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,

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

We implemented the ME model described above using a frequentist approach (using the R package 'lme4' [18]) and Bayesian approach (via the R package 'rstanarm' [14]). The ME Bayesian method utilised a prior for the four treatment coefficients.

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 $k = 1, 2, 3, 4$	representative; All treatments had equal mortality probability for all patient subgroups: P_{jk} = 37.5%, $j = A, B, C, D$ and $k = 1, 2, 3, 4$
		unrepresentative 1; All treatments had equal mortality probability for all patient subgroups: P_{jk} = 27.5%, $j = A, B, C, D$ and $k = 1, 2, 3, 4$
		unrepresentative 2; 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 $k = 1, 2, 3, 4$	representative; 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 $k = 1, 2, 3, 4$
		unrepresentative 1; 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 $k = 1, 2, 3, 4$
		unrepresentative 2; All treatments had different mortality probabilities, which was equal across patient subgroups: P_{Ak} = 30%; P_{Bk} = 35%; P_{Ck} = 40%; P_{Dk} = 45%, $k = 1, 2, 3, 4$
1.3 Small difference	All treatments had different mortality probabilities, which was equal across patient subgroups:	representative; 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>; Treatment B had larger mortality probability than treatments A, C, D for all patient subgroups: P_{Bk} = 35% $k = 1, 2, 3, 4$; and P_{jk} = 25%, $j = A, C, D$ 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%, $k = 1, 2, 3, 4$
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 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%, $j = A, B, C, D$
		<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%, $k = 1, 2, 3, 4$
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 $q = 3, \ldots, 10$; 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 $q = 3, \ldots, 10$; and P_{Akq} = 20%, P_{Bkq} = 25%, P_{Ckq} = 30%, P_{Dkq} = 35%, $k = 1, 2, 3, 4$ and $q = 1, 2$

		unrepresentative 1; 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 $q = 3, \ldots, 10$; and P_{Akq} = 10%, P_{Bkq} = 15%, P_{Ckq} = 20%, P_{Dkq} = 25%, $k = 1, 2, 3, 4$ and $q = 1, 2$
		unrepresentative 2; 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 $q = 3, \ldots, 10$ 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%, $k = 1, 2, 3, 4$	representative; All treatments had different mortality probabilities, which was equal across patient subgroups: P_{Ak} = 30%; P_{Bk} = 40%; P_{Ck} = 50%; P_{Dk} = 60%, $k = 1, 2, 3, 4$	λ_k = 25%, $k = 1, 2, 3, 4$
		unrepresentative; All treatments had different mortality probabilities, which was equal across patient subgroups: P_{Ak} = 20%; P_{Bk} = 30%; P_{Ck} = 40%; P_{Dk} = 50%, $k = 1, 2, 3, 4$	
		unrepresentative; 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%, $k = 1, 2, 3, 4$	representative; All treatments had different mortality probabilities, which was equal across patient subgroups: P_{Ak} = 35%; P_{Bk} = 30%; P_{Ck} = 45%; P_{Dk} = 40%, $k = 1, 2, 3, 4$	λ_k = 25%, $k = 1, 2, 3, 4$
		unrepresentative; All treatments had different mortality probabilities, which was equal across patient subgroups: P_{Ak} = 25%; P_{Bk} = 20%; P_{Ck} = 35%; P_{Dk} = 30%, $k = 1, 2, 3, 4$	
		unrepresentative; 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 $k = 1, 2, 3, 4$	representative; P_{jk} = 17.5%, $j = A, B, C, D$ and $k = 1, 2, 3, 4$	λ_k = 25%, $k = 1, 2, 3, 4$
		unrepresentative; P_{jk} = 7.5%, $j = A, B, C, D$ and $k = 1, 2, 3, 4$	
		unrepresentative; 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%, $k = 1, 2, 3, 4$	representative; P_{Ak} = 10%, P_{Bk} = 15%, P_{Ck} = 20%, P_{Dk} = 25%, $k = 1, 2, 3, 4$	λ_k = 25%, $k = 1, 2, 3, 4$
		unrepresentative; P_{Ak} = 5%, P_{Bk} = 10%, P_{Ck} = 15%, P_{Dk} = 20%, $k = 1, 2, 3, 4$	
		unrepresentative; 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 $k = 1, 2, 3, 4$	representative; P_{jk} = 82.5%, $j = A, B, C, D$ and $k = 1, 2, 3, 4$	λ_k = 25%, $k = 1, 2, 3, 4$
		unrepresentative; P_{jk} = 72.5%, $j = A, B, C, D$ and $k = 1, 2, 3, 4$	
		unrepresentative; 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%, $k = 1, 2, 3, 4$	representative; P_{Ak} = 75%, P_{Bk} = 80%, P_{Ck} = 85%, P_{Dk} = 90%, $k = 1, 2, 3, 4$	λ_k = 25%, $k = 1, 2, 3, 4$
		unrepresentative;	

