particular response status or response rate require additional assumptions about their outcomes. Consequently, different approaches to handling pipeline subjects may yield different PP estimates and may therefore affect interim decision-making.

Previous research and outstanding issues

Several approaches to handling pipeline subjects can be considered, including not directly incorporating pending outcomes into the current observed data at the interim analysis, treating all pipeline subjects as nonresponders or responders, assuming a fixed response rate, or using the response rate observed among subjects with ascertained outcomes. Few studies, however, have systematically compared multiple approaches under the same conditions in terms of their effects on PP estimates and interim decision-making.

Previous studies of group sequential designs with delayed responses have highlighted the statistical and operational challenges posed by pipeline subjects in interim analyses, showing that the number of pipeline subjects may affect design performance [6, 7]. Against this background, it is important to

investigate how different approaches to handling pipeline subjects affect PP estimates and interim decision-making in Bayesian interim monitoring with delayed responses.

Study objective

To investigate this general methodological issue, we considered a single-arm Bayesian phase II trial with delayed binary responses as the study setting. On that basis, we compared the performance of five approaches for handling pipeline subjects: available-case (AC), all-nonresponse (ANR), all-response (AR), fixed-rate plug-in (FRP), and observed-rate plug-in (ORP).

Methods

Study setting

We considered a single-arm Bayesian phase II trial with a binary response as the primary endpoint. Let Y_i denote the response outcome for subject i, where $Y_i = 1$ indicates a response and $Y_i = 0$ indicates no response. Let p denote the true-response probability. We assumed that

$$Y_i \mid p \sim \text{Bernoulli}(p).$$

The maximum sample size was denoted by N.

At the recruitment decision point (RDP), at which an interim early-efficacy decision was made based on PP, let n_o denote the number of subjects whose primary endpoint outcomes had been ascertained and x_o denote the number of responders among them. Thus,

$$x_o = \sum_{i=1}^{n_o} Y_i.$$

Subjects who had already been enrolled at RDP but whose primary endpoint outcomes had not yet been ascertained were defined as pipeline subjects, and their number was denoted by m.

Bayesian model

Prior and posterior distributions

We assumed a Beta prior distribution for response probability p:

$$p \sim \text{Beta}(a, b).$$

In the primary analysis, we used the uniform prior distribution $\text{Beta}(1, 1)$—that is, $a = b = 1$.

At RDP, given x_o responders among n_o subjects with ascertained outcomes, the posterior distribution of the response probability p is, by Beta–Bernoulli conjugacy,

$$p \mid x_o, n_o \sim \text{Beta}\left(a + x_o,\ b + n_o - x_o\right).$$

Therefore, under $a = b = 1$, as used in the primary analysis,

$$p \mid x_o, n_o \sim \text{Beta}\left(1 + x_o,\ 1 + n_o - x_o\right).$$

Final success criterion

Let the number of responders among N subjects at the final analysis be

$$X_N = \sum_{i=1}^{N} Y_i.$$

At the final analysis, the trial was considered successful if the posterior probability that the response probability p exceeded the clinically relevant threshold p_0 was greater than the prespecified threshold γ. Thus, the final success criterion was defined as

$$\Pr\left(p > p_0 \mid X_N, N\right) > \gamma.$$

Posterior distribution at the final analysis is

$$p \mid X_N, N \sim \text{Beta}\left(a + X_N,\ b + N - X_N\right),$$

with $a = b = 1$ in the primary analysis.

Here, γ denotes the posterior probability threshold used to determine success at the final analysis and is distinct from the PP threshold c_E used for the early-efficacy decision at RDP.

The values of p_0, γ, and c_E used in this study are specified in the section “Simulation design.”

Predictive probability

PP at RDP was defined as the probability, conditional on information available at that time, that the prespecified success criterion would be met at the final analysis if the trial were continued to the maximum sample size N.

First, using only subjects with ascertained outcomes as the current observed data, PP was defined as

$$PP\left(x_o, n_o\right) = \Pr\left[\Pr\left(p > p_0 \mid X_N, N\right) > \gamma \mid x_o, n_o\right].$$

Let Z denote the number of responders among $N - n_o$ subjects whose final outcomes are unobserved at RDP and will become available by the final analysis. The total number of responders at the final analysis can then be expressed as

$$X_N = x_o + Z.$$

Here, the $N - n_o$ subjects include m pipeline subjects who have already been enrolled at RDP but whose outcomes have not yet been ascertained, as well as subjects who will subsequently be enrolled until the maximum sample size N is reached.

Conditional on the observed data (x_o, n_o) from subjects with ascertained outcomes, the posterior predictive distribution of Z is

$$Z \mid x_o, n_o \sim \text{Beta-Binomial}\left(N - n_o,\ a + x_o,\ b + n_o - x_o\right).$$

Therefore, PP was calculated as

$$PP\left(x_o, n_o\right) = \sum_{z=0}^{N-n_o} I\left[\Pr\left(p > p_0 \mid x_o + z,\ N\right) > \gamma\right] \Pr\left(Z = z \mid x_o, n_o\right).$$

Here, $I(\cdot)$ denotes the indicator function, which takes a value of 1 when the specified condition is satisfied and 0 otherwise.

In the following sections, this PP calculation was used as the basic formulation, with the numbers of responders and subjects used in the PP calculation modified according to the approach used to handle pipeline subjects.

In this study, the early-efficacy decision based on PP at RDP was defined as

$$D = I\left(PP > c_E\right),$$

where c_E denotes the PP threshold for early efficacy. In the primary analysis, $c_E = 0.95$. The early-efficacy criterion was considered to be met when $PP > c_E$ and not met when $PP \leq c_E$. No PP threshold for futility stopping was specified in this study.

Delayed responses and pipeline subjects

When a certain observation period is required before the primary endpoint can be ascertained, some subjects may have already been enrolled but might not have an ascertained outcome for the primary endpoint at RDP. In this study, such subjects were defined as pipeline subjects.

At RDP, binary outcomes are available for n_o subjects whose primary endpoint outcomes have been ascertained, whereas the outcomes of m pipeline subjects remain pending. We compared five approaches for handling these pipeline subjects in the calculation of PP.

Figure 1 illustrates the distinction between subjects with ascertained outcomes and pipeline subjects at RDP.

AC method

Handling of pipeline subjects

Under the AC method, the pending outcomes of pipeline subjects were not incorporated into the current observed data. PP was calculated using only subjects whose primary endpoint outcomes had been ascertained at RDP.

Under the Beta–Bernoulli/Beta–Binomial model considered in this study, no additional information related to the response status of pipeline subjects was used, and subjects with ascertained outcomes and pipeline subjects were assumed to share a common response probability. Under these assumptions, explicitly marginalizing over the possible outcomes of pipeline subjects with respect to their posterior predictive distribution yields the same PP as updating the posterior distribution using only subjects with ascertained outcomes and treating pipeline subjects as part of the subjects whose outcomes remain to be observed by the final analysis.

Therefore, under these assumptions, AC can be interpreted not as simply ignoring the uncertainty associated with pipeline subjects but as implicitly marginalizing over their pending outcomes with respect to posterior predictive distribution. Appendix A details the mathematical relationship underlying this interpretation.

Calculation of PP

PP under the AC method was calculated as

$$PP_{\mathrm{AC}} = PP\left(x_o, n_o\right).$$

Specifically, posterior distribution was updated using x_o responders among the n_o subjects whose primary endpoint outcomes had been ascertained at RDP. PP was then calculated using the Beta–Binomial posterior predictive distribution for the remaining $N - n_o$ subjects whose outcomes will become available by the final analysis, including the pipeline subjects.

Assumptions and characteristics

The AC method does not require explicit imputation of the pending outcomes of pipeline subjects and only uses outcomes that have been ascertained at RDP as the current observed data. Therefore, its implementation and interpretation are relatively straightforward.

However, the validity of AC under the setting considered in this study depends on the assumption that subjects with ascertained outcomes and pipeline subjects share a common response probability and that the time to outcome ascertainment is independent of response status. That is, response delay is noninformative.

If the time to outcome ascertainment is associated with response status, subjects whose outcomes have been ascertained at RDP may not be representative of all enrolled subjects. Consequently, posterior predictive distribution based on subjects with ascertained outcomes might not adequately reflect the outcome distribution of pipeline subjects.

ANR method

Handling of pipeline subjects

Under ANR, all pipeline subjects at RDP were treated as nonresponders and incorporated into the current observed data as if their outcomes had been ascertained when calculating PP.

