Jackson *et al.*
*BMC Medical Research Methodology* (2025) 25:149
https://doi.org/10.1186/s12874-025-02537-x

BMC Medical Research Methodology

RESEARCH

Open Access

# A comparison of frequentist and Bayesian approaches to the Personalised Randomised Controlled Trial (PRACTical)—design and analysis considerations

Holly Jackson[1], Yiyun Shou[2,3,4], Nur Amira Binte Mohamed Azad[3], Jing Wen Chua[5], Rebecca Lynn Perez[2], Xinru Wang[6], Marlieke E. A. de Kraker[1] and Yin Mo[2,5,7,8,9,10*]

## Abstract

**Background** Multiple treatment options frequently exist for a single medical condition with no single standard of care (SoC), rendering a classic randomised trial comparing a specific treatment to a control treatment infeasible. A novel design, the personalised randomised controlled trial (PRACTical), allows individualised randomisation lists and borrows information across patient subpopulations to rank treatments against each other without comparison to a SoC. We evaluated standard frequentist analysis with Bayesian analyses, and developed a novel performance measure, utilising the precision in treatment coefficient estimates, for treatment ranking.

**Methods** We simulated trial data to compare four targeted antibiotic treatments for multidrug resistant bloodstream infections as an example. Four patient subgroups were simulated based on different combinations of patient and bacteria characteristics, which required four different randomisation lists with some overlapping treatments. The primary outcome was binary, using 60-day mortality. Treatment effects were derived using frequentist and Bayesian analytical approaches, with logistic multivariable regression. The performance measures were: probability of predicting the true best treatment, and novel proxy variables for power (probability of interval separation) and type I error (probability of incorrect interval separation). Several scenarios with varying treatment effects and sample sizes were compared.

**Results** The Frequentist model and Bayesian model using a strong informative prior, were both likely to predict the true best treatment ($P_{best} \geq 80\%$) and gave a large probability of interval separation (reaching a maximum of $P_{IS} = 96\%$), at a given sample size. Both methods had a low probability of incorrect interval separation ($P_{IIS} < 0.05$), for all sample sizes ($N = 500 - 5000$) in the null scenarios considered. The sample size required for probability of interval separation to reach 80% ($N = 1500 - 3000$), was larger than the sample size required for the probability of predicting the true best treatment to reach 80% ($N \leq 500$).

**Conclusions** Utilising uncertainty intervals on the treatment coefficient estimates are highly conservative, limiting applicability to large pragmatic trials. Bayesian analysis performed similarly to the frequentist approach in terms of predicting the true best treatment.

*Correspondence:
Yin Mo
mdcmy@nus.edu.sg
Full list of author information is available at the end of the article

**Keywords** PRACTical, Bayesian, Statistical simulation, Personalised randomisation, Trial design, Infectious diseases, Antibiotic treatment

## Background

With rapid advancements in medical science, novel treatments are increasingly complex and often cater to specific patient subgroups [1, 2]. This trend towards individualised care requires stringent enrolment criteria for conventional two-arm randomised controlled trials and often makes them infeasible to perform due to high costs and limited generalisability [3]. In addition, multiple alternative treatment options for the same disease are often assumed to be of similar efficacy with a lack of clear evidence showing which is best [4–6]. This means there are multiple standards of care (SoC), posing challenges in defining a single control arm for comparison. New approaches for randomised controlled trials (RCTs) are needed to overcome these barriers.

Multi-arm trials allow multiple interventions or treatment arms to be compared simultaneously [7]. Such trials offer potential efficiency gains, i.e. time, costs and participant recruitment efforts, by reducing the need for separate trials for each new intervention. However, multi-arm trials impose more stringent eligibility criteria for participation than a two-arm trial, as a patient often should be eligible for all treatments in the trial, and they frequently require a single SoC [8–10].

Magaret et al. [11] proposed a multi-arm trial design which did not require a single 'SoC' treatment. Walker et al. [12] extended this idea by proposing the "Personalised Randomised Controlled Trial" (PRACTical) design. This design allows each trial participant to be randomised amongst treatments in a "personalised" randomisation list that only contains treatments suitable for them. Hence, a patient may be enrolled into the trial, as long as they are eligible for at least two treatments in the overall 'master list' of treatment options. The analysis is likened to a network meta-analysis, where participants randomised within the same randomisation list are analogous to a single trial and the multiple different lists represent the multiple different trials. Both direct and indirect evidence are used to rank the efficacy of each treatment across the master randomisation list. The ongoing NeoSep1 trial (ISRCTN 48721236), which aims to rank antibiotic regimens for neonatal sepsis, is the first trial to use the PRACTical design.

The initial PRACTical design [13] utilised a multivariable logistic regression model which incorporated both the treatments and patient subpopulations as fixed effects, using a frequentist approach, to rank treatments. Alternatively, a Bayesian approach could be utilised to create such a multivariable logistic regression model, to allow the inclusion of prior (or historical) information gathered from previous observational studies and/or clinical trials. The main performance measures suggested in the PRACTical design [12–14] were based on point estimates, as Walker et al. [12] suggest it is more important to avoid the worst treatment, than pick a single best treatment. This approach may be well suited if a trial only includes treatments which are already approved and clinically accepted for eligible patients. However, if a trial includes novel treatments, or one would want to remove inferior treatments from practice guidelines or a platform trial, confidence intervals would be important for considering type I and type II errors.

The objective of this simulation study is to extend the analysis of the PRACTical design and compare a Bayesian approach with the previously recommended frequentist approach. In addition, we will evaluate the feasibility of a novel performance measure, which considers uncertainty intervals for the treatment coefficient estimates, against those measures which have been previously recommended.

## Methods

### Motivational example

An example for demonstrating the performance of the PRACTical design is a multicentre randomised trial comparing effectiveness of treatments for severe Gram-negative bacterial infections, such as bloodstream infections. There is no single SoC for the targeted treatment of highly multidrug resistant Gram-negative bacterial infections [6] and given the numerous deciding factors behind antibiotic choice, the PRACTical design would be appropriate in this setting.

In the PRACTical design, an individual patient will be randomised, with equal probability, between treatments to which they are eligible. This personalised randomisation list of treatments is referred to as a patient's *pattern* [13], patients who share the same pattern are part of the same subgroup. In our simulated trial, a patient may not be eligible for a particular antibiotic treatment due to allergy or intolerable side effects, and/or because the bacterial pathogen(s) cannot be effectively targeted by the treatment.

The primary outcome of the simulated trial was 60-day mortality (binary). We explored total sample sizes ranging between 500 and 5,000 patients, recruited equally among 10 sites, where the distribution of

**Table 1** Possible treatment options and randomisation lists

| Antibiotic Treatment, *j* | Pattern $S_1$ | Pattern $S_2$ | Pattern $S_3$ | Pattern $S_4$ |
|---|---|---|---|---|
| *A* | ✗ | ✓ | ✗ | ✓ |
| *B* | ✓ | ✓ | ✓ | ✓ |
| *C* | ✓ | ✓ | ✓ | ✓ |
| *D* | ✗ | ✗ | ✓ | ✓ |

patient subgroups was assumed equal across sites. The presumed mortality through our scenarios was mostly varied between 20 to 50%, similar to what is expected for patients with highly multidrug resistant Gram negative bloodstream infections [15, 16].

### Notation

Four treatments were denoted by $j = A, B, C, D$. Each patient was assigned a subgroup $k = 1, ..., K$, where $K = 4$, based on which treatments they were eligible for, at each trial site$q = 1, ..., Q$, where $Q = 10$. Each patient was then only randomised to a treatment within their pattern, $S_{kq}$, where patients in the same subgroup shared the same pattern. We assumed all treatments were available at each trial site, and each site had patients eligible to the four different patterns. Hence, each pattern, $S_k$, was the same for all sites.

The total trial sample size was denoted by $N$. Then, the sample size of each patient subgroup, $k$, was $n_k = \lambda_k \cdot N$, where $\lambda_k$ represented the frequency (%) of each subgroup,$k = 1, ..., 4$, in the trial. Patient subgroup and their site allocation were generated from multinomial distributions, where each patient had a probability of $1/Q$ and $\lambda_k = 1/K$ of being from each site and assigned to each subgroup $k$, respectively.

