Supplement 2: Bayesian Analysis

Mortality Effect of Albumin Fluid Resuscitation in Adults with Septic Shock: A Systematic Review and Dual Frequentist–Bayesian Meta-Analysis of Randomised Trials

Authors:

Henrique Mendes, MD, MRCEM[1,2]

Henry Kinkei Wen, MD[1,2]

Harry de Souza, MD[1,2]

Yuanyuan Peng, MD[1,2]

Kush Deshpande, MD, FCICM[1,2]

Siebe G. Blok, MD, MSc[3]

Pieter R. Tuinman, MD, PhD[3,4]

Ary Serpa Neto, MD, PhD, FCICM[5,6,7,8]

Marcus J. Schultz, MD, PhD[9,10,11,12]

Manu L. N. G. Malbrain, MD, PhD[13,14]

Giacomo Grasselli, MD[15,16]

Frank M.P. van Haren, MD, PhD, FCICM[1,2,17,18]

[1]St George Hospital, Sydney, Australia [2]University of New South Wales, Sydney, Australia [3]Department of Intensive Care Medicine, Amsterdam UMC, Amsterdam, The Netherlands [4]Amsterdam Cardiovascular Sciences Research Institute, Amsterdam UMC, Amsterdam, The Netherlands [5]Australian and New Zealand Intensive Care Research Centre (ANZIC-RC), School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia [6]Department of Intensive Care, Austin Hospital, Melbourne, Australia [7]Department of Critical Care, Melbourne Medical School, University of Melbourne, Austin Hospital, Melbourne, Australia [8]Department of Critical Care Medicine, Hospital Israelita Albert Einstein, Sao Paulo, Brazil [9]Medical University Wien, Vienna, Austria [10]Cantonal Hospital St. Gallen, HOCH Health Ostschweiz, St. Gallen, Switzerland [11]University of Oxford, Oxford, UK [12]Mahidol University, Bangkok, Thailand [13]First Department of Anaesthesiology and Intensive Therapy, Medical University of Lublin, Jaczewskiego Lublin, Poland [14]International Fluid Academy, Lovenjoel, Belgium [15]Department of Emergency, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy [16]Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy [17]Australian National University, Canberra, Australia [18]The George Institute for Global Health, Sydney, Australia

Bayesian Analysis

• Bayesian methods

We conducted a Bayesian hierarchical meta-analysis in R (version 4.5.1) using brms (version 2.22.0) with the cmdstanr backend (cmdStan version 2.36.0). Outcomes were modelled at the arm level as the number of events out of the total number of patients in each arm using a binomial likelihood with logit link. To account for between-trial differences in baseline risk and heterogeneity in treatment effects, models included a random intercept for each trial and a random treatment effect nested within trial, specified as a correlated random-effects structure (Treatment | Study). The correlation between the study-specific intercepts and treatment effects was assigned an LKJ(2) prior.

Model specification

The full hierarchical model is specified as follows. For arm j of study i:

$$Y_{ij} \sim \text{Binomial}(n_{ij}, p_{ij})$$

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 \cdot \text{Treatment}_{ij} + u_{0i} + u_{1i} \cdot \text{Treatment}_{ij}$$

$$(u_{0i}, u_{1i}) \sim \text{MVN}(0, \Sigma)$$

$$\Sigma = \text{diag}(\sigma_0, \sigma_1) \cdot R \cdot \text{diag}(\sigma_0, \sigma_1)$$

where Y_{ij} is the observed event count, n_{ij} the total patients, p_{ij} the probability of the event, β_0 the population intercept, β_1 the pooled log-odds ratio for Treatment, u_{0i} and u_{1i} the study-specific random intercept and treatment-effect deviations, σ_0 and σ_1 their standard deviations, and R the correlation matrix. Priors were assigned as: $\beta_1 \sim$ Normal(μ, σ) as specified below for each prior family; σ_0, $\sigma_1 \sim$ half-Student-t(3, 0, 1); R ~ LKJ(2). The population intercept β_0 received the brms default (improper flat) prior, as it is nuisance. For the fixed-effect sensitivity model, the random-effects terms (u_{0i}, u_{1i}) were removed.

Priors for the pooled treatment effect (log-odds ratio; log-OR) followed four prespecified families. The primary analysis used a weakly informative Normal(0, 1.5) prior, chosen to allow data-driven estimation while preventing pathological posteriors. Sensitivity analyses used: (i) a sceptical Normal(0, 0.354) prior (approximately OR 0.50–2.00 within the central 95% prior interval), (ii) an optimistic Normal(−0.15, 0.28) prior (approximately OR 0.50–1.49, informed by ALBIOS septic shock subgroup [RR ≈ 0.87] and pooled hyperoncotic signal), and (iii) a pessimistic Normal(+0.15, 0.28) prior (approximately OR 0.67–2.01, symmetric counterpart allowing for harm from fluid loading or intervention complexity). The sceptical, optimistic, and pessimistic priors were pre-specified in the study protocol (PROSPERO CRD420261325998) to reflect varying degrees of equipoise based on preliminary evidence. Between-study heterogeneity was regularised with Student-t(3, 0, 1) priors on SD parameters, which are constrained to be non-negative and therefore imply half-Student-t(3, 0, 1) priors on the random-effects standard deviations for the study intercepts and treatment effects. As an additional structural sensitivity analysis, we refit the random-effects model under the weak

prior using an uncorrelated random-intercept–slope specification (1 + Treatment || Study), which assumes independence (i.e., zero correlation) between the study-level random intercepts and treatment effects, to assess robustness to the correlated random-effects structure.

eTable A: Prior specifications for the pooled treatment effect (β_1, log-OR scale)

The table below summarises the four prespecified prior families. The central 95% prior interval is computed as the 2.5th–97.5th percentiles of the prior distribution on the OR scale; the approximate RR range was derived by simulation under the fitted model's baseline risk.

Prior family	Distribution	95% prior interval (OR)	Approx. RR range	Rationale
Weakly informative (primary)	Normal(0, 1.5)	0.05–18.9	~0.22–4.5	Maximal agnosticism; regularises extremes
Sceptical	Normal(0, 0.354)	0.50–2.00	~0.70–1.43	Clinical equipoise; most fluid interventions show null effects
Optimistic	Normal(−0.15, 0.28)	0.50–1.49	~0.72–1.10	Informed by ALBIOS septic shock subgroup (RR ~0.87), set slightly toward null (prior centre RR ~0.89)
Pessimistic	Normal(+0.15, 0.28)	0.67–2.01	~0.91–1.39	Symmetric counterpart; allows for harm from fluid loading

The heterogeneity prior, half-Student-t(3, 0, 1), was held constant across all sensitivity analyses. This prior places approximately 82% of its mass below 1.0 on the log-OR SD scale, consistent with empirically observed heterogeneity in mortality meta-analyses of ICU interventions, while allowing heavier tails than a half-Normal to accommodate unexpectedly large between-study variation. Sensitivity to the heterogeneity prior was not formally assessed; this is acknowledged as a limitation, though the near-zero frequentist I^2 suggests the data are not strongly informative about heterogeneity magnitude.

Sampling used Hamiltonian Monte Carlo (No-U-Turn Sampler) with 4 chains, 20,000 iterations per chain and 4,000 warm-up iterations. The protocol pre-specified a minimum of 4,000 iterations per chain (2,000 warmup); iteration counts were increased to ensure stable posterior summaries and adequate effective sample sizes for tail probability estimates [P(RR < 0.95), P(RR < 0.90)], particularly given the challenging posterior geometry of arm-level binomial-logit models with correlated random effects on a small number of trials and very small studies (Rackow 1983, n = 11). Endpoint-specific tuning parameters were applied as follows: for the primary analysis (longest follow-up), adapt_delta = 0.9999, max_treedepth = 20, and initial step_size = 0.005; for 28-day models, adapt_delta = 0.999, max_treedepth = 20, and initial step_size = 0.01. Leave-one-out sensitivity analyses were performed by refitting the primary model (weak prior; correlated random effects) while sequentially excluding each trial. These leave-one-out refits used adapt_delta = 0.999, max_treedepth = 20, and initial step_size = 0.005.

Overlap between the optimistic prior and the likelihood. The optimistic prior (Normal(−0.15, 0.28) on the log-odds scale) was informed in part by the ALBIOS septic shock subgroup, which also contributes to the likelihood, so the same data inform both the prior location and the likelihood. Three features bound the practical effect of this partial non-independence. First, the prior is only mildly informative (SD 0.28, with roughly 30% of prior mass on no benefit) and was set milder than the ALBIOS effect itself (prior centre RR ≈ 0.89 versus ALBIOS RR ≈ 0.87), reflecting a blend of the ALBIOS subgroup and the broader hyperoncotic signal rather than a direct copy. Second, the optimistic-prior posterior (RR 0.919, P(RR < 0.95) = 75.9%) was almost identical to the primary weakly informative posterior (RR 0.920, P(RR < 0.95) = 73.9%), so the overlap shifted the reported quantities by at most about two percentage points. Third, the influence of ALBIOS is exerted through its likelihood weight rather than the prior: leave-one-out removal of ALBIOS from the likelihood under the primary prior moved P(RR < 0.95) from 73.9% to 47.9%, whereas its additional appearance in the optimistic prior changed the result negligibly. The optimistic-prior posterior is therefore reported only as a sensitivity bound, and the primary inference, based on the weakly informative Normal(0, 1.5) prior, is unaffected.

Software and reproducibility

All analyses were performed in R version 4.5.1 (R Foundation for Statistical Computing) using brms version 2.22.0 with the cmdstanr backend (cmdStan version 2.36.0). The global random seed was set to 2026 for all model fits to ensure reproducibility. The complete, self-contained R script is provided at the end of this supplement, enabling independent replication from raw data entry through to all reported results, diagnostics, and figures.

Convergence and sampling quality were assessed using the potential scale reduction factor (R-hat), effective sample sizes (bulk and tail), divergent transitions, treedepth saturation, and the Energy Bayesian Fraction of Missing Information (E-BFMI). Model calibration was evaluated with posterior predictive checks (PPCs), including density overlays, empirical CDF overlays, a two-dimensional statistic plot (mean vs SD) comparing observed to replicated datasets, and per-arm 90% posterior predictive intervals for observed event counts. In addition, we computed a posterior predictive p-value for the global log-risk-ratio, defined as the proportion of replicated datasets in which the replicated global log-risk-ratio was greater than or equal to the observed global log-risk-ratio. All PPC plots and PPC summary metrics were saved per endpoint alongside results and diagnostics outputs.