Calculation of PP

PP under the ANR method was calculated as

$$PP_{\mathrm{ANR}} = PP\left(x_o, n_o + m\right).$$

Because all m pipeline subjects were treated as nonresponders, the number of responders remained x_o, whereas the number of subjects treated as having ascertained outcomes in the PP calculation increased to $n_o + m$. Accordingly, the number of subjects whose outcomes remain to be observed by the final analysis was

$$N - (n_o + m).$$

Assumptions and characteristics

The ANR method treats all pipeline subjects as nonresponders and therefore handles their pending outcomes in a manner that tends to lower the calculated PP.

If some pipeline subjects ultimately respond, treating them as nonresponders may lead to underestimating PP relative to the reference PP.

All-response (AR) method

Handling of pipeline subjects

Under AR, all pipeline subjects at RDP were treated as responders and incorporated into the current observed data as if their outcomes had been ascertained when calculating PP.

Calculation of PP

PP under the AR method was calculated as

$$PP_{\mathrm{AR}} = PP\left(x_o + m, n_o + m\right).$$

Because all m pipeline subjects were treated as responders, the number of responders increased to $x_o + m$, and the number of subjects treated as having ascertained outcomes for PP calculation increased to $n_o + m$. Accordingly, the number of remaining subjects whose outcomes will become available by the final analysis was

$$N - (n_o + m).$$

Assumptions and characteristics

The AR method treats all pipeline subjects as responders and therefore handles their pending outcomes in a manner that tends to increase the calculated PP.

If some pipeline subjects ultimately do not respond, treating them as responders may lead to overestimating PP relative to the reference PP.

FRP method

Handling of pipeline subjects

Under FRP, the response rate used for pipeline subjects at RDP was fixed at 0.5, and a pseudo-response count of $0.5m$ based on this fixed value was used in the PP calculation. A value of 0.5 was selected as a simple, data-independent fixed value to provide a benchmark for the FRP approach, rather than as a value intended to represent a clinically valid response rate across settings.

Calculation of PP

PP under the FRP method was calculated as

$$PP_{\mathrm{FRP}} = PP\left(x_o + 0.5m, n_o + m\right).$$

Specifically, for the m pipeline subjects, the pseudo-response count was set to $0.5m$ based on the fixed value of 0.5. Accordingly, the number of responses used in the PP calculation was $x_o + 0.5m$, and the number of subjects treated as having ascertained outcomes was $n_o + m$.

The pseudo-response count $0.5m$ was not rounded. When $0.5m$ was noninteger, it was incorporated as a fractional pseudo-count into the shape parameters of the Beta distribution used in the PP calculation. Thus, this distribution was not a Beta posterior distribution based solely on actually observed binary outcomes but a working Beta distribution constructed using fractional pseudo-counts based on the fixed value of 0.5.

The FRP method does not calculate PP for all possible outcomes of the pipeline subjects and average these values with respect to their posterior predictive distribution. Rather, it is a plug-in approximation in which a single pseudo-response count based on the fixed value of 0.5 is substituted into the PP calculation.

Under this approach, the number of subjects whose outcomes remained to be observed by the final analysis was

$$N - (n_o + m).$$

Assumptions and characteristics

In FRP, the prespecified fixed value of 0.5 was used for pipeline subjects regardless of the response information observed by RDP. Thus, this method is not a data-adaptive imputation approach but a plug-in approximation based on a fixed value. Its performance may depend on the discrepancy between the fixed value of 0.5 and the actual response pattern among pipeline subjects.

ORP method

Handling of pipeline subjects

In ORP, the observed response proportion among subjects with ascertained outcomes at RDP was used to calculate a pseudo-response count for the pipeline subjects, which was then incorporated into the PP calculation.

The observed response proportion among subjects with ascertained outcomes was defined as

$$\hat{p}_{\mathrm{obs}} = \frac{x_o}{n_o}.$$

Based on this observed response proportion, the pseudo-response count for the pipeline subjects was defined as

$$\tilde{x}_{\mathrm{pipe}} = m\hat{p}_{\mathrm{obs}} = m\frac{x_o}{n_o}.$$

Accordingly, the number of responders used to calculate PP under ORP was

$$x_{\mathrm{ORP}} = x_o + \tilde{x}_{\mathrm{pipe}} = x_o + m\frac{x_o}{n_o}.$$

Here, $\tilde{x}_{\mathrm{pipe}} = mx_o/n_o$ is not the posterior predictive expected number of responders among pipeline subjects under the Beta–Binomial model. Rather, it is a plug-in value constructed using the observed response proportion among subjects with ascertained outcomes at RDP.

Calculation of PP

PP under the ORP method was calculated as

$$PP_{\mathrm{ORP}} = PP\left(x_o + m\frac{x_o}{n_o}, n_o + m\right).$$

The pseudo-response count mx_o/n_o was not rounded. When noninteger, it was incorporated as a fractional pseudo-response count into the shape parameters of the Beta distribution used to calculate PP. Thus, this Beta distribution differed from a Beta posterior distribution based solely on actually observed binary outcomes and was instead treated as a working Beta distribution constructed using fractional pseudo-counts based on the observed response proportion.

ORP does not calculate PP for all possible outcomes of the pipeline subjects and average these values with respect to their posterior predictive distribution. Rather, it is a plug-in approximation in which a single pseudo-response count calculated from the observed response proportion among subjects with ascertained outcomes at RDP is substituted into the PP calculation.

The number of subjects treated as having ascertained outcomes in the PP calculation was $n_o + m$. The number of subjects whose outcomes remained to be observed by the final analysis was

$$N - (n_o + m)\,.$$

Assumptions and characteristics

In ORP, the observed response proportion among subjects with ascertained outcomes at RDP is used as a plug-in value representing the response rate among pipeline subjects. This approach allows the response information obtained from subjects with ascertained outcomes to be incorporated into the pseudo-response count for the pipeline subjects.

However, the validity of ORP under the setting considered in this study depends on the assumptions that subjects with ascertained outcomes and pipeline subjects share a common response probability and that the time to outcome ascertainment is independent of response status; that is, response delay is noninformative.

If the time to outcome ascertainment is associated with response status, the observed response proportion among subjects with ascertained outcomes at RDP might not adequately represent the response pattern among pipeline subjects, potentially affecting the resulting PP.

Summary of the five methods

Table 1 summarizes, for each method, the information used in PP calculation, the handling of pipeline subjects, and the corresponding PP calculation.

Table 1
Summary of the five methods used to handle pipeline subjects

Method	Information used	Handling of pipeline subjects	PP calculation
AC	Outcomes among subjects with ascertained outcomes	Pending outcomes are not incorporated into the current observed data	$PP\left(x_o, n_o\right)$; the remaining $N - n_o$ subjects, including the pipeline subjects, are predicted
ANR	Outcomes among subjects with ascertained outcomes and the number of pipeline subjects	All pipeline subjects are treated as nonresponders	$PP\left(x_o, n_o + m\right)$; the remaining $N - \left(n_o + m\right)$ subjects are predicted
AR	Outcomes among subjects with ascertained outcomes and the number of pipeline subjects	All pipeline subjects are treated as responders	$PP\left(x_o + m, n_o + m\right)$; the remaining $N - \left(n_o + m\right)$ subjects are predicted
FRP	Outcomes among subjects with ascertained outcomes and the number of pipeline subjects	A pseudo-response count of $0.5m$, based on a fixed value of 0.5, is used	$PP\left(x_o + 0.5m, n_o + m\right)$; the remaining $N - \left(n_o + m\right)$ subjects are predicted
ORP	Outcomes among subjects with ascertained outcomes, the observed response proportion x_o/n_o, and the number of pipeline subjects	A pseudo-response count of mx_o/n_o, based on the observed response proportion, is used	$PP\left(x_o + mx_o/n_o, n_o + m\right)$; the remaining $N - \left(n_o + m\right)$ subjects are predicted

Note: The pseudo-response counts used in the FRP and ORP methods were not rounded; noninteger values were incorporated as fractional pseudo-counts in the PP calculation.

Reference PP

Among the m pipeline subjects at RDP, let x_m denote the number of responders whose outcomes were subsequently ascertained after RDP.

Reference PP was defined as the PP calculated using the subsequently ascertained outcomes of all pipeline subjects who had already been enrolled at RDP, based on the combined data from subjects with ascertained outcomes at RDP and the pipeline subjects. The number of subjects and the number of responders used to calculate the reference PP were, respectively,

$$n_{\text{ref}} = n_o + m,$$

and

$$x_{\text{ref}} = x_o + x_m.$$

Reference PP was calculated using the same procedure described above for PP, treating $(x_{\text{ref}}, n_{\text{ref}})$ as the current observed data.

