

SUPPLEMENTARY MATERIALS

Please find here additional analyses investigating the fixed effects (FE) logistic regression model from the main text and the mixed effects (ME) logistic regression model. We explore their performance using scenarios from the main text and additional scenarios listed below and comparing performance measures from the main text and additional performance measures listed below.

METHODS

Mixed effects model

In this model, treatments and patient subgroups were included as independent variables in a hierarchical structure, such that treatments were categorised as fixed effects and the patient subgroups were treated as random effects. The ME model included a coefficient (the log odds for the risk of death), ψ_j , for each treatment, j and an intercept (the log odds for the risk of death), $\ln(\alpha_k)$, for each patient subgroup, k ,

$$\text{logit}(P_{jk}) = \ln(\alpha_k) + \psi_j.$$

We implemented the ME model described above using a frequentist approach (using the R package ‘lme4’ [25]) and Bayesian approach (via the R package ‘rstanarm’ [19]). The ME Bayesian method utilised a prior for the four treatment coefficients.

When calculating the interval separation performance measure, the ME frequentist method utilised 80% Wald confidence intervals (CIs) and the ME Bayesian method utilised 80% central credible intervals (CrIs, sometimes referred to as uncertainty intervals) [21], for each treatment coefficient.

Historical Datasets

Table SM1: Simulation Scenarios, including the true mortality probability of treatments in the historical trial.

Scenario, S	True mortality probability of treatments, P_{jk} in current trial	True mortality probability of treatments, P_{jk} in historical trial
S1 series: Various effects across treatments		
1.1: Null	All treatments had equal mortality probability for all patient subgroups: $P_{jk} = 37.5\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$	<u>representative</u> ; All treatments had equal mortality probability for all patient subgroups: $P_{jk} = 37.5\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$
		<u>unrepresentative 1</u> ; All treatments had equal mortality probability for all patient subgroups: $P_{jk} = 27.5\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$
		<u>unrepresentative 2</u> ; All treatments had different mortality probabilities, which was equal across patient subgroups: $P_{Ak} = 30\%$; $P_{Bk} = 35\%$; $P_{Ck} = 40\%$; $P_{Dk} = 45\%$, $k = 1, 2, 3, 4$
1.2: One best treatment	Treatment B had 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 $k = 1, 2, 3, 4$	<u>representative</u> ; Treatment B had 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 $k = 1, 2, 3, 4$
		<u>unrepresentative 1</u> ; Treatment B had smaller mortality probability than treatments A, C, D for all patient subgroups: $P_{Bk} = 25\%$ $k = 1, 2, 3, 4$; and $P_{jk} = 35\%$, $j = A, C, D$ and $k = 1, 2, 3, 4$
		<u>unrepresentative 2</u> ; All treatments had different mortality probabilities, which was equal across patient subgroups: $P_{Ak} = 45\%$; $P_{Bk} = 40\%$; $P_{Ck} = 35\%$; $P_{Dk} = 30\%$, $k = 1, 2, 3, 4$
1.3 Small difference	All treatments had different mortality probabilities, which was equal across patient subgroups:	<u>representative</u> ; All treatments had different mortality probabilities, which was equal across patient subgroups: $P_{Ak} = 30\%$; $P_{Bk} = 35\%$; $P_{Ck} = 40\%$; $P_{Dk} = 45\%$, $k = 1, 2, 3, 4$

	$P_{Ak} = 30\%$; $P_{Bk} = 35\%$; $P_{Ck} = 40\%$; $P_{Dk} = 45\%$, $k = 1, 2, 3, 4$	<u>unrepresentative 1</u> ; All treatments had different mortality probabilities, which was equal across patient subgroups: $P_{Ak} = 20\%$; $P_{Bk} = 25\%$; $P_{Ck} = 30\%$; $P_{Dk} = 35\%$, $k = 1, 2, 3, 4$ <u>unrepresentative 2</u> ; All treatments had equal mortality probability for all patient subgroups: $P_{jk} = 37.5\%$, $j = A, B, C, D$, $k = 1, 2, 3, 4$
1.4 One worst treatment	Treatment B had 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 $k = 1, 2, 3, 4$	<u>representative</u> ; Treatment B had 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 $k = 1, 2, 3, 4$ <u>unrepresentative 1</u> ; All treatments had equal mortality probability for all patient subgroups: $P_{jk} = 25\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$ <u>unrepresentative 2</u> ; All treatments had different mortality probabilities, which was equal across patient subgroups: $P_{Ak} = 45\%$; $P_{Bk} = 40\%$; $P_{Ck} = 35\%$; $P_{Dk} = 30\%$, $k = 1, 2, 3, 4$
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: $P_{j4} = 45\%$, $j = A, B, C, D$	<u>representative</u> ; 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: $P_{j4} = 45\%$, $j = A, B, C, D$ <u>unrepresentative 1</u> ; All treatments had different mortality probabilities, which was equal across patient subgroups $P_{Ak} = 30\%$; $P_{Bk} = 35\%$; $P_{Ck} = 40\%$; $P_{Dk} = 45\%$, $k = 1, 2, 3, 4$ <u>unrepresentative 2</u> ; All treatments had equal mortality probabilities, but it was different across the patient subgroups: Subgroup 1: $P_{j1} = 20\%$, $j = B, C$

		Subgroup 2: $P_{j2} = 25\%$, $j = A, B, C$ Subgroup 3: $P_{j3} = 30\%$, $j = B, C, D$ Subgroup 4: $P_{j4} = 35\%$, $j = A, B, C, D$
2.2: 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: $P_{A4} = 78.2\%$, $P_{B4} = 82.8\%$, $P_{C4} = 86.1\%$, $P_{D4} = 88.6\%$	<u>representative</u>; 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: $P_{A4} = 78.2\%$, $P_{B4} = 82.8\%$, $P_{C4} = 86.1\%$, $P_{D4} = 88.6\%$ <u>unrepresentative 1</u>; All treatments had different mortality probabilities, which was equal across patient subgroups $P_{Ak} = 20\%$; $P_{Bk} = 25\%$; $P_{Ck} = 30\%$; $P_{Dk} = 35\%$, $k = 1, 2, 3, 4$ <u>unrepresentative 2</u>; All treatments had equal mortality probabilities, but it was different across the patient subgroups: Subgroup 1: $P_{j1} = 27.5\%$, $j = B, C$ Subgroup 2: $P_{j2} = 38.3\%$, $j = A, B, C$ Subgroup 3: $P_{j3} = 73\%$, $j = B, C, D$ Subgroup 4: $P_{j4} = 83.9\%$, $j = A, B, C, D$
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: $P_{Akq} = 30\%$, $P_{Bkq} = 35\%$, $P_{Ckq} = 40\%$, $P_{Dkq} = 45\%$, $k = 1, 2, 3, 4$ and $q = 3, \dots, 10$; and $P_{Akq} = 20\%$, $P_{Bkq} = 25\%$, $P_{Ckq} = 30\%$, $P_{Dkq} = 35\%$, $k = 1, 2, 3, 4$ and $q = 1, 2$	<u>representative</u>; 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: $P_{Akq} = 30\%$, $P_{Bkq} = 35\%$, $P_{Ckq} = 40\%$, $P_{Dkq} = 45\%$, $k = 1, 2, 3, 4$ and $q = 3, \dots, 10$; and $P_{Akq} = 20\%$, $P_{Bkq} = 25\%$, $P_{Ckq} = 30\%$, $P_{Dkq} = 35\%$, $k = 1, 2, 3, 4$ and $q = 1, 2$

		<u>unrepresentative 1</u>; 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: $P_{Akq} = 20\%$, $P_{Bkq} = 25\%$, $P_{Ckq} = 30\%$, $P_{Dkq} = 35\%$, $k = 1, 2, 3, 4$ and $q = 3, \dots, 10$; and $P_{Akq} = 10\%$, $P_{Bkq} = 15\%$, $P_{Ckq} = 20\%$, $P_{Dkq} = 25\%$, $k = 1, 2, 3, 4$ and $q = 1, 2$
		<u>unrepresentative 2</u>; All treatments had equal mortality probability for all patient subgroups, which were the same in 8 study sites but different to the other 2 sites: $P_{jk} = 37.5\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$ and $q = 3, \dots, 10$ $P_{jk} = 27.5\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$ and $q = 1, 2$

Treatments were represented as j , where $j = A, B, C, D$. Each patient was randomised to a treatment within their pattern, S_{kq} . The group of patients who shared the same pattern, S_{kq} , 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. The three Bayesian methods used historical data, of sample size 500, which was representative and two different ways of being unrepresentative, respectively.

Additional Scenarios

Table SM2. Additional Simulation Scenarios, including the true mortality probability of treatments in the historical trial.

