SUPPLEMENTARY APPENDIX

Powering bias and clinically important treatment effects in randomized trials of critical illness

Darryl Abrams, MD, Sydney B. Montesi, MD, Sarah K. L. Moore, MD, Daniel K. Manson, MD, Kaitlin M. Klipper, MD, Meredith A. Case, MD, MBE, Daniel Brodie, MD, Jeremy R. Beitler, MD, MPH

Table of Contents

Additional Methods

Identification of Trials for Inclusion

Studies eligible for inclusion were multicenter randomized clinical trials of critically ill adult patients in which mortality was the main endpoint. For inclusion, the publication must have appeared between January 1, 2008 and December 31, 2018 in one of seven journals: *New England Journal of Medicine*, *Journal of the American Medical Association* (*JAMA*), *The Lancet*, *American Journal of Respiratory and Critical Care Medicine*, *Lancet Respiratory Medicine*, *Intensive Care Medicine*, or *Critical Care Medicine*. Trials were excluded if not designed as superiority trials or if patient-level randomization was not employed. Studies involving patients who underwent an elective procedure, were not critically ill prior to the procedure, and rapidly recovered post-procedure without high probability of life-threatening deterioration, were excluded. For final arbitration of discordance regarding what constituted critical illness, the Centers for Medicare and Medicaid Services definition of critical illness was applied, as illness or injury that "acutely impairs one or more vital organ systems such that there is a high probability of imminent or life-threatening deterioration in the patient's condition."(1)

To identify articles for inclusion, tables of contents for each journal issue were screened independently by two study physicians. To ensure studies of key topics were not missed during manual screening, a PubMed search was also conducted for each journal using the following Medical Subject Headings: *critical illness* or *shock* or *acute respiratory distress syndrome* or *respiratory failure*; and either *randomized controlled trial* or *clinical trial.* The final study list was compiled combining results from all search strategies. Discordance in studies identified for possible inclusion was resolved by review by a third study physician.

Data Collection

Data were extracted in duplicate by study physicians blinded to each other's data entry. These datasets were merged to assess for discordance, which was resolved by an independent third reviewer as needed to create the final dataset. Data sources included the main manuscript publication, online data and protocol supplements published concomitantly with the manuscript, separately published trial protocols referenced by the manuscript, and trial registration websites.

Handling of Trials with More than One Primary Comparison

Trials with more than two parallel groups or 2x2 factorial design were reviewed to identify the main prespecified comparison per protocol. When only a single comparison was prespecified or emphasized in the results, that sole comparison was included in this analysis. For factorial trials in which more than one comparison was prespecified as the main analysis of the primary endpoint, each pairwise comparison was entered as a separate trial for all analyses (**Figure S1**).

Statistical Analysis

Differences in sample size by trial type (cardiovascular, respiratory, sepsis, general critical care, other) or funding source were assessed via Wilcoxon rank-sum test or two-sample t-test, as appropriate, comparing each category of interest to all others.

To evaluate accuracy of the anticipated control group event rate, observed versus predicted control group mortality was assessed graphically with a waterfall plot. A paired t-test was used to determine whether the difference in observed versus control group mortality was non-zero.

To assess treatment effect for which the trial was powered, predicted risk difference was calculated as the difference in hypothesized treatment minus control group mortality, ascertained from the reported sample size calculation. Differences in hypothesized risk difference by trial type were assessed via two-sample t-test.

Observed risk difference was calculated directly using data extracted from each trial. Number needed to treat (NNT) was calculated as the inverse of the risk difference.

Potential clinical significance of trial results was evaluated by several methods according to the approach of Kaul and Diamond.(2) Briefly, risk difference and 95% confidence interval (CI) computed from main trial results were displayed in a forest plot. We prespecified a 5% risk difference for all trials (NNT = 20 patients) as potentially clinically important. Trials were evaluated for whether the 95% CI included this prespecified clinically important threshold. Trials also were evaluated for whether the 95% CI risk difference included the trial's own predicted treatment effect. Analyses were performed for all trials combined and separately for trials that did not meet statistical significance for benefit or harm with the intervention. Bayesian models were developed via the Markov Chain Monte Carlo method with noninformative priors, 1,000 burn-in iterations, 50,000 iterations and a thinning rate of five. Model results were used to calculate the posterior probability of observing a 5% risk difference (benefit or harm), the proposed MCID. Numeric posterior probabilities for treatment effect of each individual trial were reported. Other treatment effect thresholds for clinical importance on the absolute and relative risk scales were considered in secondary analyses.

All frequentist hypothesis testing applied a significance threshold of 0.05 without adjustment for multiple comparisons. Analyses were conducted using SAS 9.4 and PASS 14.

Additional Results

Characteristics of Included Trials

Of 657 unique publications identified on initial screen, 101 multicenter superiority trials with patient-level randomization and mortality as the main endpoint were included (**Table 1, Figure S1**). A complete list of included trials is provided later in this online supplement.

Four publications described a factorial 2x2 design with pairwise comparisons of both factors specified in their main endpoint analyses. Consistent with the analyses detailed in those publications, each factor was handled as a separate trial in the present study (**Figure S1**).

Most included trials were government-funded and conducted in Europe (**Table 1**). Twelve trials (11.9%) met the statistical significance threshold for the primary endpoint, five of which demonstrated increased mortality with the intervention.

Sample Size

The median (IQR) sample size for analysis of the main endpoint was 843 (411-1588) patients. The smallest sample size was 62 patients, and the largest was 20,127 patients.

Seven trial protocols (6.9%) incorporated an adaptive sample size design that permitted increasing enrollment targets if prespecified criteria were met (3-8). One such trial employed an event-driven design that continued enrollment until a prespecified number of deaths was reached (5). Others employed sequential design (6) or revised the target sample size if the observed overall mortality (8), control group mortality (7) or treatment effect (4) was less than expected.