Let Z_{ref} denote the number of responders among the remaining $N - n_{\text{ref}}$ subjects whose outcomes would become available by the final analysis. Its posterior predictive distribution is

$$Z_{\text{ref}} \mid x_{\text{ref}}, n_{\text{ref}} \sim \text{Beta-Binomial}\left(N - n_{\text{ref}}, a + x_{\text{ref}}, b + n_{\text{ref}} - x_{\text{ref}}\right).$$

Accordingly, reference PP was calculated as

$$PP_{\text{ref}} = \sum_{z=0}^{N-n_{\text{ref}}} I\left[\Pr\left(p > p_0 \mid x_{\text{ref}} + z,\ N\right) > \gamma\right] \Pr\left(Z_{\text{ref}} = z \mid x_{\text{ref}}, n_{\text{ref}}\right).$$

The reference PP is not the PP at the final analysis. Rather, it is the PP calculated after the outcomes of all pipeline subjects who had already been enrolled at RDP became available, using their subsequently ascertained outcomes. Because the outcomes of the pipeline subjects are not available at the actual RDP, the reference PP is not an implementable decision-making approach at RDP. In this study, it was used as a retrospective oracle benchmark for evaluating the PP estimation performance of each method and the agreement of interim decisions with those based on the reference PP.

Simulation design

Table 2 summarizes the simulation settings used in this study.

Table 2
Simulation settings

Item	Symbol	Setting
Maximum sample size	N	40
Number of subjects with ascertained outcomes at RDP	n_o	20
Number of pipeline subjects	m	5, 10
Clinical threshold	p_0	0.1–0.8 (increments of 0.1)
True-response probability	p	0.2–0.9 (increments of 0.1)
Condition for response probabilities		$p > p_0$
Prior distribution	p	$\mathrm{Beta}\,(1, 1)$
Posterior probability threshold at the final analysis	γ	0.90
PP threshold for early efficacy	c_E	0.95
Interim decision		Binary classification based on $PP > c_E$ (no futility stopping threshold)
Number of simulations per condition	B	1,000
Methods for handling pipeline subjects		AC, ANR, AR, FRP, ORP
Reference benchmark		Reference PP
Performance measures		Pearson correlation coefficient, bias, MAE, RMSE, decision agreement

Note: AC, available-case; ANR, all-nonresponse; AR, all-response; FRP, fixed-rate plug-in; ORP, observed-rate plug-in; PP, predictive probability; RDP, recruitment decision point.

The maximum sample size was set to $N = 40$, and the number of subjects with ascertained outcomes at RDP was set to $n_o = 20$. The number of pipeline subjects was set to either $m = 5$ or $m = 10$.

The clinical threshold p_0 was varied from 0.1 to 0.8, and true-response probability p was varied from 0.2 to 0.9, both in increments of 0.1. This yielded 36 response-rate scenarios satisfying $0.1 \leq p_0 < p \leq 0.9$. Each scenario was evaluated under two conditions, with $m = 5$ and $m = 10$ pipeline subjects, resulting in a total of 72 analysis conditions.

Response times and outcome ascertainment times themselves were not generated in this study. Instead, the binary outcomes of both subjects with ascertained outcomes and pipeline subjects at RDP were generated from Bernoulli distributions with a common true-response probability p. This setting represents a noninformative response-delay mechanism, in which whether a subject had an ascertained or pending outcome at RDP was independent of the eventual response outcome.

For each analysis condition, the following procedure was repeated $B = 1,000$ times.

1. Under true-response probability p, the binary outcomes of n_o subjects with ascertained outcomes and m pipeline subjects at RDP were generated from Bernoulli distributions, yielding the number of responders among subjects with ascertained outcomes, x_o, and among the pipeline subjects, x_m.
2. At RDP, the outcomes of the pipeline subjects were treated as unobserved.
3. PP was calculated using each method: AC, ANR, AR, FRP, and ORP. Realized outcomes were not generated for subjects remaining until the final analysis; instead, PP was calculated using the Beta–Binomial posterior predictive distribution corresponding to each method.
4. For pipeline subjects who had already been enrolled at RDP, reference PP was calculated based on combined data from the subjects with ascertained outcomes at RDP and the pipeline subjects, using x_m, the number of responders among the pipeline subjects based on their generated binary outcomes.
5. Based on the 1000 simulation replicates within each analysis condition, we calculated the Pearson correlation coefficient between the PP obtained using each method and the reference PP, bias, mean absolute error (MAE), root mean square error (RMSE), and decision agreement.

Because the 72 analysis conditions were prespecified fixed conditions rather than scenarios randomly sampled from a population, we summarized the results across the analysis conditions descriptively. The definitions of each performance measure and the methods used to summarize results across analysis conditions are described below.

To ensure reproducibility of the random-number generation, the random seed was set to 123 at the start of the simulation study (set.seed(123)).

We performed all simulations and statistical analyses using R version 4.4.0.

Performance measures

Let $\widehat{PP}_b$ denote the PP calculated using each method and $PP_{\mathrm{ref},b}$ denote the reference PP, where $b = 1, \ldots, B$ indexes the simulation replicates within each analysis condition. In this study,

$B = 1,000.$

Pearson correlation coefficient

To evaluate the linear association between the PP obtained using each method and the reference PP, the Pearson correlation coefficient was calculated as

$$r = \mathrm{Corr}\left(\widehat{PP}, PP_{\mathrm{ref}}\right).$$

Pearson correlation coefficient was calculated based on the $B = 1,000$ simulation replicates within each analysis condition. It ranges from − 1 to 1, with values closer to 1 indicating a stronger positive linear association, values closer to − 1 indicating a stronger negative linear association, and values closer to 0 indicating a weaker linear association.

If the variance of either the PP obtained using a given method or the reference PP was zero within an analysis condition, the Pearson correlation coefficient was mathematically undefined and was therefore treated as missing (NA) for that condition.

The 95% confidence intervals (CIs) for the Pearson correlation coefficients in representative scenarios were calculated using Fisher's z-transformation.

The distribution of Pearson correlation coefficients across the 72 analysis conditions was descriptively summarized using the median and interquartile range (IQR), including only conditions for which the correlation coefficient was defined. The number of analysis conditions for which the correlation coefficient could be calculated was also reported.

To descriptively examine correlations according to the number of pipeline subjects, the median and IQR of the Pearson correlation coefficients were calculated separately for $m = 5$ and $m = 10$, using the scenarios among the 36 response-rate scenarios for which the correlation coefficient was defined. We also reported the number of scenarios for which the correlation coefficient could be calculated.

As a supplementary descriptive analysis, AC and ORP were compared on a scenario-by-scenario basis by determining which method yielded the higher Pearson correlation coefficient, separately for $m = 5$ and $m = 10$. This comparison was restricted to scenarios in which the Pearson correlation coefficients were defined for both methods.

In addition, as an exploratory descriptive analysis of patterns according to the difference between the clinical threshold p_0 and the true-response probability p, the 36 response-rate scenarios were classified into two categories: $p - p_0 \leq 0.3$ and $p - p_0 > 0.3$. Within each category and separately for $m = 5$ and $m = 10$, the AC and ORP methods were compared by determining which method yielded the higher Pearson correlation coefficient, restricting the comparison to scenarios in which the Pearson correlation coefficients were defined for both methods.

Bias

Within each analysis condition, bias was calculated to assess the systematic estimation error of the PP obtained using each method relative to the reference PP, as follows:

$$\text{Bias} = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{PP}_b - PP_{\text{ref},b}\right).$$

Here, $\widehat{PP}_b$ and $PP_{\text{ref},b}$ denote the PP obtained using each method and the reference PP, respectively, in simulation replicate b.

A bias closer to 0 indicates less systematic deviation from the reference PP. A positive value indicates overestimation on average, whereas a negative value indicates underestimation on average.

MAE

Within each analysis condition, MAE was calculated to assess the magnitude of the estimation error of the PP obtained using each method relative to the reference PP, as follows:

$$\text{MAE} = \frac{1}{B}\sum_{b=1}^{B}\left|\widehat{PP}_b - PP_{\text{ref},b}\right|.$$

An MAE closer to 0 indicates a smaller average absolute difference between the PP obtained using each method and the reference PP.

RMSE

Within each analysis condition, RMSE was calculated to assess the magnitude of the estimation error of the PP obtained using each method relative to the reference PP, as follows:

$$\text{RMSE} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{PP}_b - PP_{\text{ref},b}\right)^2}.$$

RMSE is calculated by squaring the differences between the PP obtained using each method and the reference PP, averaging the squared differences, and then taking the square root. Consequently, RMSE assigns greater weight to larger errors than MAE. An RMSE closer to 0 indicates a smaller overall discrepancy between the PP obtained using each method and the reference PP.