Scenario, S	True mortality probability of treatments, P_{jk} , in current trial	True mortality probability of treatments, P_{jk} in historical trial	Frequency of subgroups, λ_k
S1 series: Various effects across treatments			
1.5: Large difference	All treatments had different mortality probabilities, which was equal across patient subgroups: $P_{Ak} = 30\%$; $P_{Bk} = 40\%$; $P_{Ck} = 50\%$; $P_{Dk} = 60\%$, $k = 1, 2, 3, 4$	<u>representative</u> ; All treatments had different mortality probabilities, which was equal across patient subgroups: $P_{Ak} = 30\%$; $P_{Bk} = 40\%$; $P_{Ck} = 50\%$; $P_{Dk} = 60\%$, $k = 1, 2, 3, 4$	$\lambda_k = 25\%$, $k = 1, 2, 3, 4$
		<u>unrepresentative</u> ; All treatments had different mortality probabilities, which was equal across patient subgroups: $P_{Ak} = 20\%$; $P_{Bk} = 30\%$; $P_{Ck} = 40\%$; $P_{Dk} = 50\%$, $k = 1, 2, 3, 4$	
		<u>unrepresentative</u> ; All treatments had equal mortality probability for all patient subgroups: $P_{jk} = 45\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$	
1.6: Small difference, common best treatment	All treatments had different mortality probabilities, which was equal across patient subgroups: $P_{Ak} = 35\%$; $P_{Bk} = 30\%$; $P_{Ck} = 45\%$; $P_{Dk} = 40\%$, $k = 1, 2, 3, 4$	<u>representative</u> ; All treatments had different mortality probabilities, which was equal across patient subgroups: $P_{Ak} = 35\%$; $P_{Bk} = 30\%$; $P_{Ck} = 45\%$; $P_{Dk} = 40\%$, $k = 1, 2, 3, 4$	$\lambda_k = 25\%$, $k = 1, 2, 3, 4$
		<u>unrepresentative</u> ; All treatments had different mortality probabilities, which was equal across patient subgroups: $P_{Ak} = 25\%$; $P_{Bk} = 20\%$; $P_{Ck} = 35\%$; $P_{Dk} = 30\%$, $k = 1, 2, 3, 4$	
		<u>unrepresentative</u> ; All treatments had equal mortality probability for all patient subgroups: $P_{jk} = 37.5\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$	

S3 series: Low or high average treatment mortality probabilities			
3.1: Null, low average mortality	$P_{jk} = 17.5\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$	<u>representative</u> ; $P_{jk} = 17.5\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$	$\lambda_k = 25\%$, $k = 1, 2, 3, 4$
		<u>unrepresentative</u> ; $P_{jk} = 7.5\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$	
		<u>unrepresentative</u> ; $P_{Ak} = 10\%$, $P_{Bk} = 15\%$, $P_{Ck} = 20\%$, $P_{Dk} = 25\%$, $k = 1, 2, 3, 4$	
3.2: Small difference, low average mortality	$P_{Ak} = 10\%$, $P_{Bk} = 15\%$, $P_{Ck} = 20\%$, $P_{Dk} = 25\%$, $k = 1, 2, 3, 4$	<u>representative</u> ; $P_{Ak} = 10\%$, $P_{Bk} = 15\%$, $P_{Ck} = 20\%$, $P_{Dk} = 25\%$, $k = 1, 2, 3, 4$	$\lambda_k = 25\%$, $k = 1, 2, 3, 4$
		<u>unrepresentative</u> ; $P_{Ak} = 5\%$, $P_{Bk} = 10\%$, $P_{Ck} = 15\%$, $P_{Dk} = 20\%$, $k = 1, 2, 3, 4$	
		<u>unrepresentative</u> ; $P_{jk} = 17.5\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$	
3.3: Null, high average mortality	$P_{jk} = 82.5\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$	<u>representative</u> ; $P_{jk} = 82.5\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$	$\lambda_k = 25\%$, $k = 1, 2, 3, 4$
		<u>unrepresentative</u> ; $P_{jk} = 72.5\%$, $j = A, B, C, D$ and $k = 1, 2, 3, 4$	
		<u>unrepresentative</u> ; $P_{Ak} = 75\%$, $P_{Bk} = 80\%$, $P_{Ck} = 85\%$, $P_{Dk} = 90\%$, $k = 1, 2, 3, 4$	
3.4: Small difference,	$P_{Ak} = 75\%$, $P_{Bk} = 80\%$, $P_{Ck} = 85\%$, $P_{Dk} = 90\%$, $k = 1, 2, 3, 4$	<u>representative</u> ; $P_{Ak} = 75\%$, $P_{Bk} = 80\%$, $P_{Ck} = 85\%$, $P_{Dk} = 90\%$, $k = 1, 2, 3, 4$	$\lambda_k = 25\%$, $k = 1, 2, 3, 4$
		<u>unrepresentative</u> ;	

high average mortality		$P_{Ak} = 65\%, P_{Bk} = 70\%, P_{Ck} = 75\%, P_{Dk} = 80\%, k = 1, 2, 3, 4$	
		<u>unrepresentative</u> ; $P_{jk} = 82.5\%, j = A, B, C, D$ and $k = 1, 2, 3, 4$	
S4 series: Different subgroup frequency			
4.1: Null, one least common subgroup	$P_{jk} = 37.5\%, j = A, B, C, D$ and $k = 1, 2, 3, 4$	<u>representative</u> ; $P_{jk} = 37.5\%, j = A, B, C, D$ and $k = 1, 2, 3, 4$	$\lambda_1 = 10\%, \lambda_k = 30\%, k = 2, 3, 4$
		<u>unrepresentative</u> ; $P_{jk} = 27.5\%, j = A, B, C, D$ and $k = 1, 2, 3, 4$	
		<u>unrepresentative</u> ; $P_{Ak} = 30\%, P_{Bk} = 35\%, P_{Ck} = 40\%, P_{Dk} = 45\%, k = 1, 2, 3, 4$	
4.2: Small difference, one least common subgroup	$P_{Ak} = 30\%, P_{Bk} = 35\%, P_{Ck} = 40\%, P_{Dk} = 45\%, k = 1, 2, 3, 4$	<u>representative</u> ; $P_{Ak} = 30\%, P_{Bk} = 35\%, P_{Ck} = 40\%, P_{Dk} = 45\%, k = 1, 2, 3, 4$	$\lambda_1 = 10\%, \lambda_k = 30\%, k = 2, 3, 4$
		<u>unrepresentative</u> ; $P_{Ak} = 20\%, P_{Bk} = 25\%, P_{Ck} = 30\%, P_{Dk} = 35\%, k = 1, 2, 3, 4$	
		<u>unrepresentative</u> ; $P_{jk} = 37.5\%, j = A, B, C, D$ and $k = 1, 2, 3, 4$	

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 a subgroup, k . Unless otherwise stated, treatment mortality probabilities were homogenous across all $Q = 10$ sites and all $K = 4$ subgroups. The three Bayesian methods used historical data, of sample size 500, which was representative and two different ways of being unrepresentative, respectively.

Additional Performance Measures

The following performance measures from Lee *et al.* [13] were used to further compare the different models and scenarios.

- **Bias:** referred to as the mean difference between the input or “true” treatment mortality probability values versus the predicted treatment mortality probability values calculated from the model, range -Inf to Inf, where 0 is desired.

$$\text{Bias} = E[\Delta_j] = E[\hat{P}_j - P_j].$$

- **Mean squared error (MSE):** the average squared difference between the input and predicted treatment mortality probability values, which captures both the bias and the variability amongst the iteration runs, range 0 (desired) to Inf.

$$\text{MSE} = E[\Delta_j^2].$$

- **Coverage probability:** the proportion of iterations for which the calculated 80% confidence (for frequentist methods) or credible intervals (for the Bayesian methods), CIs/CrIs, contained the true treatment coefficient value.

As this measure is a proportion it is bound between 0 and 1 (desired).

- **Near-best treatment probability ($P_{\text{near-best}}$):** proportion of simulations where the top-ranked treatment was within $\eta = 5\%$ with respect to mortality probability, of the best treatment, i.e. the chance of the difference between mortality under the top-ranked and the true best treatment being $\eta = 5\%$ or less. The condition is met when the estimated effect of the top-ranked treatment in the pattern is no more than 0.05 greater than the true minimum treatment mortality probability in the pattern (true best), range 0 to 1 (desired).

$$P_{\text{near-best}} = E\left[\sum_k \hat{\lambda}_k I(P_{j_k k} \leq \min_{j \in S_k} P_{j k} + \eta)\right]$$

- **Probability of interval separation:** proportion of iterations which correctly identified the predefined best and worst treatment(s) utilising the separation in

CIs and CrIs for the frequentist and Bayesian methods, respectively. We used four definitions of probability of interval separation due to us including 4 treatments in our simulations. The first definition, P_{IS} (i), was the proportion of iterations which concluded the predefined worst treatment as worst, we then added more flexibility to this definition by calculating the proportion of iterations which concluded either of the bottom 2 predefined treatments as worst, P_{IS} (ii), either singularly or in combination. The third definition, P_{IS} (iii), was the proportion of iterations which concluded the predefined best treatment as best. We again added more flexibility to this definition by calculating the proportion of iterations which concluded either of the top 2 predefined treatments as best, P_{IS} (iv), either alone or together. As this measure is a proportion it is bound between 0 and 1 (desired).

RESULTS

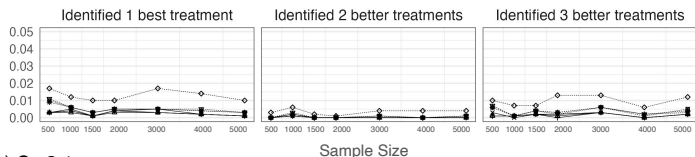
Here we compare the ME frequentist and Bayesian methods with the FE frequentist and Bayesian methods described in the main text. We explore the simulation scenarios listed in the main text and the additional scenarios listed above.