Data Available from Power Calculations

Power calculations contained sufficient information to ascertain the hypothesized absolute risk reduction in 100 of 101 trials and the hypothesized relative risk reduction in 97 of 101 trials.

Analysis According to Trial Type

Sample size: Trials evaluating respiratory interventions had a significantly smaller sample size than non-respiratory trials, while general critical care trials had a significantly larger sample size compared to other trial types (Table S1).

Control group mortality: Compared to other included trials, predicted control group mortality was significantly more accurate in cardiovascular and general ICU trials. Sepsis trials had the least accurate prediction of control group mortality (Table S2).

Predicted risk difference: General critical care trials were powered to detect a significantly smaller risk difference than non-general critical care trials (mean difference -3.0%, 95% CI -5.6% to -0.5%; p = 0.02;). The predicted risk difference for other trial types is provided in Table S3, and how they compare to observed risk difference is reported in Table S4.

Analysis According to Government vs. Non-Government Funding Source

Government funding was not associated with sample size, misestimation of control group mortality, or predicted risk difference (Table S5).

Alternative Treatment Effect Thresholds

The proportion of trials meeting alternative treatment effect size thresholds for mortality is reported in Tables S6-S8 and Figures S1-S4.

Limitations of Additional Results

Tables S1-S5 report multiple p-values without adjustment for multiplicity. As such, these findings should be viewed as hypothesis-generating.

Table S1. Sample size by trial type

Trial Type	**Sample Size, Median [IQR]**		
	Trial type of interest	All other trial types	*P*
Cardiovascular trials	1276 [506-2857]	776 [374-1341]	0.07
Respiratory trials	548 [339-843]	1051.5 [434-1947.5]	0.01
Sepsis trials	536 [350-1241]	983 [418-2026]	0.07
General critical care trials	2396.5 [1218-3914]	745 [350-1243]	< 0.01

Table S2. Accuracy of control group mortality by trial type

Trial Type	**Control Group Mortality Difference: Observed minus Predicted, Mean ± SD**		
	Trial type of interest	All other trial types	*P*
Cardiovascular trials	-1.0 ± 12.3	-8.0 ± 8.7	0.03
Respiratory trials	-8.6 ± 9.5	-6.2 ± 9.8	0.33
Sepsis trials	-10.0 ± 8.9	-5.1 ± 9.8	0.02
General critical care trials	-2.0 ± 6.2	-7.5 ± 10.1	0.01

Table S3. Predicted absolute risk reduction in mortality by trial type

Trial Type	**Predicted Risk Difference in % Mortality, Mean ± SD**		
	Trial type of interest	All other trial types	*P*
Cardiovascular trials	7.7 ± 5.3	9.8 ± 4.1	0.07
Respiratory trials	11.0 ± 3.7	8.9 ± 4.5	0.06
Sepsis trials	10.0 ± 3.5	9.0 ± 4.8	0.29
General critical care trials	6.7 ± 4.3	9.8 ± 4.3	0.02

Table S4. Difference in predicted verses observed treatment effect, absolute risk difference, by trial type

Trial Type	**Predicted Risk Difference in % Mortality, Mean ± SD**		
	Trial type of interest	All other trial types	*P*
Cardiovascular trials	6.3 ± 5.3	9.7 ± 5.5	0.02
Respiratory trials	10.1 ± 5.6	8.7 ± 5.6	0.32
Sepsis trials	10.2 ± 5.9	8.4 ± 5.4	0.14
General critical care trials	7.0 ± 2.6	9.3 ± 5.9	0.02

Table S5. Characteristics of trials by government funding

Characteristic	**Government funded**	**Non-government funded**	*P*
Sample size, median [IQR]	1004 [466-1588]	598 [342-1679]	0.11
Difference in observed vs. predicted control group mortality (%), mean ± SD	-6.2 ± 8.9	-7.5 ± 11.1	0.53
Predicted absolute risk reduction in % mortality, mean ± SD	8.9 ± 4.0	10.2 ± 4.9	0.14
Observed absolute risk difference within ± 5% of predicted, number (%) of trials	14 (22.6%)	6 (15.4%)	0.45

Table S6. Proportion of trials for which effect estimate includes specified treatment effect size, all trials (n = 101)

Threshold risk difference	95% CI includes threshold risk difference, no. (%) of trials	Threshold risk difference more probable than not*, no. (%) of trials
Absolute risk reduction		
3% reduction for death	75 (74.3%)	23 (22.8%)
5% reduction for death	49 (48.5%)	11 (10.9%)
10% reduction for death	24 (23.8%)	5 (5.0%)
15% reduction for death	14 (13.9%)	1 (1.0%)
20% reduction for death	6 (5.9%)	0
Absolute risk increase		
3% increase for death	73 (72.3%)	18 (17.8%)
5% increase for death	53 (52.5%)	11 (10.9%)
10% increase for death	27 (26.7%)	2 (2.0%)
15% increase for death	11 (10.9%)	0
20% increase for death	2 (2.0%)	0

* Trials with Bayesian posterior probability > 0.5 for observing threshold treatment effect, calculated using noninformative priors.

Table S7. Inconclusiveness of important effect size for *reduction* in mortality among trials without a statistically significant treatment *benefit* (n = 94)

Threshold absolute risk reduction	95% CI includes threshold risk decrease, no. (%) of trials	Threshold risk decrease more probable than not*, no. (%) of trials
3% reduction for death	70 (74.5%)	18 (19.1%)
5% reduction for death	44 (46.8%)	6 (6.4%)
10% reduction for death	19 (20.2)	2 (2.1%)
15% reduction for death	10 (10.6%)	0
20% reduction for death	4 (4.3%)	0