The probability of a patient from subgroup $k$, who was assigned treatment $j$, meeting the primary outcome, was denoted by $P_{jk} = P(y_{jk} = 1)$, where $y_{jk} = 1$ represented 60-day mortality and $y_{jk} = 0$ otherwise. The primary outcome was generated from a binomial distribution.

### Randomisation lists

In our simulated trial, personalised randomisation lists were drawn from a list of four possible antibiotic treatments, and patients were assigned to one of four subgroups (Table 1). In this arbitrary example, there was a minimum overlap of two treatments between patterns. The treatments in the simulated trial were then ranked using both direct and indirect evidence across the patient subgroups.

### Analysis

Similarly to network meta-analysis, we analysed all subgroups together, making use of direct comparisons (i.e. comparing treatments within each pattern), as well as indirect comparisons (i.e. comparing treatments across patterns) to produce an overall ranking of all the treatments [14]. This assumed that patients were comparable within subgroups, which was established by randomisation based on the same pattern. An additional assumption required for network meta-analysis is consistency, which refers to the agreement between direct and indirect evidence, subject to the expected variation under the random-effects model for meta-analysis [17]. Here, we assumed the log odds ratio between treatments' mortality probabilities was consistent across all subgroups.

We implemented logistic regression as proposed by Lee et al. [13], with the binary 60-day mortality outcome as the dependent variable, and treatments and patient subgroups as independent categorical variables, included as fixed effects. This fixed effects model included a coefficient (the log odds for the risk of death), $\psi_{jk\prime}$, for each treatment, $j$, for a patient in the reference subgroup, $k\prime = 1$. It also included a coefficient (the log of the odds ratio for the risk of death for each patient subgroup, $k = 2, ..., K$, in comparison to the reference subgroup, $k\prime = 1$), $ln(\alpha_k/\alpha_{k\prime})$, for each patient subgroup, $k = 2, ..., K$.

$$logit(P_{jk}) = ln(\alpha_k/\alpha_{k\prime}) + \psi_{jk\prime}$$

In addition, we further investigated a mixed effects model, where the treatments were categorised as fixed effects and the patient subgroups were treated as random effects (Additional file 1). Neither the fixed effects nor mixed effects models included a coefficient to account for the potential difference in treatment mortality probability across sites.

We investigated two different analysis methods: a frequentist approach (using the R package 'stats' [18] and a Bayesian approach (via the R package 'rstanarm' [19]). The Bayesian method used three strongly informative normal priors, based on one *representative,* and two *unrepresentative* historical datasets [20]. We chose to use both representative and unrepresentative priors because, in practice, it is unclear whether historical trial data will reliably represent a new trial. This approach allowed us to evaluate the impact of different priors on the performance of the Bayesian method. The Bayesian approach used a prior for all seven coefficients (four treatment coefficients and three subgroup coefficients). The means of the priors were calculated by fitting a logistic

**Table 2** Simulation scenarios

| Scenario, S | True mortality probability of treatments, $P_{jk}$ |
|---|---|
| S1 series: Various effects across treatments | |
| 1.1: Null | All treatments had equal mortality probabilities for all patient subgroups: $P_{jk}$= 37.5%, $j = A, B, C, D$ and $k = 1, 2, 3, 4$ |
| 1.2: One best treatment | Treatment *B* had a smaller mortality probability than treatments *A*, *C*, *D* for all patient subgroups: $P_{Bk}$= 35%, $k = 1, 2, 3, 4$; and $P_{jk}$= 45%, $j = A, C, D$ and $k = 1, 2, 3, 4$ |
| 1.3: Small difference | All 4 treatments had different mortality probabilities, which was equal across patient subgroups: $P_{Ak}$= 30%; $P_{Bk}$= 35%; $P_{Ck}$= 40%; $P_{Dk}$= 45%,$k = 1, 2, 3, 4$ |
| 1.4: One worst treatment | Treatment *B* had a larger mortality probability than treatments *A*, *C*, *D* for all patient subgroups: $P_{Bk}$= 45%$k = 1, 2, 3, 4$; and $P_{jk}$= 35%,$j = A, C, D$and$k = 1, 2, 3, 4$ |
| S2 series: Different effects across subgroups or sites | |
| 2.1: Null, Subgroup effect | All treatments had equal mortality probabilities, but it was different across the patient subgroups: Subgroup 1:$P_{j1}$= 30%,$j = B, C$ Subgroup 2:$P_{j2}$= 35%,$j = A, B, C$ Subgroup 3:$P_{j3}$= 40%,$j = B, C, D$ Subgroup 4:$P_{j4}$= 45%,$j = A, B, C, D$ |
| 2.2[a]: Small difference, subgroup effect | All treatments had different mortality probabilities, which varied across patient subgroups: Subgroup 1:$P_{A1}$= *NA*,$P_{B1}$= 25.0%,$P_{C1}$= 29.9%,$P_{D1}$= *NA* Subgroup 2:$P_{A2}$= 29.1%,$P_{B2}$= 35.4%,$P_{C2}$= 41.3%,$P_{D2}$= *NA* Subgroup 3:$P_{A3}$= *NA*,$P_{B3}$= 71.1%,$P_{C3}$= 76.0%,$P_{D3}$= 79.9% Subgroup 4:$P_{A4}$= 78.2%,$P_{B4}$= 82.8%,$P_{C4}$= 86.1%,$P_{D4}$= 88.6% |
| 2.3: Small difference, site effect | All treatments had different mortality probabilities, which were constant across patient subgroups, which were the same in 8 study sites but different to the other 2 sites: $q = 3, ..., 10$: $P_{Akq}$= 30%,$P_{Bkq}$= 35%,$P_{Ckq}$= 40%,$P_{Dkq}$= 45%,$k = 1, 2, 3, 4$, $q = 1, 2$: $P_{Akq}$= 20%,$P_{Bkq}$= 25%,$P_{Ckq}$= 30%,$P_{Dkq}$= 35%,$k = 1, 2, 3, 4$, |

Treatments were represented as *j*, where $j = A, B, C, D$. Each patient was randomised to a treatment within their pattern,$s_k$. The group of patients who shared the same pattern,$s_k$, was referred to as a subgroup, *k*. Unless otherwise stated, treatment mortality probabilities were homogenous across all $Q = 10$ sites and all $K = 4$ subgroups

[a] The use of decimals allows us to keep the treatment difference (the log odds ratio for mortality probability between treatments) roughly equal across the patient subgroups

regression model to the historical dataset and the standard deviation of the priors was assigned '1'. This standard deviation value was selected based on preliminary simulations which showed that it was strongly informative, but flexible enough to allow the data to influence the posterior distribution when an unrepresentative prior was utilised.

Here, historical datasets referred to simulated data representing a previous study of sample size 500; 'representative' referred to the case where the underlying distribution of the simulated historical dataset was identical to the simulated current trial dataset; 'unrepresentative' referred to the case where the underlying distribution of the simulated historical dataset was different from the simulated current trial dataset. (Table SM1 in Additional file 1).

### Simulation scenarios

We simulated 1000 trial datasets based on each of the following scenarios (Table 2):

a) Various mortality probabilities per antibiotic, which was constant across patient subgroups (Scenarios S1);
b) Various mortality probabilities across patient subgroups, with constant log odds ratios between treatments across subgroups (Scenarios S2.1 to 2.2);
c) Various treatment mortality probabilities across trial sites (Scenarios S2.3).

Baseline mortality probability (probability of mortality if no treatment was received) was assumed to be higher than the worst performing treatment per subgroup for all scenarios. Additional scenarios, varying average mortality probabilities and patient subgroup frequencies, can be found in Additional file 1 (Table SM2).

### Performance measures

We utilised the following performance measures, listed in Turner et al. [14] to compare the four analysis methods across scenarios. Additional performance measures

from Lee et al. [13] were also reported in Additional file 1. Here, the expectation,$E$, was taken with respect to the 1,000 trial iterations and the hat symbol, $\widehat{\ }$, was used to denote variables estimated from the simulated data.