Posterior results were primarily reported on the RR scale as median risk ratios (RRs) with central 95% credible intervals (2.5th–97.5th percentiles), obtained via posterior marginal standardisation. For each posterior draw, we generated predicted risks under Control and under Experimental for all study arms using the inverse-logit of the linear predictor, averaged predicted risks across study arms within each treatment condition, and formed the RR as the ratio of the mean predicted risk under Experimental to the mean predicted risk under Control. We also reported the mean odds ratio (OR) with 95% credible interval from the population-level logit coefficient for Treatment for consistency with parallel frequentist analyses.

The posterior probability of benefit was summarised as follows, consistent with the co-primary Bayesian endpoint specified in the protocol: P(RR < 0.95) as the probability of clinically meaningful benefit (≥5% relative risk reduction); P(RR < 1) as the probability of any mortality benefit; and P(RR < 0.90) as the probability of a larger, unambiguous clinical effect (≥10% relative risk reduction). Posterior probabilities were interpreted along a graduated framework rather than against a single threshold: values above 0.95 were considered strong evidence, 0.90–0.95 moderate evidence, 0.80–0.90 suggestive evidence, and below 0.80 inconclusive, consistent with Bayesian interpretive frameworks applied to clinical trial evidence (Brophy, Trials 2020; Spiegelhalter, Abrams & Myles, 2004). This graduated interpretation was applied consistently across all prior specifications.

Between-study heterogeneity in treatment effects was summarised by the posterior distribution of τ, defined here as the posterior distribution of the random treatment-effect standard deviation (on the log-OR scale), reported with posterior median and central 95% credible interval. Prediction intervals (PIs) on the RR scale were obtained by simulating a single new-study random intercept and random treatment effect per posterior draw from the fitted hierarchical random-effects distribution (using the posterior draws of the random-effects standard deviations and their correlation), converting to risks via the inverse-logit, forming the RR for that simulated new study, and taking the 2.5th and 97.5th percentiles of the resulting simulated RR distribution. This PI summarises the dispersion of true effects expected in similar future settings. The posterior correlation between study intercepts and treatment effects was summarised as the posterior median correlation with central 95% credible interval.

The full posterior distribution for the pooled treatment effect was reported graphically using a combined cumulative distribution function (CDF) and posterior density plot, with annotated threshold probabilities at RR = 0.90, 0.95, and 1.0.

All analyses were performed using the R code provided in this supplement. The primary model specification, prior families, fixed-effect model, and Bayesian leave-one-out analyses were pre-specified in the study protocol (PROSPERO CRD420261325998); the uncorrelated random-effects (correlation constrained to zero) refit was performed as an additional structural sensitivity analysis as described above.

- Protocol compliance and deviations

All prespecified analytical components were executed as planned in the registered protocol (PROSPERO CRD420261325998, version 2.0, 26 February 2026). One deviation from the minimum computational specification was made: the protocol specified a minimum of 4,000 iterations per chain with 2,000 warmup; this was increased to 20,000 iterations per chain with 4,000 warmup to ensure adequate effective sample sizes for tail probability estimates [P(RR < 0.95), P(RR < 0.90)] given the challenging posterior geometry of arm-level binomial-logit models with correlated random effects on k = 7 trials including one very small study (Rackow 1983, n = 11). This deviation is conservative (increased precision, unchanged model) and does

not alter the analytical framework. No other deviations from the prespecified protocol were made.

• Input data

eTable B: Arm-level input data for the longest available follow-up mortality analysis (7 RCTs)

Study	Experimental Events	Experimental Total	Control Events	Control Total
Rackow et al. 1983	5	7	3	4
SAFE (Finfer et al.) 2004	70	209	90	229
EARSS (Charpentier et al.) 2011	96	399	105	399
ALBIOS (Caironi et al.) 2014	243	558	281	563
ICARUS-ED (Williams et al.) 2025	35	197	30	189
Cusack et al. 2025	15	50	12	50
ARISS (Sakr et al.) 2026	91	210	96	209

Follow-up timepoints: Rackow = study-period (hospital) mortality; SAFE = 28-day; EARSS = 28-day; ALBIOS = 90-day; ICARUS-ED = 30-day; Cusack = ICU mortality; ARISS = 90-day. Sources and extraction details are provided in Supplement 1.

eTable C: Arm-level input data for the 28-day mortality analysis (4 RCTs)

ALBIOS excluded (28-day shock subgroup not separately reported); Cusack excluded (in-hospital mortality only; 28-day event counts not reported); Rackow excluded (study-period mortality only; non-calendar-fixed endpoint).

Study	Experimental Events	Experimental Total	Control Events	Control Total
SAFE (Finfer et al.) 2004	70	209	90	229
EARSS (Charpentier et al.) 2011	96	399	105	399
ICARUS-ED (Williams et al.) 2025	35	197	30	189
ARISS (Sakr et al.) 2026	66	213	80	210

Follow-up timepoints: SAFE = 28-day; EARSS = 28-day; ICARUS-ED = 30-day; ARISS = 28-day. Note: ARISS denominators differ slightly from the 90-day analysis due to differing numbers lost to follow-up at each timepoint.

• Extended Results

eTable 1: Bayesian output of longest available follow-up mortality (primary outcome, 7 RCTs)

Scenario	OR (mean)	95% CrI (OR)	RR (median)	95% CrI (RR)	P(RR<1) (%)	P(RR>1) (%)	P(RR<0.95) (%)	P(RR<0.90) (%)
Main: RE weak prior	0.873	(0.715, 1.103)	0.920	(0.836, 1.020)	94.7%	5.3%	73.9%	32.5%
Sensitivity: RE sceptical	0.881	(0.730, 1.101)	0.924	(0.842, 1.021)	94.5%	5.5%	72.1%	28.9%
Sensitivity: RE optimistic	0.869	(0.724, 1.060)	0.919	(0.839, 1.011)	95.9%	4.1%	75.9%	32.5%
Sensitivity: RE pessimistic	0.902	(0.750, 1.143)	0.932	(0.849, 1.030)	92.2%	7.8%	65.6%	22.9%
Sensitivity: RE corr = 0	0.875	(0.718, 1.108)	0.922	(0.838, 1.024)	94.1%	5.9%	72.3%	30.5%
Sensitivity: Fixed-effect	0.859	(0.745, 0.991)	0.907	(0.828, 0.994)	98.2%	1.8%	84.1%	43.8%
LOO (drop Rackow 1983)	0.874	(0.715, 1.104)	0.913	(0.827, 1.012)	95.8%	4.2%	77.9%	39.0%
LOO (drop SAFE 2004)	0.901	(0.707, 1.229)	0.936	(0.842, 1.060)	87.3%	12.7%	60.6%	23.1%
LOO (drop EARSS 2011)	0.876	(0.680, 1.208)	0.924	(0.830, 1.045)	91.2%	8.8%	69.2%	31.6%
LOO (drop ALBIOS 2014)	0.928	(0.712, 1.262)	0.953	(0.839, 1.092)	77.0%	23.0%	47.9%	18.5%
LOO (drop ICARUS-ED 2025)	0.850	(0.681, 1.103)	0.909	(0.823, 1.014)	96.1%	3.9%	80.3%	41.6%
LOO (drop Cusack 2025)	0.856	(0.692, 1.083)	0.914	(0.832, 1.006)	96.8%	3.2%	79.4%	37.3%
LOO (drop ARISS 2026)	0.874	(0.682, 1.184)	0.919	(0.823, 1.042)	92.2%	7.8%	72.1%	35.8%

Across all prior specifications, the posterior distribution consistently favoured a reduction in longest follow-up mortality with albumin-containing resuscitation. Under the primary model (correlated random effects, weakly informative prior), the marginal-standardised risk ratio (RR) was 0.920 (95% CrI 0.836–1.020), with a 94.7% posterior probability of any benefit [P(RR<1)] and a 73.9% probability of clinically meaningful benefit [P(RR<0.95)]. While the credible interval includes 1.0, the high probability of benefit indicates the data provide suggestive-to-moderate evidence of a beneficial effect under the graduated interpretive framework. Sensitivity analyses confirmed robustness: the direction of effect remained unchanged under sceptical (RR 0.924; P[RR<1] = 94.5%), optimistic (RR 0.919; P[RR<1] = 95.9%), and pessimistic (RR 0.932; P[RR<1] = 92.2%) priors. The uncorrelated random-effects specification yielded nearly identical estimates (RR 0.922; P[RR<1] = 94.1%), and the fixed-effect model strengthened the posterior signal (RR 0.907; P[RR<1] = 98.2%; P[RR<0.95] = 84.1%), consistent with the near-zero heterogeneity observed in the frequentist analysis. Leave-one-out analyses showed the pooled estimate was not driven by any single trial, though excluding ALBIOS 2014 attenuated the estimate most substantially (RR 0.953; P[RR<1] = 77.0%), reflecting its status as

the largest contributing trial. Excluding ICARUS-ED 2025 (P[RR<1] = 96.1%) or Cusack 2025 (P[RR<1] = 96.8%) had minimal impact, confirming stability of the primary inference.

eTable 2: Bayesian output of 28-day mortality (or closest ≤30 days, 4 RCTs)

ALBIOS excluded (28-day shock subgroup not separately reported); Cusack excluded (in-hospital mortality only; 28-day event counts not reported); Rackow excluded (study-period mortality only; non-calendar-fixed endpoint).