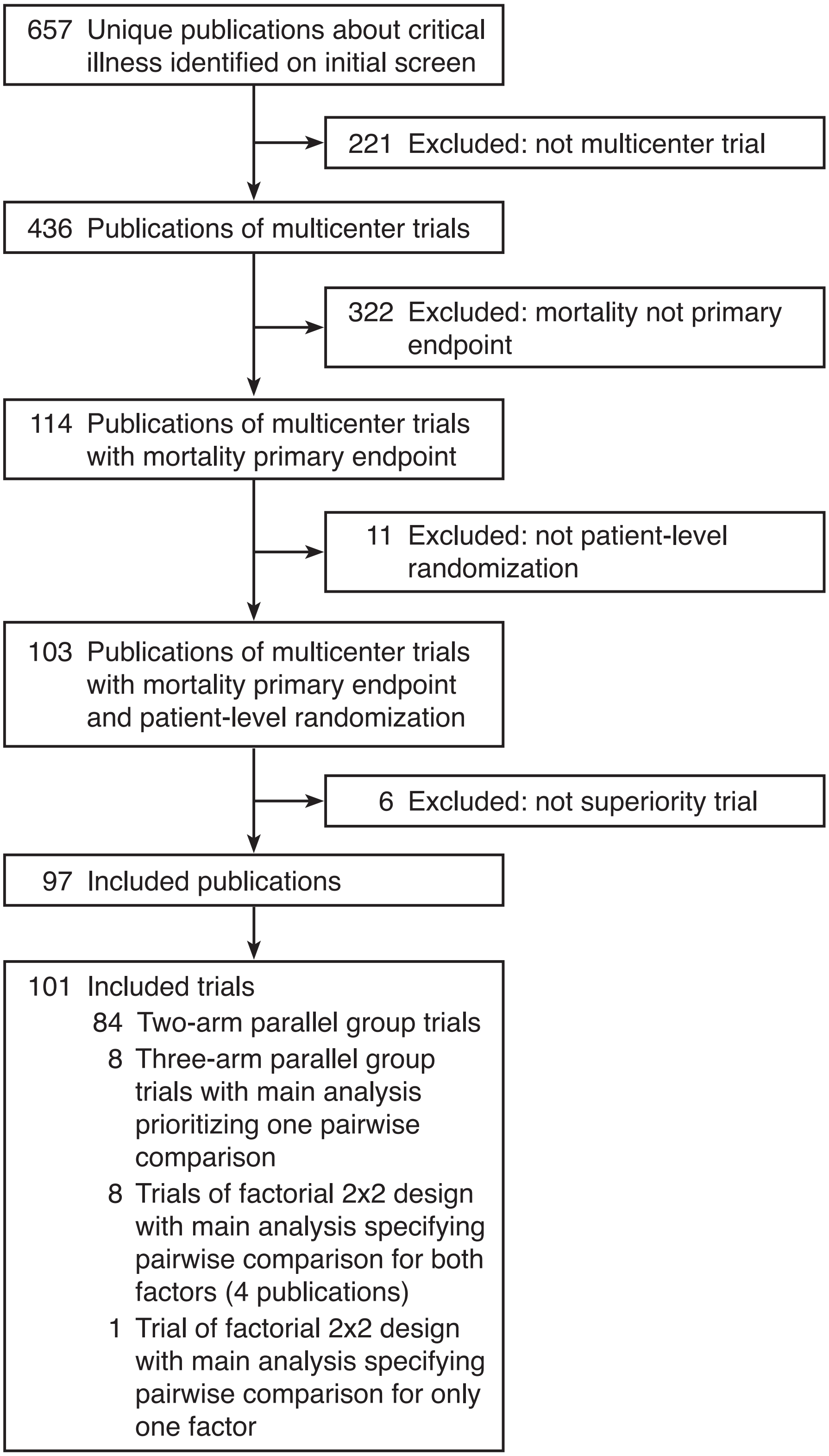
* Trials with Bayesian posterior probability > 0.5 for observing threshold treatment effect, calculated using noninformative priors.

Table S8. Inconclusiveness of important effect size for *increase* in mortality among trials without a statistically significant treatment *harm* (n = 96)

Threshold absolute risk increase	95% CI includes threshold risk increase	Threshold risk increase more probable than not*
3% increase for death	68 (71.9%)	14 (14.6%)
5% increase for death	49 (51.0%)	7 (7.3%)
10% increase for death	23 (24.0%)	0
15% increase for death	9 (9.4%)	0
20% increase for death	1 (1.0%)	0

* Trials with Bayesian posterior probability > 0.5 for observing threshold treatment effect, calculated using noninformative priors.

Figure S1. Screening and inclusion of potentially eligible studies.


657 Unique publications about critical illness identified on initial screen
221 Excluded: not multicenter trial
436 Publications of multicenter trials
322 Excluded: mortality not primary endpoint
114 Publications of multicenter trials with mortality primary endpoint
11 Excluded: not patient-level randomization
103 Publications of multicenter trials with mortality primary endpoint and patient-level randomization
6 Excluded: not superiority trial
97 Included publications
101 Included trials
84 Two-arm parallel group trials
8 Three-arm parallel group trials with main analysis prioritizing one pairwise comparison
8 Trials of factorial 2x2 design with main analysis specifying pairwise comparison for both factors (4 publications)
1 Trial of factorial 2x2 design with main analysis specifying pairwise comparison for only one factor