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 $k = 1, 2, 3, 4$	λ_1 = 10%, λ_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%, $k = 1, 2, 3, 4$	λ_1 = 10%, λ_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.* [7] 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.

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

- **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.

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

- **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.

- **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).

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

- **Probability of interval separation:** proportion of iterations which correctly identified the predefined best and worst treatment(s). We used four definitions of probability of interval separation due to us including 4 treatments in our

simulations. The first definition, P_{IS} (A), 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} (B), either singularly or in combination. The third definition, P_{IS} (C), 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} (D), either alone or together.

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).

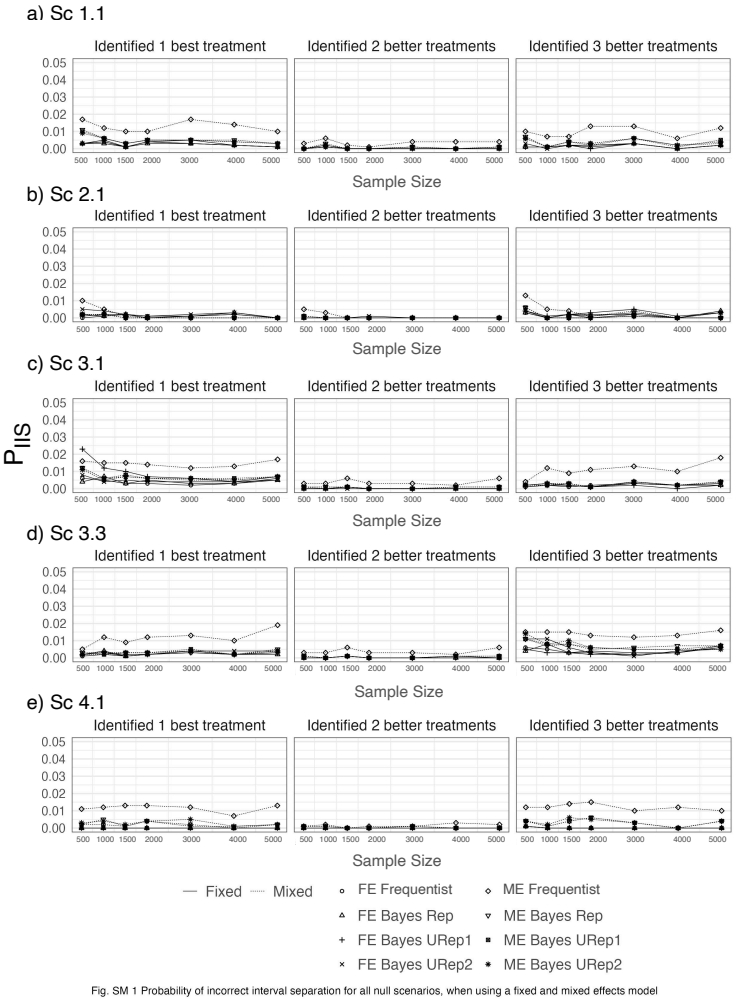

Fig. SM 1 Probability of incorrect interval separation for all null scenarios, when using a fixed and mixed effects model using frequentist and Bayesian approaches for sample sizes ranging between N = 500-5,000.