Scenario	OR (mean)	95% CrI (OR)	RR (median)	95% CrI (RR)	P(RR<1) (%)	P(RR>1) (%)	P(RR<0.95) (%)	P(RR<0.90) (%)
Main: RE weak prior	0.855	(0.619, 1.211)	0.89	(0.774, 1.024)	94.9%	5.1%	82.2%	56.5%
Sensitivity: RE sceptical	0.872	(0.671, 1.163)	0.896	(0.782, 1.024)	94.4%	5.6%	80.4%	52.3%
Sensitivity: RE optimistic	0.853	(0.665, 1.100)	0.89	(0.781, 1.014)	96.0%	4.0%	83.4%	56.6%
Sensitivity: RE pessimistic	0.912	(0.714, 1.251)	0.91	(0.797, 1.037)	92.2%	7.8%	74.2%	43.6%
Sensitivity: RE corr = 0	0.854	(0.617, 1.209)	0.89	(0.774, 1.022)	95.2%	4.8%	82.4%	56.4%
Sensitivity: Fixed-effect	0.842	(0.693, 1.023)	0.883	(0.767, 1.017)	95.9%	4.2%	84.6%	60.4%
LOO (drop SAFE 2004)	0.889	(0.535, 1.566)	0.906	(0.766, 1.071)	87.6%	12.4%	71.3%	47.0%
LOO (drop EARSS 2011)	0.847	(0.498, 1.528)	0.879	(0.741, 1.040)	93.3%	6.7%	81.7%	60.6%
LOO (drop ICARUS-ED 2025)	0.812	(0.518, 1.320)	0.863	(0.746, 0.998)	97.7%	2.3%	90.3%	71.5%
LOO (drop ARISS 2026)	0.903	(0.560, 1.534)	0.92	(0.781, 1.080)	84.3%	15.7%	64.9%	39.4%

For 28-day mortality, the posterior estimates were directionally concordant with the primary longest follow-up analysis but with greater uncertainty, reflecting the smaller number of contributing trials (k = 4) and reduced sample size. Under the primary model, the RR was 0.89 (95% CrI 0.774–1.024), with a 94.9% probability of any benefit and 82.2% probability of clinically meaningful benefit [P(RR<0.95)]. These probabilities represent suggestive-to-moderate evidence under the graduated framework. Prior sensitivity was modest: the probability of benefit ranged from 92.2% (pessimistic) to 96.0% (optimistic). The fixed-effect model again yielded stronger posterior probabilities (P[RR<1] = 95.9%; P[RR<0.95] = 84.6%), consistent with the pattern observed in the primary analysis. Leave-one-out analyses demonstrated stability: the posterior was most sensitive to exclusion of ARISS 2026 (P[RR<1] dropping to 84.3%) and SAFE 2004 (P[RR<1] = 87.6%), and least affected by exclusion of ICARUS-ED 2025 (P[RR<1] = 97.7%). No single trial exclusion shifted the posterior median RR above 1.0. The wider credible intervals compared with the longest follow-up analysis are consistent with reduced statistical information rather than a materially different treatment effect.

• Diagnostics and posterior predictive checks

Model diagnostics and fit — Across fitted models, convergence was acceptable (R-hat values approximately 1.00 with adequate effective sample sizes across parameters; eTables 3–4). Divergent transitions were absent across all longest follow-up models and across all 28-day models except the leave-one-out refit excluding ARISS 2026, which had one divergent transition, indicating that the posterior was sampled without pathological geometry in all specifications. Treedepth saturation was not universal: it was minimal in several models but prominent in others (notably the fixed-effect models, consistent with reduced sampling efficiency for those specifications. Accordingly, primary inferences are supported by stable sampling behaviour, while interpretation of the fixed-effect and certain LOO refits should note the reduced sampling efficiency where treedepth saturation occurred (eTables 3–4).

Random-effects correlation — The posterior correlation between study baseline risk (random intercept) and treatment effect (random slope) was imprecisely estimated, with wide credible intervals spanning both negative and positive values for both endpoints (eTable 6). For longest follow-up, the posterior median correlation was −0.192 (95% CrI −0.874 to 0.726); for 28-day mortality, it was −0.182 (95% CrI −0.881 to 0.736). This pattern is consistent with weak identifiability of an intercept–slope correlation in a small set of trials, and implies that the pooled treatment effect is not strongly driven by any inferred "baseline-risk effect modification." In practical terms, allowing a correlated random-effects structure did not generate a clearly supported non-zero correlation; the pooled treatment effect should therefore be interpreted primarily through the marginal pooled RR estimates rather than the baseline-risk dependence (eTable 6).

eTable 3: Diagnostics summary for the longest follow-up mortality models

Model	Divergent Transitions	Treedepth Saturation	Min E-BFMI	Max Rhat	Min ESS (bulk)	Min ESS (tail)
RE weak prior	0	261	0.7945	1.0003	15,391	23,835
RE sceptical prior	0	1	0.7879	1.0002	14,758	24,291
RE optimistic prior	0	10	0.8053	1.0004	15,754	25,284
RE pessimistic prior	0	513	0.7892	1.0002	12,607	21,989
RE corr = 0 (weak prior)	0	241	0.7857	1.0003	13,969	22,694
FE model	0	9,815	0.9644	1.0002	21,498	25,278
LOO (drop Rackow 1983)	0	12	0.8075	1.0003	17,107	24,830
LOO (drop SAFE 2004)	0	39	0.8034	1.0006	15,387	21,697
LOO (drop EARSS 2011)	0	566	0.7885	1.0005	15,169	24,159
LOO (drop ALBIOS 2014)	0	7	0.7978	1.0002	15,191	23,079
LOO (drop ICARUS-ED 2025)	0	88	0.7722	1.0002	16,464	26,744
LOO (drop Cusack 2025)	0	1	0.8048	1.0003	15,403	23,659
LOO (drop ARISS 2026)	0	1	0.7926	1.0002	15,182	22,491

The longest follow-up models demonstrated acceptable sampling behaviour across all specifications, supporting the reliability of the reported posterior summaries. No divergent

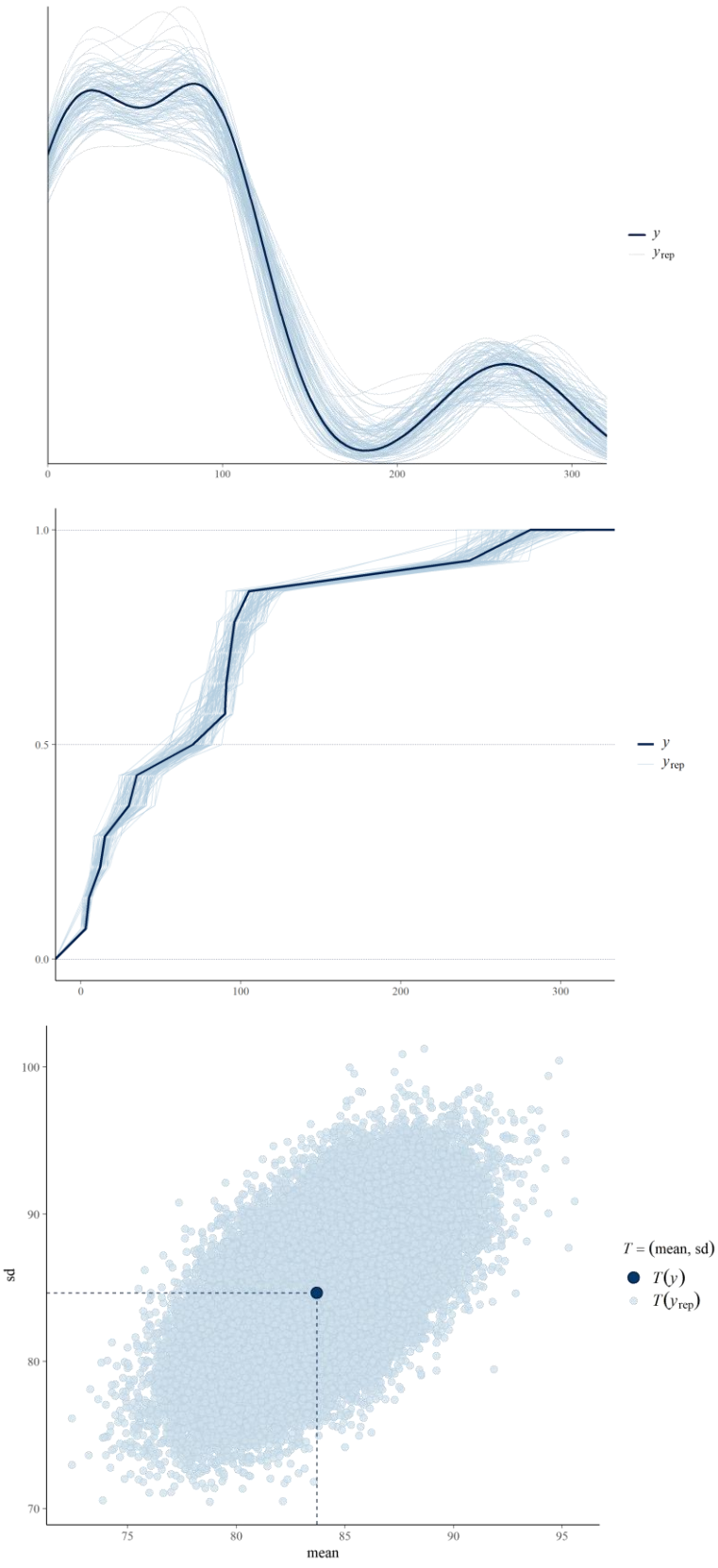
transitions were observed in any model. Where treedepth saturation is present (notably in the fixed-effect model with 9,815 saturated transitions and some LOO refits such as drop-EARSS with 566), it should be interpreted as reduced sampling efficiency rather than as invalidating the model—particularly when R-hat remains near 1.00 and effective sample sizes remain adequate (minimum bulk ESS > 12,600 across all models). E-BFMI values ranged from 0.77 to 0.96, all above the conventional 0.30 threshold. Overall, these diagnostics support the primary longest follow-up inference while identifying which sensitivity specifications are computationally more demanding and therefore should be interpreted with more caution (eTable 3).

eTable 4: Diagnostics summary for the 28-day mortality models

Model	Divergent Transitions	Treedepth Saturation	Min E-BFMI	Max Rhat	Min ESS (bulk)	Min ESS (tail)
RE weak prior	0	89	0.8163	1.0003	16,734	23,610
RE sceptical prior	0	98	0.8169	1.0003	17,144	22,876
RE optimistic prior	0	14	0.8263	1.0002	17,102	24,055
RE pessimistic prior	0	64	0.8019	1.0002	16,902	26,031
RE corr = 0 (weak prior)	0	112	0.8182	1.0003	16,148	20,624
FE model	0	24,114	0.9999	1.0002	24,347	27,980
LOO (drop SAFE 2004)	0	47	0.8311	1.0004	17,724	20,322
LOO (drop EARSS 2011)	0	142	0.8434	1.0002	17,714	20,811
LOO (drop ICARUS-ED 2025)	0	2,539	0.8173	1.0003	17,236	20,627
LOO (drop ARISS 2026)	1	17	0.8407	1.0002	16,630	19,692

