SYSTEMATIC REVIEW

Unexplained mortality differences between septic shock trials: a systematic analysis of population characteristics and control-group mortality rates

Harm-Jan de Grooth1,2*, Jonne Postema2, Stephan A. Loer2, Jean-Jacques Parienti3,4, Heleen M. Oudemans-van Straaten1 and Armand R. Girbes1

© 2018 The Author(s)

Abstract

Purpose: Although the definition of septic shock has been standardized, some variation in mortality rates among clinical trials is expected. Insights into the sources of heterogeneity may influence the design and interpretation of septic shock studies. We set out to identify inclusion criteria and baseline characteristics associated with between-trial differences in control group mortality rates.

Methods: We conducted a systematic review of RCTs published between 2006 and 2018 that included patients with septic shock. The percentage of variance in control-group mortality attributable to study heterogeneity rather than chance was measured by $I^2$. The association between control-group mortality and population characteristics was estimated using linear mixed models and a recursive partitioning algorithm.

Results: Sixty-five septic shock RCTs were included. Overall control-group mortality was 38.6%, with significant heterogeneity ($I^2 = 93\%, P < 0.0001$) and a 95% prediction interval of 13.5–71.7%. The mean mortality rate did not differ between trials with different definitions of hypotension, infection or vasopressor or mechanical ventilation inclusion criteria. Population characteristics univariately associated with mortality rates were mean Sequential Organ Failure Assessment score ($\beta = 0.57, P = 0.007$), mean serum creatinine ($\beta = 0.48, P = 0.007$), the proportion of patients on mechanical ventilation ($\beta = 0.61, P < 0.001$), and the proportion with vasopressors ($\beta = 0.57, P = 0.002$). Combinations of population characteristics selected with a linear model and recursive partitioning explained 41 and 42%, respectively, of the heterogeneity in mortality rates.

Conclusions: Among 65 septic shock trials, there was a clinically relevant amount of heterogeneity in control group mortality rates which was explained only partly by differences in inclusion criteria and reported baseline characteristics.

Keywords: Septic shock, Heterogeneity, Clinical trials, Methodology, Meta-research, Machine learning

*Correspondence: h.degrooth@vumc.nl
1 Department of Intensive Care, VU University Medical Center, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands
Full author information is available at the end of the article
Introduction
The fundamental criteria from the consensus definitions of septic shock are used to select patients for inclusion in clinical studies [1–4]. While the mortality rate of septic shock was found to be 46% (95% confidence interval (CI) 43–50%) in a meta-analysis of observational cohorts [5], randomized controlled trials report more diverse numbers. For example, two high-profile septic shock trials published a year apart reported control group mortality rates as disparate as 16% [6] and 80% [7]. Despite the seemingly wide range of mortality rates there has not yet been a systematic inquiry into its patterns and possible causes.

Identifying the correct patient population to benefit from a specific therapy has been recognized as an essential condition for improving critical care research [8–10]. Yet large unexplained mortality differences among trials that all aim to include septic shock patients may hamper reproducibility and generalizability. Insights into the magnitude and sources of between-trial heterogeneity are therefore valuable in the design, reporting, and interpretation of septic shock trials. For example, incorrect prediction of baseline mortality rates has been identified as a major reason for negative critical care trials, as a discrepancy between expected and observed event rates often leads to underpowered studies [11].

We sought to quantify between-trial heterogeneity and identify inclusion criteria and population characteristics associated with differences in control group mortality rates.

Methods
After a systematic search to identify all trials published in the past decade that aimed to include patients with septic shock, we used linear mixed models to estimate the total heterogeneity in control group mortality rates and its association with reported baseline characteristics. Using both a multivariate linear model and a machine learning algorithm, we estimated the proportion of heterogeneity that can be explained by population characteristics.

The review protocol was prospectively registered [12] and adheres to the PRISMA checklist [13], which is included in the electronic supplementary material (ESM). Study screening, application of the inclusion- and exclusion criteria and data-extraction were performed independently by two reviewers (HJdG and JP). Conflicting entries were resolved by consensus.

Inclusion criteria and search strategy
PubMed, Embase, and the Cochrane Central Register of Controlled Trials were queried using the search term ["septic shock" AND (random* or rct)]. Embase was additionally queried using the search term "septic shock" with the randomized controlled trial filter activated. The queries were limited to publications from 1 January 2006 and the queries were last performed on 20 January 2018.

We limited the search to trials published between 2006 and 2018 as a compromise between the number of eligible studies and secular trends in clinical practice, research practice, and reporting standards. Publications from 2006 and later had sufficient lead time to incorporate the 2004 update of the Surviving Sepsis Campaign guidelines [4].

Eligible for inclusion were parallel-group randomized controlled trials with adult patients in septic shock according to the published consensus definitions or Surviving Sepsis Campaign guidelines [1, 2, 4]. Trials were excluded if the report was not written in English, if it was only available in abstract, if no baseline characteristics were reported, or if no mortality outcome was reported. Trials that aimed to include a specific subcategory of septic shock patients (e.g., “septic shock patients requiring renal replacement therapy”) were also excluded, as these would be a major source of between-trial heterogeneity.

Identification of the control group and variables of interest
Because the nature of the randomized intervention could contribute to heterogeneity, we focused on the control groups. For each trial, we identified the control group as defined by the authors as ‘control group’, ‘usual care group’, or a variation thereof. When no control group could be identified (in a comparison of two usual care therapies) we defined the control group as the means of the two groups in terms of sample size, mortality, and baseline characteristics. A sensitivity analysis was performed towards this construct by analyzing whether trials with and without specifically defined control groups differed in terms of mean mortality or the amount of between-trial heterogeneity.