The analysis methods explained above were used to evaluate the best treatment, creating an overall ranking across the subgroups. Let $\widehat{j}_k = argmin_{j \in S_k} \widehat{P}_{jk}$ indicate the estimated best treatment in pattern $S_k$. This was the treatment with the lowest estimated mortality rate $\widehat{P}_{jk}$, which would be recommended to patients in subgroup $k$ [13].

- Reduction in mortality ratio (RMR): the mean mortality reduction by choosing the top-ranked treatment, $\widehat{j}_k$, to treat a future population compared to picking a treatment at random, divided by the mean mortality reduction with the true best treatment compared to picking a treatment at random, range -Inf to 1 (desired). Let the observed frequency of each subgroup $k$, be written as $\widehat{\lambda}_k$.

  $\text{RMR} = E\left[\sum_k \widehat{\lambda}_k (ave_{j \in S_k} P_{jk} - P_{\widehat{j}_k k})\right] / E\left[\sum_k \widehat{\lambda}_k (ave_{j \in S_k} P_{jk} - min_{j \in S_k} P_{jk})\right]$
- Best treatment probability ($\text{P}_{\text{best}}$): proportion of simulations where the top-ranked treatment, $\widehat{j}_k$ , was truly the best treatment. Let $I(.)$ represent an indicator function, range 0 (not true best treatment across all trial iterations) to 1 (desired).
  $\text{P}_{\text{best}} = E\left[\sum_k \widehat{\lambda}_k I\left(P_{\widehat{j}_k k} = min_{j \in S_k} P_{jk}\right)\right]$
- Improvement over random choice (IORC): proportion of simulations where the top-ranked treatment, $\widehat{j}_k$, was truly better than a randomly chosen treatment, range 0 (no difference) to 1 (desired).
  $\text{IORC} = E\left[\sum_k \widehat{\lambda}_k I(P_{\widehat{j}_k k} < ave_{j \in S_k} P_{jk})\right]$

In addition, we proposed a proxy variable for power and type I error. This was based on a comparison of the treatment coefficients and their 80% Wald confidence intervals (CIs) for the frequentist method, or 80% central credible intervals (CrIs, sometimes referred to as uncertainty intervals), for the Bayesian methods [19], for each treatment. When comparing the CIs/CrIs of a pair of treatments (

$A$ and $B$) if they were separated (ie did not overlap), we concluded them to be sufficiently different (where the treatment with the largest coefficient was selected as worst), unless there existed a third treatment ($C$) whose CIs/CrIs overlapped with the CIs/CrIs of the first two treatments ($A$ and $B$), and in this case we concluded equivalence between all three treatments. The same reasoning was applied if there were a higher number of treatments. The names we have given to these proxy variables depended on the underlying truth of the scenario investigated.

The size of the CIs/CrIs used will have an impact on how these measures perform; increasing the size of the intervals would produce more overlap and thus, a more conservative measure, while decreasing their size could produce less overlap and hence, a more powerful measure. We chose 80% CIs/CrIs as this balanced sample size feasibility.

#### Probability of incorrect interval separation

For the null scenarios, probability of incorrect interval separation was used as a proxy variable for type I error. It was estimated by the proportion of iterations which *incorrectly* identified any of the treatments as best, singularly or in combination with any of the other treatments. As this measure is a proportion it is bound between 0 (desired) and 1.

Each graph in Fig. 1 shows four treatment options (ordered on the y axis) with their estimated treatment coefficient and 80% CIs/CrIs (indicated on the x axis). The first row represents the underlying true (null) scenario. The second row illustrates iterations that would not commit a $\text{P}_{\text{IIS}}$ error, as all four treatments have overlapping CIs/CrIs and thus, the correct conclusion of all four treatments being equivalent is reached. The final three rows illustrate iterations that would commit a $\text{P}_{\text{IIS}}$ error, classified into three types: $\text{P}_{\text{IIS}}$ (i), (ii), (iii) depending on if the trial incorrectly concludes 1, 2 or 3 best treatments, respectively. The three types of probability of incorrect interval separation have varying degrees of repercussions. For example, it is worse to incorrectly conclude a single best treatment (i), therefore recommending the removal of three perfectly adequate treatments from the SoC, than concluding 3 equally best treatments (iii), thus removing only one treatment from the SoC guidelines.

We considered that a probability of incorrect interval separation below some level $\alpha$ (for example 5%), implied the probability of concluding a difference in treatments when one does not exist (type I error) is controlled at the $\alpha$ level, as such, there is no need for multiplicity adjustment.

#### Probability of interval separation

Probability of interval separation was used as a proxy variable for power. It was estimated by the proportion of iterations which correctly identified differences between the treatments. This measure is context specific, depending on if the focus is on correctly identifying the predefined worst treatment(s), the true best treatment(s), or the true ranking of all included treatments. We chose to focus on finding the predefined worst treatment(s) in

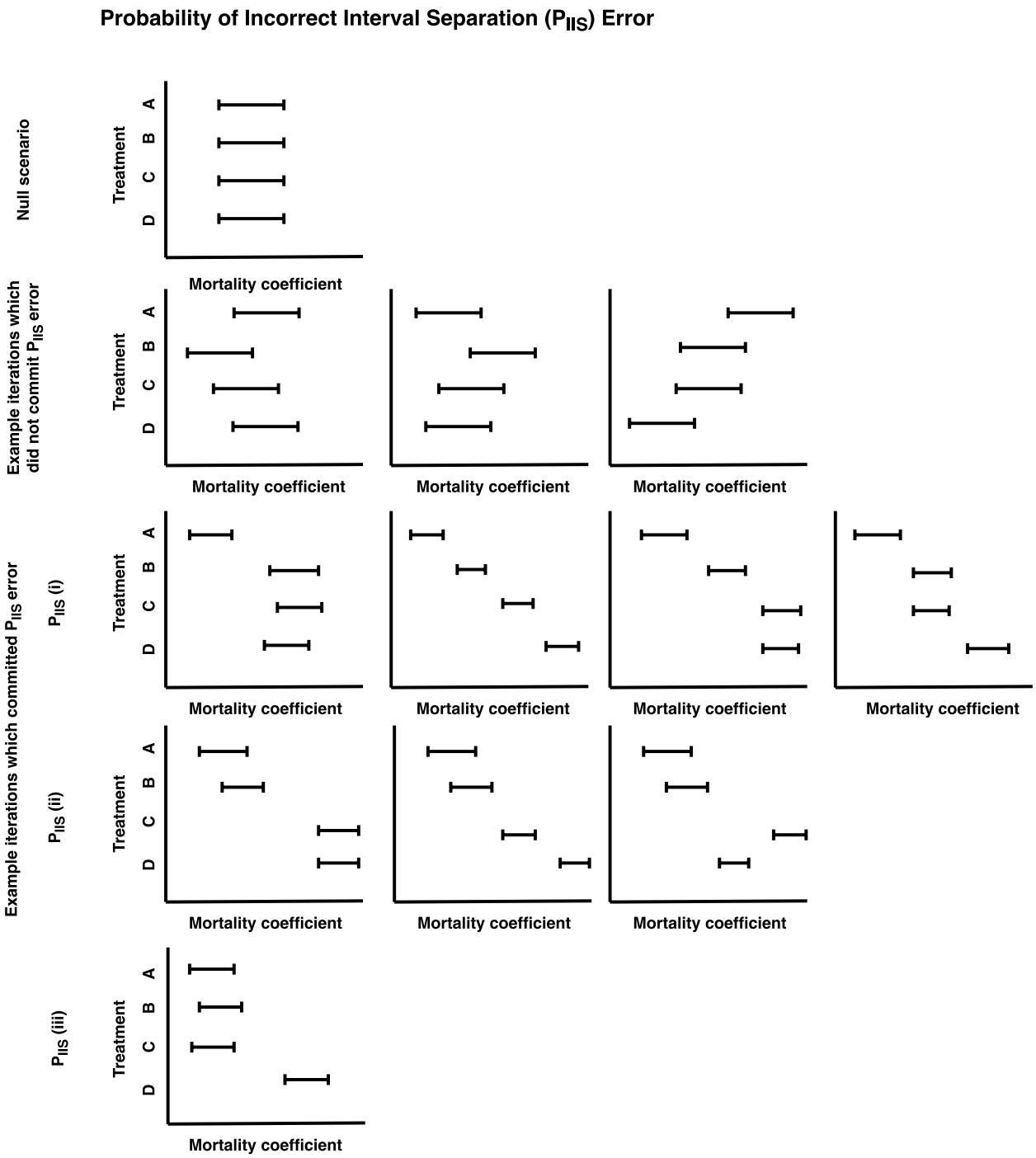


**Fig. 1** Probability of incorrect interval separation ($P_{IIS}$)

order to remove them from practice guidelines. As this measure is a proportion, it is bound between 0 and 1 (desired).