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 where predicted as best from 30% (using FE frequentist method) to 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 (35% for treatment A and 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).

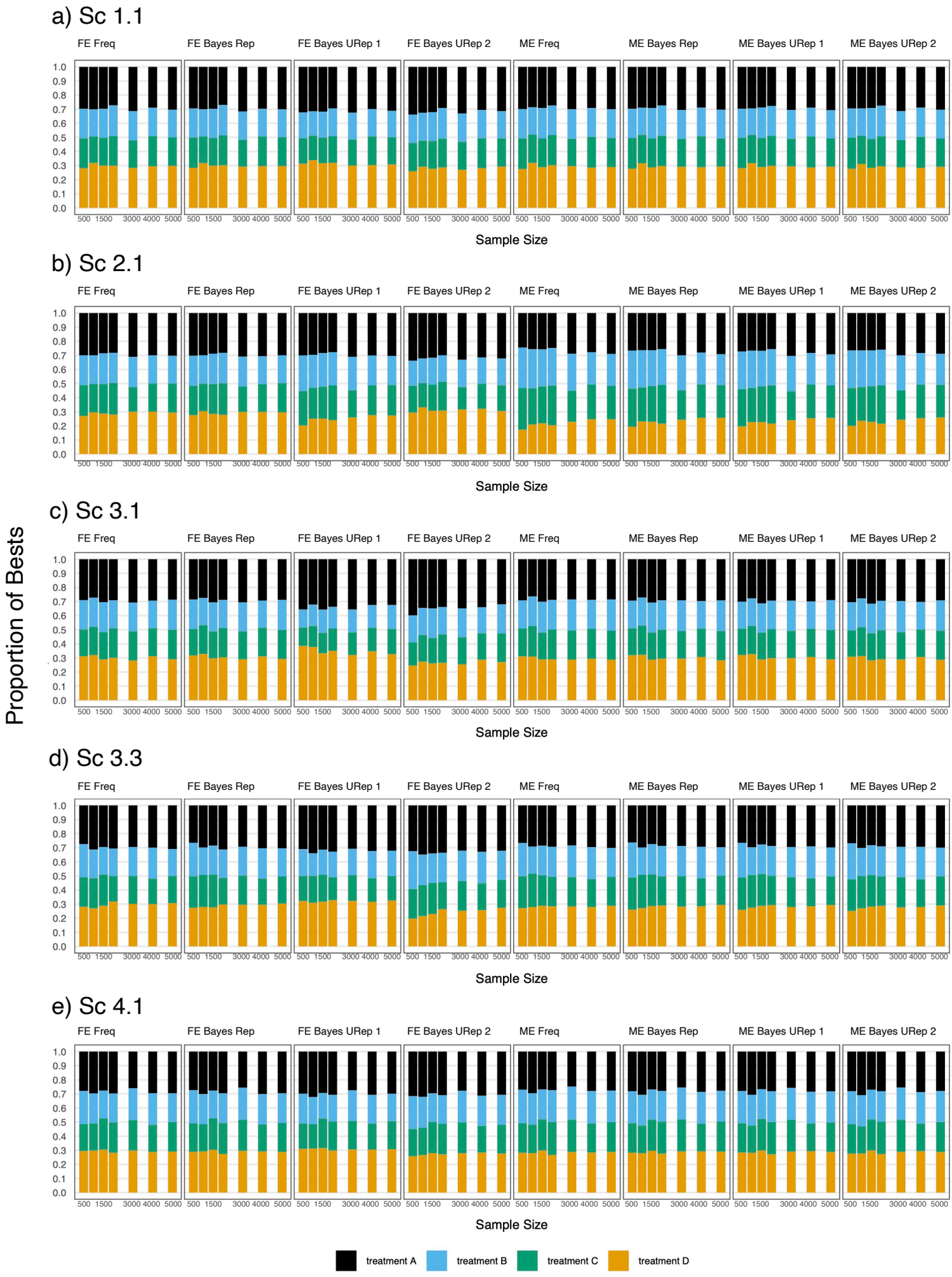

Fig. SM 2 Proportion of scenarios where each treatment (A–D) were incorrectly predicted as best for five null scenarios, when using a fixed and mixed effects using frequentist and Bayesian approaches for sample sizes ranging between N = 500–5,000.

Reduction in mortality ratio