For each trial, we recorded the type of intervention, single- or multicenter design, and the primary endpoint. Trials were graded according to the Jadad scale [14]. For the control group in each trial, we recorded the sample size, the reported baseline characteristics, and the mortality rates.

Estimation of heterogeneity in mortality rates and associations with population characteristics
We used 28-day mortality throughout all analyses. For trials that did not report this outcome, we estimated 28-day mortality based on reported hospital, ICU, or 90-day mortality using linear regression with data from trials that reported both 28-day and another mortality measure.

To analyze mortality rates across trials we used a random-effects meta-regression model with the log odds of
mortality as dependent variable and a random intercept for each study. Each trial was weighted by the inverse of the sampling variance of the mortality rates. A maximum likelihood estimator was used to estimate the mean mortality (random effects pooled estimate), the between-study standard deviation due to heterogeneity ($\tau$), and the percentage of variation due to heterogeneity rather than change ($I^2$). To quantify between-trial heterogeneity, we report the 95% prediction interval (mean mortality $\pm 1.96 \tau$), which represents the distribution of estimated future mortality rates based on observed mortalities weighted by sampling variance (trial size) and corrected for random chance [15]. In the absence of between-study heterogeneity, the 95% prediction interval is equal to the 95% confidence interval, but when significant heterogeneity is present the prediction interval estimates the bandwidth of expected mortality rates from similar studies [15, 16]. In other words, the 95% prediction interval can be thought of as the estimate of true between-study distribution of mortality rates. The prediction interval can therefore be used to guide power calculations for future studies [16].

The between-trial heterogeneity in mortality rates was calculated for subcategories of trials employing different inclusion criteria: confirmed or suspected infection; confirmed infection only; different definitions of hypotension; mandatory hyperlactatemia; mandatory vasoconstrictor therapy; and mandatory mechanical ventilation. Differences in mortality rates between subcategories were calculated by addition of dummy variables to the mixed-effects model.

To estimate the association between study and population characteristics and mortality, these variables were added to the model as covariates. Residuals were checked for normality with Q–Q plots, and the goodness of fit of the log-linear model was compared with quadratic and power models by selecting the model with the lowest Akaike information criterion (AIC). To facilitate comparisons between variables, we report standardized regression coefficients ($\beta$) and the proportion of between-trial variability in mortality explained by the population variable (unadjusted $R^2$) for all univariate analyses.

**Predicting mortality rates using a linear model and recursive partitioning**

We then constructed a comprehensive model to predict between-study differences in mortality. Population variables that were reported by at least 25% of the included trials with a univariate regression $R^2 \geq 0.10$ were included as regressors in a multivariate model and removed in a stepwise manner for $P$ values $\geq 0.05$. The threshold $R^2$ of 0.10 was a compromise between the number of variables and the limited number of observations. This model selection process was not prospectively protocolized as the number of eligible variables could not be estimated a priori. Multiple imputation (generating 20 datasets) with predictive mean matching was used for missing observations (i.e., missing population characteristics). The imputation methods are further described in section 7 of the ESM.

As a complementary approach to predict 28-day mortality rates from population characteristics, we constructed a regression tree model based on recursive partitioning (a machine learning algorithm) [17, 18] for its ability to handle partially missing observations (obviating the need for imputation) and its robustness to non-linear relations. We set up the model to predict 28-day mortality based on all inclusion criteria and population characteristics. In short, the recursive partitioning algorithm selected the most informative variable, which was then ‘split’ at the value that best differentiates low from high mortality. The algorithm then selected the most informative variable for each of the two resulting subgroups, and split it again. When a splitting variable was missing for a specific trial, a surrogate variable (the variable most closely correlated to the splitting variable) was used. After multiple splits, this recursive partitioning resulted in a regression tree (similar to a decision tree) with subgroups of trials ranked from low to high expected mortality. $R^2$ represents the variance in mortality explained by the decision tree. Overfitting was examined using the cross-validated error.

For all analyses, $P < 0.05$ was considered significant. The analyses were performed in R version 3.4.2 using the metafor, mice and rpart packages [19–21].

**Results**

**Characteristics of the included trials**

The search resulted in 65 trials that met all inclusion and exclusion criteria (eFigure 1 in the ESM), representing a total of 8634 control group patients [6, 7, 22–84]. A list of excluded trials is available in the ESM. The trial characteristics are presented in Table 1.

Twenty trials (31%) did not report 28-day mortality but only hospital mortality, ICU mortality, or 90-day mortality. Using trials that reported multiple mortality measures, 28-day mortality was estimated as a linear function of hospital mortality, ICU mortality, or 90-day mortality ($R^2$ values 0.99, 0.98, and 0.98, respectively). The estimates and validation plots are presented in eTable 1 and eFigure 2 of the ESM.

In 14 trials (21%) the control group could not be identified because two usual care therapies were compared. For these trials, the control group characteristics and mortality rates were defined as the means of the two treatment
none of these 14 trials reported significant mortality differences between the treatment groups.

The distribution of mortality rates
The control group mortality rates ranged between 13.8 and 84.6%, with a random-effects estimated mean mortality rate of 38.6%. There was significant heterogeneity among trials ($I^2 = 93\%, \tau = 0.710, p < 0.0001$), and the 95% prediction interval was 13.5–71.7%.