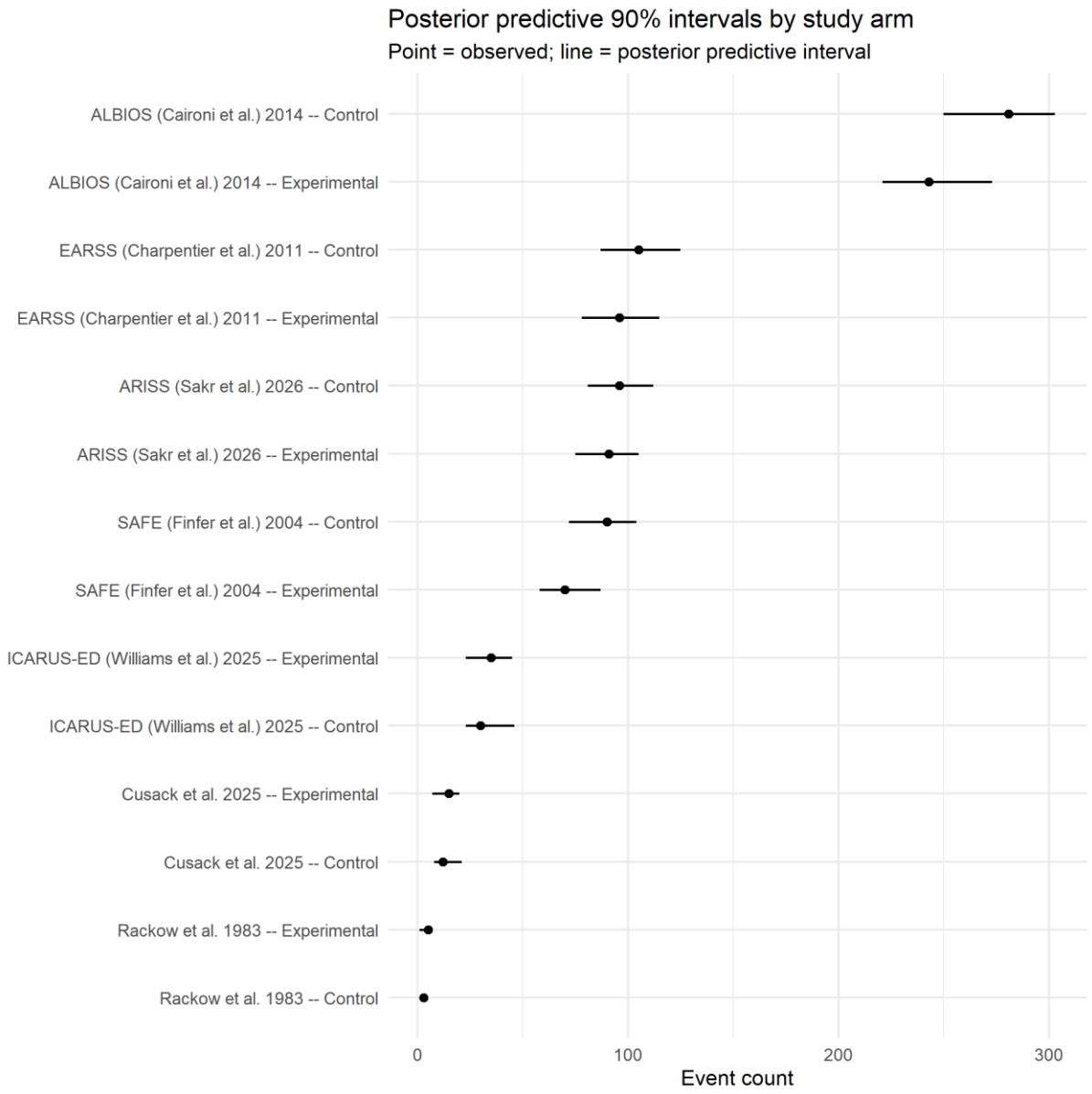
The 28-day models show acceptable convergence overall. No divergent transitions were observed across all specifications except the LOO refit excluding ARISS 2026, which had one divergent transition. Treedepth saturation was most prominent in the fixed-effect model (24,114) and the LOO refit excluding ICARUS-ED (2,539), reflecting more challenging posterior geometry for these specifications. R-hat values remained near 1.00 across all models (maximum 1.0004), and minimum effective sample sizes were adequate (bulk ESS > 16,100; tail ESS > 19,600). E-BFMI values ranged from 0.82 to 1.00. These diagnostics support the 28-day inference, with the caveat that the fixed-effect and certain LOO specifications exhibit reduced sampling efficiency (eTable 4).

eFigure 1: Visual PPC checks for the longest follow-up mortality models



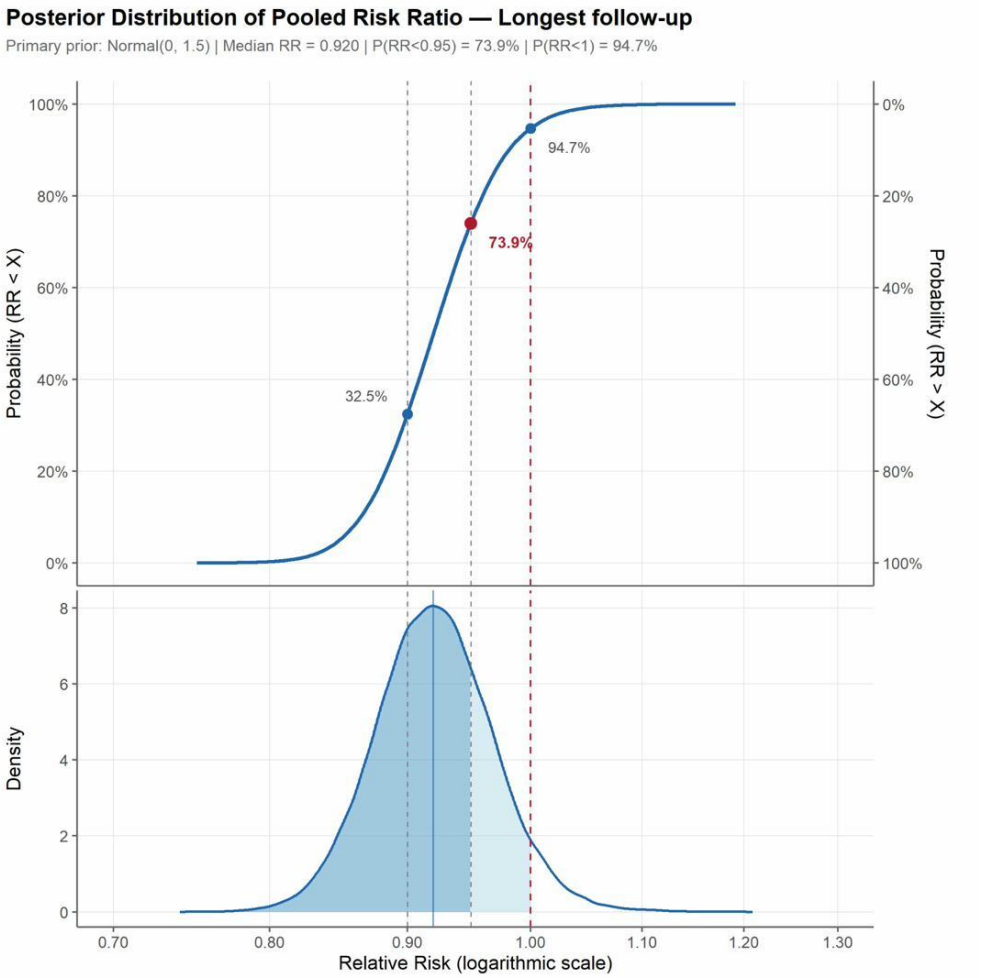
These posterior predictive checks compare the observed arm-level event-count distribution with replicated datasets generated from the fitted model (density and ECDF overlays), assess whether replicated summary statistics reproduce the observed mean–SD relationship, and display arm-level posterior predictive 90% intervals alongside observed counts. Visual agreement between observed and replicated summaries suggests no systematic lack of fit for the longest follow-up models on these diagnostics. Consistent with this, the posterior predictive p-value for the global log risk-ratio statistic is 0.487, indicating no evidence that the observed statistic is extreme under the fitted model (eFigure 1).

eFigure 2: Posterior predictive intervals (longest follow-up mortality)



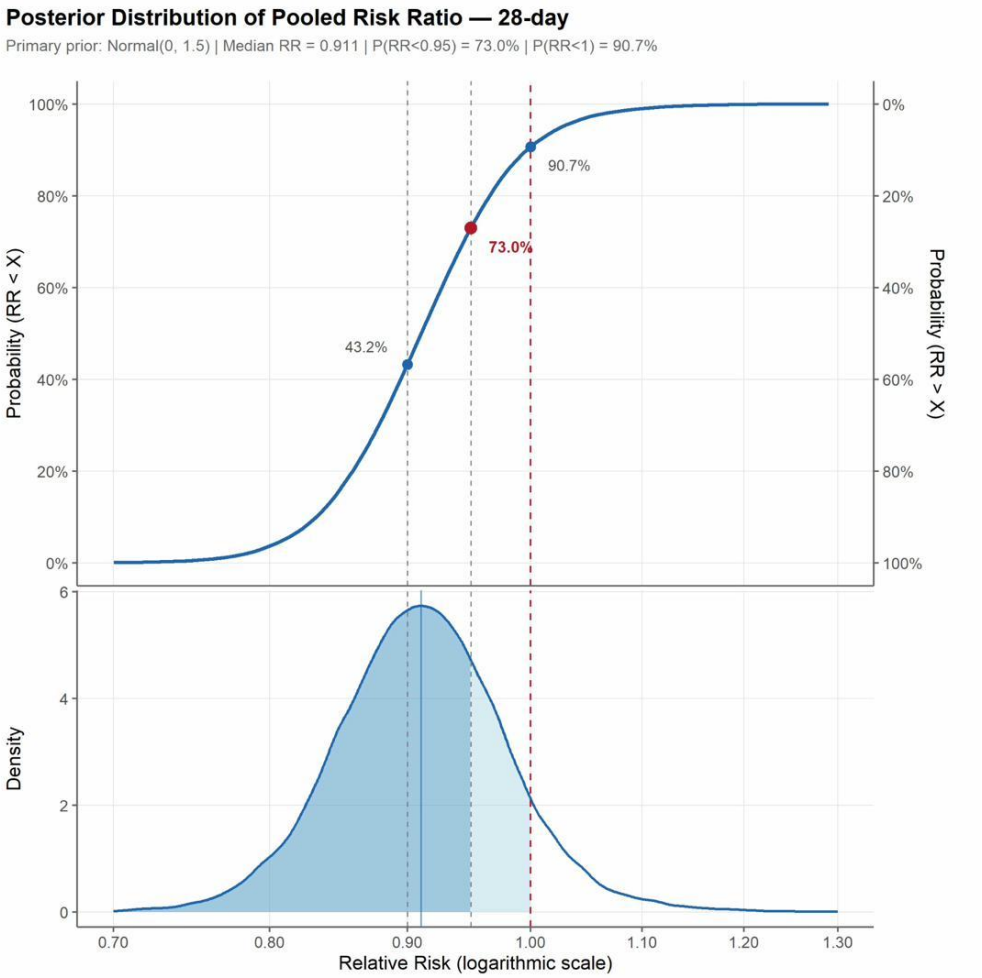
Arm-level posterior predictive 90% intervals generally cover the observed event counts (14/14 arms within intervals; 100% coverage), indicating that the model can plausibly generate the observed arm-level outcomes across studies. Wider intervals reflect smaller sample sizes and greater uncertainty in expected event counts. Any arms with observed counts near interval boundaries should be interpreted as localised tension rather than definitive misfit, particularly when global PPC summaries remain well calibrated (eFigures 1 and 2).