In the simulation scenario where there were three worst treatments (Scenario 1.2, Fig. 2a), simulation iterations which concluded that the single predefined best treatment was indeed best, were considered to have correctly identified interval separation. In the simulation scenarios where all treatment mortality probabilities were different (Scenarios 1.3, 2.2–2.3, Fig. 2b), we considered two instances where an iteration would correctly identify interval separation. In the first, PIS (i),

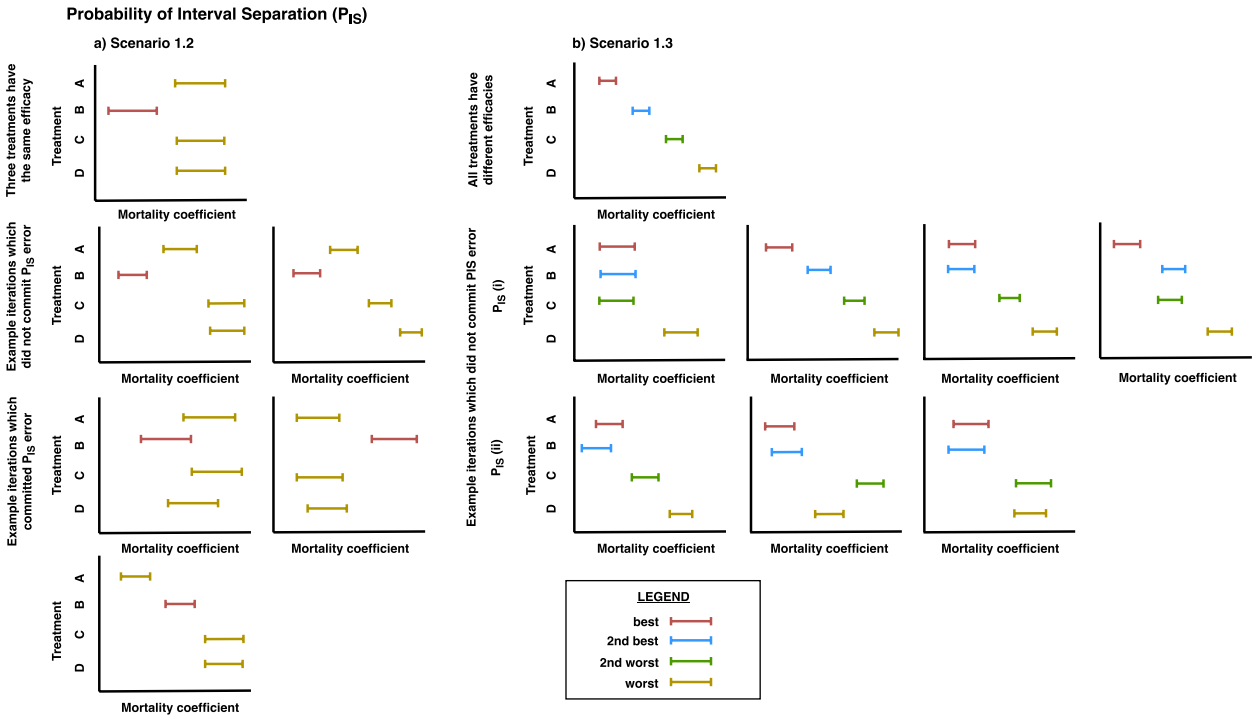


**Fig. 2** Probability of interval separation, in scenario 1.2 (**a**) and 1.3 (**b**)

if the trial correctly identified the true worst treatment and, to allow for more flexibility in our measure, in the second, PIS (ii), if the trial correctly identified at least one of the true bottom two treatments.

Each graph shows four treatment options (ordered on the y axis) with their estimated coefficient and 80% CIs/CrIs indicated on the x axis. Each treatment's true ranking is colour-coded shown in the legend. The first row demonstrates the underlying true scenarios. For Scenario 1.2, the second row illustrates iterations that would not commit $P_{IS}$ error, as the true best treatment is correctly identified. The third and fourth rows illustrate iterations that would commit $P_{IS}$ error, as the true best treatment is not correctly identified. For Scenario 1.3, the second and third rows each illustrate iterations that would not commit $P_{IS}$ under each definition (i-ii): $P_{IS}$ (i), iterations which concluded the predefined worst treatment as worst and $P_{IS}$ (ii), iterations which concluded either or both of the bottom two predefined treatments as worst.

The null scenarios would give zero for the performance measures: reduction in mortality ratio, best treatment probability, improvement over random choice and probability of interval separation, therefore, these performance measures are not presented for the null scenarios. However, the null scenarios remain informative for the probability of incorrect interval separation.

All simulations and analyses were performed in the R environment version 4.1.1 [18]. All analyses and plots can be reproduced using codes at https://github.com/moyinNUHS/practical_BayesSI.

## Results

We found that the Bayesian approach performed similarly to the frequentist approach for all scenarios, in all performance measures. However, as we utilised three different priors based on the three different historical datasets, the performance of the Bayesian approaches varied across scenarios. There were some scenarios when using an unrepresentative prior outperformed the representative prior (e.g. scenarios 1.3–1.4). However, all methods performed similarly, and the only noticeable differences were at small sample sizes ($N=500$), where the Bayesian approach performed better. For all methods, in all scenarios, all the performance measures improved with increased sample size. The performance measures: reduction in mortality ratio, best treatment probability, and improvement over random choice were consistently closer to 1 than the probability of interval separation, for all scenarios and all methods for equivalent sample sizes.

### Scenario 1.1 and 2.1: null scenarios

All methods consistently produced small probabilities of incorrect interval separation, (i.e. $<0.05$) concluding

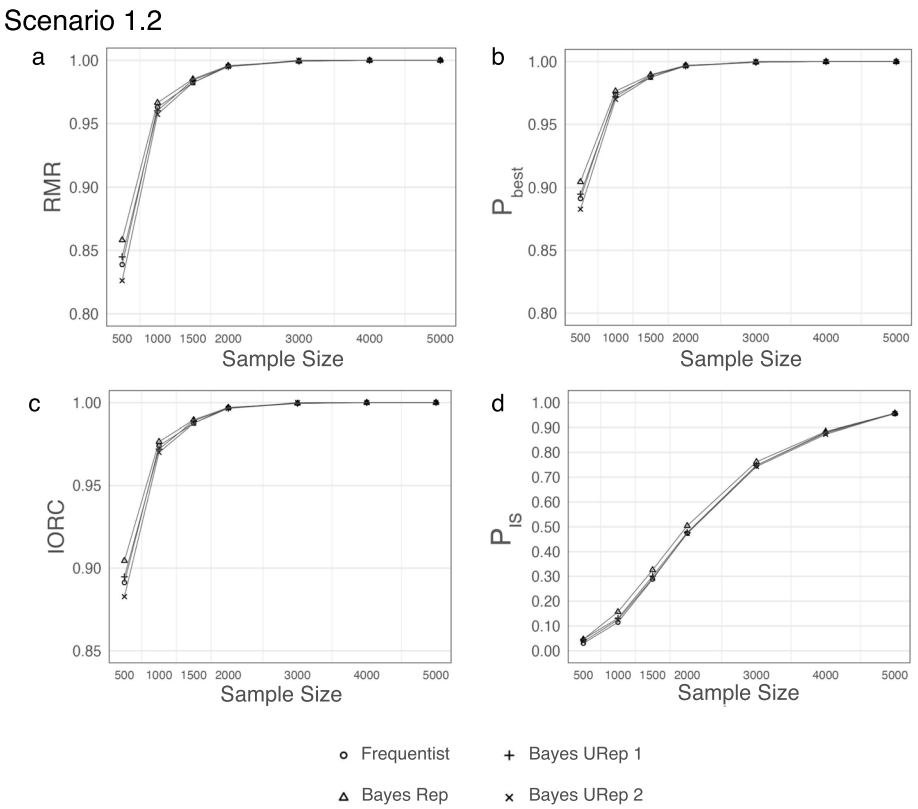


**Fig. 3** Results for scenario 1.2

one, two or three 'best treatment(s)' in all the null scenarios (Figure SM 1, Table SM 3, in Additional file 1). The treatments which were included in only two patterns (treatment *A* and *D*) were more likely to be incorrectly predicted as 'best', when utilising only the point estimates to rank the treatments (Figure SM 2, Table SM 3, in Additional file 1).