Figure 1 shows the mortality rates of trials categorized by inclusion criteria. The mean mortality rate did not differ between trials with different definitions of hypotension, infection (confirmed vs. suspected), or vasopressor or mechanical ventilation inclusion criteria. There were no significant differences in mean mortality rate or in heterogeneity between large vs. small trials, monocenter vs. multicenter trials, unblinded vs blinded trials, high-quality trials vs. low-quality trials, or trials with vs. without a specifically defined control group (eTable 2 in the ESM).

The exclusion criteria employed in the trials were too diverse for statistical analysis, but the total number of exclusion criteria (ranging from 0 to 30) was inversely associated with the mortality rate ($\beta = -0.375, R^2 = 0.14, P = 0.007$).

The heatmap in Fig. 2 provides an overview of the between-trial differences in mortality rates and population characteristics. The log-linear associations between the mortality rate and reported control group baseline characteristics are presented in Table 2 (goodness-of-fit statistics are reported in eTable 3 in the ESM). There was no significant decrease in mortality over the period 2006–2018, with only ($R^2$) 4% of heterogeneity explained by the year of publication (Table 2, eFigure 3). Baseline variables that were univariately associated with mortality were: mean Sequential Organ Failure Assessment (SOFA) score, the proportion of patients on mechanical ventilation, the proportion of patients on vasopressors, and mean serum creatinine. Regression plots of selected associations are shown in eFigure 3 of the ESM.

Predicting mortality rates from population characteristics
Details of the variable selection process for the multivariate model are available in section 7 of the ESM. Significant independent variables in the final multivariate model were: baseline mean SOFA score ($\beta = 0.39$, standardized standard error (SSE) = 0.17, $P = 0.019$), the proportion of patients on mechanical ventilation ($\beta = 0.42$, SSE = 0.18, $P = 0.019$), and mean serum creatinine ($\beta = 0.31$, SSE = 0.10, $P = 0.0015$). The multivariate model $R^2$ was 0.41 with significant residual heterogeneity ($I^2 = 82\%, \tau = 0.544, P < 0.0001$). Figure 3 shows the predicted and actual mortality rates of the included trials.

The recursive partitioning algorithm resulted in a regression tree with the following variables as informative determinants of the mortality rate: mean age (split at 64.8 years); the proportion of patients with a respiratory infection (split at 54.5%); the proportion of patients on mechanical ventilation (split at 74.3%); and the proportion of male patients (splits at 63.8 and 53.8%). The $R^2$ value of the regression tree was 0.42. The cross-validated relative error decreases to below the root (split 0) value, which indicates that the tree was not overfitted. The results from the regression tree analysis are further described in eFigures 4 and 5 of the ESM (section 7).

Discussion
In this analysis of 65 septic shock trials published in the past decade, we found a statistically significant and clinically relevant amount of heterogeneity in control group mortality rates. The mean mortality rate was 38.6% with

| Table 1 Characteristics of included trials |
|-------------------------------------------|
| **No. (%) or median (IQR)**                |
| Number of included trials                 | 65 |
| Control group sample size: median (IQR)   | 34 (20–100) |
| Multicenter trials: n (%)                 | 28 (43) |
| **Trial country: n (%)**                  |
| France                                    | 12 (18) |
| China                                     | 9 (14)  |
| Italy                                     | 8 (12)  |
| USA                                       | 6 (9)   |
| India                                     | 3 (5)   |
| The Netherlands                           | 3 (5)   |
| UK                                        | 3 (5)   |
| Other countries (1 each)                  | 13 (20) |
| Multinational trials                      | 9 (14)  |
| **Trial intervention: n (%)**             |
| Drug                                      | 44 (68) |
| Treatment bundle                          | 14 (21) |
| Device                                    | 7 (11)  |
| **Primary endpoint: n (%)**               |
| Mortality                                 | 21 (32) |
| Other                                     | 32 (49) |
| Not specified                             | 12 (18) |
| Jadad scale: median (IQR)                 | 3 (2–4) |
| Jadad scale components: n (%)             | 65 (100) |
| Randomization                             | 65 (100) |
| Randomization appropriate                 | 45 (69) |
| Blinding                                  | 23 (35) |
| Blinding appropriate                      | 19 (29) |
| Description of withdrawals and dropouts   | 42 (65) |

IQR interquartile range
estimated 95% prediction limits of 13.5–71.7%, revealing a wide range in underlying mortality rates after discounting the effects of random change and small trials.

In contrast to findings from large observational studies that the mortality of sepsis has decreased in the past decade, we found only a small nonsignificant decline in the period 2006–2018 [85, 86]. Different inclusion definitions of septic shock did not affect mean mortality rates, but a higher total number of exclusion criteria was associated with lower mortality. We used three statistical methods to analyze the association between population characteristics and mortality.

The univariate associations reflect how the reader of a trial report could interpret the population characteristics in relation to the mortality rate, and shows that the proportion of ventilated patients, mean SOFA score, and the proportion of patients on vasopressor support were most informative (i.e. have highest standardized regression coefficients).

The multivariate linear model (with missing observations imputed) shows which combinations of characteristics were predictive of mortality if all trials hypothetically reported the same variables. A combination of three independently significant characteristics (mean SOFA score, proportion of ventilated patients, and mean creatinine) explained only 41% of the heterogeneity in mortality rates across trials.