eFigure 3: Posterior cumulative distribution of pooled risk ratio (longest follow-up, primary prior)



The upper panel displays the cumulative distribution function of the posterior pooled risk ratio (RR) under the primary weakly informative prior, with the left y-axis showing P(RR < X) and the right y-axis showing P(RR > X). Annotated threshold probabilities are displayed at RR = 0.90, 0.95, and 1.0. The lower panel shows the corresponding posterior density, with shading distinguishing the regions corresponding to clinically meaningful benefit (RR < 0.95, darker) and any benefit (0.95 ≤ RR < 1.0, lighter). The x-axis is presented on a logarithmic scale (eFigure 3).

eFigure 4: Posterior cumulative distribution of pooled risk ratio (28-day mortality, primary prior)



The same posterior CDF and density display is shown for the 28-day mortality endpoint. The wider posterior distribution reflects the reduced information from fewer contributing trials (k = 4). Threshold probabilities at RR = 0.90, 0.95, and 1.0 are annotated on the CDF panel (eFigure 4).

eTable 5: Heterogeneity (τ) and prediction intervals (RR scale)

Endpoint	τ (median)	95% CrI (τ)	95% PI (RR)
Longest follow-up (up to 90 days)	0.102	0.005–0.447	0.705–1.222
28-day mortality (or closest ≤30 days)	0.141	0.006–0.735	0.559–1.409

Between-trial heterogeneity (tau) was estimated to be modest but with wide credible intervals, indicating that the data support anything from very low heterogeneity to meaningfully variable treatment effects across studies. For the longest follow-up analysis, τ = 0.102 (95% CrI 0.005–0.447); for 28-day mortality, τ = 0.141 (95% CrI 0.006–0.735). Importantly, the prediction intervals on the RR scale span both potential benefit and potential harm (longest follow-up: 0.705–1.222; 28-day: 0.559–1.409), implying that, if a new trial were conducted, the true effect could plausibly differ in direction from the pooled estimate. These results reinforce that inference should not rely only on the pooled RR, but should explicitly acknowledge uncertainty in the effect across settings (eTable 5).

eTable 6: Posterior correlation between study intercepts and treatment effects

Endpoint	Posterior median correlation	95% CrI
Longest follow-up (up to 90 days)	−0.192	−0.874 to 0.726
28-day mortality (or closest ≤30 days)	−0.182	−0.881 to 0.736

The posterior correlation between trial baseline risk and treatment effect was not well identified, with wide credible intervals spanning negative and positive correlations. This indicates insufficient evidence to conclude that higher- (or lower-) baseline-risk studies systematically experience larger treatment effects. Consequently, the main interpretation should emphasise the marginal pooled RR estimates, while treating any apparent intercept–slope correlation as exploratory and likely prior- and data-limited rather than as a stable, clinically actionable effect-modification signal (eTable 6).