The scenarios in the main text focus on the situation where the total patient population incorporates each patient subgroup equally. Additional scenarios 4.1 and 4.2 explore the situation where the first patient subgroup (whose pattern only includes two treatments B and C) was less likely to occur. This change in patient subgroup frequency (scenario 1.1 vs 4.1 and scenario 1.3 vs 4.2) had very little impact on all performance measures.

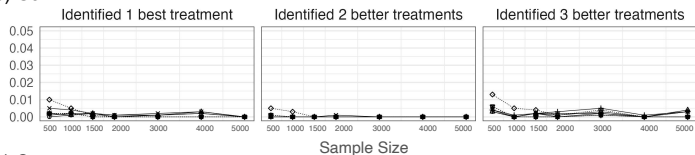
Probability of Incorrect Interval Separation

Across the sample sizes investigated, the ME frequentist approach often gave the largest probability of incorrect interval separation, the three ME Bayesian methods all produced very similar moderate probabilities of interval separation and the four FE methods consistently produced the smallest probabilities of interval separation, sometimes reaching 0. Even though the ME frequentist approach often produced the largest probability of incorrect interval separation, it was still always below 0.05 (Figure SM 1, Table SM 3).

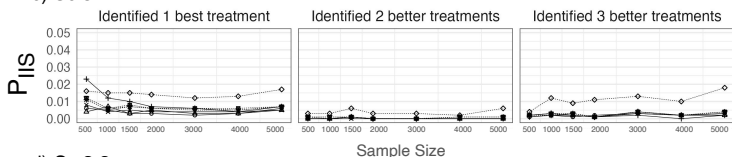
a) Sc 1.1



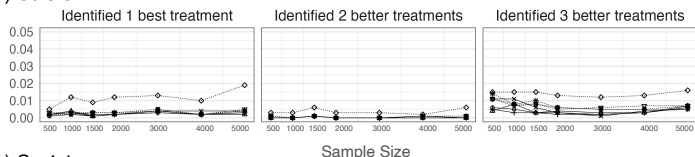
b) Sc 2.1



c) Sc 3.1



d) Sc 3.3



e) Sc 4.1

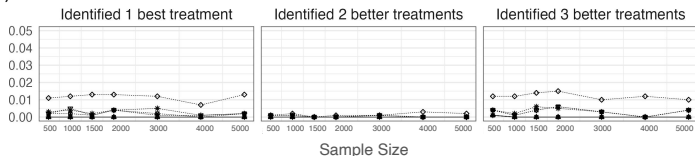
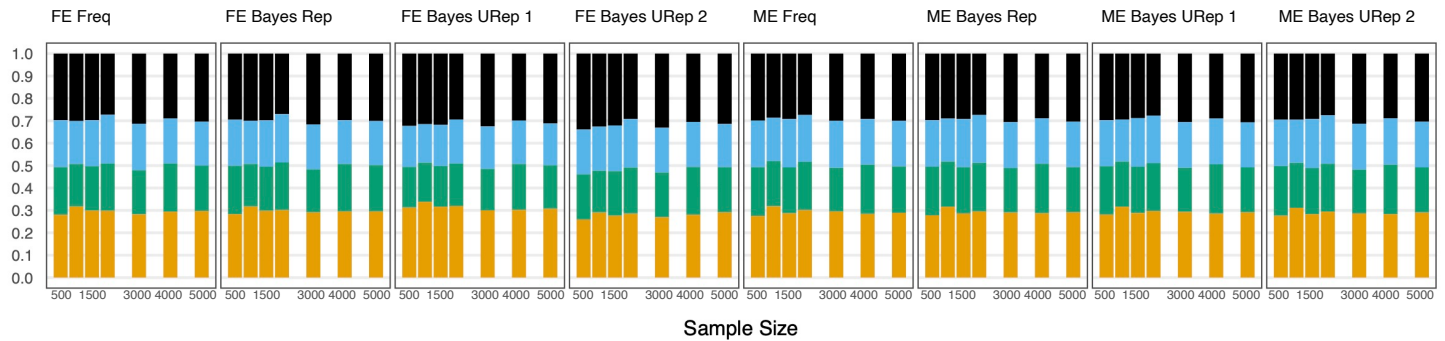


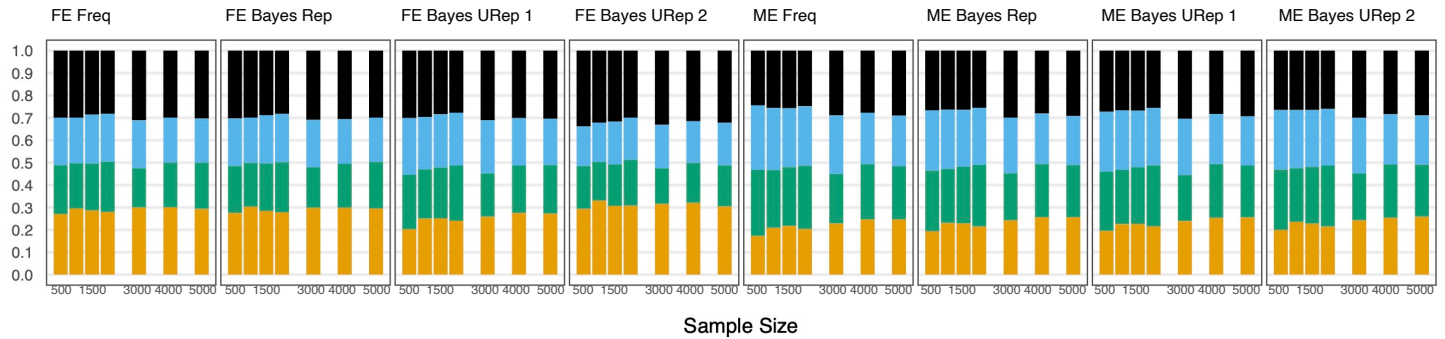
Fig. SM 1 Probability of incorrect interval separation for all null scenarios, when using a fixed and mixed effects model using frequentist and Bayesian approaches for sample sizes ranging between $N = 500$ -5,000.

When utilising only the point estimates to rank the treatments, the four ME methods had a more even spread of predicting each treatment as best than the four FE methods when the treatment mortality differed by patient subgroup (scenario 2.1). When the average mortality was low and high (scenario 3.1 and 3.3) the FE Bayes unrepresentative 1 and 2 were most likely to produce an uneven spread of predicting each treatment as best, which was more pronounced when the average mortality was low. The first unrepresentative prior had equal treatment mortality probabilities, but the probability was 10% lower than in the current trial, this increased the proportion of simulations where treatments *A* and *D* were predicted as best from roughly 30% (using FE frequentist method) to roughly 35% using the FE Bayes unrepresentative 1 in scenario 3.1. However, the second unrepresentative prior included treatment *A* as best and treatment *D* as worst, which was reflected in the proportion of simulations they are each predicted as best (roughly 35% for treatment *A* and roughly 25% for treatment *D*) in scenario 3.1. The ME Bayesian unrepresentative prior methods did not have these same issues, the proportion of simulations where each treatment was predicted as best was similar to the ME frequentist method (Figure SM 2, Table SM3).