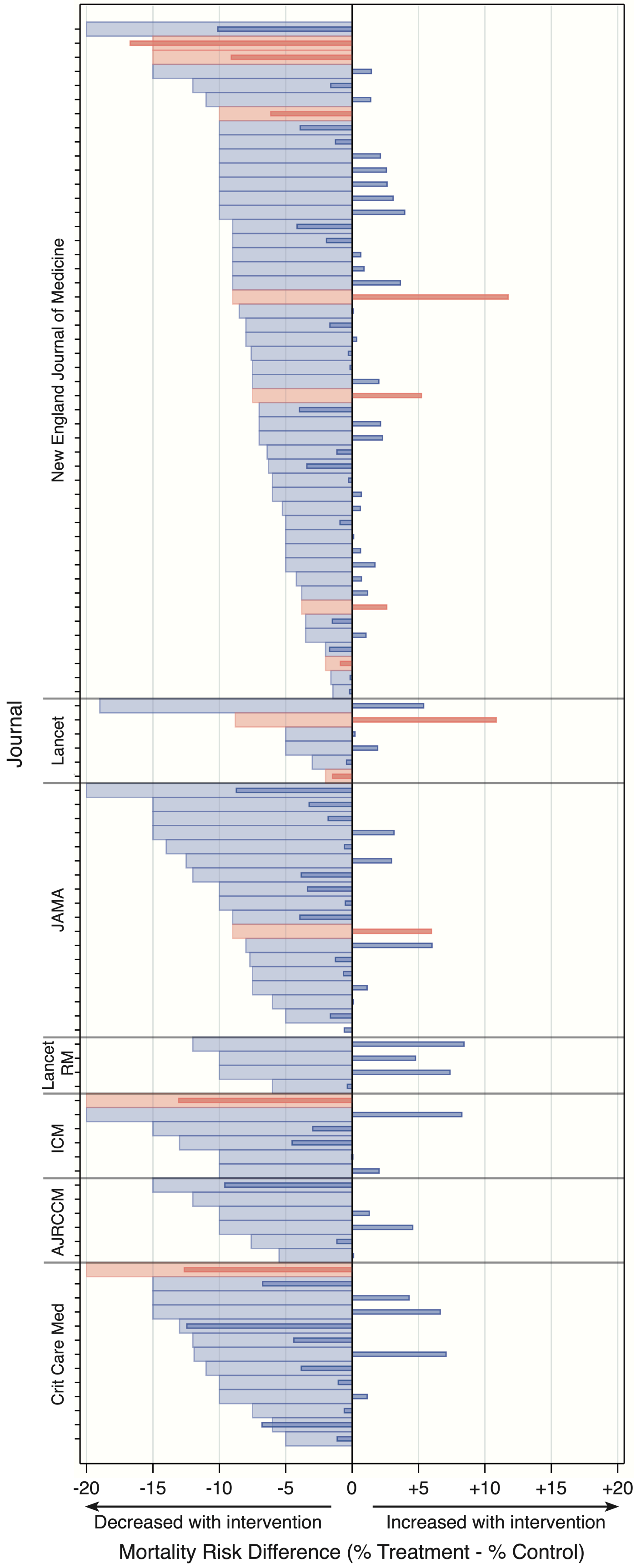


Figure S2. Butterfly plot of predicted and observed mortality risk difference, grouped by journal. Data are grouped by journal and then sorted by sample size. One trial (sample size 1,439 patients) contained insufficient information in its reported power and sample size determination to identify predicted risk difference.

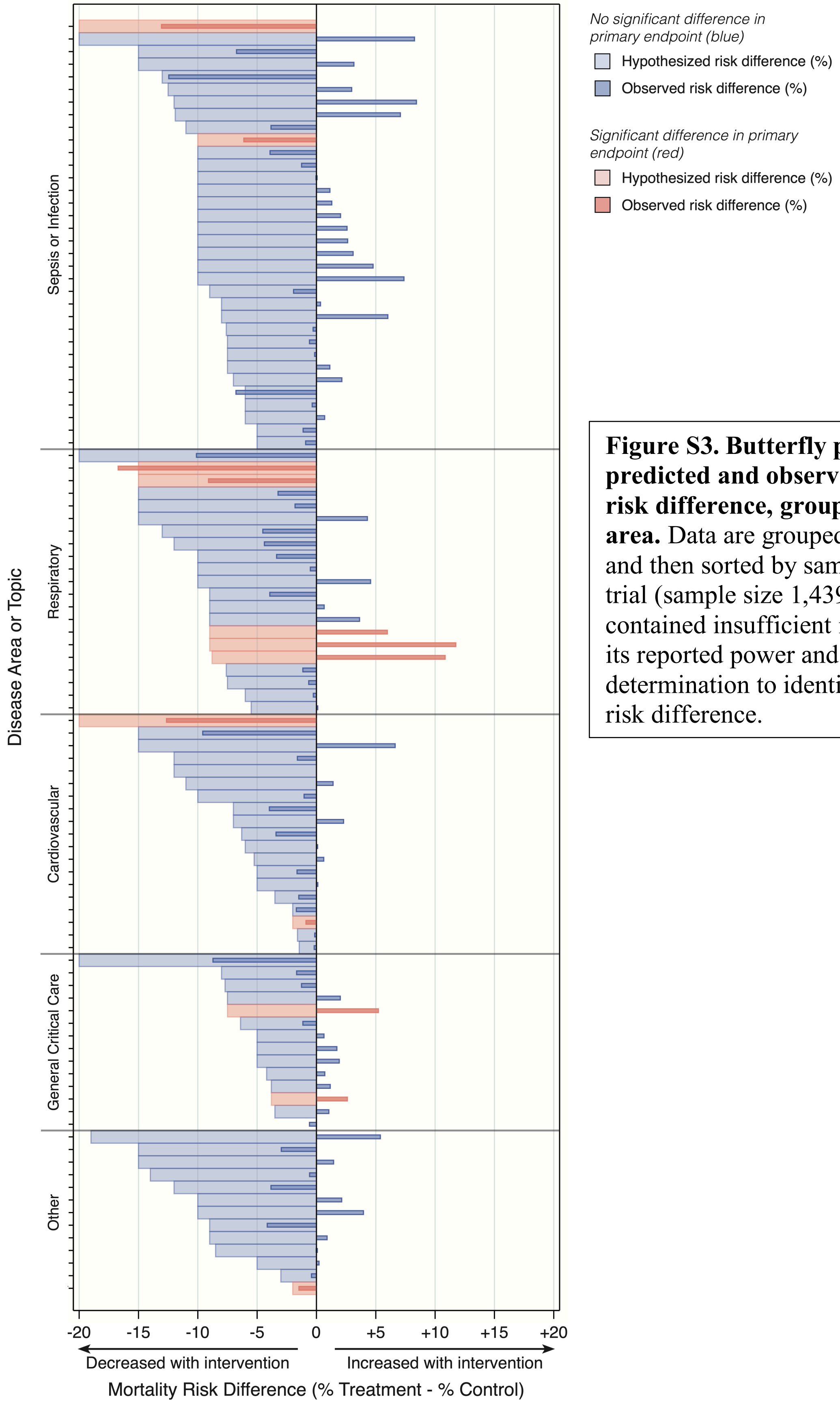


Figure S3. Butterfly plot of predicted and observed mortality risk difference, grouped by disease area. Data are grouped by disease area and then sorted by sample size. One trial (sample size 1,439 patients) contained insufficient information in its reported power and sample size determination to identify predicted risk difference.