• R Script for longest follow-up and 28-day mortality analysis

The complete R script is provided below. The script is self-contained and includes data entry, model specification, prior definitions, posterior extraction with marginal standardisation, leave-one-out sensitivity analyses, posterior predictive checks, posterior CDF/density plot generation, and all diagnostics. Key specifications: arm-level binomial-logit model via brms/cmdstanr; 4 chains × 20,000 iterations (4,000 warmup); adapt_delta ≥ 0.999 (0.9999 for longest follow-up); max_treedepth = 20; priors as specified in the study protocol (PROSPERO CRD420261325998).

```
# Load required packages
suppressPackageStartupMessages({
  library(brms)
  library(cmdstanr)
  library(dplyr)
  library(tidyr)
  library(posterior)
  library(ggplot2)
  library(bayesplot)
})

# Check cmdstan installation
stopifnot(!identical(cmdstanr::cmdstan_version(), FALSE))

# ==================== DATA ====================
# Primary outcome: longest available follow-up (up to 90 days), 7 RCTs
# Rackow 1983:    Hospital mortality (longest FU). Septic shock subgroup only (albumin vs saline).
# SAFE 2004:      28-day (only available for shock subgroup).
# EARSS 2011:     28-day. Abstract only; n per arm estimated from 798 randomised 1:1.
# ALBIOS 2014:    90-day. Septic shock subgroup.
# ICARUS-ED 2025: 30-day. Refractory hypotension subgroup.
# Cusack 2025:    ICU mortality (longest FU).
# ARISS 2026:     90-day. Septic shock (Sepsis-3).

data_longest <- data.frame(
  Study = c(
    "Rackow et al. 1983",
    "SAFE (Finfer et al.) 2004",
    "EARSS (Charpentier et al.) 2011",
    "ALBIOS (Caironi et al.) 2014",
    "ICARUS-ED (Williams et al.) 2025",
    "Cusack et al. 2025",
    "ARISS (Sakr et al.) 2026"
  ),
  `Experimental Events` = c(5, 70, 96, 243, 35, 15, 91),
  `Experimental Total`  = c(7, 209, 399, 558, 197, 50, 210),
  `Control Events`      = c(3, 90, 105, 281, 30, 12, 96),
  `Control Total`       = c(4, 229, 399, 563, 189, 50, 209),
  check.names = FALSE
)

# Sensitivity: 28-day mortality (or closest <= 30 days), 4 RCTs
# Excludes Rackow (study-period / non-calendar-fixed endpoint),
# ALBIOS (28-day shock subgroup not separately reported), and
# Cusack (28-day event counts not separately reported; only ICU survival).
# ICARUS-ED contributes 30-day (closest <= 30 days).
# ARISS denominators differ slightly from 90-day due to fewer lost to follow-up.

data_28d <- data.frame(
  Study = c(
    "SAFE (Finfer et al.) 2004",
    "EARSS (Charpentier et al.) 2011",
    "ICARUS-ED (Williams et al.) 2025",
    "ARISS (Sakr et al.) 2026"
  ),
  `Experimental Events` = c(70, 96, 35, 66),
  `Experimental Total`  = c(209, 399, 197, 213),
  `Control Events`      = c(90, 105, 30, 80),
  `Control Total`       = c(229, 399, 189, 210),
  check.names = FALSE
)

# ==================== HELPER FUNCTIONS ====================
```

```
prepare_brms_data <- function(study_data) {
  study_data %>%
    pivot_longer(
      cols = matches("^(Experimental|Control)"),
      names_to = c("Treatment", ".value"),
      names_pattern = "(Experimental|Control) (Events|Total)"
    ) %>%
    mutate(
      Study = factor(Study),
      Treatment = factor(Treatment, levels = c("Control", "Experimental"))
    )
}

get_results_row <- function(model, label) {
  draws <- posterior::as_draws_df(model)
  log_or <- draws$b_TreatmentExperimental
  or_mean <- exp(mean(log_or))
  or_ci <- quantile(exp(log_or), probs = c(0.025, 0.975))

  # Marginal standardisation for RR
  model_data <- model$data
  new_control <- model_data
  new_control$Treatment <- factor("Control", levels = levels(model_data$Treatment))
  new_experimental <- model_data
  new_experimental$Treatment <- factor("Experimental", levels = levels(model_data$Treatment))

  p_control <- posterior_linpred(model, newdata = new_control, transform = TRUE, allow_new_levels = TRUE)
  p_experimental <- posterior_linpred(model, newdata = new_experimental, transform = TRUE, allow_new_levels = TRUE)

  if (is.matrix(p_control)) {
    p_control_means <- rowMeans(p_control)
    p_experimental_means <- rowMeans(p_experimental)
  } else {
    p_control_means <- p_control
    p_experimental_means <- p_experimental
  }

  rr_draws <- p_experimental_means / p_control_means
  rr_med <- median(rr_draws)
  rr_ci <- quantile(rr_draws, probs = c(0.025, 0.975))

  # Protocol-specified posterior probability thresholds
  p_rr_lt_1   <- mean(rr_draws < 1)
  p_rr_gt_1   <- mean(rr_draws > 1)
  p_rr_lt_095 <- mean(rr_draws < 0.95)
  p_rr_lt_090 <- mean(rr_draws < 0.90)

  data.frame(
    Scenario        = label,
    `OR (mean)`     = round(or_mean, 3),
    `95% CrI (OR)`  = sprintf("(%.3f, %.3f)", or_ci[1], or_ci[2]),
    `RR (median)`   = round(rr_med, 3),
    `95% CrI (RR)`  = sprintf("(%.3f, %.3f)", rr_ci[1], rr_ci[2]),
    `P(RR<1) (%)`   = sprintf("%.1f%%", p_rr_lt_1 * 100),
    `P(RR>1) (%)`   = sprintf("%.1f%%", p_rr_gt_1 * 100),
    `P(RR<0.95) (%)` = sprintf("%.1f%%", p_rr_lt_095 * 100),
    `P(RR<0.90) (%)` = sprintf("%.1f%%", p_rr_lt_090 * 100),
    check.names = FALSE
  )
}

get_tau_and_pi <- function(model, model_name, endpoint_name) {
  draws <- as_draws_df(model)

  # tau: SD of random treatment effect
  tau_draws <- draws$sd_Study__TreatmentExperimental
  tau_median <- median(tau_draws)
  tau_cri <- quantile(tau_draws, probs = c(0.025, 0.975))

  # Prediction interval for RR in a new study
  # Seed the PI simulation for reproducibility across reruns
  inv_logit <- function(x) 1 / (1 + exp(-x))
  b0 <- draws$b_Intercept
  b1 <- draws$b_TreatmentExperimental
  sd0 <- draws$sd_Study__Intercept
```

```
  sd1 <- draws$sd_Study__TreatmentExperimental
  rho <- draws$cor_Study__Intercept__TreatmentExperimental

  if (is.null(rho)) {
    rho <- rep(0, length(b0))
  }

  set.seed(42)
  z0 <- rnorm(length(b0))
  z1 <- rnorm(length(b0))
  u0 <- sd0 * z0
  u1 <- sd1 * (rho * z0 + sqrt(pmax(0, 1 - rho^2)) * z1)

  p_control <- inv_logit(b0 + u0)
  p_experimental <- inv_logit(b0 + b1 + u0 + u1)
  rr_new <- p_experimental / p_control
  pred_interval_rr <- quantile(rr_new, probs = c(0.025, 0.975))

  tau_pi_df <- data.frame(
    Model = model_name,
    Endpoint = endpoint_name,
    Tau_Median = round(tau_median, 3),
    Tau_95CrI_Lower = round(tau_cri[1], 3),
    Tau_95CrI_Upper = round(tau_cri[2], 3),
    PI_RR_Lower = round(pred_interval_rr[1], 3),
    PI_RR_Upper = round(pred_interval_rr[2], 3),
    stringsAsFactors = FALSE
  )

  # Overwrite (not append) for reproducibility
  filename <- paste0("tau_pi_", gsub("[- ]", "", endpoint_name), ".csv")
  write.csv(tau_pi_df, filename, row.names = FALSE)

  cat("\n---", model_name, "---\n")
  cat("Heterogeneity (tau) - Posterior Median & 95% CrI:\n")
  cat("  Median:", round(tau_median, 3), "\n")
  cat("  95% CrI: [", round(tau_cri[1], 3), ", ", round(tau_cri[2], 3), "]\n\n", sep = "")
  cat("95% Prediction Interval for Risk Ratio in a new study (RR scale):\n")
  cat("  PI: [", round(pred_interval_rr[1], 3), ", ", round(pred_interval_rr[2], 3), "]\n", sep = "")
  invisible(tau_pi_df)
}

get_correlation <- function(model, model_name, endpoint_name) {
  draws <- as_draws_df(model)
  cor_draws <- draws$cor_Study__Intercept__TreatmentExperimental

  if (is.null(cor_draws)) {
    cor_median <- 0
    cor_cri <- c(0, 0)
    interpretation <- "Model specified without correlation parameter (corr=0)."
  } else {
    cor_median <- median(cor_draws)
    cor_cri <- quantile(cor_draws, probs = c(0.025, 0.975))

    interpretation <- if (cor_cri[1] > 0.1) {
      "Substantial positive correlation likely."
    } else if (cor_cri[2] < -0.1) {
      "Substantial negative correlation likely."
    } else {
      "Correlation is not substantial (CrI includes zero and is near null)."
    }
  }

  cor_df <- data.frame(
    Model = model_name,
    Endpoint = endpoint_name,
    Cor_Median = round(cor_median, 3),
    Cor_95CrI_Lower = round(cor_cri[1], 3),
    Cor_95CrI_Upper = round(cor_cri[2], 3),
    Interpretation = interpretation,
    stringsAsFactors = FALSE
  )

  # Overwrite (not append) for reproducibility
  filename <- paste0("correlation_", gsub("[- ]", "", endpoint_name), ".csv")
```

```
  write.csv(cor_df, filename, row.names = FALSE)

  cat("\n---", model_name, "---\n")
  cat("Correlation (Intercept vs. Treatment Effect) - Posterior Median & 95% CrI:\n")
  cat("  Median:", round(cor_median, 3), "\n")
  cat("  95% CrI: [", round(cor_cri[1], 3), ", ", round(cor_cri[2], 3), "]\n", sep = "")
  cat("  Interpretation:", interpretation, "\n")
  invisible(cor_df)
}

# Backend-agnostic diagnostics
get_diagnostics_row <- function(brmsfit, label) {
  f <- brmsfit$fit
  if (inherits(f, "CmdStanMCMC")) {
    sdiag <- f$sampler_diagnostics()
    vnames <- dimnames(sdiag)[[3]]; idx <- function(nm) match(nm, vnames)

    div_trans <- sum(sdiag[, , idx("divergent__")], na.rm = TRUE)
    td_arr <- sdiag[, , idx("treedepth__"), drop = FALSE]
    max_td <- max(td_arr, na.rm = TRUE)
    td_sat <- sum(td_arr == max_td, na.rm = TRUE)
    ebfmi <- suppressWarnings(min(cmdstanr::nuts_energy_information(f)$E_BFMI, na.rm = TRUE))
  } else if (inherits(f, "stanfit")) {
    sp <- rstan::get_sampler_params(f, inc_warmup = FALSE)
    div_trans <- sum(vapply(sp, function(x) sum(x[, "divergent__"]), numeric(1)))
    max_td <- max(vapply(sp, function(x) max(x[, "treedepth__"], na.rm = TRUE), numeric(1)), na.rm = TRUE)
    td_sat <- sum(vapply(sp, function(x) sum(x[, "treedepth__"] >= max_td), numeric(1)))
    ebfmi <- suppressWarnings(min(rstan::get_bfmi(f), na.rm = TRUE))
  } else stop("Unknown backend in brmsfit$fit (expected CmdStanMCMC or stanfit).")

  summ <- posterior::summarise_draws(posterior::as_draws_df(brmsfit))

  data.frame(
    Model = label,
    `Divergent Transitions` = div_trans,
    `Treedepth Saturation` = td_sat,
    `Min E-BFMI` = round(ebfmi, 4),
    `Max Rhat` = round(max(summ$rhat, na.rm = TRUE), 4),