#### Scenario 1.2: one best treatment

When there was one true best treatment and three similarly bad treatments (scenario 1.2, Fig. 3a-c), reduction in mortality ratio, best treatment probability, and improvement over random choice were large (roughly 0.86, 0.90 and 0.90, respectively, for Bayes representative method with a sample size of 500). However, even though all methods often correctly predicted the true best treatment, the probability of interval separation was much lower due to the overlapping CIs/CrIs (Fig. 3d). For example, the Bayes representative method only reached a probability of interval separation of 0.80 with a sample size of roughly 3,250 patients (Table SM 4 in Additional file 2).

Reduction in mortality ratio (a), best treatment probability (b), improvement over random choice (c), probability of interval separation (d) in scenario 1.2, when using logistic regression comparing frequentist and Bayesian approaches for sample sizes ranging between $N = 500–5{,}000$.

#### Scenario 1.3: small treatment difference

When all the treatments had a small difference in mortality probability (scenario 1.3, Fig. 4a-c), reduction in mortality ratio, best treatment probability, and improvement over random choice were smaller than in scenario 1.2 where there was one best treatment (for a sample size of 500, the Bayes unrepresentative prior 1 method produced 0.77, 0.83, and 0.86, respectively). The Bayesian approach using the representative prior only reached a probability of interval separation (to correctly identify the true worst treatment) of 0.40 even with a sample size of roughly 5,000 patients (Fig. 4d). Whereas it reached a probability of 0.80 for a sample size of 4,250 patients, for identifying one of the true bottom two treatments as worst (Fig. 4e, Table SM 4 in Additional file 2).

Reduction in mortality ratio (a), best treatment probability (b), improvement over random choice (c), probability of interval separation based on worst treatment (i) (d), probability of interval separation based on bottom two treatments (ii) (e) in scenario 1.3, when using logistic regression comparing frequentist and

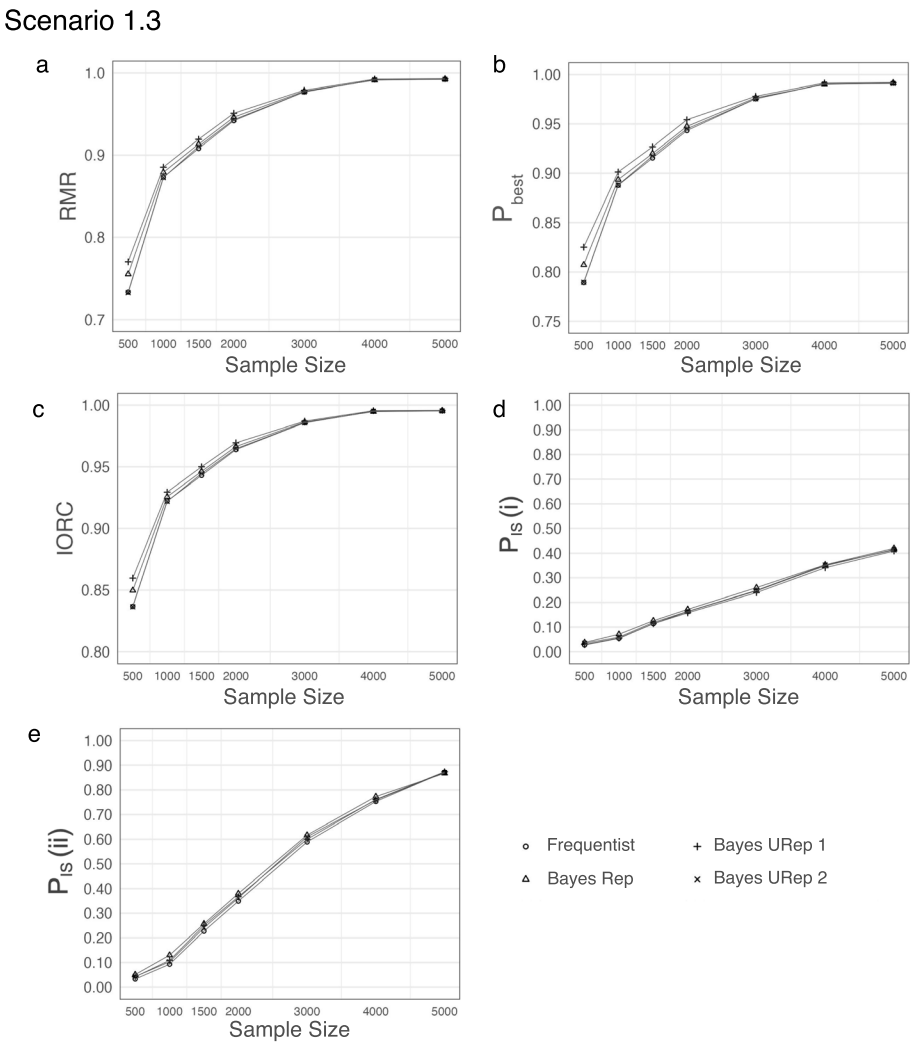


**Fig. 4** Results for scenario 1.3

Bayesian approaches for sample sizes ranging between $N=500$–5,000.

#### Scenario 1.4: one worst treatment

When there was one true worst treatment and three similarly good treatments (scenario 1.4, Fig. 5a-c), reduction in mortality ratio, best treatment probability, and improvement over random choice were the largest of all scenarios investigated (roughly 0.98, 0.99 and 0.99, respectively, for a sample size of 500, for the Bayes unrepresentative prior 1 method). To secure a probability of interval separation of 80% for the Bayes unrepresentative prior 1 method, roughly 3,250 patients were needed, see Fig. 5d. This scenario also produced the largest probability of interval separation (Table SM 4 in Additional file 2).

Reduction in mortality ratio (a), best treatment probability (b), improvement over random choice (c), probability of interval separation (d) in scenario 1.4, when using logistic regression comparing frequentist and Bayesian approaches for sample sizes ranging between $N=500$–5,000.