Figure S4. Trial results according to clinically important difference in mortality on absolute and relative scales, grouped by journal. The prespecified threshold used for MCID was a 5% absolute risk difference (number needed to treat = 20) or 20% relative risk difference (risk ratio ≤ 0.8 or ≥ 1.2) for either benefit or harm with treatment. Thresholds are indicated by the shaded areas: blue for benefit and red for harm. * denotes statistical significance according to the trial's main analysis.

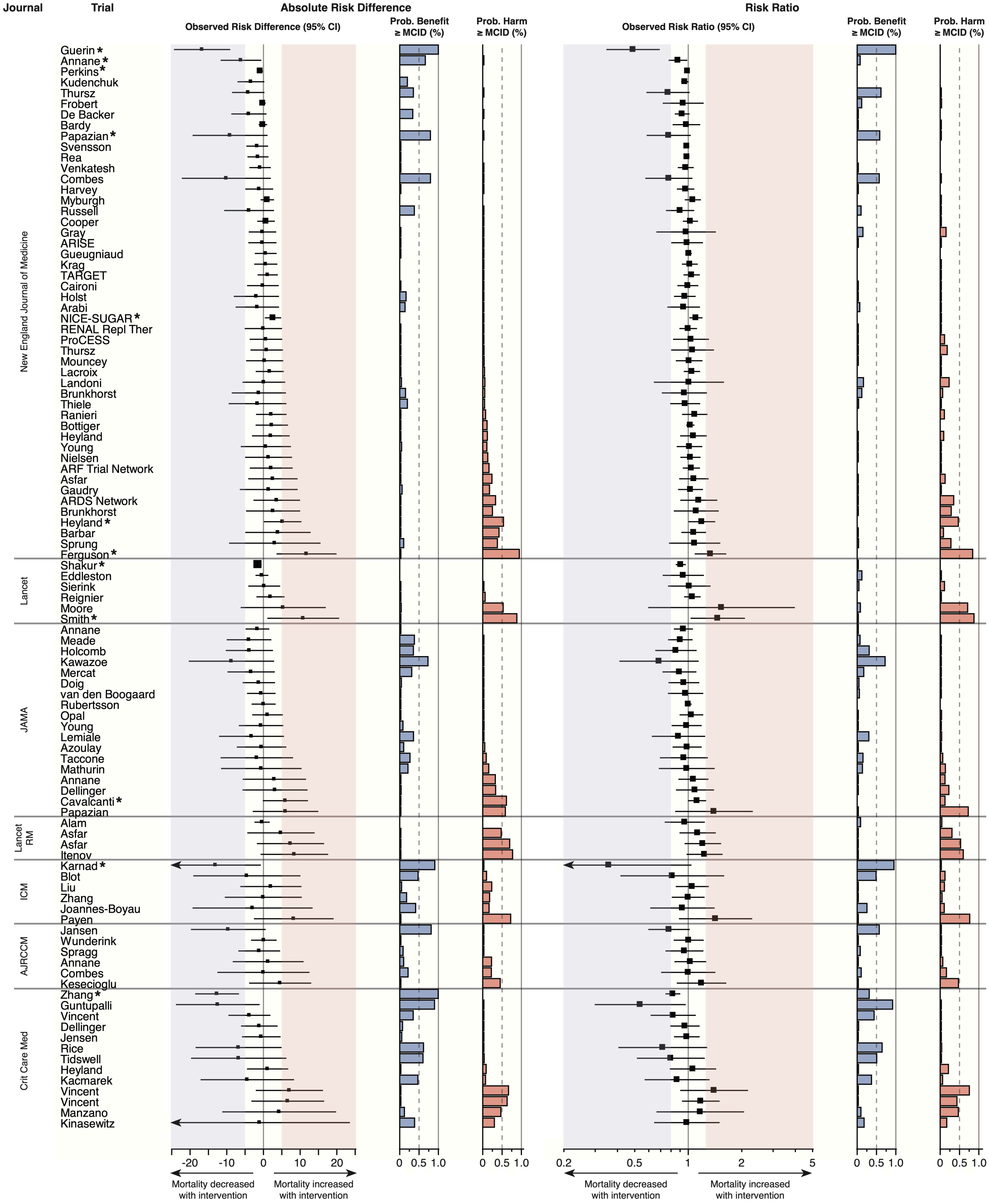


Figure S5. Trial results according to clinically important difference in mortality on absolute and relative scales, grouped by disease area. The prespecified threshold used for MCID was a 5% absolute risk difference (number needed to treat = 20) or 20% relative risk difference (risk ratio ≤ 0.8 or ≥ 1.2) for either benefit or harm with treatment. Thresholds are indicated by the shaded areas: blue for benefit and red for harm. * denotes statistical significance according to the trial's main analysis.