The recursive partitioning algorithm, which is not limited by dependence on multiple imputation and the assumption of linearity, shows which characteristics were most informative, given that different trials report different characteristics. The resulting regression tree explained only 42% of the heterogeneity in mortality.

The linear model and the regression tree arrived at different predictor variables because the linear model is biased towards more informative linear associations, while the regression tree allows for nonlinear relations and is biased towards variables with less missing data.

In all, these results indicate that there are clinically significant between-trial differences in control group mortality rates, and that these differences are not associated with differences in inclusion criteria and only weakly associated with reported baseline characteristics. Visual inspection of the heatmap (Fig. 2) shows that there are no unambiguous patterns in the relation between population characteristics and mortality rates. This heterogeneity is reflected in our finding that different statistical methods result in different predictive variables.

Possible sources of residual heterogeneity
Residual heterogeneity among trials may be caused by population differences in nutrition and socio-economic status, heterogenous exclusion criteria, incomplete reporting, between-trial differences in variable definitions, the timing of randomization, and differences in post-randomization co-interventions and standards of care.

We found that no single measure of chronic comorbidity was reported in more than 40% of the included trials and that characteristics of causative pathogens were
| Control-group mortality rates | Chronic comorbidities | Severity of illness scores | Characteristics of acute illness | Infection site characteristics | Culture results |
|--------------------------------|-----------------------|---------------------------|--------------------------------|-------------------------------|-----------------|
| Torraco 2014 (76) n=13         |                       |                           |                                |                               |                 |
| Ngaoonwan 2017 (58) n=37       |                       |                           |                                |                               |                 |
| Morelli 2013 (77) n=77         |                       |                           |                                |                               |                 |
| Cicarelli 2007 (28) n=15       |                       |                           |                                |                               |                 |
| Juffermans 2010 (41) n=4       |                       |                           |                                |                               |                 |
| Morelli 2010 (56) n=20         |                       |                           |                                |                               |                 |
| Morelli 2008 (53) n=20         |                       |                           |                                |                               |                 |
| Morelli 2009 (63) n=20         |                       |                           |                                |                               |                 |
| Lin 2006 (44) n=116            |                       |                           |                                |                               |                 |
| Puskarich 2014 (63) n=15       |                       |                           |                                |                               |                 |
| Chakraborty 2007 (26) n=25     |                       |                           |                                |                               |                 |
| Singh 2017 (73) n=37           |                       |                           |                                |                               |                 |
| Hyvernat 2016 (40) n=59        |                       |                           |                                |                               |                 |
| Morelli 2008 (54) n=16         |                       |                           |                                |                               |                 |
| Garg 2016 (34) n=18            |                       |                           |                                |                               |                 |
| Hajie 2017 (37) n=10           |                       |                           |                                |                               |                 |
| Babhio 2012 (255) n=23         |                       |                           |                                |                               |                 |
| Patel 2010 (60) n=126          |                       |                           |                                |                               |                 |
| Richard 2016 (66) n=20         |                       |                           |                                |                               |                 |
| Forceville 2007 (333) n=29     |                       |                           |                                |                               |                 |
| Chen 2015 (277) n=41           |                       |                           |                                |                               |                 |
| Livigni 2014 (455) n=93        |                       |                           |                                |                               |                 |
| Donnino 2016 (311) n=45        |                       |                           |                                |                               |                 |
| McIntyre 2008 (49) n=20        |                       |                           |                                |                               |                 |
| Schortgen 2012 (700) n=100     |                       |                           |                                |                               |                 |
| Russell 2008 (68) n=382        |                       |                           |                                |                               |                 |
| Trof 2012 (777) n=38           |                       |                           |                                |                               |                 |
| Shum 2014 (722) n=8            |                       |                           |                                |                               |                 |
| van Genderen 2015 (79) n=15    |                       |                           |                                |                               |                 |
| Morelli 2007 (52) n=10         |                       |                           |                                |                               |                 |
| COITSS 2010 (29) n=254         |                       |                           |                                |                               |                 |
| Huh 2011 (38) n=65            |                       |                           |                                |                               |                 |
| Asfar 2014 (24) n=388          |                       |                           |                                |                               |                 |
| Schmoei 2006 (69) n=20         |                       |                           |                                |                               |                 |
| Merouani 2008 (51) n=20        |                       |                           |                                |                               |                 |
| Annane 2013 (233) n=203        |                       |                           |                                |                               |                 |
| Annane 2007 (222) n=166        |                       |                           |                                |                               |                 |
| Quenot 2015 (64) n=31          |                       |                           |                                |                               |                 |
| Yu 2015 (83) n=22              |                       |                           |                                |                               |                 |
| Hjortrup 2016 (38) n=76        |                       |                           |                                |                               |                 |
| Lv 2017 (18) n=60             |                       |                           |                                |                               |                 |
| Sprung 2008 (74) n=248         |                       |                           |                                |                               |                 |
| Palazz 2009 (59) n=66          |                       |                           |                                |                               |                 |
| Li 2016 (43) n=83             |                       |                           |                                |                               |                 |
| Seguini 2006 (71) n=10         |                       |                           |                                |                               |                 |
| Li 2015 (42) n=24             |                       |                           |                                |                               |                 |
| van Haren 2012 (80) n=12        |                       |                           |                                |                               |                 |
| Dubin 2010 (32) n=10           |                       |                           |                                |                               |                 |
| Zhou 2017 (84) n=180          |                       |                           |                                |                               |                 |
| Gordon 2016 (36) n=200         |                       |                           |                                |                               |                 |
| Lu 2015 (47) n=52              |                       |                           |                                |                               |                 |
| Mouncey 2015 (57) n=621        |                       |                           |                                |                               |                 |
| Ranieri 2012 (85) n=834        |                       |                           |                                |                               |                 |
| Gordon 2014 (35) n=30          |                       |                           |                                |                               |                 |
| Venkatesh 2018 (81) n=1840     |                       |                           |                                |                               |                 |
| Trzeciak 2014 (78) n=23        |                       |                           |                                |                               |                 |
| Stephens 2008 (75) n=83        |                       |                           |                                |                               |                 |
| Russell 2017 (67) n=19         |                       |                           |                                |                               |                 |
| Dong 2015 (30) n=45            |                       |                           |                                |                               |                 |
| Paven 2015 (61) n=113          |                       |                           |                                |                               |                 |
| Memis 2009 (50) n=20           |                       |                           |                                |                               |                 |
| ProCESS 2014 (62) n=456        |                       |                           |                                |                               |                 |
| Loisa 2006 (46) n=23           |                       |                           |                                |                               |                 |
| ARSE 2014 (7) n=707           |                       |                           |                                |                               |                 |