Bias, MAE, and RMSE were calculated within each analysis condition based on $B = 1,000$ simulation replicates. Across the 72 prespecified analysis conditions, the bias, MAE, and RMSE values obtained for each condition were descriptively summarized for each method using the mean and standard deviation (SD).

Decision agreement

Within each analysis condition, the interim decision based on the PP obtained using each method was defined as

$$\widehat{D}_b = I\left(\widehat{PP}_b > c_E\right),$$

and the interim decision based on the reference PP was defined as

$$D_{\mathrm{ref},b} = I\left(PP_{\mathrm{ref},b} > c_E\right).$$

Here, c_E denotes the PP threshold for early efficacy, which was set to $c_E = 0.95$ in this study. Accordingly, interim decisions were classified into two categories: “early efficacy” and “not early efficacy.” No PP threshold for futility stopping was specified.

Decision agreement was calculated as follows:

$$\mathrm{Agreement} = \frac{1}{B}\sum_{b=1}^{B} I\left(\widehat{D}_b = D_{\mathrm{ref},b}\right).$$

Decision agreement represents the proportion of the $B = 1,000$ simulation replicates within each analysis condition in which the interim decision based on the PP obtained using each method agreed with the interim decision based on the reference PP. Values closer to 1 indicate more frequent agreement between the two interim decisions.

To descriptively compare decision agreement according to the number of pipeline subjects, the decision agreement values calculated for each scenario were averaged across the 36 response-rate scenarios for each method, separately for $m = 5$ and $m = 10$.

Summary across analysis conditions

The 72 analysis conditions specified in this study (36 response-rate scenarios × 2 numbers of pipeline subjects) were prespecified fixed conditions rather than scenarios randomly sampled from a population. Therefore, we descriptively summarized the results across the analysis conditions.

For bias, MAE, and RMSE, the values calculated in each analysis condition were summarized for each method using the mean and SD across all 72 analysis conditions.

For the Pearson correlation coefficient, we calculated the median and IQR across the 72 analysis conditions for each method, including only conditions for which the correlation coefficient was defined. We additionally reported the number of analysis conditions for which the correlation coefficient could be calculated. To descriptively examine correlations according to the number of pipeline subjects, we also calculated the median and IQR separately for $m = 5$ and $m = 10$, including only scenarios for which the correlation coefficient was defined. The number of scenarios for which the correlation coefficient could be calculated was reported as well.

For decision agreement, we averaged the values calculated within each analysis condition across the 36 response-rate scenarios for each method—separately for $m = 5$ and $m = 10$—and summarized them descriptively. To describe the change according to the number of pipeline subjects, the difference between the mean decision agreement for $m = 10$ and that for $m = 5$ was calculated for each method.

As a supplementary descriptive analysis, AC and ORP were compared on a scenario-by-scenario basis by determining which method yielded the higher Pearson correlation coefficient, separately for $m = 5$ and $m = 10$. This comparison was restricted to scenarios in which the Pearson correlation coefficients were defined for both methods.

In addition, as an exploratory descriptive analysis of patterns according to the difference between the clinical threshold p_0 and the true-response probability p, the response-rate scenarios were classified into two categories: $p - p_0 \leq 0.3$ and $p - p_0 > 0.3$. Within each category and separately for $m = 5$ and $m = 10$, AC and ORP were compared in the same manner, restricting the comparison to scenarios in which the Pearson correlation coefficients were defined for both methods.

Results

Overview of the simulation study

In this study, 36 response-rate scenarios were specified based on combinations of the clinical threshold p_0 and the true-response probability p satisfying $0.1 \leq p_0 < p \leq 0.9$. Each scenario was evaluated under two conditions with $m = 5$ and $m = 10$ pipeline subjects, resulting in a total of 72 analysis conditions.

Maximum sample size was set to $N = 40$, and the number of subjects with ascertained outcomes at RDP was set to $n_o = 20$. A total of 1000 simulation replicates were performed for each analysis condition. $\mathrm{Beta}(1, 1)$ prior distribution was used.

We first present the results for the performance of the five methods in a representative scenario. This is followed by the distribution of Pearson correlation coefficients across the 72 analysis conditions, the impact of the number of pipeline subjects, overall estimation accuracy, and decision agreement.

Representative scenario

As a representative scenario, we considered the analysis condition with a clinical threshold of $p_0 = 0.5$, a true-response probability of $p = 0.7$, and $m = 5$ pipeline subjects.

Table 3 presents the Pearson correlation coefficients and their 95% CIs between the PP obtained using each method and the reference PP in the representative scenario.

Table 3
Pearson correlation coefficients between PP obtained using each method and reference PP in the representative scenario ($p_0 = 0.5, p = 0.7, m = 5$)

Method	Pearson correlation (r)	95% CI
AC	0.876	0.860–0.889
ORP	0.862	0.845–0.877
FRP	0.870	0.854–0.884
AR	0.814	0.792–0.834
ANR	0.632	0.593–0.668

In this representative scenario, all five methods showed positive correlations with the reference PP. The Pearson correlation coefficients were $r = 0.876$ (95% CI: 0.860–0.889) for AC, $r = 0.870$ (95% CI: 0.854–0.884) for FRP, $r = 0.862$ (95% CI: 0.845–0.877) for ORP, $r = 0.814$ (95% CI: 0.792–0.834) for AR, and $r = 0.632$ (95% CI: 0.593–0.668) for ANR.

Pearson correlation coefficients across the 72 analysis conditions

Table 4 presents a descriptive summary of the Pearson correlation coefficients for each method across the 72 analysis conditions.

The median Pearson correlation coefficient across the 72 analysis conditions was 0.727 (IQR: 0.586–0.813) for AC, 0.712 (IQR: 0.277–0.811) for ORP, 0.642 (IQR: 0.113–0.761) for FRP, 0.413 (IQR: 0.086–0.652) for AR, and 0.351 (IQR: 0.096–0.616) for ANR.

The Pearson correlation coefficient could be calculated for 64 of the 72 analysis conditions for AC, 72 for ORP, 71 for FRP, 72 for AR, and 68 for ANR. In analysis conditions for which the correlation coefficient could not be calculated, the variance of either the PP obtained using the corresponding method or the reference PP was zero across the simulation replicates. Therefore, the Pearson correlation coefficient was mathematically undefined.

Table 4
Median and interquartile range of Pearson correlation coefficients across the 72 analysis conditions

Method	Median r	IQR	Valid/72
AC	0.727	0.586–0.813	64/72
ORP	0.712	0.277–0.811	72/72
FRP	0.642	0.113–0.761	71/72
AR	0.413	0.086–0.652	72/72
ANR	0.351	0.096–0.616	68/72

Correlation by number of pipeline subjects

Table 5 presents a descriptive summary of the Pearson correlation coefficients according to the number of pipeline subjects. Median and IQR were calculated using scenarios for which the Pearson correlation coefficient was defined.

For $m = 5$, the median Pearson correlation coefficient was 0.815 (IQR: 0.704–0.863) for AC and 0.807 (IQR: 0.516–0.852) for ORP. For $m = 10$, the corresponding median values were 0.647 (IQR: 0.434–0.739) for AC and 0.661 (IQR: 0.247–0.739) for ORP. For all five methods, the median Pearson correlation coefficient was lower for $m = 10$ than for $m = 5$.

For $m = 5$, the Pearson correlation coefficient could be calculated for 32 of the 36 scenarios for AC, 36 for ORP, 36 for FRP, 36 for AR, and 35 for ANR. The corresponding numbers for $m = 10$ were 32, 36, 35, 36, and 33, respectively.

Table 5
Median and interquartile range of scenario-specific Pearson correlation coefficients by number of pipeline subjects

Method	$m = 5$, median$r$ [IQR]	Valid/36	$m = 10$, median$r$ [IQR]	Valid/36
AC	0.815 [0.704–0.863]	32/36	0.647 [0.434–0.739]	32/36
ORP	0.807 [0.516–0.852]	36/36	0.661 [0.247–0.739]	36/36
FRP	0.742 [0.387–0.808]	36/36	0.511 [0.056–0.675]	35/36
AR	0.586 [0.316–0.737]	36/36	0.151 [- 0.017–0.440]	36/36
ANR	0.604 [0.429–0.677]	35/36	0.200 [0.086–0.333]	33/36

Impact of the number of pipeline subjects

Figure 2 shows the Pearson correlation coefficients between the PP obtained using each method and the reference PP across the response-rate scenarios for $m = 5$ and $m = 10$, together with the median and IQR among scenarios for which the Pearson correlation coefficient was defined.