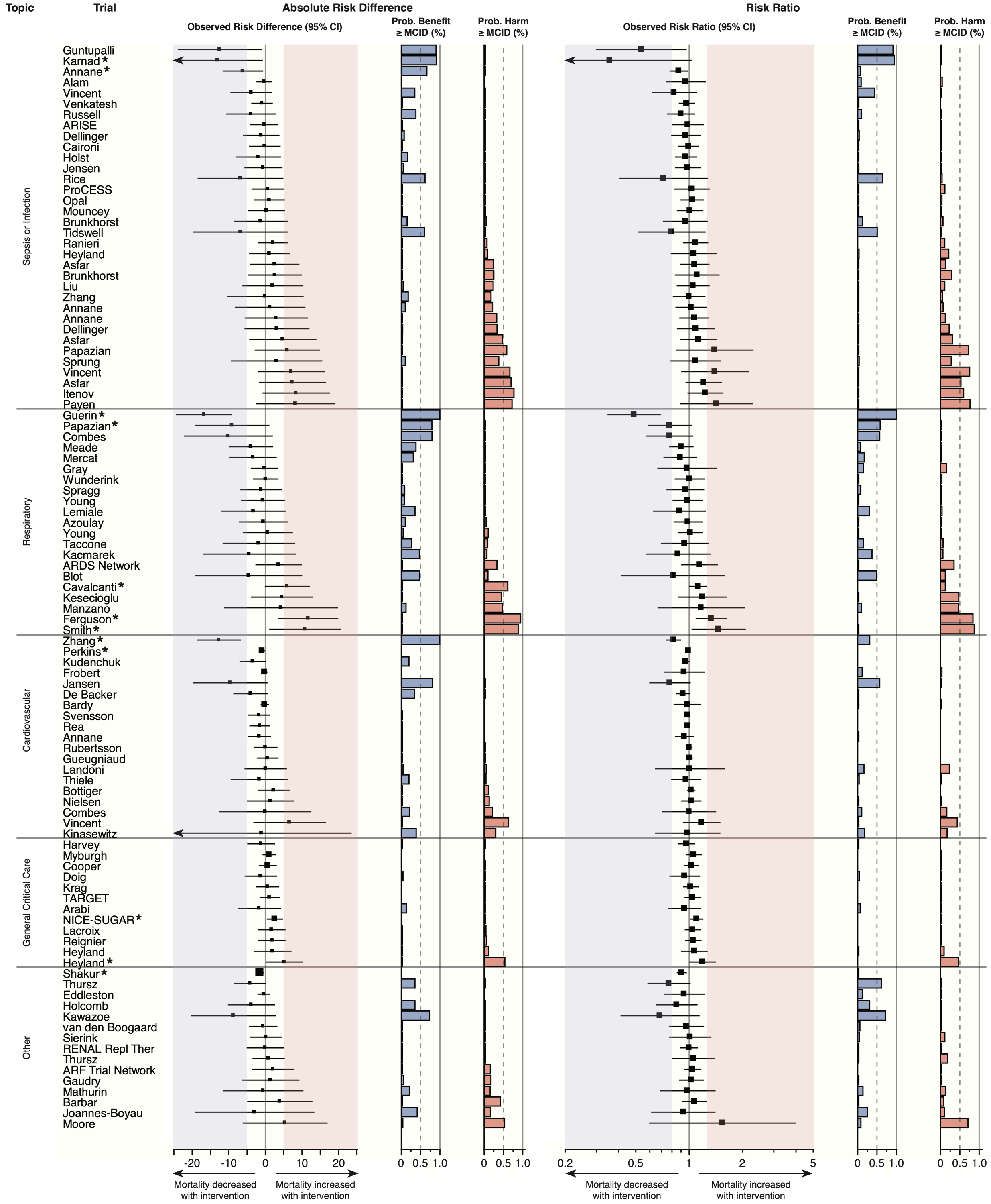


Figure S6. Percent of trials in which the intervention was more probable than not to confer at least the specified absolute risk reduction. For each specified absolute risk reduction (x-axis), bar height indicates the percent of trials (y-axis) for which the Bayesian posterior probability of observing that effect size exceeds 0.5.