Fig. 2. Heatmap of included trials (n = 65) and associated baseline characteristics, ranked by decreasing mortality rates. White tiles represent the mean value across trials, while red and blue tiles are indicative of higher and lower than average values, respectively. Gray tiles (N/A) are variables that were not reported. The 28-day mortality rate ranged between 13.8 and 84.6%, with a mean of 38.6%. APACHE Acute Physiology and Chronic Health Evaluation, SAPS Simplified Acute Physiology Score, SOFA Sequential Organ Failure Assessment score, MAP mean arterial pressure, CVP central venous pressure, CNS central nervous system. (Asterisk) Variables with a significant univariate association with 28-day mortality.
Table 2  Univariate associations between mortality rates and reported mean or median population characteristics

| Trials reporting variable (% of n = 56) | Mean (SD) | Standardized regression coefficient β (R²) | P value |
|----------------------------------------|-----------|--------------------------------------------|---------|
| **Publication year**                   |           |                                            |         |
|                                        | 65 (100)  | 2013.3 (3.58)                              | − 0.19 (0.04) | 0.197 |
| **Age, years**                         | 64 (98)   | 62.9 (3.80)                                | 0.18 (0.03) | 0.160 |
| **Male patients %**                    | 63 (97)   | 60.5 (5.80)                                | 0.02 (0.00) | 0.927 |
| **Comorbidity characteristics**       |           |                                            |         |
| Charlson Comorbidity Index             | 5 (8)     | 1.90 (1.11)                                | 0.52 (0.27) | 0.183 |
| From long-term care facility %         | 6 (9)     | 5.8 (5.6)                                  | 0.44 (0.20) | 0.312 |
| McCabe class I %                       | 6 (9)     | 34.1 (15.2)                                | − 0.40 (0.16) | 0.374 |
| McCabe class II %                      | 6 (9)     | 14.7 (12.9)                                | 0.02 (0.00) | 0.948 |
| McCabe class III %                     | 4 (6)     | 16.2 (15.0)                                | 0.71 (0.50) | 0.120 |
| Diabetes mellitus %                    | 23 (36)   | 24.4 (6.88)                                | 0.01 (0.00) | 0.856 |
| Heart failure or coronary disease %    | 26 (40)   | 20.7 (8.7)                                 | 0.33 (0.11) | 0.133 |
| Chronic obstructive pulmonary disease %| 25 (39)   | 15.1 (6.3)                                 | 0.04 (0.00) | 0.911 |
| Chronic renal disease %                | 21 (33)   | 7.6 (5.0)                                  | 0.06 (0.00) | 0.773 |
| Chronic liver disease %                | 17 (26)   | 5.5 (2.8)                                  | 0.25 (0.06) | 0.320 |
| Cancer %                               | 20 (31)   | 21.2 (8.1)                                 | 0.19 (0.03) | 0.426 |
| **Severity of illness scores**         |           |                                            |         |
| APACHE II score                        | 33 (51)   | 22.5 (3.65)                                | 0.21 (0.05) | 0.376 |
| APACHE III score                       | 1 (2)     | −                                          | −        | −       |
| APACHE IV score                        | 1 (2)     | −                                          | −        | −       |
| SAPS II score                          | 24 (37)   | 55.7 (4.42)                                | 0.36 (0.13) | 0.079 |
| SAPS III score                         | 3 (4)     | 77.6 (1.91)                                | 0.01 (0.00) | 0.644 |
| SOFA score                             | 37 (58)   | 9.59 (2.47)                                | 0.57 (0.33) | 0.007**|
| **Characteristics of acute illness**   |           |                                            |         |
| Medical (non-surgical) %               | 22 (34)   | 69.7 (13.1)                                | 0.26 (0.07) | 0.314 |
| Time from diagnosis to randomization, hours | 13 (20)   | 13.77 (8.84)                              | 0.47 (0.22) | 0.069 |
| Mechanical ventilation %               | 33 (51)   | 78.1 (28.3)                                | 0.61 (0.38) | 0.0005***|
| Heart rate, 1/min                      | 39 (60)   | 104 (8.8)                                 | 0.13 (0.02) | 0.435 |
| Mean arterial pressure, mmHg           | 43 (66)   | 70.7 (6.65)                                | 0.06 (0.00) | 0.561 |
| Central venous pressure, mmHg          | 22 (34)   | 11.2 (2.21)                                | 0.17 (0.03) | 0.425 |
| Vasopressor support %                  | 38 (58)   | 84.6 (30.0)                                | 0.57 (0.32) | 0.0019**|
| Serum lactate, mmol/l                  | 52 (80)   | 4.00 (1.28)                                | − 0.13 (0.02) | 0.389 |
| Serum creatinine, µmol/l               | 26 (40)   | 168 (31.1)                                 | 0.48 (0.23) | 0.007**|
| Fluids before randomization, ml        | 19 (30)   | 3209 (1637)                                | 0.31 (0.10) | 0.194 |
| **Infection site characteristics**     |           |                                            |         |
| Respiratory %                          | 53 (82)   | 42.6 (13.7)                                | 0.27 (0.08) | 0.087 |
| Abdominal %                            | 51 (78)   | 24.0 (15.0)                                | 0.06 (0.00) | 0.686 |
| Urogenital %                           | 41 (63)   | 11.3 (5.7)                                 | − 0.27 (0.07) | 0.094 |
| Central nervous system %               | 19 (30)   | 1.2 (1.6)                                  | 0.03 (0.00) | 0.885 |
| Skin and soft tissue %                 | 26 (43)   | 6.8 (3.6)                                  | − 0.09 (0.01) | 0.803 |
| bloodstream %                          | 32 (49)   | 12.9 (8.2)                                 | − 0.11 (0.01) | 0.487 |
| **Pathogen characteristics**           |           |                                            |         |
| Gram-negative %                        | 25 (39)   | 32.0 (16.1)                                | 0.41 (0.17) | 0.0573 |
| Gram-positive %                        | 22 (34)   | 24.6 (7.12)                                | − 0.41 (0.17) | 0.083 |
| Other pathogen %                       | 22 (34)   | 44.0 (23.3)                                | − 0.13 (0.02) | 0.473 |
| Culture negative %                     | 18 (28)   | 29.4 (8.3)                                 | − 0.38 (0.14) | 0.085 |