For all five methods, the median Pearson correlation coefficient was lower for $m = 10$ than for $m = 5$. The decrease in the median correlation coefficient from $m = 5$ to $m = 10$ was 0.168 for AC, 0.146 for ORP, 0.231 for FRP, 0.435 for AR, and 0.404 for ANR.

Comparison of AC and ORP methods

As a supplementary descriptive analysis, the scenario-specific Pearson correlation coefficients for the AC and ORP methods were compared separately for $m = 5$ and $m = 10$. The heatmaps in Fig. 2 show the scenario-specific Pearson correlation coefficients.

Among the 36 response-rate scenarios, the Pearson correlation coefficients were defined for both methods in 32 scenarios. For $m = 5$, AC yielded a higher Pearson correlation coefficient in 22 of the 32 scenarios (68.8%), whereas ORP yielded a higher coefficient in 10 of the 32 scenarios (31.3%). For $m = 10$, the corresponding numbers were 23 of 32 scenarios (71.9%) for AC and 9 of 32 scenarios (28.1%) for ORP.

As an exploratory descriptive analysis, we further examined the pattern according to the magnitude of $p - p_0$. Among the 21 scenarios with $p - p_0 \leq 0.3$, the Pearson correlation coefficients were defined for both methods in all scenarios. For $m = 5$, AC yielded a higher Pearson correlation coefficient in 15 scenarios, whereas ORP yielded a higher coefficient in six scenarios. For $m = 10$, the corresponding numbers were 16 scenarios for AC and five scenarios for ORP.

Among the 15 scenarios with $p - p_0 > 0.3$, the PP obtained using AC was constant across the simulation replicates in four scenarios for each of $m = 5$ and $m = 10$. Therefore, the Pearson correlation coefficient was undefined in these scenarios. Among the remaining 11 scenarios in which the Pearson correlation coefficients were defined for both methods, AC yielded a higher Pearson correlation coefficient in seven scenarios and ORP in four scenarios for both $m = 5$ and $m = 10$.

Overall, in this supplementary descriptive comparison restricted to scenarios in which the Pearson correlation coefficients were defined for both methods, AC yielded a higher Pearson correlation coefficient than ORP in a greater number of scenarios for both $m = 5$ and $m = 10$. However, for $m = 10$, the median of the scenario-specific Pearson correlation coefficients was slightly higher for ORP (0.661) than for AC (0.647). Thus, the scenario-wise comparison and the comparison of median correlation coefficients provide different descriptive summaries of the correlation patterns.

Overall estimation accuracy

We calculated bias, MAE, and RMSE to evaluate the estimation performance of each method relative to the reference PP.

Because the 72 analysis conditions were prespecified fixed conditions rather than randomly sampled scenarios, the results across analysis conditions were summarized descriptively. Table 6 and Fig. 3 present the mean and SD of bias, MAE, and RMSE for each method across the 72 analysis conditions (36 response-rate scenarios × 2 numbers of pipeline subjects).

Table 6
Overall estimation accuracy across the 72 analysis conditions

Method	Bias, mean (SD)	MAE, mean (SD)	RMSE, mean (SD)
AC	-0.018 (0.022)	0.059 (0.063)	0.098 (0.088)
ORP	-0.008 (0.018)	0.055 (0.063)	0.103 (0.098)
FRP	-0.045 (0.166)	0.104 (0.147)	0.153 (0.178)
AR	0.098 (0.140)	0.098 (0.140)	0.157 (0.184)
ANR	-0.276 (0.274)	0.276 (0.274)	0.344 (0.287)

Note: Values are the mean and standard deviation (SD) of the condition-specific performance measures across the 72 prespecified analysis conditions. These values represent descriptive summaries of the analysis conditions examined in this study.

Across the 72 analysis conditions, the mean bias (SD) was − 0.008 (0.018) for ORP, − 0.018 (0.022) for AC, − 0.045 (0.166) for FRP, 0.098 (0.140) for AR, and − 0.276 (0.274) for ANR. Thus, the mean bias was closest to 0 for ORP. The mean bias was positive for AR and negative for AC, ORP, FRP, and ANR.

The mean MAE (SD) was 0.055 (0.063) for ORP, 0.059 (0.063) for AC, 0.104 (0.147) for FRP, 0.098 (0.140) for AR, and 0.276 (0.274) for ANR. The lowest mean MAE was observed for ORP.

The mean RMSE (SD) was 0.098 (0.088) for AC, 0.103 (0.098) for ORP, 0.153 (0.178) for FRP, 0.157 (0.184) for AR, and 0.344 (0.287) for ANR. The lowest mean RMSE was observed for AC.

Decision agreement

We evaluated the agreement between the interim decisions based on the PP obtained using each method and those based on the reference PP.

Table 7 presents the mean decision agreement across the 36 response-rate scenarios, descriptively summarized according to the number of pipeline subjects, together with the change from $m = 5$ to $m = 10$.

For $m = 5$, the mean decision agreement was 0.938 for ORP, 0.914 for AC, 0.886 for FRP, 0.860 for AR, and 0.780 for ANR. For $m = 10$, the corresponding values were 0.919, 0.872, 0.846, 0.824, and 0.638, respectively. Thus, ORP had the highest mean decision agreement for both $m = 5$ and $m = 10$.

For all five methods, the mean decision agreement was lower for $m = 10$ than for $m = 5$. The changes from $m = 5$ to $m = 10$ were − 0.042 for AC, − 0.019 for ORP, − 0.039 for FRP, − 0.036 for AR, and − 0.143 for ANR. The decrease in mean decision agreement was smallest for ORP and largest for ANR.

Table 7
Mean decision agreement by number of pipeline subjects

Method	m = 5	m = 10	Change (m = 10 - m = 5)
AC	0.914	0.872	-0.042
ORP	0.938	0.919	-0.019
FRP	0.886	0.846	-0.039
AR	0.860	0.824	-0.036
ANR	0.780	0.638	-0.143

Discussion

Main findings

This study investigated the methodological issue of handling pipeline subjects in Bayesian interim monitoring with delayed binary responses, using a single-arm Bayesian phase II trial as the study setting. We compared five approaches for handling pipeline subjects whose outcomes remained pending at RDP. The results showed that the association with the reference PP, estimation error, and agreement in interim decision-making differed according to the method used to handle pipeline subjects.

The performance of the methods varied according to the performance measure considered. Across the 72 analysis conditions, ORP had the mean bias closest to 0, the lowest mean MAE, and the highest mean decision agreement. The AC method, meanwhile, had the lowest mean RMSE. Regarding the distribution of scenario-specific Pearson correlation coefficients, AC and ORP generally showed relatively high correlations with the reference PP compared with the other methods. When examined according to the number of pipeline subjects, the median Pearson correlation coefficient was 0.815 for AC and 0.807 for ORP when $m = 5$, whereas the corresponding values were 0.647 and 0.661, respectively, when $m = 10$.

In the supplementary descriptive comparison restricted to scenarios in which the Pearson correlation coefficients were defined for both AC and ORP, AC yielded a higher Pearson correlation coefficient than ORP in a greater number of scenarios for both $m = 5$ and $m = 10$. A similar pattern was observed in the exploratory descriptive comparisons stratified by $p - p_0 \leq 0.3$ and $p - p_0 > 0.3$. However, these scenario-wise comparisons provide a different descriptive summary from the distribution of correlation coefficients across scenarios. For example, when $m = 10$, the median correlation coefficient was

slightly higher for ORP than for AC. These findings should therefore not be interpreted as establishing an overall advantage of one method over the other based on Pearson correlation alone.

The ANR method showed a different pattern. It had the largest negative mean bias among the five methods, indicating the greatest average underestimation of the reference PP under the examined conditions. It also had the largest mean MAE and mean RMSE. These findings are consistent with the way ANR handles pending outcomes by treating all pipeline subjects as nonresponders, which shifts the information used in PP calculation in a direction that tends to lower PP.

Overall, these findings indicate that approaches to handling pipeline subjects under delayed responses have different performance characteristics depending on the considered performance measure. The ORP method had the mean bias closest to 0, the lowest mean MAE, and the highest mean decision agreement, whereas the AC method had the lowest mean RMSE. Both methods also showed relatively high correlations with the reference PP across the examined scenarios. Thus, instead of drawing conclusions from a single performance measure, the methods should be evaluated by considering multiple aspects of performance, including association with the reference PP, the direction and magnitude of estimation error, and agreement in interim decision-making.