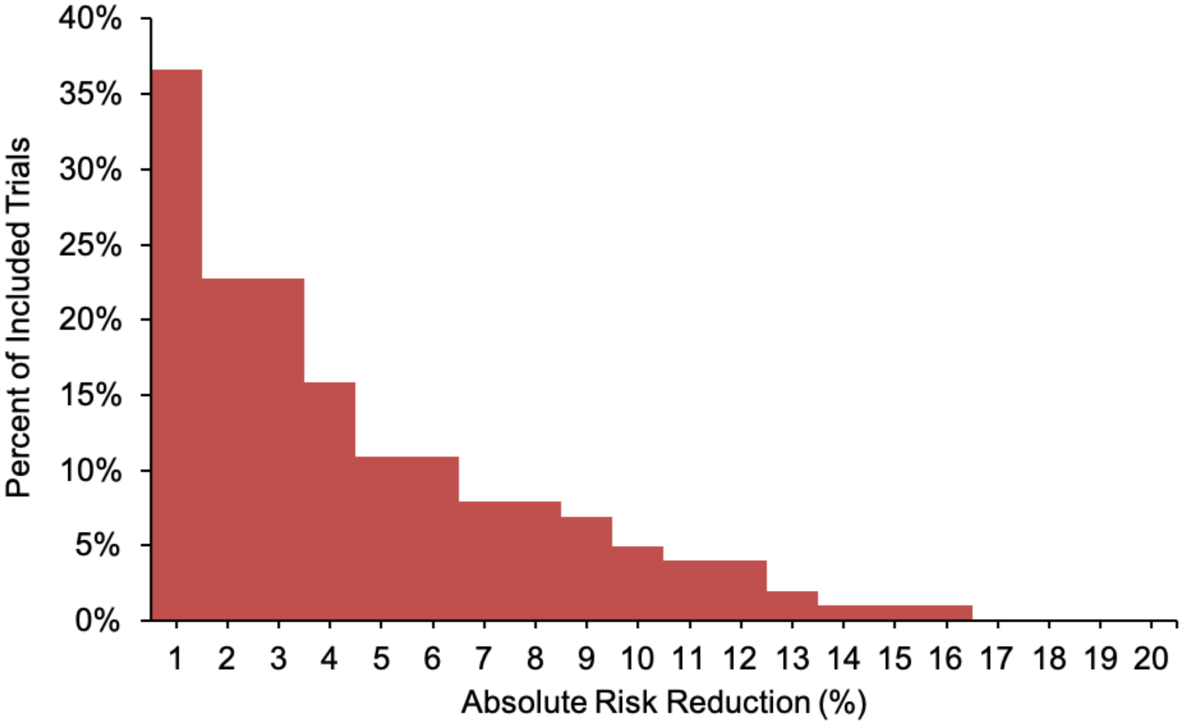


Figure S7. Percent of trials in which the intervention was more probable than not to confer at least the specified absolute risk increase. For each specified absolute risk increase (x-axis), bar height indicates the percent of trials (y-axis) for which the Bayesian posterior probability of observing that effect size exceeds 0.5.

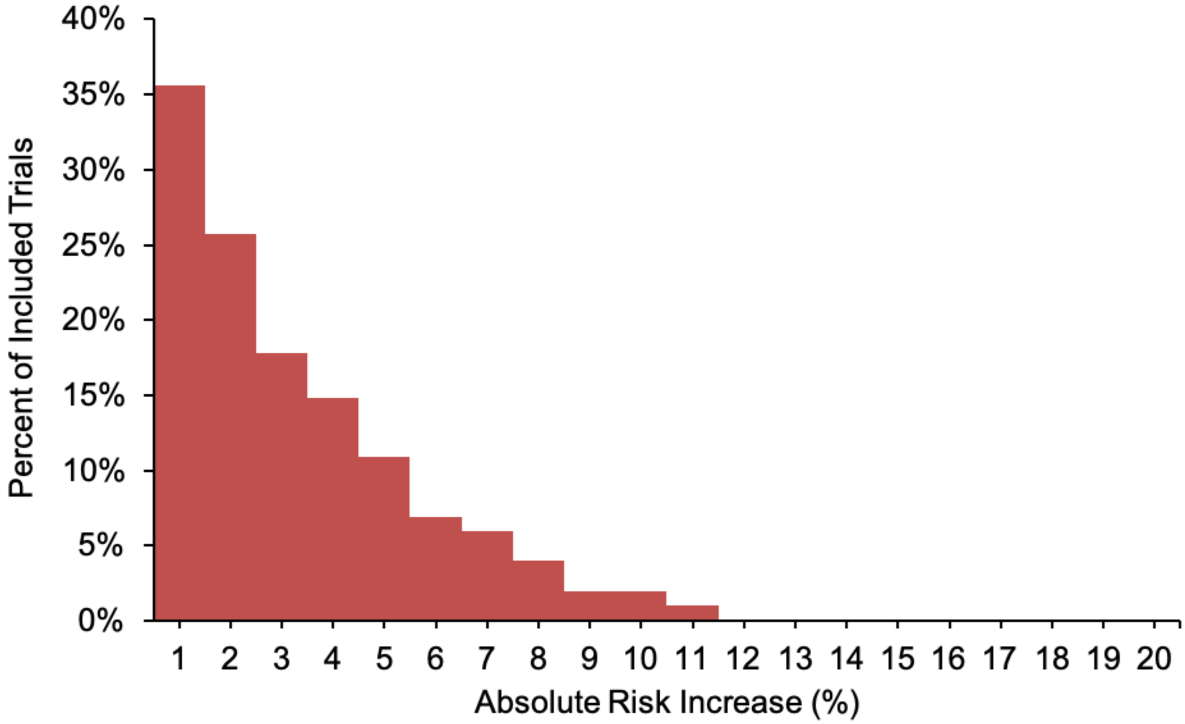


Figure S8. Percent of trials in which the intervention was more probable than not to confer at least the specified relative risk reduction. For each specified relative risk reduction (x-axis), bar height indicates the percent of trials (y-axis) for which the Bayesian posterior probability of observing that effect size exceeds 0.5.