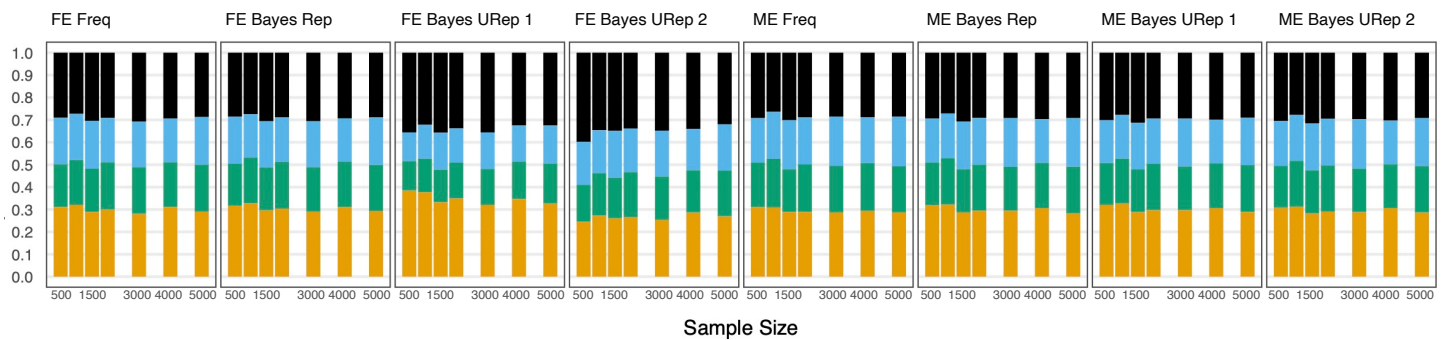
a) Sc 1.1



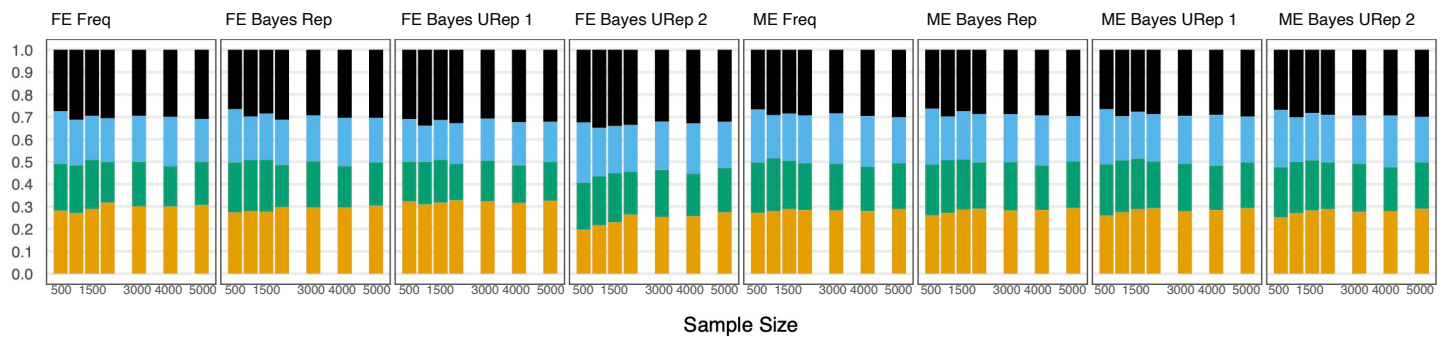
b) Sc 2.1



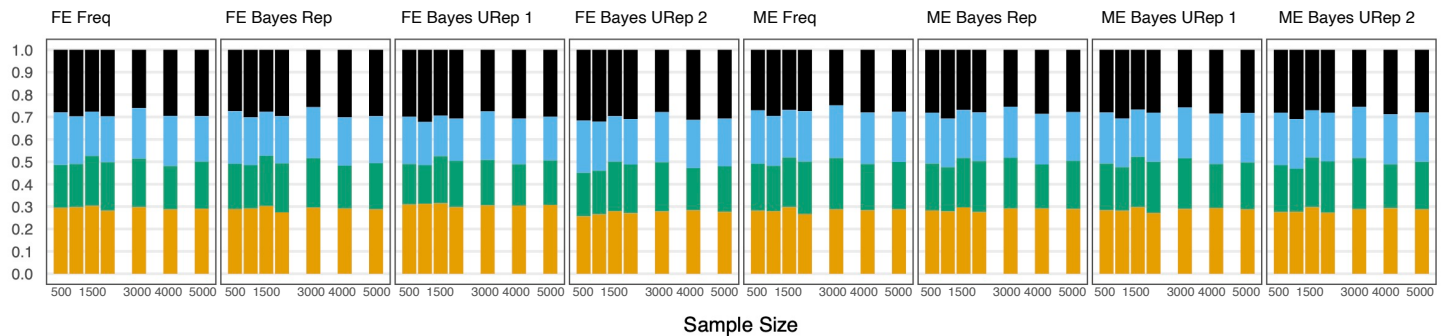
c) Sc 3.1



d) Sc 3.3



e) Sc 4.1



treatment A treatment B treatment C treatment D

Fig. SM 2 Proportion of scenarios where each treatment (A–D) were incorrectly predicted as best for five null scenarios, when using a fixed and mixed effects using frequentist and Bayesian approaches for sample sizes ranging between $N = 500$ – $5,000$.

Table SM3. Results for all null scenarios (1.1, 2.1, 3.1, 3.3, 4.1).

Scenario	Sample size	Method	P _{IIS} (i)	P _{IIS} (ii)	P _{IIS} (iii)	Prob A selected best	Prob B selected best	Prob C selected best	Prob D selected bests
1.1	500	FE Freq	0.003	0	0.001	0.297	0.21	0.213	0.28
		FE Bayes Rep	0.003	0	0.001	0.295	0.206	0.216	0.283
		FE Bayes UR1	0.003	0	0.001	0.323	0.183	0.182	0.312
		FE Bayes UR2	0.003	0	0.003	0.339	0.2	0.202	0.259
		ME Freq	0.017	0.003	0.01	0.299	0.208	0.218	0.275
		ME Bayes Rep	0.011	0	0.007	0.297	0.208	0.217	0.278
		ME Bayes UR1	0.01	0	0.006	0.297	0.206	0.216	0.281
		ME Bayes UR2	0.009	0	0.006	0.295	0.207	0.222	0.276
1.1	1000	FE Freq	0.004	0.001	0.001	0.301	0.192	0.19	0.317
		FE Bayes Rep	0.003	0.002	0.001	0.3	0.193	0.19	0.317
		FE Bayes UR1	0.005	0.001	0.001	0.315	0.172	0.175	0.338
		FE Bayes UR2	0.004	0.001	0	0.326	0.197	0.187	0.29
		ME Freq	0.012	0.006	0.007	0.286	0.194	0.202	0.318
		ME Bayes Rep	0.006	0.003	0.001	0.29	0.192	0.202	0.316
		ME Bayes UR1	0.006	0.002	0.001	0.294	0.188	0.202	0.316
		ME Bayes UR2	0.006	0.002	0.001	0.295	0.193	0.201	0.311
1.1	1500	FE Freq	0.001	0	0.002	0.297	0.206	0.198	0.299
		FE Bayes Rep	0.001	0	0.002	0.298	0.207	0.197	0.298
		FE Bayes UR1	0.001	0	0.002	0.318	0.184	0.182	0.316
		FE Bayes UR2	0.001	0	0.002	0.321	0.204	0.199	0.276
		ME Freq	0.01	0.002	0.007	0.292	0.216	0.205	0.287
		ME Bayes Rep	0.003	0	0.004	0.292	0.215	0.207	0.286
		ME Bayes UR1	0.003	0	0.004	0.288	0.217	0.207	0.288
		ME Bayes UR2	0.003	0	0.004	0.292	0.219	0.206	0.283
1.1	2000	FE Freq	0.004	0	0.002	0.273	0.218	0.211	0.298
		FE Bayes Rep	0.003	0	0.001	0.27	0.216	0.213	0.301
		FE Bayes UR1	0.004	0	0	0.294	0.197	0.19	0.319
		FE Bayes UR2	0.004	0	0.002	0.292	0.217	0.206	0.285
		ME Freq	0.01	0.001	0.013	0.274	0.209	0.216	0.301
		ME Bayes Rep	0.005	0	0.002	0.274	0.214	0.217	0.295
		ME Bayes UR1	0.005	0	0.003	0.277	0.212	0.214	0.297
		ME Bayes UR2	0.005	0	0.003	0.275	0.217	0.214	0.294
1.1	3000	FE Freq	0.003	0	0.003	0.313	0.208	0.197	0.282
		FE Bayes Rep	0.003	0	0.003	0.316	0.201	0.192	0.291
		FE Bayes UR1	0.005	0	0.003	0.325	0.19	0.185	0.3
		FE Bayes UR2	0.003	0	0.003	0.331	0.2	0.2	0.269
		ME Freq	0.017	0.004	0.013	0.3	0.21	0.195	0.295
		ME Bayes Rep	0.005	0.001	0.006	0.306	0.205	0.199	0.29
		ME Bayes UR1	0.005	0.001	0.006	0.306	0.204	0.197	0.293

		ME Bayes UR2	0.005	0.001	0.006	0.313	0.205	0.196	0.286
1.1	4000	FE Freq	0.002	0	0	0.29	0.201	0.215	0.294
		FE Bayes Rep	0.002	0	0	0.297	0.196	0.212	0.295
		FE Bayes UR1	0.002	0	0	0.299	0.194	0.205	0.302
		FE Bayes UR2	0.002	0	0	0.305	0.201	0.214	0.28
		ME Freq	0.014	0.004	0.006	0.292	0.204	0.22	0.284
		ME Bayes Rep	0.005	0	0.002	0.289	0.203	0.221	0.287
		ME Bayes UR1	0.004	0	0.001	0.29	0.204	0.221	0.285
		ME Bayes UR2	0.004	0	0.002	0.289	0.207	0.221	0.283
1.1	5000	FE Freq	0.001	0	0.002	0.304	0.196	0.203	0.297
		FE Bayes Rep	0.001	0	0.002	0.301	0.197	0.207	0.295
		FE Bayes UR1	0.001	0	0.002	0.312	0.187	0.194	0.307
		FE Bayes UR2	0.001	0	0.002	0.314	0.193	0.202	0.291
		ME Freq	0.01	0.004	0.012	0.3	0.204	0.207	0.289
		ME Bayes Rep	0.003	0.001	0.003	0.304	0.203	0.202	0.291
		ME Bayes UR1	0.003	0.001	0.005	0.307	0.201	0.201	0.291
		ME Bayes UR2	0.003	0.001	0.004	0.304	0.204	0.202	0.29
2.1	500	FE Freq	0	0	0.003	0.299	0.212	0.219	0.27
		FE Bayes Rep	0.002	0	0.003	0.302	0.213	0.21	0.275
		FE Bayes UR1	0.001	0	0.004	0.301	0.253	0.244	0.202
		FE Bayes UR2	0.005	0	0.004	0.338	0.177	0.192	0.293
		ME Freq	0.01	0.005	0.013	0.244	0.288	0.296	0.172
		ME Bayes Rep	0.002	0.001	0.006	0.266	0.27	0.271	0.193
		ME Bayes UR1	0.002	0.001	0.006	0.273	0.267	0.265	0.195
		ME Bayes UR2	0.002	0.001	0.004	0.265	0.266	0.27	0.199
2.1	1000	FE Freq	0.001	0	0	0.299	0.204	0.202	0.295
		FE Bayes Rep	0.001	0	0	0.299	0.202	0.196	0.303
		FE Bayes UR1	0.002	0	0.001	0.296	0.234	0.22	0.25
		FE Bayes UR2	0.004	0	0	0.321	0.176	0.173	0.33
		ME Freq	0.005	0.003	0.005	0.256	0.277	0.259	0.208
		ME Bayes Rep	0.002	0	0	0.263	0.265	0.242	0.23
		ME Bayes UR1	0.002	0	0	0.266	0.266	0.243	0.225
		ME Bayes UR2	0.002	0	0	0.265	0.26	0.24	0.235
2.1	1500	FE Freq	0.002	0	0.002	0.285	0.219	0.21	0.286
		FE Bayes Rep	0.002	0	0.002	0.288	0.216	0.212	0.284
		FE Bayes UR1	0.002	0	0.002	0.284	0.238	0.227	0.251
		FE Bayes UR2	0.002	0	0.002	0.317	0.191	0.186	0.306
		ME Freq	0.001	0	0.004	0.257	0.263	0.263	0.217
		ME Bayes Rep	0	0	0	0.264	0.254	0.254	0.228
		ME Bayes UR1	0	0	0	0.268	0.253	0.253	0.226
		ME Bayes UR2	0	0	0	0.265	0.254	0.254	0.227
2.1	2000	FE Freq	0.001	0	0.002	0.282	0.214	0.225	0.279
		FE Bayes Rep	0	0	0.001	0.282	0.216	0.224	0.278