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. Interestingly, it was FE Bayes unrepresentative 1 which performed best when the true effects of all treatments were different but the best and worst treatments were available in fewer patterns (scenarios 1.3 and 1.5). However, it performed worst when the best and worst treatments were available in more patterns (scenario 1.6). 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 and the FE Bayesian with unrepresentative prior 1 performed best.

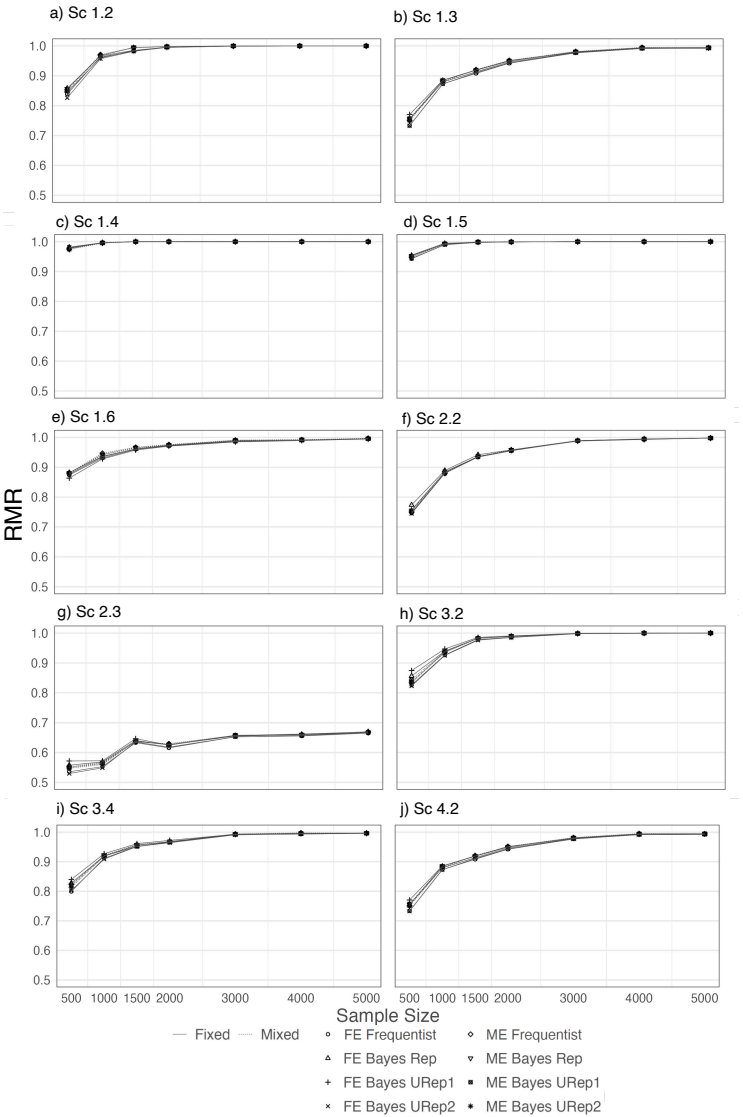

Fig. SM 3 Reduction in mortality ratio for ten scenarios, when using a fixed and mixed effects model using frequentist and Bayesian approaches for sample sizes ranging between N=500- 5,000.