#### Scenario 2.2: small treatment difference and subgroup effect

When the mortality probability of the four treatments differed across subgroups, where the difference was a constant log odds ratio across subgroups, the reduction in mortality ratio, best treatment probability and improvement over random choice were 0.78, 0.83 and 0.86, respectively for a sample size of 500 for the Bayes representative method (Fig. 6a-c). The Bayes method utilising the unrepresentative prior 1 produced a probability of interval separation, reaching 45% for a sample size of $N=5{,}000$ when only predicting the true worst treatment and 80% for $N=4{,}000$ when predicting at least one of

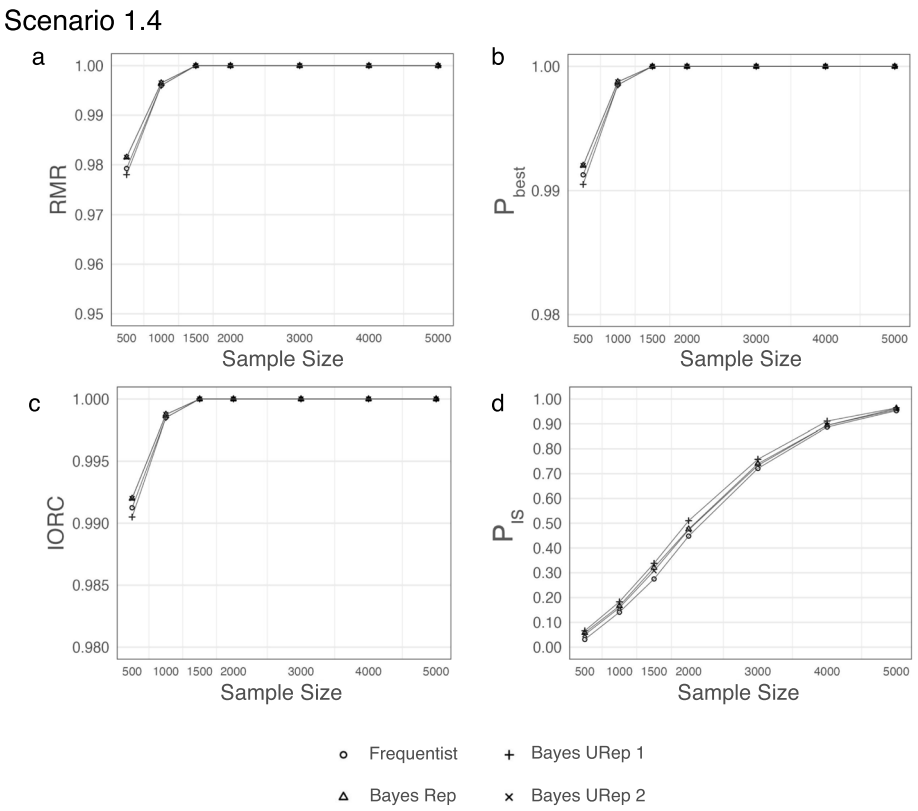


**Fig. 5** Results for scenario 1.4

the two true bottom treatments (Fig. 6d-e, Table SM 4 in Additional file 2).

Reduction in mortality ratio (a), best treatment probability (b), improvement over random choice (c), probability of interval separation based on worst treatment (i) (d), probability of interval separation based on bottom two treatments (ii) (e) in scenario 2.2, when using logistic regression comparing frequentist and Bayesian approaches for sample sizes ranging between $N = 500$–5,000.

#### Scenario 2.3: Small treatment difference and site effect

None of the methods performed well when the mortality probability of the four treatments differed across sites. The reduction in mortality ratio, best treatment probability and improvement over random choice were all substantially lower than in previous scenarios, only reaching 0.66, 0.74 and 0.80, respectively, when the sample size was 5,000 using the Bayes unrepresentative prior 1 method (Fig. 7a-c). Similarly, the probability of interval separation for the Bayes unrepresentative prior 1 method, only increased to 0.40 when predicting the single true worst treatment and 0.69 when predicting at least one of the two true bottom treatments, both for a sample size of 5,000 (Fig. 7d-e, Table SM 4 in Additional file 2).

Reduction in mortality ratio (a), best treatment probability (b), improvement over random choice (c), probability of interval separation based on worst treatment (i) (d), probability of interval separation based on bottom two treatments (ii) (e) in scenario 2.3, when using logistic regression comparing frequentist and Bayesian approaches for sample sizes ranging between $N = 500$–5,000.

Additional results were obtained comparing the fixed effects model with the mixed effects model utilising additional scenarios and performance measures (Figures SM 1–10 and Tables SM 3–4 in Additional files 1 and 2). Across the majority of scenarios, fixed effect Bayesian methods produced the largest reduction in mortality ratio (Figure SM 3), best treatment probability (Figure SM 4) and improvement over random choice (Figure SM 5). Although, the difference between methods was small. However, Figure SM 7 shows that the mixed effects frequentist method often outperformed the fixed effects frequentist and Bayesian methods in terms of probability of interval separation, when the treatment mortality probability did not vary by patient subgroup (scenarios 1.1–1.4). However, when the treatment mortality probability did vary by patient subgroup (scenario 2.2) the mixed effects frequentist and Bayesian methods consistently

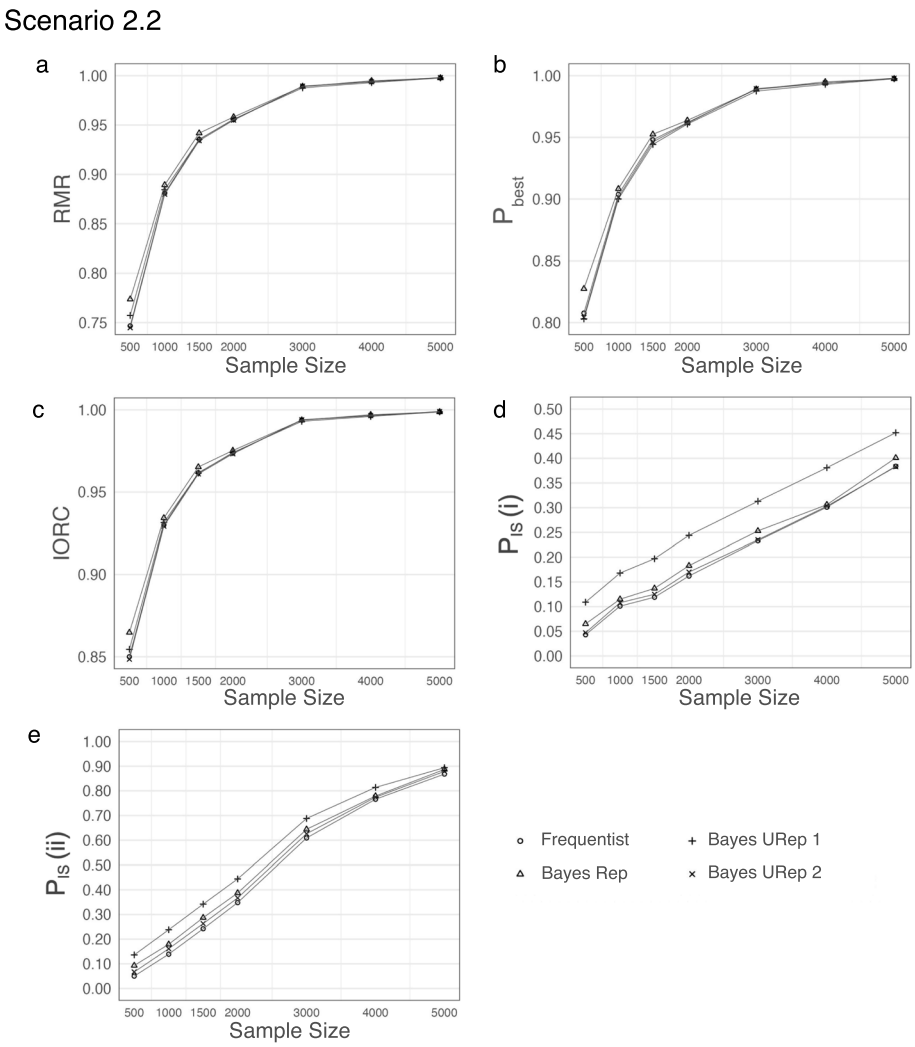


**Fig. 6** Results for scenario 2.2

produced zero probability of interval separation. This is likely due to the unstable estimates (large CI/CrIs) produced by the mixed effects model, probably caused by the small number of subgroups (4) included as random effects in the model [21]. Hence, the fixed effect Bayesian methods performed best in this scenario, although, the difference between fixed effect methods was small.

## Discussion

In this simulation study, we compared frequentist and Bayesian methods in various real-world scenarios, and proposed using interval separation to evaluate the novel PRACTical design. We showed that the Bayesian approach with a strongly informative prior was only better than the frequentist approach at a small sample size where the difference in performance was small. This was the case for most performance measures, in most scenarios. In terms of performance measures indicative of power, the previously proposed measures (reduction in mortality ratio, best treatment probability and improvement over random choice) were much larger than our novel measure, probability of interval separation (sample size of 500 to 1000 versus above 3,500 needed to produce power above 80%, respectively), for all scenarios investigated.