ORP method

The favorable performance of ORP in terms of mean bias, mean MAE, and decision agreement may be related to its methodological feature of using the observed response proportion among subjects with ascertained outcomes at RDP to construct a pseudo-response count for pipeline subjects and incorporating this information into PP calculation. Specifically, the observed response proportion x_o/n_o is used as a plug-in value for the response rate among pipeline subjects, and mx_o/n_o is incorporated into PP calculation as a fractional pseudo-response count. This approach allows the response information available at RDP to be incorporated into the handling of pipeline subjects, in contrast to methods that uniformly treat all pipeline subjects as either responders or nonresponders.

However, ORP's performance depends on how well the observed response proportion among subjects with ascertained outcomes at RDP represents the response pattern among pipeline subjects. Under the setting considered in this study, subjects with ascertained outcomes and pipeline subjects were assumed to share a common response probability, and response delay was assumed to be noninformative; that is, the time to outcome ascertainment was independent of response status. Under these assumptions, the observed response proportion may provide a useful data-adaptive plug-in value for handling pipeline subjects. If response delay is informative, however, subjects with ascertained outcomes at RDP may not be representative of pipeline subjects, and the plug-in value based on the observed response proportion may therefore lead to systematic deviation from the reference PP.

We should also emphasize that ORP is a plug-in approximation rather than an explicit marginalization over the possible outcomes of pipeline subjects. In particular, the pseudo-response count mx_o/n_o is

based directly on the observed response proportion and should not be interpreted as the posterior predictive expected number of responses among pipeline subjects under the Beta–Binomial model.

AC method

The AC method does not explicitly impute the pending outcomes of pipeline subjects and uses only subjects whose outcomes have been ascertained by RDP as the current observed data. In this study, AC had the lowest mean RMSE among the five methods. Regarding the distribution of scenario-specific Pearson correlation coefficients, the AC method, together with ORP, also showed relatively high correlations with the reference PP compared with the other methods.

One possible explanation for the relatively favorable performance of AC is its relationship to explicit marginalization under the model and assumptions considered in this study. Under the Beta–Bernoulli/Beta–Binomial model, when subjects with ascertained outcomes and pipeline subjects share a common response probability, response delay is noninformative, and no additional partial information on pipeline subjects is used, the PP obtained by explicitly marginalizing over the possible outcomes of pipeline subjects with respect to their posterior predictive distribution is identical to that obtained by updating the posterior distribution using only subjects with ascertained outcomes and treating pipeline subjects as part of the subjects whose outcomes remain to be observed in the final analysis. Thus, under these assumptions, AC can be interpreted as implicitly marginalizing over the pending outcomes of pipeline subjects rather than simply ignoring their uncertainty. This mathematical relationship is demonstrated in Appendix A.

The AC method has the practical advantages of requiring no explicit imputation model for the pending outcomes of pipeline subjects and being relatively straightforward to implement and explain. It should not, however, be interpreted as an assumption-free approach. The interpretation described above depends on the model and assumptions considered in this study. These include a common response probability for subjects with ascertained outcomes and pipeline subjects, noninformative response delay, and the absence of additional partial information on pipeline subjects, such as follow-up time.

If response delay is informative, if response probabilities differ according to relevant subject characteristics, or if additional partial information on pipeline subjects is available and informative, the assumptions underlying this interpretation may no longer hold, and the equivalence with explicit marginalization described above is not generally guaranteed. For example, if the time required to ascertain the primary endpoint differs systematically between responders and nonresponders, subjects whose outcomes have been ascertained at RDP may not be representative of all enrolled subjects. In such settings, the posterior predictive distribution based only on subjects with ascertained outcomes may not adequately represent the outcome distribution of pipeline subjects, potentially resulting in the systematic deviation of the PP obtained using AC from the reference PP.

Impact of the number of pipeline subjects

In this study, for all five methods, both the median Pearson correlation coefficient with the reference PP and the mean decision agreement were lower for $m = 10$ than for $m = 5$. In particular, AR and ANR showed relatively large decreases in the median Pearson correlation coefficient, and ANR also showed a marked decrease in mean decision agreement. These findings suggest that, as the number of pipeline subjects at RDP increases, the association between the PP obtained using each method and the reference PP may weaken, and disagreement in interim decision-making may become more frequent.

The AC and ORP methods showed relatively high correlations with the reference PP and relatively high decision agreement compared with the other methods, even when $m = 10$. Nevertheless, for both methods, the median Pearson correlation coefficient and mean decision agreement were lower for $m = 10$ than for $m = 5$. This indicates that neither method completely avoided the reduction in these performance measures associated with a larger number of pipeline subjects under the conditions examined in this study.

These findings suggest that the number of pipeline subjects present at RDP may be an important consideration when planning the timing of interim analysis, in addition to the number of subjects whose outcomes have been ascertained. In actual clinical trials, the number of pipeline subjects at RDP may depend on factors such as the rate of patient accrual and the time required to ascertain the primary endpoint. Therefore, when planning an interim analysis, it may be useful to consider the number of pipeline subjects expected to be present at RDP and to assess the robustness of the planned interim decision rule under a plausible number of pipeline subjects.

Practical interpretation

In practice, the choice of method for handling pipeline subjects should account for the number of pipeline subjects at RDP, the plausibility of the assumptions regarding response delay, and the proximity of PP to the decision threshold.

When response delay is noninformative, and subjects with ascertained outcomes and pipeline subjects can be reasonably assumed to share a common response probability, the AC method may be practical because it does not require an explicit imputation model for the pending outcomes of pipeline subjects and is relatively simple and transparent to implement. Under the Beta–Bernoulli/Beta–Binomial model considered in this study, and when no additional partial information on pipeline subjects is used, AC can also be interpreted as implicitly marginalizing over their pending outcomes with respect to posterior predictive distribution. By contrast, when the observed response proportion among subjects with ascertained outcomes at RDP can reasonably be considered representative of the response pattern among pipeline subjects, the ORP method, which uses a fractional pseudo-response count based on the observed response proportion, may also be a useful option.

In this study, both the median Pearson correlation coefficient and the mean decision agreement were lower for $m = 10$ than for $m = 5$ for all five methods. This suggests that greater attention to the

handling of pending outcomes may be warranted when a larger number of pipeline subjects is expected to be present at RDP.

In actual trials, the primary method for handling pipeline subjects and the interim decision rule should be prespecified. In addition, sensitivity analyses using alternative approaches may be considered to assess the robustness of interim decisions to the handling of pipeline subjects. Such sensitivity analyses may be particularly informative when the number of pipeline subjects is large or when PP is close to the decision threshold. This is because differences in the handling of pending outcomes may affect whether the prespecified interim decision criterion is met.

Limitations

This study has several limitations.

First, we used the reference PP—calculated using the subsequently ascertained outcomes of all pipeline subjects who had already been enrolled at RDP—as the benchmark for comparison. Because the outcomes of pipeline subjects are not available at the actual RDP, the reference PP cannot be implemented as a decision-making criterion at that time. Accordingly, the reference PP should be interpreted as a retrospective oracle benchmark for evaluating the PP estimation performance of each method and the agreement of interim decisions with those based on the reference PP, rather than as an implementable interim monitoring approach.

Second, we fixed the maximum sample size at $N = 40$, the number of subjects with ascertained outcomes at RDP at $n_o = 20$, and the number of pipeline subjects at either $m = 5$ or $m = 10$. On that basis, we considered 36 prespecified response-rate scenarios satisfying $p > p_0$. Therefore, similar results might not necessarily be obtained under other response-rate settings, including conditions in which $p \leq p_0$. In addition, the 72 analysis conditions were prespecified fixed conditions rather than scenarios randomly sampled from a population. Accordingly, the distributions of performance measures across analysis conditions should be interpreted as descriptive summaries of the conditions examined in this study. Furthermore, the scenario-specific comparisons between AC and ORP and the exploratory comparisons stratified by $p - p_0$ were descriptive analyses within prespecified scenarios and were not intended for statistical hypothesis testing. Moreover, the findings may not be directly generalizable to larger trials, those in which interim analyses are conducted at different time points, or those with multiple interim analyses.

Third, we assumed that the time to outcome ascertainment was unrelated to the eventual response outcome; that is, response delay was noninformative. Specifically, rather than generating response times or outcome ascertainment times themselves, we generated the binary outcomes of subjects with ascertained outcomes and pipeline subjects from Bernoulli distributions with a common true-response probability. In this situation, whether a subject had an ascertained or pending outcome at RDP was independent of the eventual response outcome. Consequently, we did not directly evaluate the effects of

patient accrual rate, assessment period, response-time distribution, or overall trial duration on the performance of the methods.