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).

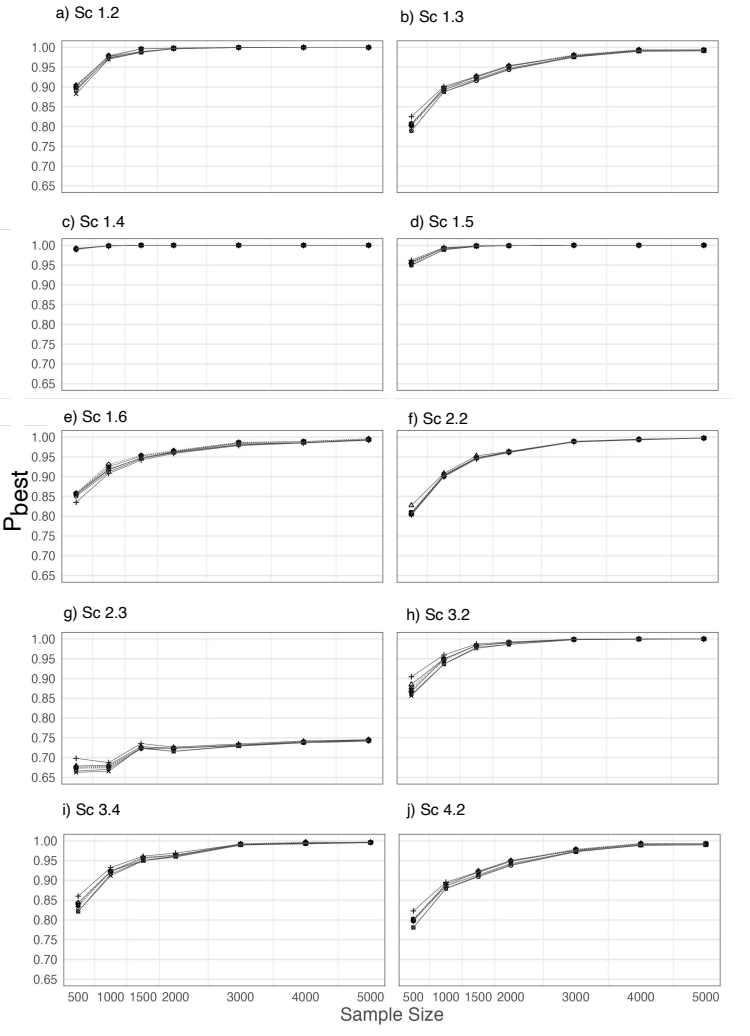

Fig. SM 4 Best treatment probability for all non-null scenarios, when using a fixed and mixed effects model using frequentist and Bayesian approaches for sample sizes ranging between N = 500-5,000.

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).