Univariate associations between control group mortality rate and commonly reported mean baseline characteristics. Associations were estimated using a weighted random-effects model on the log-odds scale. Some baseline characteristics were reported by a minority of trials, which resulted in low power to detect a significant association. R² can be interpreted as the proportion of heterogeneity that is explained by the population characteristic for the n trials that report that characteristic.

APACHE Acute Physiology and Chronic Health Evaluation score, SAPS Simplified Acute Physiology score, SOFA Sequential Organ Failure Assessment score.
reported in only 28–39% of trials. This compromised the power of our analysis to detect associations across all trials, but, more importantly, it also prevents readers of trial reports from evaluating and comparing populations among trials and from judging to what extent a trial population corresponds to the population under their care.

Another source of heterogeneity is the imprecise definition of many variables. It is unclear whether a variable like ‘pre-existing kidney disease’ in one trial has the same meaning as ‘chronic renal insufficiency’ in another trial. Minor variations in variable definitions and data capture methods have been shown to lead to significantly different septic shock populations and to inter-observer variability in severity-of-illness scoring systems [5, 87, 88]. The importance of this ‘fine print’ in defining a population does not receive due attention in the methods section of most trials.

The time of inclusion may be an additional source of heterogeneity. Patients recruited later after the diagnosis of septic shock have not responded to treatment in an earlier phase and are therefore likely to have a worse prognosis. Only 13 trials reported the time from diagnosis to randomization, and for those trials it explained 22% of the heterogeneity.

While we have focused on inclusion criteria and baseline characteristics, the prognosis of septic shock may be largely influenced by post-randomization standards of care and co-interventions. Unfortunately, co-interventions and (control group) treatment standards are often described as ‘according to the Surviving Sepsis Campaign guidelines’ or not discussed at all in trial reports. Variables describing important post-randomization interventions, such as red blood cell transfusions, vasopressor dose, or fluid balance were recently found to be reported in only 33, 17, and 13% of large septic shock trials, respectively [89].

We did not analyze the association between trial countries and the mortality rate because many countries are represented by a single trial in the present sample. Nevertheless, between-country differences in standards of care or access to early healthcare may account for part of the residual heterogeneity. Large international observational studies are a more appropriate instrument for the investigation of differences in mortality rates among countries.

Implications for investigators and clinicians
Clinicians demand of clinical trials that they are relevant, reproducible, and generalizable to a clearly defined patient population. The results of this study indicate that many of the baseline characteristics upon which clinicians rely to gauge the applicability of trial results to their practice are in fact only weakly or not at all associated with mortality outcomes across trials.

The association between the number of exclusion criteria and mortality suggests that many seemingly inconsequential criteria together may have a significant effect on the composition of a trial population. Investigators should therefore be aware of this phenomenon in the design phase of a trial, as it affects the generalizability and external validity of trial results.

The wide prediction limits of control-group mortality have consequences for sample size calculations. Detecting a relative risk reduction of 25% with 80% power requires 245 patients if mortality is estimated to be 71.7%, while it requires 795 patients if control group mortality is 38.6% or 2980 patients if mortality is 13.5%. In practice, misestimation of the mortality rate by more than 7.5% occurred in 65% of critical care trials [11]. We therefore suggest that sample size calculations should not be based on the mean of reported control-group mortality rates in the literature but should be robust towards a wider range of expected event rates.