		FE Bayes UR1	0	0	0.003	0.277	0.236	0.247	0.24
		FE Bayes UR2	0.001	0.001	0	0.299	0.189	0.205	0.307
		ME Freq	0	0	0.001	0.248	0.266	0.283	0.203
		ME Bayes Rep	0	0	0	0.255	0.255	0.276	0.214
		ME Bayes UR1	0	0	0	0.256	0.256	0.273	0.215
		ME Bayes UR2	0	0	0	0.26	0.253	0.273	0.214
2.1	3000	FE Freq	0.001	0	0.002	0.311	0.214	0.176	0.299
		FE Bayes Rep	0.001	0	0.003	0.309	0.212	0.181	0.298
		FE Bayes UR1	0.001	0	0.005	0.31	0.238	0.194	0.258
		FE Bayes UR2	0.002	0	0.002	0.331	0.194	0.16	0.315
		ME Freq	0	0	0.004	0.289	0.262	0.221	0.228
		ME Bayes Rep	0	0	0.001	0.299	0.249	0.21	0.242
		ME Bayes UR1	0	0	0.001	0.304	0.251	0.206	0.239
		ME Bayes UR2	0	0	0.001	0.3	0.249	0.209	0.242
2.1	4000	FE Freq	0.002	0	0	0.299	0.201	0.2	0.3
		FE Bayes Rep	0.003	0	0	0.306	0.199	0.197	0.298
		FE Bayes UR1	0.002	0	0.001	0.301	0.212	0.212	0.275
		FE Bayes UR2	0.003	0	0	0.315	0.186	0.179	0.32
		ME Freq	0	0	0	0.277	0.23	0.247	0.246
		ME Bayes Rep	0	0	0	0.28	0.226	0.238	0.256
		ME Bayes UR1	0	0	0	0.283	0.224	0.24	0.253
		ME Bayes UR2	0	0	0	0.284	0.224	0.239	0.253
2.1	5000	FE Freq	0	0	0.003	0.303	0.197	0.207	0.293
		FE Bayes Rep	0	0	0.004	0.299	0.198	0.208	0.295
		FE Bayes UR1	0	0	0.003	0.304	0.207	0.217	0.272
		FE Bayes UR2	0	0	0.003	0.322	0.19	0.184	0.304
		ME Freq	0	0	0	0.29	0.226	0.238	0.246
		ME Bayes Rep	0	0	0	0.292	0.219	0.233	0.256
		ME Bayes UR1	0	0	0	0.293	0.22	0.232	0.255
		ME Bayes UR2	0	0	0	0.289	0.22	0.233	0.258
3.1	500	FE Freq	0.006	0	0.001	0.29	0.208	0.191	0.311
		FE Bayes Rep	0.004	0	0.002	0.285	0.211	0.187	0.317
		FE Bayes UR1	0.023	0	0.001	0.356	0.128	0.132	0.384
		FE Bayes UR2	0.008	0	0.002	0.398	0.192	0.165	0.245
		ME Freq	0.016	0.003	0.004	0.292	0.199	0.199	0.31
		ME Bayes Rep	0.012	0.001	0.002	0.294	0.196	0.192	0.318
		ME Bayes UR1	0.012	0.001	0.002	0.301	0.192	0.187	0.32
		ME Bayes UR2	0.011	0.001	0.002	0.305	0.2	0.188	0.307
3.1	1000	FE Freq	0.005	0	0.002	0.273	0.206	0.201	0.32
		FE Bayes Rep	0.007	0	0.003	0.274	0.194	0.205	0.327
		FE Bayes UR1	0.012	0	0.002	0.322	0.151	0.15	0.377
		FE Bayes UR2	0.004	0	0.003	0.346	0.192	0.19	0.272
		ME Freq	0.015	0.003	0.012	0.264	0.21	0.217	0.309

		ME Bayes Rep	0.006	0.001	0.003	0.271	0.2	0.208	0.321
		ME Bayes UR1	0.006	0.001	0.002	0.278	0.195	0.201	0.326
		ME Bayes UR2	0.005	0	0.002	0.278	0.205	0.205	0.312
3.1	1500	FE Freq	0.003	0.001	0.001	0.304	0.214	0.193	0.289
		FE Bayes Rep	0.003	0.001	0.002	0.306	0.206	0.192	0.296
		FE Bayes UR1	0.01	0.001	0.002	0.357	0.165	0.146	0.332
		FE Bayes UR2	0.005	0	0.002	0.349	0.209	0.182	0.26
		ME Freq	0.015	0.006	0.009	0.301	0.219	0.191	0.289
		ME Bayes Rep	0.007	0.001	0.003	0.307	0.213	0.193	0.287
		ME Bayes UR1	0.008	0.001	0.003	0.313	0.208	0.191	0.288
		ME Bayes UR2	0.007	0.001	0.003	0.316	0.209	0.193	0.282
3.1	2000	FE Freq	0.003	0	0.002	0.291	0.198	0.212	0.299
		FE Bayes Rep	0.005	0	0.001	0.289	0.198	0.21	0.303
		FE Bayes UR1	0.007	0	0.001	0.338	0.153	0.16	0.349
		FE Bayes UR2	0.004	0	0.001	0.339	0.194	0.203	0.264
		ME Freq	0.014	0.003	0.011	0.289	0.21	0.212	0.289
		ME Bayes Rep	0.006	0	0.001	0.291	0.209	0.206	0.294
		ME Bayes UR1	0.006	0	0.001	0.294	0.202	0.207	0.297
		ME Bayes UR2	0.006	0	0.001	0.295	0.208	0.207	0.29
3.1	3000	FE Freq	0.002	0	0.003	0.308	0.203	0.209	0.28
		FE Bayes Rep	0.003	0	0.004	0.306	0.205	0.199	0.29
		FE Bayes UR1	0.006	0	0.002	0.356	0.163	0.161	0.32
		FE Bayes UR2	0.004	0	0.003	0.348	0.205	0.193	0.254
		ME Freq	0.012	0.003	0.013	0.285	0.22	0.208	0.287
		ME Bayes Rep	0.005	0	0.004	0.292	0.218	0.196	0.294
		ME Bayes UR1	0.006	0	0.004	0.294	0.214	0.195	0.297
		ME Bayes UR2	0.006	0	0.003	0.297	0.22	0.195	0.288
3.1	4000	FE Freq	0.003	0	0.002	0.294	0.195	0.2	0.311
		FE Bayes Rep	0.003	0	0.002	0.293	0.193	0.204	0.31
		FE Bayes UR1	0.004	0	0	0.325	0.161	0.168	0.346
		FE Bayes UR2	0.004	0	0.002	0.34	0.185	0.189	0.286
		ME Freq	0.013	0.002	0.01	0.288	0.205	0.214	0.293
		ME Bayes Rep	0.005	0.001	0.002	0.297	0.195	0.202	0.306
		ME Bayes UR1	0.006	0.001	0.002	0.299	0.195	0.2	0.306
		ME Bayes UR2	0.005	0.001	0.002	0.303	0.195	0.196	0.306
3.1	5000	FE Freq	0.006	0	0.003	0.287	0.213	0.21	0.29
		FE Bayes Rep	0.005	0	0.002	0.289	0.212	0.207	0.292
		FE Bayes UR1	0.007	0	0.002	0.325	0.17	0.179	0.326
		FE Bayes UR2	0.005	0	0.003	0.32	0.206	0.205	0.269
		ME Freq	0.017	0.006	0.018	0.285	0.222	0.206	0.287
		ME Bayes Rep	0.007	0.001	0.004	0.292	0.217	0.209	0.282
		ME Bayes UR1	0.007	0.001	0.004	0.29	0.212	0.209	0.289
		ME Bayes UR2	0.007	0.001	0.004	0.292	0.214	0.208	0.286