    `Min ESS (bulk)` = suppressWarnings(min(summ$ess_bulk, na.rm = TRUE)),
    `Min ESS (tail)` = suppressWarnings(min(summ$ess_tail, na.rm = TRUE)),
    check.names = FALSE
  )
}

ppp_lr_ratio <- function(fit, dat, ndraws = 1000) {
  y_rep <- posterior_predict(fit, ndraws = ndraws)
  obs_eE <- sum(dat$Events[dat$Treatment == "Experimental"]); obs_tE <- sum(dat$Total[dat$Treatment == "Experimental"])
  obs_eC <- sum(dat$Events[dat$Treatment == "Control"]); obs_tC <- sum(dat$Total[dat$Treatment == "Control"])
  # Continuity correction (0.5) to avoid log(0) if an arm has zero events
  obs_log_rr <- log(((obs_eE + 0.5) / (obs_tE + 0.5)) / ((obs_eC + 0.5) / (obs_tC + 0.5)))

  rep_log_rr <- numeric(ndraws)
  for (i in 1:ndraws) {
    rep_eE <- sum(y_rep[i, dat$Treatment == "Experimental"]); rep_tE <- obs_tE
    rep_eC <- sum(y_rep[i, dat$Treatment == "Control"]); rep_tC <- obs_tC
    rep_log_rr[i] <- log(((rep_eE + 0.5) / (rep_tE + 0.5)) / ((rep_eC + 0.5) / (rep_tC + 0.5)))
  }
  data.frame(Metric = "Global log-risk-ratio", PPP = round(mean(rep_log_rr >= obs_log_rr), 3))
}

run_ppcs <- function(fit, dat, outdir, prefix, ndraws_overlay = 100, prob_interval = 0.90) {
  dir.create(outdir, showWarnings = FALSE, recursive = TRUE)
  old_opt <- getOption("bayesplot_ppc_ndraws", NULL)
  options(bayesplot_ppc_ndraws = ndraws_overlay)
  on.exit(options(bayesplot_ppc_ndraws = old_opt), add = TRUE)

  p <- pp_check(fit, type = "dens_overlay", ndraws = ndraws_overlay)
  ggsave(file.path(outdir, sprintf("%s_ppc_dens_overlay.png", prefix)), p, width = 8, height = 6, dpi = 300)

  p <- pp_check(fit, type = "ecdf_overlay", ndraws = ndraws_overlay)
  ggsave(file.path(outdir, sprintf("%s_ppc_ecdf_overlay.png", prefix)), p, width = 8, height = 6, dpi = 300)

  p <- pp_check(fit, type = "stat_2d", stat = c("mean", "sd"))
  ggsave(file.path(outdir, sprintf("%s_ppc_stat2d_mean_sd.png", prefix)), p, width = 8, height = 6, dpi = 300)
```

```
 pp <- posterior_predict(fit)
 qs <- apply(pp, 2, function(x) quantile(x, probs = c((1 - prob_interval) / 2, 1 - (1 - prob_interval) / 2)))

 ppc_arm <- dat
 ppc_arm$ppc_lo <- qs[1, ]; ppc_arm$ppc_hi <- qs[2, ]
 ppc_arm$within <- as.integer(ppc_arm$Events >= ppc_arm$ppc_lo & ppc_arm$Events <= ppc_arm$ppc_hi)
 ppc_arm$Arm <- paste(ppc_arm$Study, ppc_arm$Treatment, sep = " -- ")

 p <- ggplot(ppc_arm, aes(x = reorder(Arm, Events), y = Events)) +
  geom_linerange(aes(ymin = ppc_lo, ymax = ppc_hi), linewidth = 0.6) +
  geom_point(size = 1.8) + coord_flip() +
  labs(x = NULL, y = "Event count",
       title = sprintf("Posterior predictive %d%% intervals by study arm", round(prob_interval * 100)),
       subtitle = "Point = observed; line = posterior predictive interval") +
  theme_minimal(base_size = 11)

 ggsave(file.path(outdir, sprintf("%s_ppc_intervals_by_arm.png", prefix)), p, width = 8, height = 8, dpi = 300)

 ppp_lr_tab <- ppp_lr_ratio(fit, dat)
 ppc_intervals_tab <- data.frame(
  `Total Observations` = nrow(dat),
  `Observed Rates within Interval` = sum(ppc_arm$within),
  `Percent within Interval` = round(mean(ppc_arm$within) * 100, 2),
  check.names = FALSE
 )

 ppc_metrics <- data.frame(
  Metric = c(ppp_lr_tab$Metric, "Coverage (per-arm)"),
  Value = c(ppp_lr_tab$PPP,
         sprintf("%d/%d (%.2f%%)",
                 ppc_intervals_tab$`Observed Rates within Interval`,
                 ppc_intervals_tab$`Total Observations`,
                 ppc_intervals_tab$`Percent within Interval`)),
  stringsAsFactors = FALSE
 )

 write.csv(ppc_metrics, file.path(outdir, sprintf("%s_ppc_metrics.csv", prefix)), row.names = FALSE)
 message("\nPosterior predictive metrics (", prefix, "):"); print(ppc_metrics)
 invisible(list(ppc_metrics = ppc_metrics, per_arm = ppc_arm))
}

# ==================== POSTERIOR CDF PLOT ====================
# Generates a two-panel figure:
#   Top:    Cumulative distribution function P(RR < X) with annotated
#           thresholds at RR = 0.90, 0.95, and 1.0
#   Bottom: Posterior density of RR with shaded regions
#
# Modelled on the uploaded reference image (CDF + density, log-scale x-axis).

generate_posterior_cdf_plot <- function(model, endpoint_label, outdir = ".", prefix = "cdf") {

 dir.create(outdir, showWarnings = FALSE, recursive = TRUE)

 # Extract marginal-standardised RR draws
 model_data <- model$data
 new_control <- model_data
 new_control$Treatment <- factor("Control", levels = levels(model_data$Treatment))
 new_experimental <- model_data
 new_experimental$Treatment <- factor("Experimental", levels = levels(model_data$Treatment))

 p_control <- posterior_linpred(model, newdata = new_control, transform = TRUE, allow_new_levels = TRUE)
 p_experimental <- posterior_linpred(model, newdata = new_experimental, transform = TRUE, allow_new_levels = TRUE)

 if (is.matrix(p_control)) {
  p_control_means <- rowMeans(p_control)
  p_experimental_means <- rowMeans(p_experimental)
 } else {
  p_control_means <- p_control
  p_experimental_means <- p_experimental
 }

 rr_draws <- p_experimental_means / p_control_means

 # Key posterior probabilities
 p_lt_090 <- mean(rr_draws < 0.90)
```

```
p_lt_095 <- mean(rr_draws < 0.95)
p_lt_100 <- mean(rr_draws < 1.00)
rr_median <- median(rr_draws)

# Build data for plotting on log-RR scale
log_rr <- log(rr_draws)
dens <- density(log_rr, n = 1024)

# x-axis range (RR scale)
xmin <- 0.70; xmax <- 1.30
log_xmin <- log(xmin); log_xmax <- log(xmax)

# CDF: sort RR draws and compute empirical CDF
rr_sorted <- sort(rr_draws)
ecdf_y <- seq_along(rr_sorted) / length(rr_sorted)

# Restrict to plotting range
in_range <- rr_sorted >= xmin & rr_sorted <= xmax
rr_plot <- rr_sorted[in_range]
ecdf_plot <- ecdf_y[in_range]

# Density data restricted to range
dens_in <- dens$x >= log_xmin & dens$x <= log_xmax
dens_x_rr <- exp(dens$x[dens_in])
dens_y <- dens$y[dens_in]

# ---- Build the plot ----

# Custom theme
theme_cdf <- theme_minimal(base_size = 11) +
  theme(
    panel.grid.minor = element_blank(),
    panel.grid.major = element_line(colour = "grey90", linewidth = 0.3),
    axis.line = element_line(colour = "grey40"),
    axis.ticks = element_line(colour = "grey40"),
    plot.title = element_text(size = 12, face = "bold"),
    plot.subtitle = element_text(size = 9, colour = "grey40")
  )

# --- Top panel: CDF ---
df_cdf <- data.frame(RR = rr_plot, CDF = ecdf_plot)

p_cdf <- ggplot(df_cdf, aes(x = RR, y = CDF)) +
  geom_line(colour = "#2166AC", linewidth = 1.0) +
  # Reference lines at RR = 0.90, 0.95, 1.0
  geom_vline(xintercept = 0.90, linetype = "dashed", colour = "grey50", linewidth = 0.4) +
  geom_vline(xintercept = 0.95, linetype = "dashed", colour = "grey50", linewidth = 0.4) +
  geom_vline(xintercept = 1.00, linetype = "dashed", colour = "#B2182B", linewidth = 0.5) +
  # Annotated points
  annotate("point", x = 0.90, y = p_lt_090, colour = "#2166AC", size = 2.5) +
  annotate("point", x = 0.95, y = p_lt_095, colour = "#B2182B", size = 3) +
  annotate("point", x = 1.00, y = p_lt_100, colour = "#2166AC", size = 2.5) +
  # Labels
  annotate("text", x = 0.95 + 0.015, y = p_lt_095 - 0.04,
           label = sprintf("%.1f%%", p_lt_095 * 100),
           hjust = 0, size = 3.2, fontface = "bold", colour = "#B2182B") +
  annotate("text", x = 1.00 + 0.015, y = p_lt_100 - 0.04,
           label = sprintf("%.1f%%", p_lt_100 * 100),
           hjust = 0, size = 3.0, colour = "grey30") +
  annotate("text", x = 0.90 - 0.015, y = p_lt_090 + 0.04,
           label = sprintf("%.1f%%", p_lt_090 * 100),
           hjust = 1, size = 3.0, colour = "grey30") +
  scale_x_continuous(
    trans = "log",
    limits = c(xmin, xmax),
    breaks = seq(0.70, 1.30, by = 0.10),
    labels = sprintf("%.2f", seq(0.70, 1.30, by = 0.10))
  ) +
  scale_y_continuous(
    labels = scales::percent,
    limits = c(0, 1),
    breaks = seq(0, 1, by = 0.20),
    sec.axis = sec_axis(~ 1 - ., name = "Probability (RR > X)", labels = scales::percent,
                        breaks = seq(0, 1, by = 0.20))
  ) +
```

```
  labs(y = "Probability (RR < X)", x = NULL) +
  theme_cdf +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank(),
        plot.margin = margin(10, 15, 0, 10))

 # --- Bottom panel: Density ---
 df_dens <- data.frame(RR = dens_x_rr, Density = dens_y)

 # Shaded regions: RR < 0.95 (darker) and 0.95-1.0 (lighter)
 df_shade_strong <- df_dens[df_dens$RR <= 0.95, ]
 df_shade_mild   <- df_dens[df_dens$RR > 0.95 & df_dens$RR <= 1.0, ]

 p_dens <- ggplot(df_dens, aes(x = RR, y = Density)) +
  geom_area(data = df_shade_strong, aes(x = RR, y = Density),
            fill = "#4393C3", alpha = 0.50) +
  geom_area(data = df_shade_mild, aes(x = RR, y = Density),
            fill = "#92C5DE", alpha = 0.35) +
  geom_line(colour = "#2166AC", linewidth = 0.7) +
  geom_vline(xintercept = 0.90, linetype = "dashed", colour = "grey50", linewidth = 0.4) +
  geom_vline(xintercept = 0.95, linetype = "dashed", colour = "grey50", linewidth = 0.4) +
  geom_vline(xintercept = 1.00, linetype = "dashed", colour = "#B2182B", linewidth = 0.5) +
  # Median line
  geom_vline(xintercept = rr_median, linetype = "solid", colour = "#2166AC", linewidth = 0.5, alpha = 0.6) +
  scale_x_continuous(
    trans = "log",
    limits = c(xmin, xmax),
    breaks = seq(0.70, 1.30, by = 0.10),
    labels = sprintf("%.2f", seq(0.70, 1.30, by = 0.10))
  ) +
  labs(x = "Relative Risk (logarithmic scale)", y = "Density") +
  theme_cdf +
  theme(plot.margin = margin(0, 15, 10, 10))

 # --- Combine ---
 combined <- tryCatch({
  library(patchwork)
  p_cdf / p_dens + plot_layout(heights = c(3, 2)) +
    plot_annotation(
      title = sprintf("Posterior Distribution of Pooled Risk Ratio — %s", endpoint_label),
      subtitle = sprintf(
        "Primary prior: Normal(0, 1.5) | Median RR = %.3f | P(RR<0.95) = %.1f%% | P(RR<1) = %.1f%%",
        rr_median, p_lt_095 * 100, p_lt_100 * 100
      ),
      theme = theme(
        plot.title = element_text(size = 13, face = "bold"),