Reproducibility and generalizability also require a common phenomenological structure with respect to diagnostic definitions, inclusion criteria, patient characteristics, concomitant treatment, and outcomes. A recent review of large septic shock trials found that only half of the information deemed necessary for evaluation of the control group was reported in the investigated trials [89]. In the present study, we now find that many of the reported characteristics are not associated with control-group mortality rates, possibly due to variations in variable definitions.

The third consensus definitions for sepsis and septic shock were partly developed to harmonize the inclusion criteria for clinical studies [3]. We were unable to analyze a subset of trials with populations that might fit the Sepsis-3 septic shock definition, as none of the included trials employed both delta SOFA score and vasopressor inclusion criteria. We do note that SOFA score is independently associated with mortality rates, although baseline SOFA explains only 33% ($R^2$) of the variation in mortality rates in the 37 trials that report it. Furthermore, we found significant heterogeneity within subsets of trials employing similar inclusion criteria (Fig. 2).

We suggest that an international consensus is necessary to standardize variable definitions, data collection, and reporting of patient characteristics and outcomes for sepsis trials, as has been proposed before [89–92]. The feasibility of harmonizing study protocols has been demonstrated in three large trials investigating early goal-directed therapy [93]. The present results indicate that SOFA score, the proportion of ventilated patients, and creatinine independently reflect baseline risk across trials and should therefore be reported for each trial.
The results from this study also support the practice of data sharing, as we have shown that aggregated population characteristics are less informative than expected. Sharing individual patient data will not only increase the power to detect treatment effects across multiple studies but can also be used to test the generalizability of trial results vis-à-vis large cohorts with septic shock.

Strengths and limitations

This study was performed with a prospectively registered protocol and analysis plan. We chose to include only trials published between 2006 and 2018 to minimize the influence of long-term secular trends in septic shock diagnosis, treatment, and mortality [94, 95]. The search strategy was broad and comprehensive, but we excluded 40 trial reports not written in English, which compromised power and generalizability. We excluded trials that recruited only septic shock patients with specific organ dysfunction (such as kidney or liver failure) to rule out this source of between-trial heterogeneity.

For 20 trials, 28-day mortality was estimated using another reported mortality rate. Although the prediction equations were very precise (R² values ≥ 0.98), we cannot rule out the possibility that this influenced the results. Excluding these 20 trials would have eroded the power of the study.

Importantly, using study-level data means that, to avoid the ecological fallacy, we cannot make inferences about predictive characteristics at the individual patient level, although several predictor variables are known to be individually associated with mortality (e.g. high SOFA score as a risk factor [96, 97]). The fact that there was substantial variation in the reporting of baseline variables was an important finding in itself, but also limited our power to detect associations across trials. A more in-depth investigation into the heterogeneity among trial populations would require individual patient data, but we think that obtaining such data would lead to significant selection bias.

Conclusion

Septic shock is a syndrome with various etiologies, biochemical characteristics, and phenotypes [9, 98]. Onto this inherently heterogeneous syndrome, a layer of investigator-induced heterogeneity is added when trials employ different inclusion criteria, report different variables, and use different variable definitions. This compounded complexity causes heterogeneity among trial populations that may go unnoticed. We have shown that control-group mortality rates are very dissimilar across trials, and that the majority of this heterogeneity remains unexplained after accounting for reported population characteristics. The lack of standardized reporting
limits the usefulness of the variables explaining the mortality differences found in this study. In all, the substantial between-trial heterogeneity limits the reproducibility and generalizability of septic shock research and may inhibit the discovery of beneficial therapies for specific (sub)populations. The findings of this study therefore strongly support the argument for profound standardization and harmonization of septic shock trial reporting as well as data-sharing policies to test the external validity of trial populations.

Electronic supplementary material

The online version of this article (https://doi.org/10.1007/s00134-018-5134-8) contains supplementary material, which is available to authorized users.

Author details
1 Department of Intensive Care, VU University Medical Center, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands. 2 Department of Anaesthesiology, VU University Medical Center, Amsterdam, The Netherlands. 3 Unité de Biostatistique et de Recherche Clinique, Centre Hospitalier Universitaire de Caen, Caen, France. 4 EA2656 Groupe de Recherche sur l’Adaptation Microbienne (GRAM 2.0), Université Caen Normandie, Caen, France.

Author contributions
HJdG and AG designed the study. HJdG and JP performed the study screening and extracted the data. HJdG performed the statistical analyses. SL, HO, and AG provided technical, material, and administrative support. HJdG drafted the manuscript. JP, SL, JJP, HO and AG revised the manuscript for important intellectual content.

Compliance with ethical standards

Conflicts of interest

All authors declare that they have no conflicts of interest.