3.3	500	FE Freq	0.001	0	0.006	0.275	0.234	0.211	0.28
		FE Bayes Rep	0.002	0	0.004	0.265	0.238	0.224	0.273
		FE Bayes UR1	0.002	0	0.005	0.31	0.19	0.179	0.321
		FE Bayes UR2	0.002	0	0.011	0.325	0.268	0.211	0.196
		ME Freq	0.005	0.003	0.015	0.267	0.236	0.227	0.27
		ME Bayes Rep	0.003	0.001	0.011	0.263	0.248	0.23	0.259
		ME Bayes UR1	0.002	0.001	0.011	0.265	0.247	0.23	0.258
		ME Bayes UR2	0.002	0.001	0.014	0.269	0.255	0.226	0.25
3.3	1000	FE Freq	0.002	0	0.005	0.313	0.203	0.215	0.269
		FE Bayes Rep	0.004	0	0.008	0.298	0.195	0.228	0.279
		FE Bayes UR1	0.003	0	0.003	0.339	0.161	0.192	0.308
		FE Bayes UR2	0.003	0	0.011	0.349	0.216	0.221	0.214
		ME Freq	0.012	0.003	0.015	0.292	0.193	0.236	0.279
		ME Bayes Rep	0.003	0	0.007	0.298	0.195	0.237	0.27
		ME Bayes UR1	0.002	0	0.008	0.297	0.197	0.232	0.274
		ME Bayes UR2	0.003	0	0.008	0.302	0.199	0.23	0.269
3.3	1500	FE Freq	0.001	0.001	0.003	0.295	0.196	0.223	0.286
		FE Bayes Rep	0.001	0.001	0.003	0.285	0.206	0.233	0.276
		FE Bayes UR1	0.002	0.001	0.003	0.314	0.177	0.193	0.316
		FE Bayes UR2	0.001	0.001	0.006	0.341	0.21	0.22	0.229
		ME Freq	0.009	0.006	0.015	0.285	0.21	0.218	0.287
		ME Bayes Rep	0.003	0.001	0.008	0.275	0.215	0.225	0.285
		ME Bayes UR1	0.003	0.001	0.008	0.277	0.209	0.228	0.286
		ME Bayes UR2	0.003	0.001	0.01	0.282	0.212	0.225	0.281
3.3	2000	FE Freq	0.002	0	0.003	0.306	0.194	0.184	0.316
		FE Bayes Rep	0.002	0	0.004	0.312	0.202	0.19	0.296
		FE Bayes UR1	0.002	0	0.002	0.328	0.181	0.164	0.327
		FE Bayes UR2	0.002	0	0.003	0.336	0.209	0.192	0.263
		ME Freq	0.012	0.003	0.013	0.293	0.213	0.211	0.283
		ME Bayes Rep	0.003	0	0.005	0.287	0.216	0.208	0.289
		ME Bayes UR1	0.003	0	0.006	0.288	0.211	0.21	0.291
		ME Bayes UR2	0.003	0	0.006	0.291	0.212	0.211	0.286
3.3	3000	FE Freq	0.003	0	0.002	0.295	0.205	0.199	0.301
		FE Bayes Rep	0.004	0	0.003	0.293	0.205	0.208	0.294
		FE Bayes UR1	0.003	0	0.002	0.308	0.187	0.183	0.322
		FE Bayes UR2	0.004	0	0.001	0.321	0.215	0.212	0.252
		ME Freq	0.013	0.003	0.012	0.284	0.224	0.21	0.282
		ME Bayes Rep	0.004	0	0.006	0.288	0.214	0.217	0.281
		ME Bayes UR1	0.005	0	0.005	0.295	0.214	0.212	0.279
		ME Bayes UR2	0.004	0	0.005	0.294	0.215	0.216	0.275
3.3	4000	FE Freq	0.002	0	0.003	0.3	0.219	0.182	0.299
		FE Bayes Rep	0.002	0	0.003	0.304	0.215	0.187	0.294
		FE Bayes UR1	0.002	0	0.003	0.323	0.193	0.169	0.315
		FE Bayes UR2	0.004	0.001	0.004	0.329	0.223	0.192	0.256

		ME Freq ME Bayes Rep ME Bayes UR1 ME Bayes UR2	0.01 0.002 0.002 0.002	0.002 0.001 0.001 0.001	0.013 0.007 0.005 0.005	0.296 0.293 0.291 0.294	0.227 0.225 0.226 0.23	0.199 0.198 0.199 0.197	0.278 0.284 0.284 0.279
3.3	5000	FE Freq FE Bayes Rep FE Bayes UR1 FE Bayes UR2	0.003 0.002 0.003 0.004	0 0 0 0	0.006 0.007 0.006 0.005	0.309 0.304 0.322 0.322	0.191 0.2 0.178 0.206	0.194 0.194 0.176 0.199	0.306 0.302 0.324 0.273
		ME Freq ME Bayes Rep ME Bayes UR1 ME Bayes UR2	0.019 0.005 0.004 0.004	0.006 0.001 0.001 0.001	0.016 0.007 0.007 0.005	0.301 0.297 0.298 0.3	0.206 0.202 0.206 0.204	0.206 0.208 0.204 0.207	0.287 0.293 0.292 0.289
		FE Freq FE Bayes Rep FE Bayes UR1 FE Bayes UR2	0 0 0 0	0 0 0 0	0.001 0.001 0.001 0.001	0.279 0.274 0.299 0.316	0.235 0.235 0.211 0.233	0.192 0.203 0.181 0.195	0.294 0.288 0.309 0.256
		ME Freq ME Bayes Rep ME Bayes UR1 ME Bayes UR2	0.011 0.002 0.002 0.003	0.001 0.001 0.001 0.001	0.012 0.004 0.004 0.004	0.27 0.281 0.28 0.281	0.239 0.227 0.228 0.234	0.21 0.21 0.209 0.21	0.281 0.282 0.283 0.275
		FE Freq FE Bayes Rep FE Bayes UR1 FE Bayes UR2	0 0 0 0	0 0 0 0	0 0 0 0	0.297 0.301 0.322 0.321	0.213 0.214 0.193 0.219	0.193 0.194 0.173 0.195	0.297 0.291 0.312 0.265
		ME Freq ME Bayes Rep ME Bayes UR1 ME Bayes UR2	0.012 0.005 0.002 0.004	0.002 0.001 0.001 0.001	0.012 0.001 0.001 0.002	0.295 0.307 0.306 0.309	0.223 0.217 0.218 0.222	0.203 0.198 0.195 0.193	0.279 0.278 0.281 0.276
		FE Freq FE Bayes Rep FE Bayes UR1 FE Bayes UR2	0 0 0 0	0 0 0 0	0 0 0 0	0.276 0.277 0.294 0.296	0.198 0.196 0.181 0.203	0.223 0.225 0.211 0.222	0.303 0.302 0.314 0.279
		ME Freq ME Bayes Rep ME Bayes UR1 ME Bayes UR2	0.013 0.001 0.001 0.002	0 0 0 0	0.014 0.004 0.004 0.006	0.268 0.268 0.267 0.27	0.213 0.215 0.211 0.211	0.222 0.221 0.225 0.222	0.297 0.296 0.297 0.297
4.1	2000	FE Freq FE Bayes Rep FE Bayes UR1 FE Bayes UR2	0 0 0 0	0 0 0 0	0 0 0 0	0.297 0.296 0.307 0.309	0.205 0.211 0.189 0.203	0.217 0.219 0.207 0.218	0.281 0.274 0.297 0.27
		ME Freq ME Bayes Rep ME Bayes UR1	0.013 0.004 0.004	0.001 0 0	0.015 0.006 0.006	0.274 0.279 0.281	0.225 0.218 0.219	0.235 0.228 0.229	0.266 0.275 0.271

		ME Bayes UR2	0.004	0	0.005	0.281	0.217	0.23	0.272
4.1	3000	FE Freq	0	0	0	0.26	0.226	0.217	0.297
		FE Bayes Rep	0	0	0	0.256	0.228	0.221	0.295
		FE Bayes UR1	0	0	0	0.275	0.217	0.203	0.305
		FE Bayes UR2	0	0	0	0.278	0.224	0.22	0.278
		ME Freq	0.012	0.001	0.01	0.248	0.235	0.23	0.287
		ME Bayes Rep	0.001	0.001	0.003	0.254	0.228	0.227	0.291
		ME Bayes UR1	0.002	0.001	0.003	0.257	0.227	0.227	0.289
		ME Bayes UR2	0.005	0.001	0.003	0.255	0.228	0.229	0.288
4.1	4000	FE Freq	0	0	0	0.295	0.224	0.194	0.287
		FE Bayes Rep	0	0	0	0.301	0.216	0.192	0.291
		FE Bayes UR1	0	0	0	0.307	0.204	0.185	0.304
		FE Bayes UR2	0	0	0	0.313	0.215	0.189	0.283
		ME Freq	0.007	0.003	0.012	0.28	0.23	0.207	0.283
		ME Bayes Rep	0.001	0	0	0.286	0.226	0.197	0.291
		ME Bayes UR1	0	0	0	0.285	0.225	0.197	0.293
		ME Bayes UR2	0.001	0	0	0.288	0.223	0.197	0.292
4.1	5000	FE Freq	0	0	0	0.296	0.203	0.211	0.29
		FE Bayes Rep	0	0	0	0.296	0.21	0.207	0.287
		FE Bayes UR1	0	0	0	0.299	0.195	0.199	0.307
		FE Bayes UR2	0	0	0	0.307	0.213	0.204	0.276
		ME Freq	0.013	0.002	0.01	0.277	0.224	0.212	0.287
		ME Bayes Rep	0.002	0	0.004	0.278	0.218	0.215	0.289
		ME Bayes UR1	0.002	0	0.004	0.282	0.221	0.21	0.287
		ME Bayes UR2	0.002	0	0.004	0.28	0.22	0.213	0.287