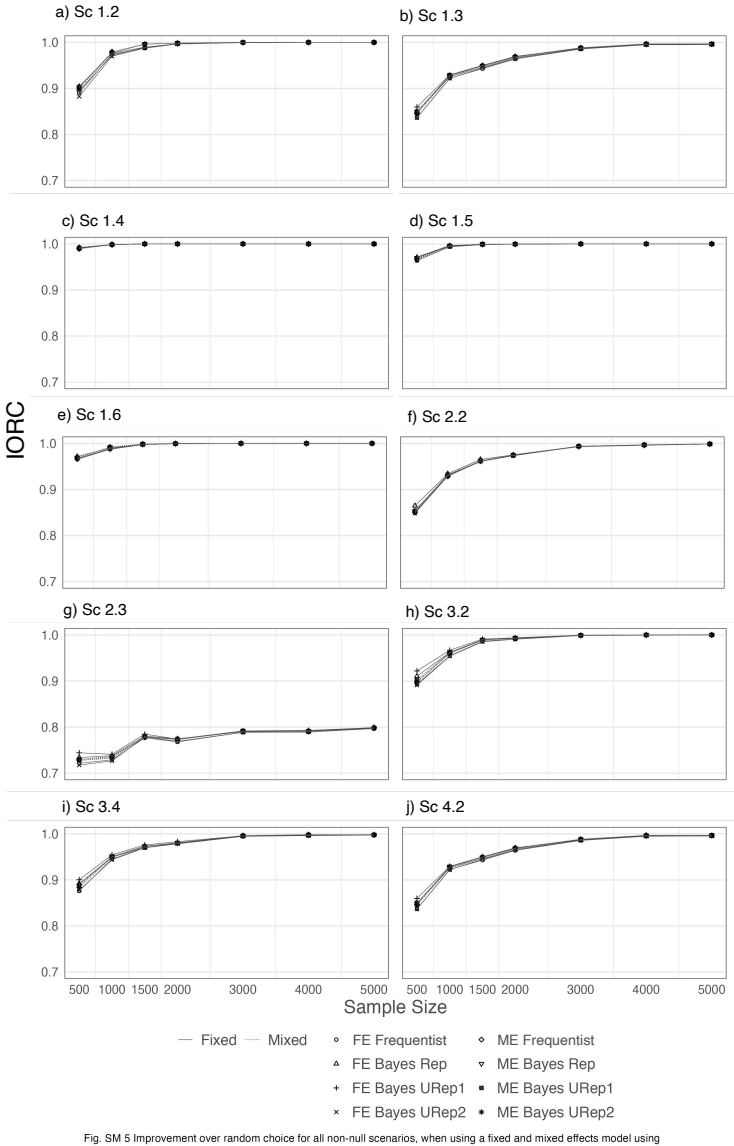

Fig. SM 5 Improvement over random choice for all non-null scenarios, when using a fixed and mixed effects model using frequentist and Bayesian approaches for sample sizes ranging between N = 500-5,000.

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).