Surprisingly, in some scenarios the Bayesian approach utilising a strongly informative prior calculated using unrepresentative historical data, performed better than utilising a prior calculated using representative historical data. This was highly dependent on *how* the historical data was unrepresentative. However, across all performance measures and scenarios, the difference in methods was very small. Furthermore, the exploration of non-informative and weakly informative priors (utilising a standard deviation of '5' for the priors), also

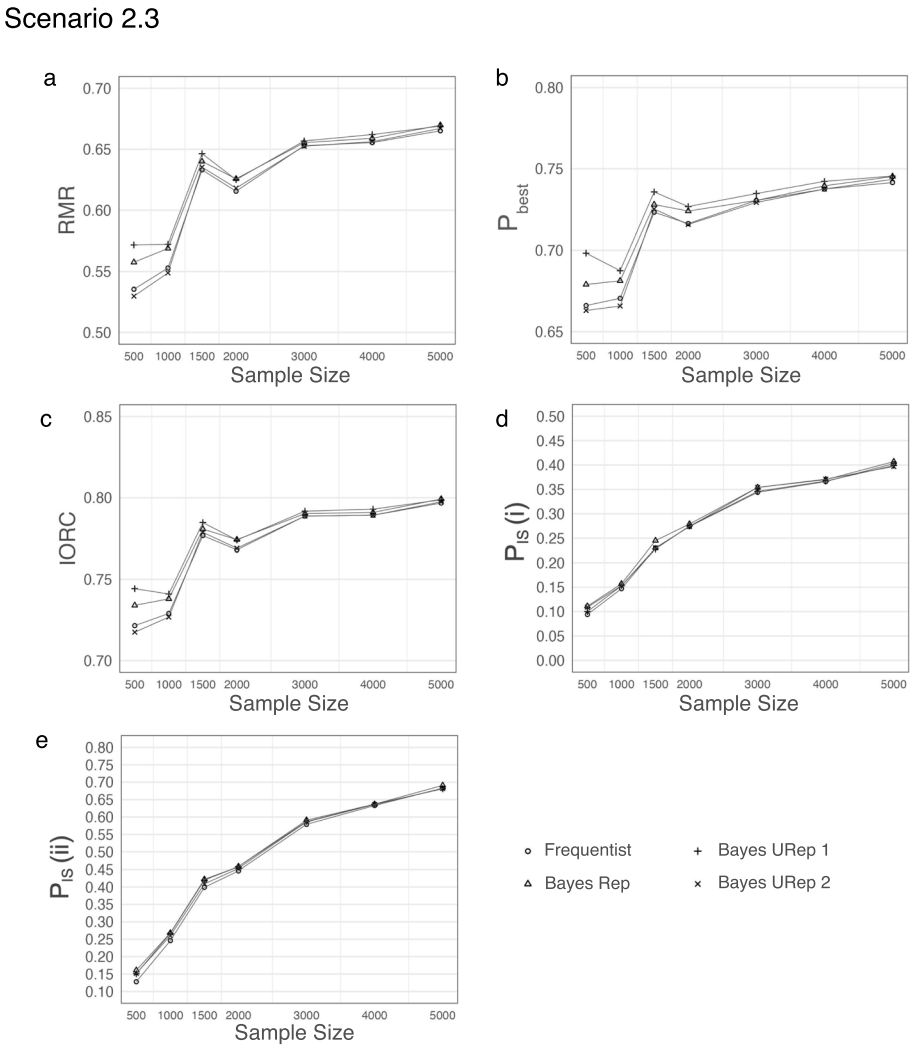


**Fig. 7** Results for scenario 2.3

produced very similar results to the frequentist method (data not shown).

The probability of interval separation was less prone to incorrect conclusions about treatment efficacy when the treatments were actually equal, compared to previously proposed measures. These earlier measures, which relied on point estimates derived from logistic regression, tended to bias away from the null [22], especially in sparse datasets [23]. In such cases, less common treatments were more likely to be inaccurately predicted as the best option. In contrast, the CIs/CrIs for less common treatments were wider, compared to those for more common treatments, leading to a more frequent conclusion that all treatments were equally effective. The size of the CI/CrIs could be adapted in practice depending on clinician judgement and the particular setting of the trial, which could increase or decrease the conservativeness of the interval separation measure.

Choosing the best performance measure depends on the goal for the planned clinical trial. The probability of interval separation is straightforward to interpret, and it facilitates the conclusions of treatment equivalence *or* superiority. Furthermore, the probability of interval separation could allow the dropping of treatments at interim analyses in platform trials. Firstly, if the interval separation measure indicates a single best treatment, then the platform trial could be stopped early for efficacy. Secondly, if it indicates a single (or multiple) worst treatment(s), then specific treatment arm(s) could be dropped for futility, while allowing the trial to continue. Finally, as trials are usually set-up to influence clinical and regulatory decisions, it is important to control the probability of incorrect interval separation

(our proxy variable for type I error). The probability of incorrect interval separation was below 0.05 across all null scenarios, for all sample sizes and all methods investigated. Therefore, utilising the interval separation measure allowed a conclusion of equivalence between treatments, which was not possible using the previously proposed measures.

A caveat in the PRACTical design, is that none of the analytical methods performed well when treatment mortality probability varied by site (scenario 2.3). The impact of this heterogeneity in the PRACTical design is especially important because there is an additional layer, in that randomisation lists may also vary by site, not just patient subgroups. Our simulation provides a useful tool for investigators to explore the relationship between site heterogeneity and sample size at the trial design stage.

Our study includes several strengths. We explored a large number of scenarios, which one could expect to encounter in real life multicentre trials, and over a large variety of sample sizes. We compared multiple approaches (frequentist and Bayesian), and several analysis models (incorporating the patient subgroups as a fixed and random effect) within the PRACTical framework. In addition, we considered the use of historical data which was both representative and unrepresentative of the current trial for the Bayesian approach.

A limitation of our work is that we have not performed a theoretical analysis. Although our simulations explored a wide-ranging parameter space which was realistic in clinical settings, these simulations are still limited to a finite set of points, which means we cannot extrapolate results beyond the scenarios tested. Theoretical analysis of the methods explored in this study are beyond the scope of this manuscript and are left as an avenue of further work. An additional limitation of our work is the arbitrary selection of four patient subgroups and treatments. We did not investigate the impact of the number of treatments included in each pattern as this was already explored by Lee et al. [13] and Turner et al. [14]. They found that including fewer treatments per pattern, hence, including less overlap between patterns and within-pattern information gained, led to lower power and increased bias. Lastly, further simulations could be explored such as varying the standard deviation used in the prior distributions in the Bayesian approaches, the optimal width of the CIs/CrIs to maximise the probability of interval separation, while maintaining a probability of incorrect interval separation below 0.05 adapted to the specific motivating example and incorporating clinical judgement, utilising real world data, and including a site effect in the models.

## Conclusions

Bayesian and frequentist analysis methods could adequately predict the true best treatment, even when we introduced different treatment mortality probabilities per patient subgroup. Where the Bayesian methods performed best at smaller sample sizes, regardless of representativeness of historical data. Our novel proposed performance measures probability of interval separation and probability of incorrect interval separation, which quantified the precision in treatment estimates, allowed a conclusion of several '*best*' treatment(s) or of equivalence between all included treatments, but would require large sample sizes. Both the frequentist and the Bayesian analysis approach, utilising uncertainty intervals, are robust to allow the PRACTical design to be used in a pragmatic adaptive platform trial.

### Abbreviations

| | |
|---|---|
| IORC | Improvement over random choice |
| $P_{best}$ | Best treatment probability |
| $P_{IS}$ | Probability of interval separation |
| $P_{IIS}$ | Probability of incorrect interval separation |
| PRACTical | Personalised randomised controlled trial |
| RMR | Reduction in mortality ratio |
| RCT | Randomised controlled trial |
| SoC | Standard of care |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-025-02537-x.