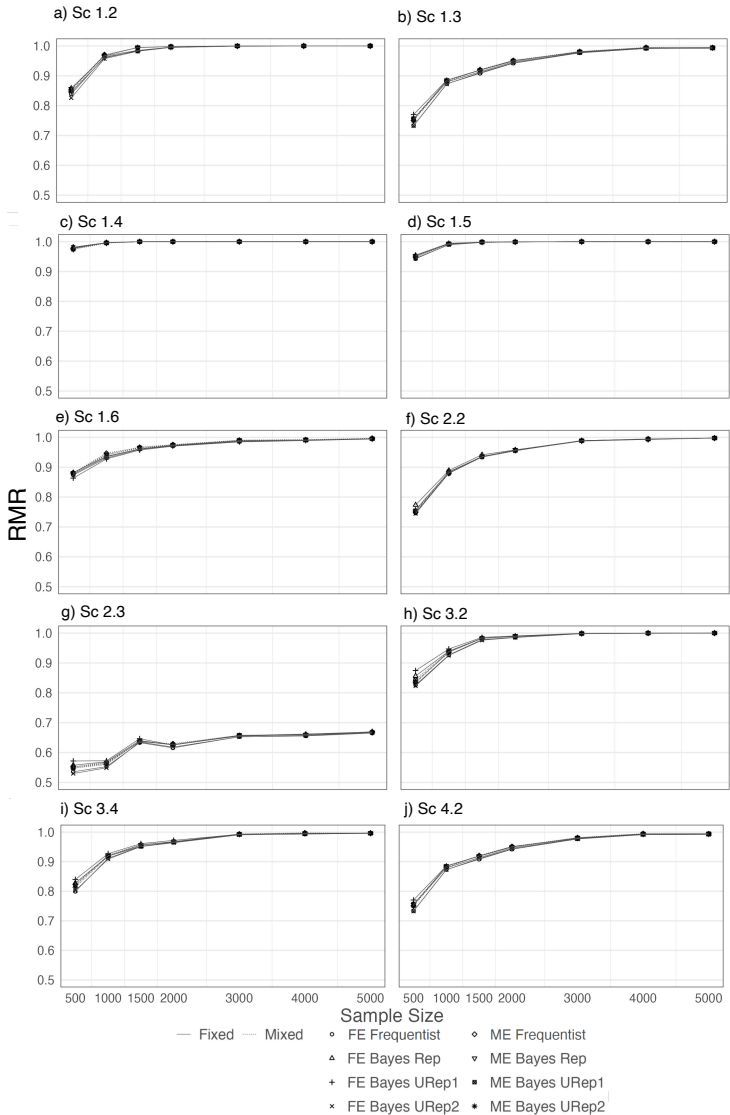
P_{IIS} , probability of incorrect interval separation; P_{IIS} (i) incorrectly predicting 1 best treatment; P_{IIS} (ii) incorrectly predicting 2 best treatments; P_{IIS} (iii) incorrectly predicting 3 best treatments; Prob, Probability;

Reduction in mortality ratio

The FE Bayesian method consistently performed better than the ME methods across the scenarios, although the difference between methods was never more than 0.04. The ME frequentist method performed better than the FE Bayesian method only in one scenario, where there was a small difference in treatment mortality probability, but the best and worst treatments were more common across the patterns (scenario 1.6). The ME frequentist method and the FE Bayesian method with a representative prior produced a reduction in mortality ratio of roughly 0.95 and 0.93, respectively for a sample size of N=1,000 (Figure SM 3).

When the true effects of all treatments were different, a large difference in treatment mortality probabilities (scenario 1.5) achieved a higher reduction in mortality ratio compared with a small difference in mortality probabilities (scenarios 1.3 and 1.6) see Figure SM 3. When comparing the two scenarios with a small difference in treatment mortality probabilities (1.3 and 1.6), it was the scenario where the best and worst treatments were available in more patterns (1.6) which produced the larger reduction in mortality ratio, see Figure SM 3.

When the average mortality probability increased (scenario 3.4 vs 1.3) the reduction in mortality ratio, increased. However, when the average mortality probability decreased (scenario 3.2 vs 3.4) reduction in mortality ratio increased further, see Figure SM 3. The order of which methods performed best stayed similar across the scenarios, where the FE Bayesian with unrepresentative prior 1 often performed best.



Best Treatment Probability

The change you see in reduction in mortality ratio across the scenarios was mirrored in the best treatment probability measure. The best treatment probability was larger than the reduction in mortality ratio, across many scenarios (Figure SM 4).

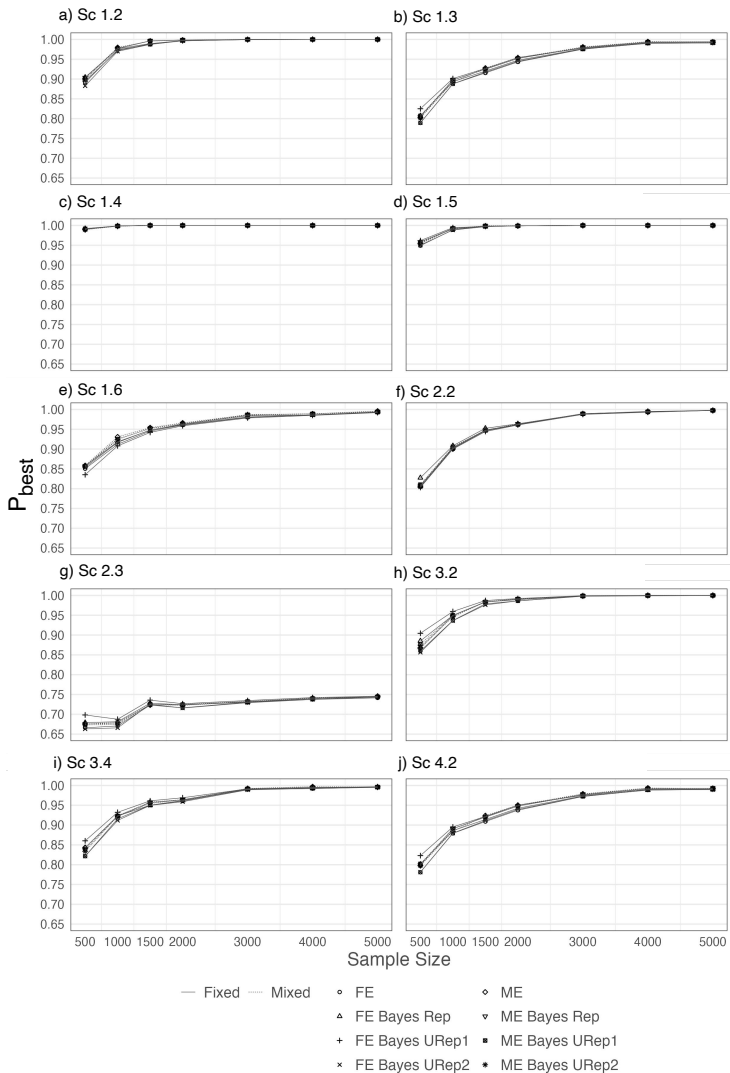


Fig. SM 4 Best treatment probability for all non-null scenarios, when using a fixed and mixed effects model using frequentist and Bayesian approaches for sample sizes ranging between $N = 500$ - $5,000$.

Improvement Over Random Choice

The change you see in reduction in mortality ratio across the scenarios was mirrored in the improvement over random choice measure. The improvement over random choice was larger than the best treatment probability, across many scenarios (Figure SM 5).