Open Access

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by‑nc/4.0/), which permits any noncommercial use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Received: 13 December 2017 Accepted: 17 February 2018

Published online: 15 March 2018

References

1. Bone RC, Balk RA, Cerra FB et al (1992) Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine. Chest 101:1644–1655
2. Levy MM, Fink MP, Marshall JC et al (2003) 2001 SCSM/ESICM/ACCP/ATS/SIS international sepsis definitions conference. Intensive Care Med 29:530–538. https://doi.org/10.1007/s00134-003-1662-4
3. Singer M, Deutschman CS, Seymour CW et al (2016) The third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA 315:801–810. https://doi.org/10.1001/jama.2016.0287
4. Dellinger RP, Carlet JM, Masur H et al (2004) Surviving sepsis campaign guidelines for management of severe sepsis and septic shock. Intensive Care Med 30:536–555. https://doi.org/10.1007/s00134-004-2210-z
5. Shankar-Hari M, Phillips GS, Levy ML et al (2016) Developing a new definition and assessing new clinical criteria for septic shock: for the third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA 315:775–787. https://doi.org/10.1001/jama.2016.0289
6. ARISE Investigators, ANZICS Clinical Trials Group, Peake SL et al (2014) Goal-directed resuscitation for patients with early septic shock. N Engl J Med 371:1496–1506. https://doi.org/10.1056/NEJMoa1404380
7. Morelli A, Emmer C, Westphal M et al (2013) Effect of heart rate control with esmolol on hemodynamic and clinical outcomes in patients with septic shock: a randomized clinical trial. JAMA 310:1683–1691. https://doi.org/10.1001/jama.2013.278477
8. Vincent JL, Hall JB, Suter CM (2015) Ten big mistakes in intensive care medicine. Intensive Care Med 41:505–507. https://doi.org/10.1007/s00134-014-3570-7
9. Marshall JC (2014) Why have clinical trials in sepsis failed? Trends Mol Med 20:195–203. https://doi.org/10.1016/molmed.2014.01.007
10. Opal SM, Dellinger RP, Vincent J-L et al (2014) The next generation of sepsis clinical trial designs: what is next after the demise of recombinant human activated protein C? Crit Care Med 42:1714–1721. https://doi.org/10.1097/CCM.0000000000000325
11. Harhay MO, Wagner J, Ratcliffe SJ et al (2014) Outcomes and statistical power in adult critical care randomized trials. Am J Respir Crit Care Med 189:1469–1478. https://doi.org/10.1164/rcrm.2014-0056CP
12. de Grooth HJ, Oudemans-van Straaten HM, Gerbes AB (2016) Heterogeneity in a decade of septic shock RCTs: a systematic analysis of mortality outcomes. PROSPERO 2016 CRD42016049330. https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=49330. Accessed 3 Nov 2016
13. Moher D, Liberati A, Tetzlaff J et al (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 339:b2535
14. Jadad AR, Moore RA, Carroll D et al (1996) Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 17:1–12. https://doi.org/10.1016/0197-2456(95)00134-4
15. Deeks JJ, Higgins JPT, Altman DG (2011) Section 9.5: heterogeneity. Cochrane handbook for systematic reviews of interventions version 5.1.0 (updated March 2011)
16. InHout J, Ioanidis JPA, Rovers MM, Goeman JJ (2016) Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 6:e010247. https://doi.org/10.1136/bmjopen-2015-010247
17. Viechtbauer W (2010) Conducting meta-analyses in R with the metafor package. J Stat Softw. https://doi.org/10.18637/jss.v036.i03
18. Van Buuren S, Groothuis-Oudshoorn K (2011) mice: Multivariate imputation by chained equations in R. J Stat Softw 45:1–67. https://doi.org/10.18637/jss.v045.i03
19. Therneau T, Atkinson B, Ripley B (2017) rpart: recursive partitioning and regression trees. CRC Press, Boca Raton, Florida, USA
20. Atkinson EJ, Therneau TM (2017) An introduction to recursive partitioning using the RPART routines. Mayo Found, Rochester, Minnesota, USA
21. Viechtbauer W (2010) Conducting meta-analyses in R with the metafor package. J Stat Softw. https://doi.org/10.18637/jss.v036.i03
22. Anstis N, Vignon P, Renault A et al (2007) Comparison of norepinephrine vs dopamine in severe sepsis. Am J Respir Crit Care Med 175:801–810. https://doi.org/10.1164/rccm.200605‑4678OC
23. Meziari M, D’Alonzo A, Meziari F et al (2014) Comparison of norepinephrine vs dopamine in severe sepsis. Am J Respir Crit Care Med 190:1583–1593. https://doi.org/10.1164/rccm.201401‑0056OC
24. Asfar P, Meziari M, D’Alonzo A et al (2014) Comparison of norepinephrine vs dopamine in severe sepsis. Am J Respir Crit Care Med 189:1091–1097. https://doi.org/10.1164/rccm.201311‑2020OC
25. Asfar P, Meziari M, Hamel J-F et al (2014) High versus low blood-pressure target in patients with septic shock. N Engl J Med 370:1583–1593. https://doi.org/10.2105/NEJMoa1312173
26. Bahroul M, Chaari A, Ben Mbarek MN et al (2012) Use of heparin and hydrochloride for catecholamine weaning in septic shock. Am J Ther 19:e8–e17. https://doi.org/10.1097/MTJ.0b013e3181e9b630
27. Chakraborty A, Mathur S, Dhunna R (2007) Comparison of norepinephrine and dopamine in the management of septic shock using impedance cardiography. Indian J Crit Care Med 11:186. https://doi.org/10.1001/jama.2013.278477
28. Chen C, Kollef MH (2015) Targeted fluid minimization following initial resuscitation in septic shock: a pilot study. Chest 148:1462–1469. https://doi.org/10.1378/chest.15‑1525
29. Cicaelli DD, Veira JE, Bensehor FEM (2007) Early dexamethasone treatment for septic shock patients: a prospective randomized clinical trial. São Paulo Med J 125:237–241