In actual clinical trials, nonresponse may be identified early because of early disease progression, whereas confirmation of a response may require a longer follow-up period. When the time to outcome ascertainment is associated with response status in this manner, subjects whose outcomes have been ascertained at RDP may not be representative of all enrolled subjects, and treating pipeline subjects and subjects yet to be enrolled as exchangeable might not be appropriate. Furthermore, the observed response proportion among subjects with ascertained outcomes may not adequately represent the response pattern among pipeline subjects. Therefore, under settings with informative response delay or with different patient accrual rates, assessment periods, response-time distributions, or trial durations, the performance of the methods, including AC and ORP, may differ from that observed in this study. Further investigation under such settings is warranted.

Fourth, we used only a $\mathrm{Beta}\,(1,1)$ prior distribution for response probability p. The choice of prior distribution may affect the posterior distribution and PP, and this influence may be nonnegligible, particularly in phase II trials with limited sample sizes. Therefore, future studies should investigate the performance of the methods under alternative prior distributions.

Fifth, we primarily evaluated the association with an estimation error relative to the reference PP, as well as agreement in interim decision-making with decisions based on the reference PP. However, a comprehensive evaluation of adaptive interim monitoring methods should also consider overall trial operating characteristics, including the type I error rate, power, probability of early stopping, expected sample size, and expected trial duration.

Sixth, the Pearson correlation coefficient measures the linear association between two quantities and does not directly assess absolute agreement between PP values. Therefore, a high correlation coefficient may be observed even in the presence of systematic overestimation or underestimation relative to the reference PP. In addition, in some analysis conditions, the PP obtained using a given method or the reference PP had zero variance across the simulation replicates, and the Pearson correlation coefficient was therefore mathematically undefined. We excluded these conditions from the calculation of the median and IQR of the Pearson correlation coefficients, and the number of analysis conditions for which the correlation coefficient could be calculated was reported. Given these characteristics of the Pearson correlation coefficient, we did not evaluate method performance based on correlation alone; rather, it was assessed using multiple performance measures, including bias, MAE, RMSE, and decision agreement.

Conclusion

Under the examined conditions—a single-arm Bayesian phase II trial with a maximum sample size of $N = 40$, $n_o = 20$ subjects with ascertained outcomes at RDP, $m = 5$ or $m = 10$ pipeline subjects, and noninformative response delay—PP estimation performance and agreement in interim decision-

making differed according to the method used to handle pipeline subjects. The ORP method had the mean bias closest to 0, lowest mean MAE, and highest mean decision agreement, whereas AC had the lowest mean RMSE. Regarding the distribution of scenario-specific Pearson correlation coefficients, AC and ORP generally showed relatively high correlations with the reference PP compared with the other methods. These findings indicate that the methods' performance varied according to the performance measure. This highlights the importance of evaluating multiple aspects of performance—including association with the reference PP, estimation error, and agreement in interim decision-making—rather than relying on a single performance measure.

When response delay is noninformative, and subjects with ascertained outcomes and pipeline subjects can reasonably be assumed to share a common response probability, AC may provide a simple, transparent, and practical option that does not require an explicit imputation model for pending outcomes. Moreover, under the Beta–Bernoulli/Beta–Binomial model considered in this study, when no additional partial information on pipeline subjects is used, the AC method can be interpreted as implicitly marginalizing over their pending outcomes with respect to posterior predictive distribution. By contrast, when the observed response proportion among subjects with ascertained outcomes at RDP can reasonably be considered representative of the response pattern among pipeline subjects, ORP may also provide a useful option.

When a larger number of pipeline subjects is present at RDP, when response delay may be informative, or when PP is close to the decision threshold, the primary method for handling pipeline subjects and the interim decision rule should be prespecified. Sensitivity analyses using alternative approaches may also be considered to assess the robustness of interim decisions to the handling of pipeline subjects.

Future studies should further clarify the applicability of methods for handling pipeline subjects in Bayesian adaptive clinical trials with delayed responses by considering informative response delay, multiple interim analyses, alternative prior distributions, and overall trial operating characteristics.

Abbreviations

AC
available-case method
ANR
all-nonresponse method
AR
all-response method
CI
confidence interval
FRP
fixed-rate plug-in method
IQR

interquartile range
MAE
mean absolute error
ORP
observed-rate plug-in method
PP
predictive probability
RDP
recruitment decision point

Declarations

Ethics approval and consent to participate

Not applicable. This study was based on computer simulations and did not involve human participants, human data, or human biological materials.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author Contribution

MT conceived and designed the study, developed the simulation framework, performed the statistical analyses, and drafted the manuscript. YI contributed to the study design and interpretation of the results and critically revised the manuscript. AT contributed to the interpretation of the findings and critically revised the manuscript. AK supervised the study, contributed to the methodological development, and critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgments

Not applicable.

Availability of data and materials

No empirical datasets were used in this study. All analyzed data were generated through computer simulations. The R code used to generate and analyze the simulated data is available from the corresponding author upon reasonable request.

References

1. Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. Stat Med. 1995;14:357–79.
2. Berry DA. Bayesian clinical trials. Nat Rev Drug Discov. 2006;5:27–36.
3. Lee JJ, Liu DD. A predictive probability design for Phase II cancer clinical trials. Clin Trials. 2008;5:93–106.
4. Yin G, Chen N, Lee JJ. Phase II trial design with Bayesian adaptive randomization and predictive probability. J R Stat Soc Ser C Appl Stat. 2012;61:219–35.
5. Hampson LV, Jennison C. Group sequential tests for delayed responses (with discussion). J R Stat Soc B. 2013;75:3–54.
6. Schüürhuis S, Konietschke F, Kunz CU. A two-stage group-sequential design for delayed treatment responses with the possibility of trial restart. Stat Med. 2024;43:2368–88.
7. Schüürhuis S, Wassmer G, Kieser M, Pahlke F, Kunz CU, Herrmann C. Two-stage group-sequential designs with delayed responses—what is the point of applying corresponding methods? BMC Med Res Methodol. 2024;24:242.