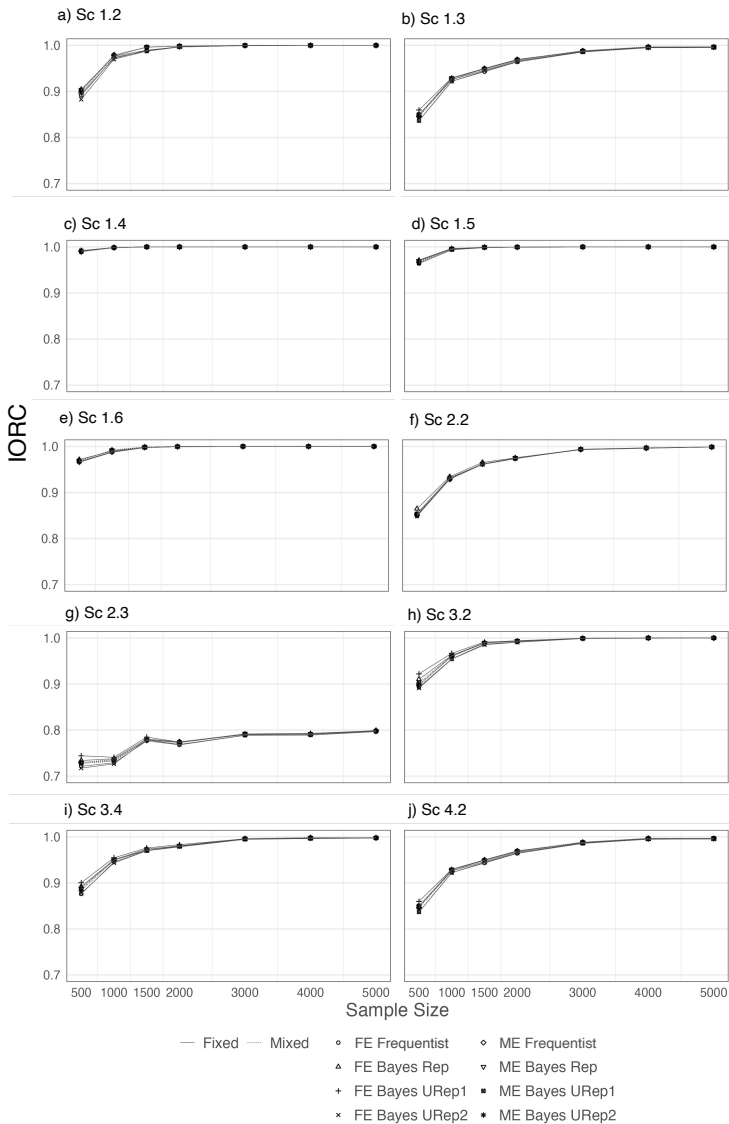


Fig. SM 5 Improvement over random choice for all non-null scenarios, when using a fixed and mixed effects model using frequentist and Bayesian approaches for sample sizes ranging between $N = 500$ -5,000.

Near-best Treatment Probability

All the methods, for all scenarios, produced similar near-best treatment probabilities (Figure SM 6) to that given by the performance measure, improvement over random choice (Figure SM 5).

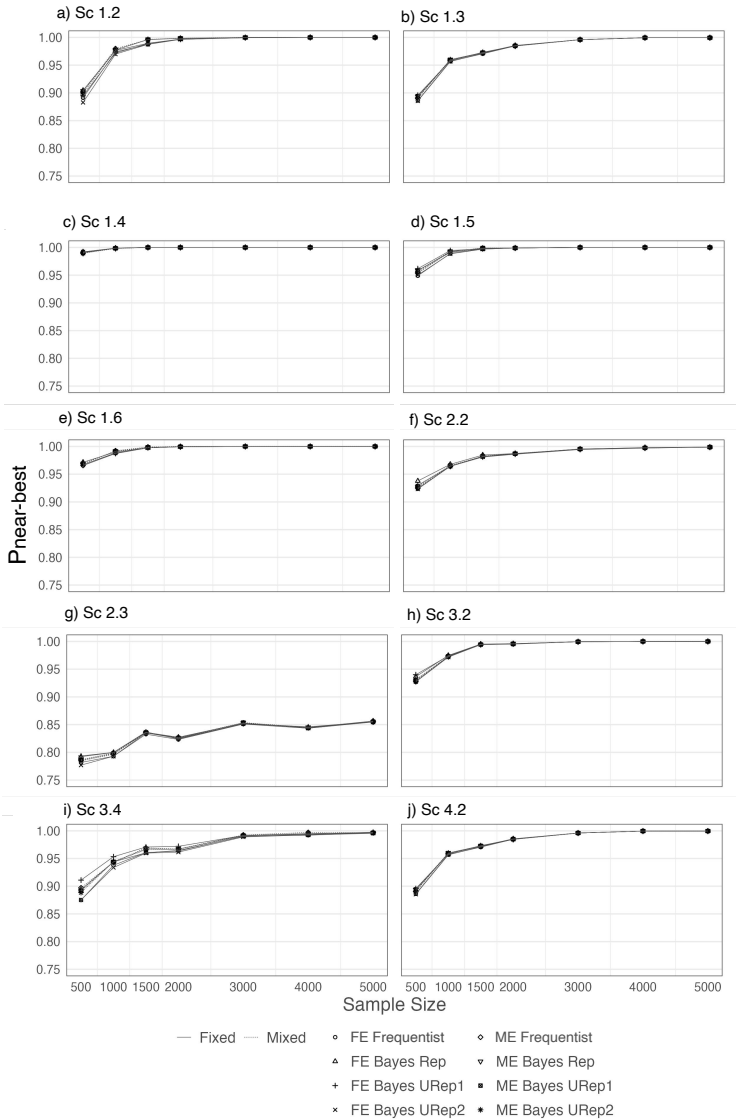


Fig. SM 6 Near-best treatment probability for all non-null scenarios, when using a fixed and mixed effects model using frequentist and Bayesian approaches for sample sizes ranging between $N = 500$ - $5,000$.

Probability of Interval Separation

For all scenarios where the treatment mortality probability did not differ by patient subgroup the ME frequentist method consistently produced the largest probability of interval separation. The three ME Bayesian methods all produced very similar moderate probabilities of interval separation and the FE methods consistently produced the smallest probabilities of interval separation. For example in scenario 1.2, the ME frequentist and Bayesian representative methods reached a probability of interval separation of 80% with a sample size of roughly 2,300 and 2,900 patients, respectively. However, the FE methods needed a sample size of roughly 3,500 patients for a similar probability (Figure SM 7).

Whether one was trying to correctly predict the best or worst treatment for the probability of interval separation, influenced the performance of each FE method. The FE Bayes representative was better at predicting the worst treatment(s) and the FE Bayes unrepresentative 1 was better at predicting the best treatment(s) (Figure SM 7).

When the mortality probabilities of all treatments were different, a large difference (scenario 1.5) achieved a higher probability of interval separation compared with a small difference in treatment mortality probabilities (scenarios 1.3 and 1.6), see Figure SM 7. Conversely to the reduction in mortality ratio, when the difference in treatment mortality probabilities was small and the best and worst treatments were available in fewer patterns (scenario 1.3) the probability of interval separation was larger than the scenario where the best and worst treatments were available in all patterns (scenario 1.6), see Figure SM 7.

When the average mortality probability increased (scenario 3.4 vs 1.3) the probability of interval separation was larger when identifying the true worst treatment(s) than the best treatment(s). However, when the average mortality decreased (scenario 3.2 vs 1.3) the probability of interval separation was larger when identifying the best treatment(s) (Figure SM 7), translating to better reduction in mortality ratio, best treatment probability, improvement over random choice and near-best treatment.

When the treatment mortality probability did differ by patient subgroup (scenario 2.2) all four ME methods produced a probability of interval separation of 0, for all sample sizes investigated (Figure SM 7).

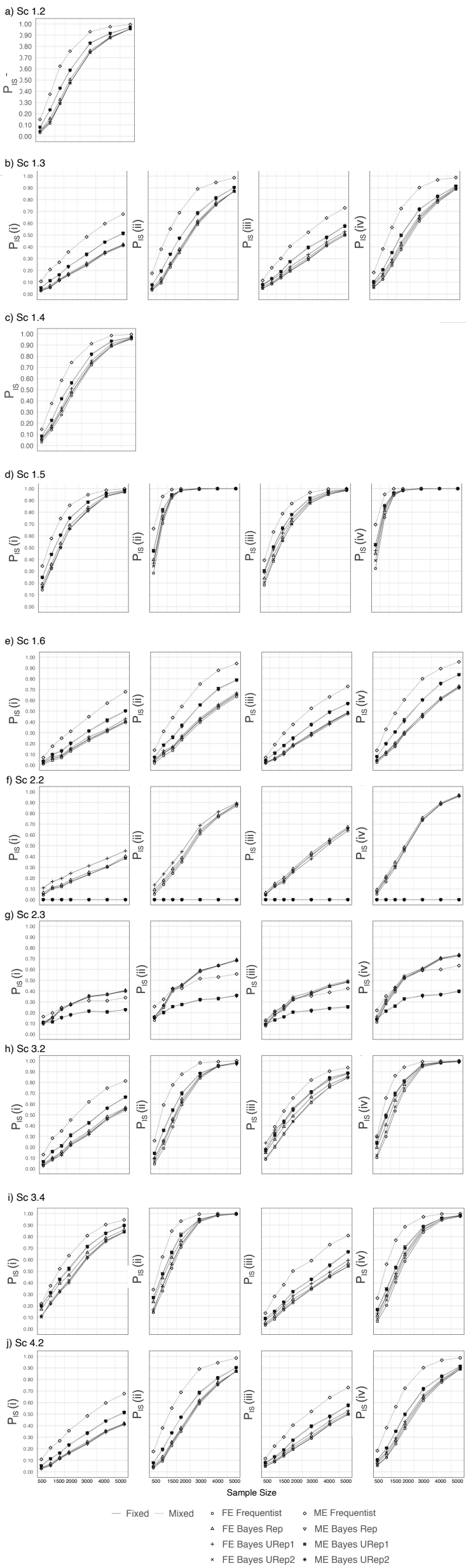


Fig. SM 7 Probability of interval separation for all non-null scenarios, when using a fixed and mixed effects model using frequentist and Bayesian approaches for sample sizes ranging between $N = 500$ -5,000.