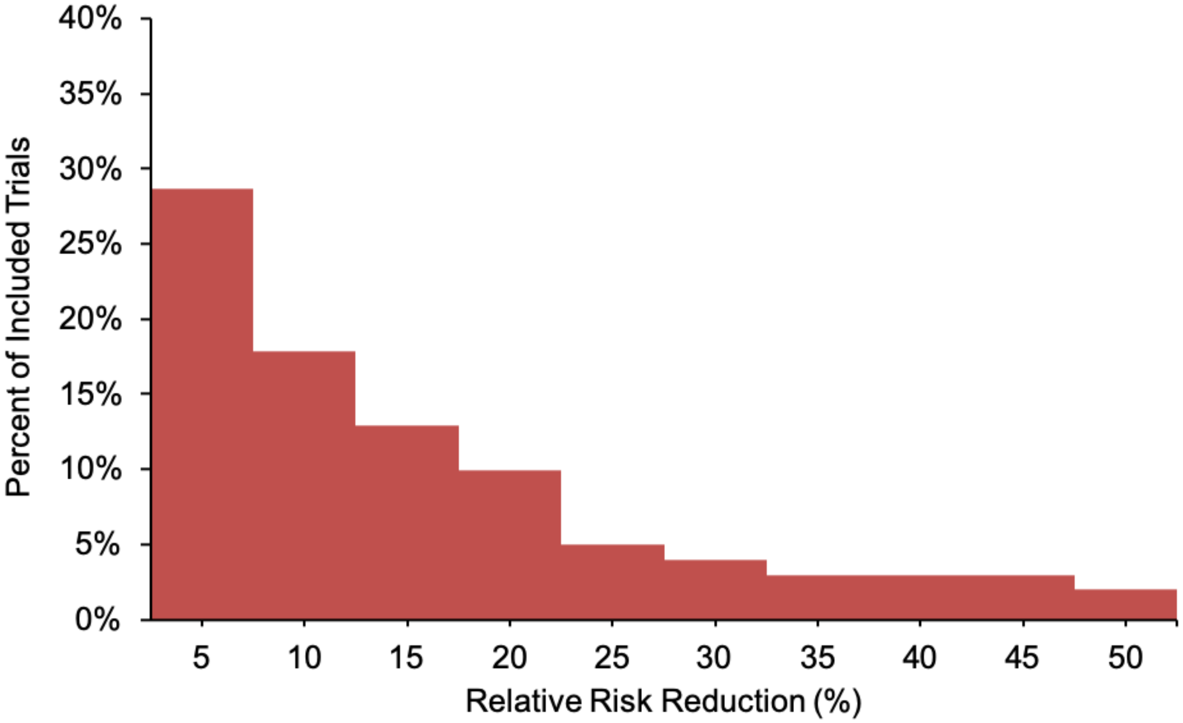
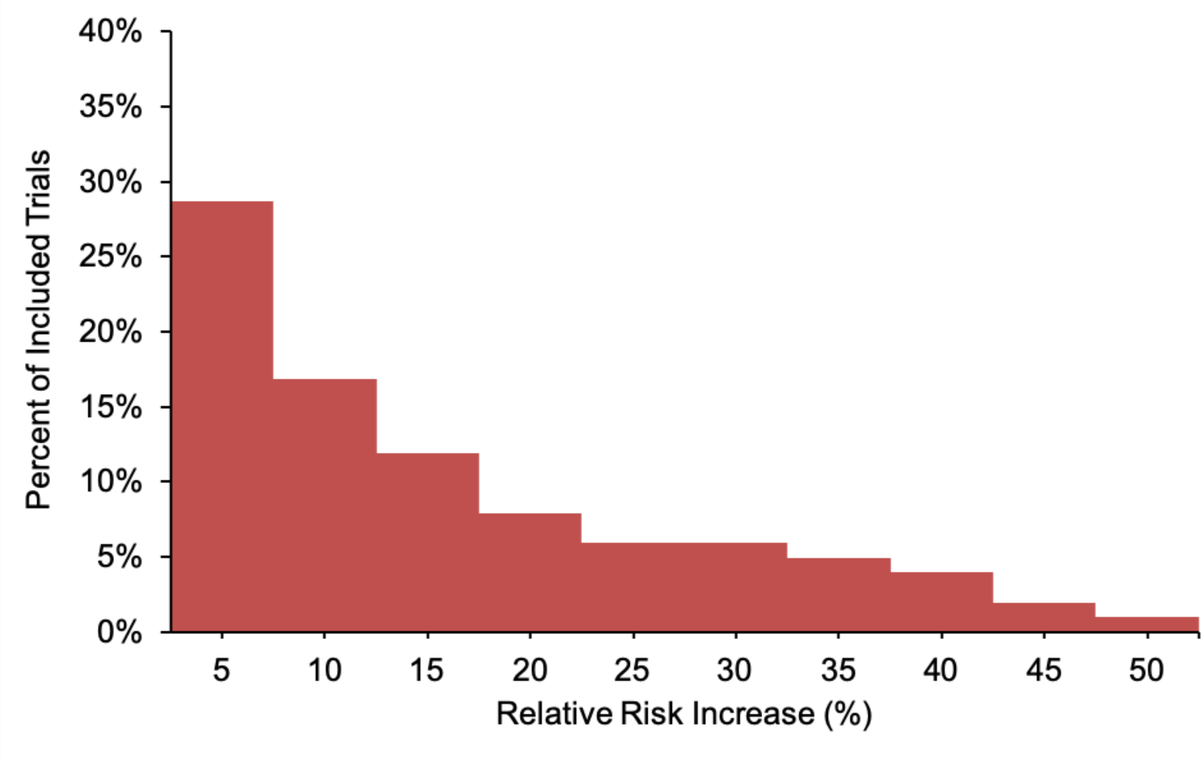


Figure S9. Percent of trials in which the intervention was more probable than not to confer at least the specified relative risk increase. For each specified relative risk increase (x-axis), bar height indicates the percent of trials (y-axis) for which the Bayesian posterior probability of observing that effect size exceeds 0.5.



Supplement References

1. Lustbader DR, Nelson JE, Weissman DE *et al*: Physician reimbursement for critical care services integrating palliative care for patients who are critically ill. *Chest* 2012; 141:787-792.
2. Kaul S, Diamond GA: Trial and error. How to avoid commonly encountered limitations of published clinical trials. *J Am Coll Cardiol* 2010; 55:415-427.
3. Brunkhorst FM, Engel C, Bloos F *et al*: Intensive insulin therapy and pentastarch resuscitation in severe sepsis. *N Engl J Med* 2008; 358:125-139.
4. Caironi P, Tognoni G, Masson S *et al*: Albumin replacement in patients with severe sepsis or septic shock. *N Engl J Med* 2014; 370:1412-1421.
5. Writing Group for the Alveolar Recruitment for Acute Respiratory Distress Syndrome Trial I, Cavalcanti AB, Suzumura EA *et al*: Effect of Lung Recruitment and Titrated Positive End-Expiratory Pressure (PEEP) vs Low PEEP on Mortality in Patients With Acute Respiratory Distress Syndrome: A Randomized Clinical Trial. *JAMA* 2017; 318:1335-1345.
6. De Backer D, Biston P, Devriendt J *et al*: Comparison of dopamine and norepinephrine in the treatment of shock. *N Engl J Med* 2010; 362:779-789.
7. Holcomb JB, Tilley BC, Baraniuk S *et al*: Transfusion of plasma, platelets, and red blood cells in a 1:1:1 vs a 1:1:2 ratio and mortality in patients with severe trauma: the PROPPR randomized clinical trial. *JAMA* 2015; 313:471-482.
8. Ranieri VM, Thompson BT, Barie PS *et al*: Drotrecogin alfa (activated) in adults with septic shock. *N Engl J Med* 2012; 366:2055-2064.