        plot.subtitle = element_text(size = 9, colour = "grey40")
      )
    )
 }, error = function(e) {
  message("patchwork not available; saving panels separately.")
  list(cdf = p_cdf, density = p_dens)
 })

 outfile <- file.path(outdir, sprintf("%s_posterior_cdf_%s.png", prefix, gsub("[- ]", "", endpoint_label)))

 if (inherits(combined, "list")) {
  ggsave(file.path(outdir, sprintf("%s_cdf_panel.png", prefix)), combined$cdf, width = 8, height = 5, dpi = 300)
  ggsave(file.path(outdir, sprintf("%s_density_panel.png", prefix)), combined$density, width = 8, height = 4, dpi = 300)
  message("Saved CDF and density panels separately (patchwork unavailable).")
 } else {
  ggsave(outfile, combined, width = 8, height = 8, dpi = 300)
  message("Saved combined posterior CDF plot: ", outfile)
 }

 # Return key summary
 invisible(data.frame(
  Endpoint = endpoint_label,
  RR_median = round(rr_median, 3),
  P_RR_lt_090 = round(p_lt_090, 4),
  P_RR_lt_095 = round(p_lt_095, 4),
  P_RR_lt_100 = round(p_lt_100, 4)
 ))
}

# ==================== MAIN ANALYSIS FUNCTION ====================
```

```
run_analysis <- function(data, endpoint_name, ctrl_main, ctrl_loo) {

  cat("\n", strrep("=", 70), "\n")
  cat("ANALYZING:", endpoint_name, "MORTALITY\n")
  cat(strrep("=", 70), "\n\n")

  # Prepare data
  brms_data <- prepare_brms_data(data)

  # ---- Protocol v2 prior specifications ----
  # Formulas
  form_RE       <- Events | trials(Total) ~ Treatment + (Treatment | Study)
  form_RE_nocor <- Events | trials(Total) ~ Treatment + (Treatment || Study)
  form_FE       <- Events | trials(Total) ~ Treatment

  # Priors for correlated RE models
  prior_weak <- c(
    set_prior("normal(0, 1.5)",    class = "b"),
    set_prior("student_t(3,0,1)",  class = "sd"),
    set_prior("lkj_corr(2)",       class = "cor")
  )

  prior_scept <- c(
    set_prior("normal(0, 0.354)",  class = "b"),
    set_prior("student_t(3,0,1)",  class = "sd"),
    set_prior("lkj_corr(2)",       class = "cor")
  )

  prior_optim <- c(
    set_prior("normal(-0.15, 0.28)", class = "b"),
    set_prior("student_t(3,0,1)",    class = "sd"),
    set_prior("lkj_corr(2)",         class = "cor")
  )

  prior_pess <- c(
    set_prior("normal(0.15, 0.28)", class = "b"),
    set_prior("student_t(3,0,1)",   class = "sd"),
    set_prior("lkj_corr(2)",        class = "cor")
  )

  # Uncorrelated RE model
  prior_weak_nocor <- c(
    set_prior("normal(0, 1.5)",    class = "b"),
    set_prior("student_t(3,0,1)",  class = "sd")
  )

  # Fixed-effect model
  prior_FE <- set_prior("normal(0, 1.5)", class = "b")

  family_logit <- binomial(link = "logit")

  # Protocol v2 specifies 4 chains, 4000 iterations per chain, 2000 warmup as
  # minimum. Increased to 20,000 iterations (4,000 warmup) to ensure stable
  # posterior summaries and adequate effective sample sizes for tail
  # probabilities [P(RR<0.95), P(RR<0.90)], particularly given the
  # challenging posterior geometry of arm-level binomial-logit models with
  # correlated random effects on small k and very small trials (Rackow n=11).
  chains <- 4
  iter   <- 20000
  warmup <- 4000
  seed   <- 2026

  # ---- Fit models ----

  cat("Fitting primary model (RE, weak prior)...\n")
  fit_RE_weak <- brm(
    form_RE, data = brms_data, family = family_logit,
    prior = prior_weak, chains = chains, iter = iter, warmup = warmup,
    seed = seed, backend = "cmdstanr", control = ctrl_main,
    threads = threading(2), silent = 2, refresh = 0
  )

  cat("Fitting sensitivity model (RE, sceptical prior)...\n")
  fit_RE_scept <- brm(
```

```
  form_RE, data = brms_data, family = family_logit,
  prior = prior_scept, chains = chains, iter = iter, warmup = warmup,
  seed = seed, backend = "cmdstanr", control = ctrl_main,
  threads = threading(2), silent = 2, refresh = 0
)

cat("Fitting sensitivity model (RE, optimistic prior)...\n")
fit_RE_optim <- brm(
  form_RE, data = brms_data, family = family_logit,
  prior = prior_optim, chains = chains, iter = iter, warmup = warmup,
  seed = seed, backend = "cmdstanr", control = ctrl_main,
  threads = threading(2), silent = 2, refresh = 0
)

cat("Fitting sensitivity model (RE, pessimistic prior)...\n")
fit_RE_pess <- brm(
  form_RE, data = brms_data, family = family_logit,
  prior = prior_pess, chains = chains, iter = iter, warmup = warmup,
  seed = seed, backend = "cmdstanr", control = ctrl_main,
  threads = threading(2), silent = 2, refresh = 0
)

cat("Fitting sensitivity model (RE, correlation constrained to 0)...\n")
fit_RE_nocor <- brm(
  form_RE_nocor, data = brms_data, family = family_logit,
  prior = prior_weak_nocor, chains = chains, iter = iter, warmup = warmup,
  seed = seed, backend = "cmdstanr", control = ctrl_main,
  threads = threading(2), silent = 2, refresh = 0
)

cat("Fitting sensitivity model (Fixed-effect)...\n")
fit_FE <- brm(
  form_FE, data = brms_data, family = family_logit,
  prior = prior_FE, chains = chains, iter = iter, warmup = warmup,
  seed = seed, backend = "cmdstanr", control = ctrl_main,
  threads = threading(2), silent = 2, refresh = 0
)

# ---- Extract heterogeneity and correlation (primary model) ----
tau_pi_results <- get_tau_and_pi(
  fit_RE_weak,
  paste(endpoint_name, "- Primary Model (RE, weak prior)"),
  endpoint_name
)

correlation_results <- get_correlation(
  fit_RE_weak,
  paste(endpoint_name, "- Primary Model (RE, weak prior)"),
  endpoint_name
)

# ---- Leave-one-out sensitivity (weak prior) ----
cat("Running leave-one-out sensitivity analyses (weak prior)...\n")
loo_fits <- list(); loo_rows <- list()

study_names <- as.character(data$Study)
for (i in seq_along(study_names)) {
  s <- study_names[i]
  dat_s <- prepare_brms_data(data[-i, , drop = FALSE])

  f <- brm(
    form_RE, data = dat_s, family = family_logit,
    prior = prior_weak, chains = chains, iter = iter, warmup = warmup,
    seed = seed, backend = "cmdstanr", control = ctrl_loo,
    threads = threading(2), silent = 2, refresh = 0
  )

  loo_fits[[s]] <- f
  loo_rows[[s]] <- get_results_row(f, paste("Leave-one-out (drop", s, ")"))
}

# ---- Posterior predictive checks (primary model) ----
ppc_dir <- paste0("ppc_", gsub("[- ]", "", endpoint_name))
ppc_out <- run_ppcs(
  fit_RE_weak, brms_data, outdir = ppc_dir, prefix = endpoint_name,
```

```
  ndraws_overlay = 100, prob_interval = 0.90
)

# ---- Posterior CDF plot (primary model) ----
cdf_dir <- paste0("figures_", gsub("[- ]", "", endpoint_name))
cdf_summary <- generate_posterior_cdf_plot(
  fit_RE_weak, endpoint_label = endpoint_name,
  outdir = cdf_dir, prefix = "albumin"
)

# ---- Assemble results table ----
res_tab <- dplyr::bind_rows(
  get_results_row(fit_RE_weak,   "Main: RE weak prior"),
  get_results_row(fit_RE_scept,  "Sensitivity: RE sceptical"),
  get_results_row(fit_RE_optim,  "Sensitivity: RE optimistic"),
  get_results_row(fit_RE_pess,   "Sensitivity: RE pessimistic"),
  get_results_row(fit_RE_nocor,  "Sensitivity: RE corr=0"),
  get_results_row(fit_FE,        "Sensitivity: Fixed-effect"),
  dplyr::bind_rows(loo_rows)
)

print(res_tab)
write.csv(res_tab, paste0("results_", gsub("[- ]", "", endpoint_name), "_with_rr.csv"), row.names = FALSE)

# ---- Diagnostics table ----
diag_main   <- get_diagnostics_row(fit_RE_weak,   "RE weak prior")
diag_scept  <- get_diagnostics_row(fit_RE_scept,  "RE sceptical prior")
diag_optim  <- get_diagnostics_row(fit_RE_optim,  "RE optimistic prior")
diag_pess   <- get_diagnostics_row(fit_RE_pess,   "RE pessimistic prior")
diag_nocor  <- get_diagnostics_row(fit_RE_nocor,  "RE corr=0 (weak prior)")
diag_fe     <- get_diagnostics_row(fit_FE,        "FE model")

loo_diagnostics <- dplyr::bind_rows(lapply(names(loo_fits), function(s) {
  get_diagnostics_row(loo_fits[[s]], paste("LOO (drop", s, ")"))
}))

diagnostics_tab <- dplyr::bind_rows(
  diag_main, diag_scept, diag_optim, diag_pess, diag_nocor, diag_fe, loo_diagnostics
)

print(diagnostics_tab)
write.csv(diagnostics_tab, paste0("diagnostics_", gsub("[- ]", "", endpoint_name), ".csv"), row.names = FALSE)

# ---- Save complete bundle ----
saveRDS(list(
  data_long      = brms_data,
  fit_RE_weak    = fit_RE_weak,
  fit_RE_scept   = fit_RE_scept,
  fit_RE_optim   = fit_RE_optim,
  fit_RE_pess    = fit_RE_pess,
  fit_RE_nocor   = fit_RE_nocor,
  fit_FE         = fit_FE,
  loo            = loo_fits,
  ppc            = ppc_out,
  results        = res_tab,
  diagnostics    = diagnostics_tab,
  tau_pi         = tau_pi_results,
  correlation    = correlation_results,
  cdf_summary    = cdf_summary
), file = paste0("fits_", gsub("[- ]", "", endpoint_name), "_with_rr.rds"))

cat("\n", strrep("=", 70), "\n")
cat("COMPLETED:", endpoint_name, "ANALYSIS\n")
cat(strrep("=", 70), "\n\n")
}

# ==================== CONTROL SETTINGS ====================
# Protocol v2: adapt_delta >= 0.999; max_treedepth = 20
# Endpoint-specific tuning calibrated from ESM-2 experience:
#   - Longest FU (7 trials, includes Rackow n=11): more aggressive adapt_delta
#     (0.9999) and smaller step_size to handle difficult posterior geometry
#   - 28-day (4 trials): standard adapt_delta (0.999) sufficient
#   - LOO refits: smaller step_size for stability with k-1 trials

ctrl_longest_main <- list(adapt_delta = 0.9999, max_treedepth = 20, step_size = 0.005)
```

```
ctrl_longest_loo  <- list(adapt_delta = 0.999,  max_treedepth = 20, step_size = 0.005)
ctrl_28d_main     <- list(adapt_delta = 0.999,  max_treedepth = 20, step_size = 0.01)
ctrl_28d_loo      <- list(adapt_delta = 0.999,  max_treedepth = 20, step_size = 0.005)

# ========== EXECUTE ==========

# Primary analysis: longest available follow-up (up to 90 days), 7 RCTs
run_analysis(data_longest, "Longest follow-up", ctrl_longest_main, ctrl_longest_loo)

# Sensitivity analysis: 28-day mortality (or closest <= 30 days), 4 RCTs
run_analysis(data_28d, "28-day", ctrl_28d_main, ctrl_28d_loo)

cat("\n", strrep("=", 70), "\n")
cat("ALL ANALYSES COMPLETED SUCCESSFULLY\n")
cat(strrep("=", 70), "\n")
```