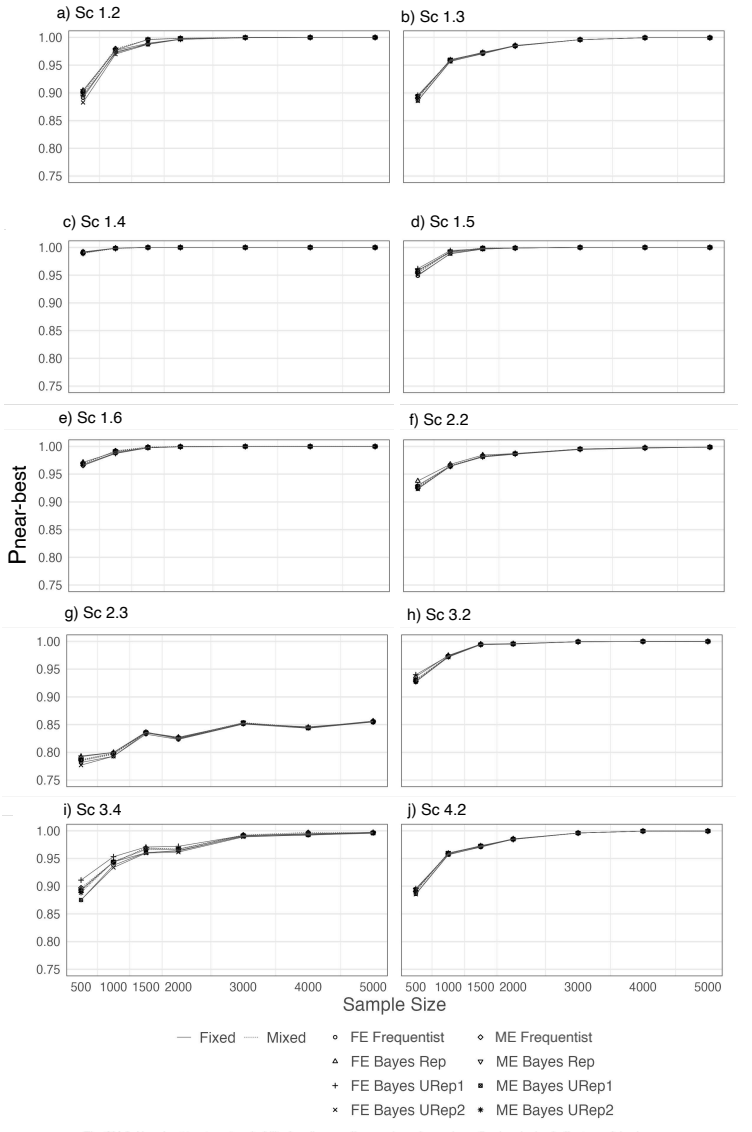

Fig. SM 6 Near-best treatment probability for all non-null scenarios, when using a fixed and mixed effects model using frequentist and Bayesian approaches for sample sizes ranging between N = 500-5,000.

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 best treatment(s) and the FE Bayes unrepresentative 1 was better at predicting the worst 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).

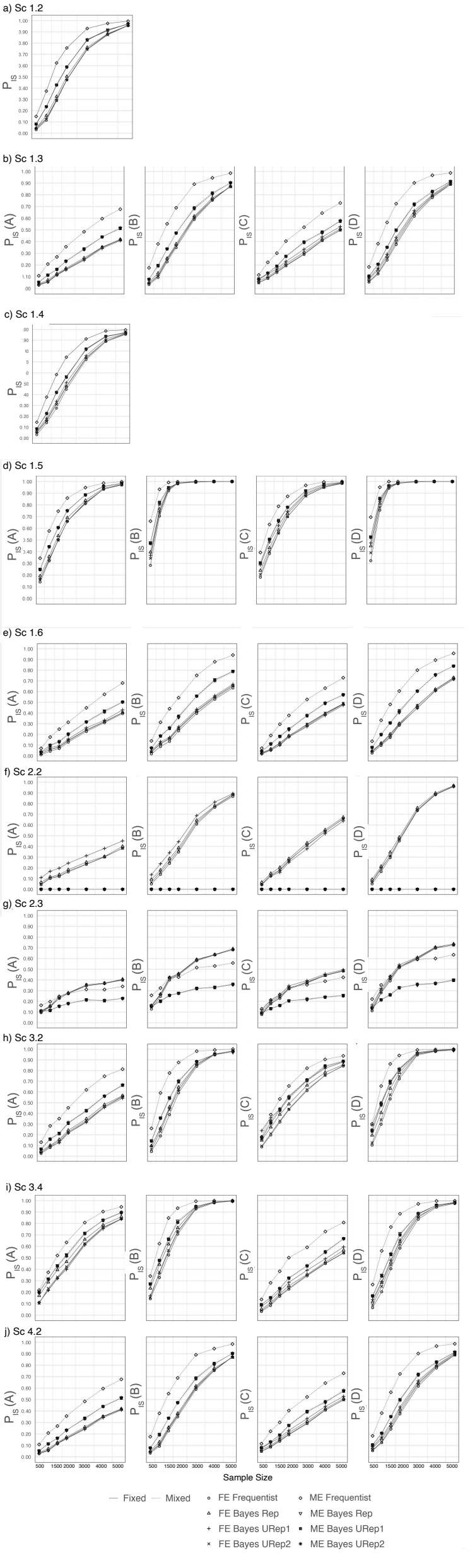

Fig. SM 7 Probability of interval separation for all non-null scenarios, when using a fixed and mixed effects model using frequentist and Bayesian approaches for sample sizes ranging between N = 500-5,000.