Additional file 1: A comparison of frequentist versus Bayesian approaches to the Personalised Randomised Controlled Trial (PRACTical) - Design and Analysis Considerations. Additional analysis models and results, for additional simulation scenarios and performance measures.

Additional file 2: A comparison of frequentist versus Bayesian approaches to the Personalised Randomised Controlled Trial (PRACTical) - Design and Analysis Considerations. Additional results, for additional simulation scenarios and performance measures.

### Acknowledgements

Not applicable.

### Authors' contributions

Y.M. conceptualised the study. H.J., Y.S., N.A.B.M.A, J.W.C., R.P., X.W. and Y.M. were involved in coding, analysing, and interpreting the results of the study. R.P. and X.W. prepared all the figures. H.J., Y.S., M.E.A.dK. and Y.M. wrote the main manuscript text. All authors reviewed the final article and approved the submitted version.

### Funding

HJ and MdK are part of the ECRAID-Base consortium; the ECRAID-Base project has received funding from the European Union's Horizon 2020 Research and Innovation programme, under Grant Agreement number 965313. YS receives funding from the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2023). YM received funding from National Medical Research Council, Singapore (TA-MOH-001177) and the Wellcome Trust via the Good Clinical Trial Collaborative.

### Data availability

All datasets generated and analysed during the current study are available, in the repository, https://github.com/moyinNUHS/practical_BayesSI.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

**Author details**
[1]Infection Control Program, Geneva University Hospitals and Faculty of Medicine, World Health Organization Collaborating Center, Geneva, Switzerland. [2]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore. [3]Lloyd's Register Foundation Institute for the Public Understanding of Risk, National University of Singapore, Singapore, Singapore. [4]Australian National University, Canberra, Australia. [5]National University Hospital, National University Health System, Singapore, Singapore. [6]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore. [7]Mahidol-Oxford Research Unit, Mahidol University, Tungphyathai, Thailand. [8]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [9]Infectious Diseases Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [10]ADVANCE-ID Network, Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore.

Received: 30 August 2024 Accepted: 24 March 2025
Published online: 29 May 2025

**References**

1. Goetz LH, Schork NJ. Personalized medicine: motivation, challenges, and progress. Fertil Steril. 2018;109(6):952–63. https://doi.org/10.1016/j.fertnstert.2018.05.006. PMID: 29935653; PMCID: PMC6366451.
2. Cinti C, Trivella MG, Joulie M, Ayoub H, Frenzel M. The roadmap toward personalized medicine: challenges and opportunities. J Pers Medi. 2024;14(6):546.
3. Janiaud P, Serghiou S, Ioannidis JP. New clinical trial designs in the era of precision medicine: an overview of definitions, strengths, weaknesses, and current use in oncology. Cancer Treat Rev. 2019;73:20–30.
4. Sheu CC, Chang YT, Lin SY, Chen YH, Hsueh PR. Infections caused by carbapenem-resistant Enterobacteriaceae: an update on therapeutic options. Front Microbiol. 2019;30(10):80. https://doi.org/10.3389/fmicb.2019.00080. PMID: 30761114; PMCID: PMC6363665.
5. Torres A, Niederman MS, Chastre J, Ewig S, Fernandez-Vandellos P, Hanberger H, Kollef M, Li Bassi G, Luna CM, Martin-Loeches I, Paiva JA, Read RC, Rigau D, François Timsit J, Welte T, Wunderink R. Summary of the international clinical guidelines for the management of hospital-acquired and ventilator-acquired pneumonia. ERJ Open Res. 2018;4(2):00028–2018. https://doi.org/10.1183/23120541.00028-2018. PMID: 29977898; PMCID: PMC6018155.
6. Carrara E, Savoldi A, Piddock LJV, et al. Clinical management of severe infections caused by carbapenem-resistant gram-negative bacteria: a worldwide cross-sectional survey addressing the use of antibiotic combinations. Clin Microbiol Infect. 2022;28:66–72. https://doi.org/10.1016/j.cmi.2021.05.002.
7. Juszczak E, Altman DG, Hopewell S, et al. Reporting of multi-arm parallel-group randomized trials: extension of the CONSORT 2010 statement. JAMA. 2019;321:1610–20. https://doi.org/10.1001/jama.2019.3087.
8. Ghosh P, Liu L, Senchaudhuri P, et al. Design and monitoring of multi-arm multi-stage clinical trials. Biometrics. 2017;73:1289–99. https://doi.org/10.1111/biom.12687.
9. Noor NM, Love SB, Isaacs T, et al. Uptake of the multi-arm multi-stage (MAMS) adaptive platform approach: a trial-registry review of late-phase randomised clinical trials. BMJ Open. 2022;12: e055615. https://doi.org/10.1136/bmjopen-2021-055615.
10. Foltynie T, Gandhi S, Gonzalez-Robles C, et al. Towards a multi-arm multi-stage platform trial of disease modifying approaches in Parkinson's disease. Brain. 2023;146:2717–22. https://doi.org/10.1093/brain/awad063.
11. Magaret A, Angus DC, Adhikari NKJ, et al. Design of a multi-arm randomized clinical trial with no control arm. Contemp Clin Trials. 2016;46:12–7. https://doi.org/10.1016/j.cct.2015.11.003.
12. Walker AS, White IR, Turner RM, et al. Personalised randomised controlled trial designs-a new paradigm to define optimal treatments for carbapenem-resistant infections. Lancet Infect Dis. 2021;21:e175–81. https://doi.org/10.1016/S1473-3099(20)30791-X.
13. Lee KM, Turner RM, Thwaites GE, et al. The personalised randomized controlled trial: evaluation of a new trial design. Stat Med. 2023;42:1156–70. https://doi.org/10.1002/sim.9663.
14. Turner RM, Lee KM, Walker AS, Ellis S, Sharland M, Bielicki JA, Stöhr W, White IR. Determining sample size in a personalized randomized controlled (PRACTical) trial. Stat Med. 2024;43(21):4098–112. https://doi.org/10.1002/sim.10168.
15. Li X, Ye H. Clinical and mortality risk factors in bloodstream infections with carbapenem-resistant Enterobacteriaceae. Can J Infect Dis Med Microbiol. 2017;2017:6212910. https://doi.org/10.1155/2017/6212910. Epub 2017 Dec 12. PMID: 29379527; PMCID: PMC5742906.
16. Falagas ME, Tansarli GS, Karageorgopoulos DE, Vardakas KZ. Deaths attributable to carbapenem-resistant Enterobacteriaceae infections. Emerg Infect Dis. 2014;20(7):1170–5. https://doi.org/10.3201/eid2007.121004. PMID:24959688;PMCID:PMC4073868.
17. White IR, Barrett JK, Jackson D, et al. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. Res Synth Methods. 2012;3:111–25. https://doi.org/10.1002/jrsm.1045.
18. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2021. https://www.R-project.org/.
19. Goodrich B, Gabry J, Ali I & Brilleman S. rstanarm: Bayesian applied regression modeling via Stan. R package version 2.32.1. 2024. https://mc-stan.org/rstanarm.
20. van de Schoot R, Kaplan D, Denissen J, Asendorpf JB, Neyer FJ, van Aken MAG. A gentle introduction to Bayesian analysis: applications to developmental research. Child Dev. 2014;85(3):842–60. https://doi.org/10.1111/cdev.12169. Epub 2013 Oct 9. PMID: 24116396; PMCID: PMC4158865.
21. Barker D, D'Este C, Campbell MJ, McElduff P. Minimum number of clusters and comparison of analysis methods for cross sectional stepped wedge cluster randomised trials with binary outcomes: a simulation study. Trials. 2017;18(1):119. https://doi.org/10.1186/s13063-017-1862-2. PMID:28279222;PMCID:PMC5345156.
22. Nemes S, Jonasson JM, Genell A, et al. Bias in odds ratios by logistic regression modelling and sample size. BMC Med Res Methodol. 2009;9:56. https://doi.org/10.1186/1471-2288-9-56.
23. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. BMJ. 2016;352:i1981. https://doi.org/10.1136/bmj.i1981.

## Publisher's Note