Figures

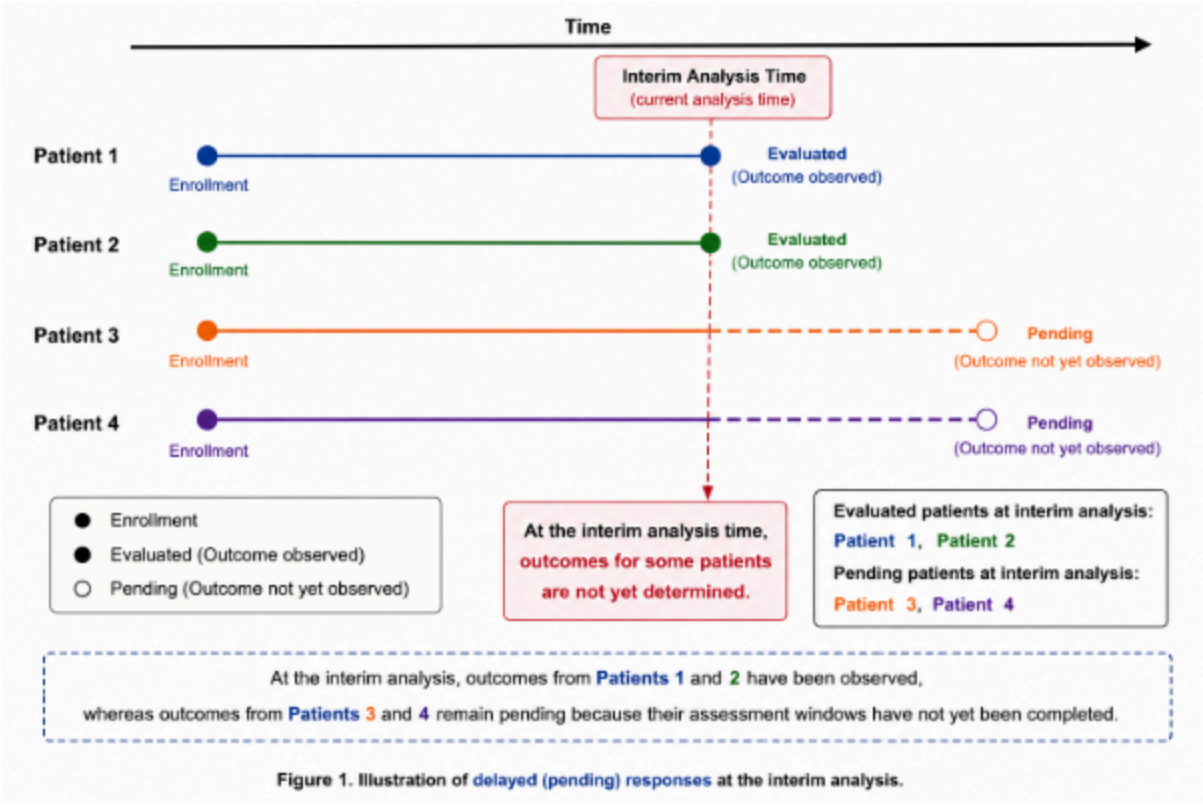


Figure 1

Conceptual illustration of delayed responses and pipeline subjects at RDP

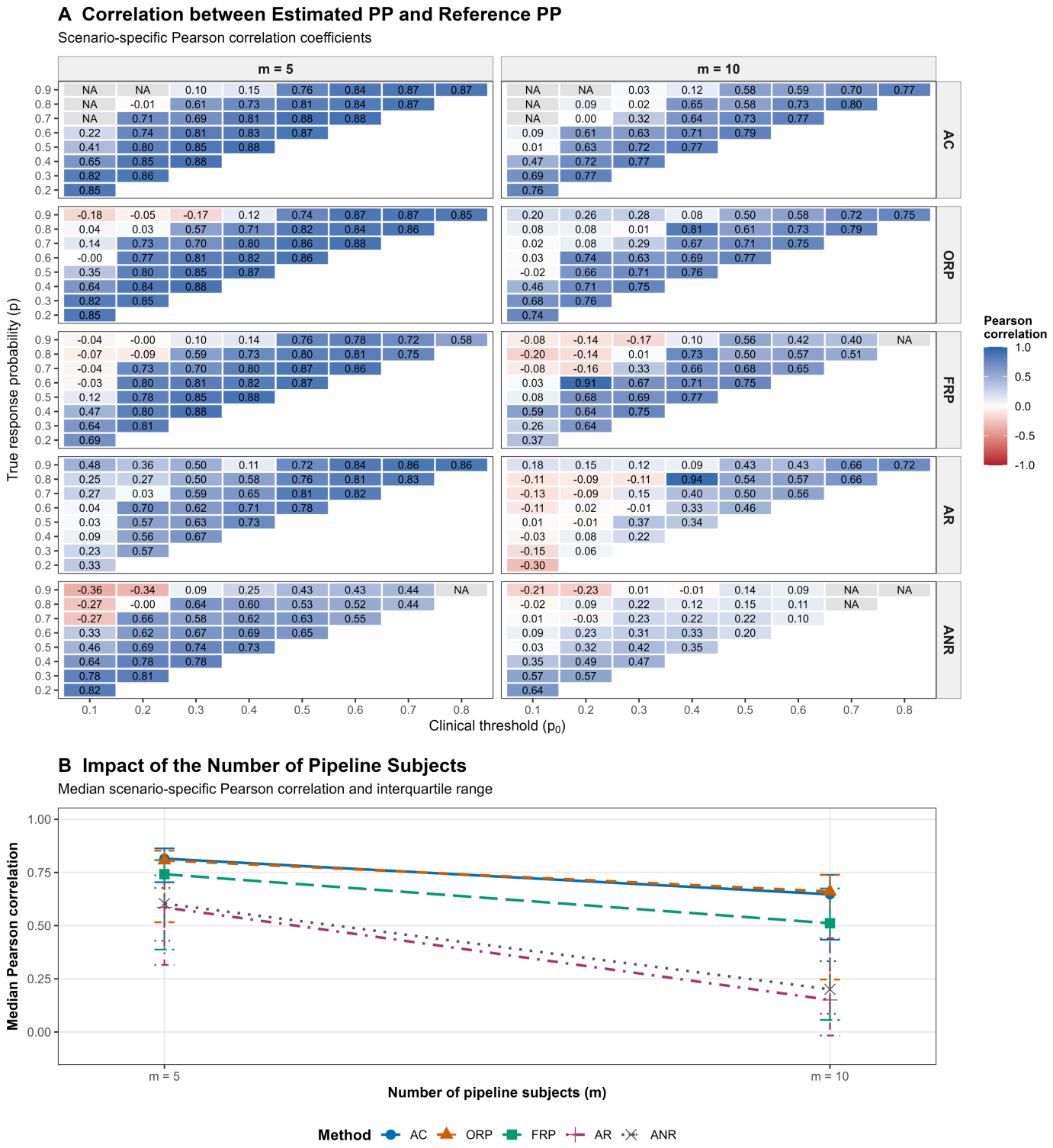


Figure 2

Pearson correlation coefficients between PP obtained using each method and reference PP according to pipeline subject number. The heatmaps show the Pearson correlation coefficients for each response-rate scenario for m = 5 and m = 10. The lower panel shows the median and IQR among scenarios for which the Pearson correlation coefficient was defined.

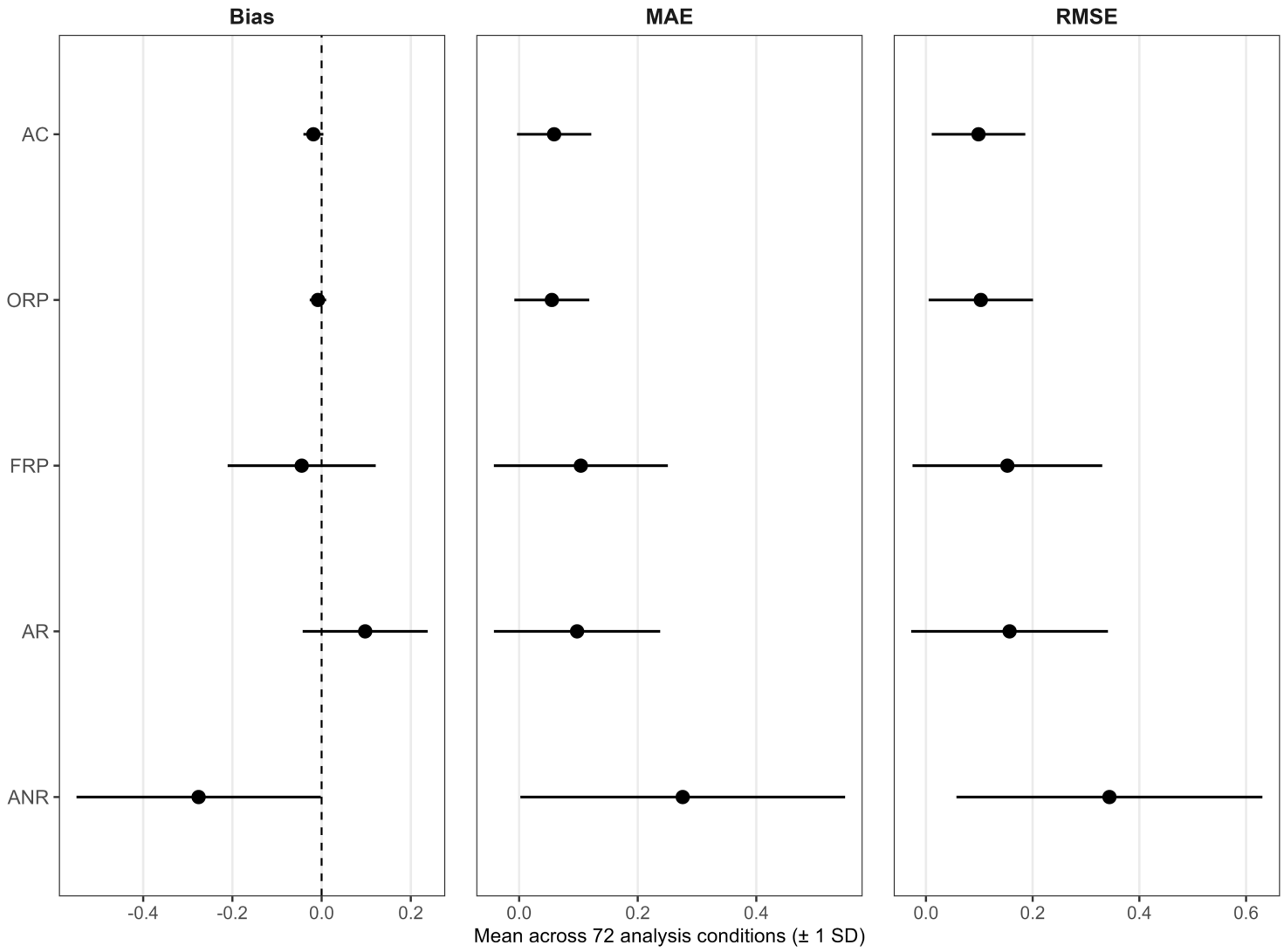


Figure 3

Overall estimation accuracy across the 72 analysis conditions. Points represent the mean bias, MAE, and RMSE for each method across the 72 prespecified analysis conditions, and horizontal lines represent mean ± 1 SD. The dashed line in the bias panel indicates bias = 0.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx