# Heterogeneity in the Effect of Early Goal-Directed Therapy for Septic Shock: A Secondary Analysis of Two Multicenter International Trials

Faraaz Ali Shah, MD, MPH[1,2]
Victor B. Talisa, PhD[3]
Chung-Chou H. Chang, PhD[4,5]
Sofia Triantafyllou, PhD[6]
Lu Tang, PhD[5]
Florian B. Mayr, MD, MPH[2,3]
Alisa M. Higgins, PhD[7]
Sandra L. Peake, MD, PhD[7]
Paul Mouncey, MS[8]
David A Harrison, PhD[8]
Kimberley M. DeMerle, MD, MS[1]
Jason N. Kennedy, MS[3]
Gregory F. Cooper, MD, PhD[9]
Rinaldo Bellomo, MD[7,10,11,12,13]
Kathy Rowan, PhD[8]
Donald M. Yealy, MD[14]
Christopher W. Seymour, MD, MSc[3]
Derek C. Angus, MD, MPH[3]
Sachin P. Yende, MD, MS[2,3]

**OBJECTIVES:** The optimal approach for resuscitation in septic shock remains unclear despite multiple randomized controlled trials (RCTs). Our objective was to investigate whether previously uncharacterized variation across individuals in their response to resuscitation strategies may contribute to conflicting average treatment effects in prior RCTs.

**DESIGN:** We randomly split study sites from the Australian Resuscitation of Sepsis Evaluation (ARISE) and Protocolized Care for Early Septic Shock (ProCESS) trials into derivation and validation cohorts. We trained machine learning models to predict individual absolute risk differences (iARDs) in 90-day mortality in derivation cohorts and tested for heterogeneity of treatment effect (HTE) in validation cohorts and swapped these cohorts in sensitivity analyses. We fit the best-performing model in a combined dataset to explore roles of patient characteristics and individual components of early goal-directed therapy (EGDT) to determine treatment responses.

**SETTING:** Eighty-one sites in Australia, New Zealand, Hong Kong, Finland, Republic of Ireland, and the United States.

**PATIENTS:** Adult patients presenting to the emergency department with severe sepsis or septic shock.

**INTERVENTIONS:** EGDT vs. usual care.

**MEASUREMENTS AND MAIN RESULTS:** A local-linear random forest model performed best in predicting iARDs. In the validation cohort, HTE was confirmed, evidenced by an interaction between iARD prediction and treatment ($p < 0.001$). When patients were grouped based on predicted iARDs, treatment response increased from the lowest to the highest quintiles (absolute risk difference [95% CI], −8% [−19% to 4%] and relative risk reduction, 1.34 [0.89–2.01] in quintile 1 suggesting harm from EGDT, and 12% [1–23%] and 0.64 [0.42–0.96] in quintile 5 suggesting benefit). Sensitivity analyses showed similar findings. Preintervention albumin contributed the most to HTE. Analyses of individual EGDT components were inconclusive.

**CONCLUSIONS:** Treatment response to EGDT varied across patients in two multicenter RCTs with large benefits for some patients while others were harmed. Patient characteristics, including albumin, were most important in identifying HTE.

**KEYWORDS:** heterogeneity of treatment effect; machine learning; precision medicine; resuscitation; sepsis

Over 50 randomized controlled trials (RCTs) have tested resuscitation strategies in sepsis (1), but the optimal approach remains uncertain. Early goal-directed therapy (EGDT), a multicomponent, 6-hour intervention consisting of protocolized administration of fluids, vasopressors,

This article has an accompanying editorial.



DOI: 10.1097/CCM.0000000000006463

## KEY POINTS

**Question:** Do individual patients with septic shock respond differently to hemodynamic resuscitation strategies?

**Findings:** We analyzed two multicenter clinical trials of protocolized early goal-directed therapy (EGDT) for septic shock. Both trials had failed to demonstrate a significant difference in mortality with EGDT compared with usual care. Using computational methods that model treatment response at an individual patient level, we showed that the effect of EGDT compared with usual care varied considerably, ranging from harm (mean mortality in the quintile for whom EGDT was predicted to be least effective = 31% [EGDT] vs. 23% [usual care], relative risk reduction [RRR], 1.34 [95% CI, 0.89–2.01]) to benefit (21% vs. 31%; RRR, 0.64 [95% CI, 0.42–0.96]). Pre-randomization patient characteristics most predictive of treatment response included albumin. Analyses exploring which components of EGDT (and usual care) were most explanatory of heterogeneous treatment responses were inconclusive.

**Meaning:** EGDT appears to have quite heterogenous treatment effects across individuals, possibly explaining inconsistent results across previous studies. Future studies should investigate individualized treatment effects to identify heterogeneity in already completed resuscitation trials. Additionally, future trials should plan data collection and statistical analyses in a manner to facilitate uncovering treatment heterogeneity and should incorporate these findings during execution of the trials.

inotropes, and blood products, demonstrated an absolute mortality risk reduction of 16% compared with usual care in a landmark single-center trial (2). However, three subsequent multicenter, international RCTs (3–5) did not reproduce the previously observed benefits. Currently, the EGDT protocol is not followed routinely in clinical practice, but several components, such as the volume and rate of fluid administration and vasopressor use, remain the cornerstone of resuscitation. Importantly, the optimal way each of these components is delivered remains a source of ongoing debate and inquiry (6–8).

There has been increasing awareness that the average treatment effect (ATE) estimated in an RCT may not reflect the varying individual treatment effect (ITE) experienced by each patient (9–11). Although an individual patient data meta-analysis (IPDMA) of prior EGDT trials using conventional statistical approaches, which may have been underpowered, did not uncover subgroups that benefitted (12), heterogeneity of treatment effect (HTE) may exist and contribute to inconsistent results. An added complexity in resuscitation is the wide variation in "dose" of the intervention (e.g., amount of fluid or vasopressor) across patients compared with the assessment of HTE in fixed-dose interventions.

We sought to explore HTE within large RCTs of EGDT and determine patient characteristics or features about intervention delivery that may explain HTE.

## METHODS

### Trial Characteristics

We performed a secondary analysis of two multicenter RCTs of EGDT: Australian Resuscitation of Sepsis Evaluation (ARISE) (4) and Protocolized Care for Early Septic Shock (ProCESS) (3). We did not include data from the Protocolised Management in Sepsis (ProMISe) (5) trial because it lacked pre-intervention variables found to be important in predicting treatment responses in preliminary analyses (13).

Both trials randomized adults presenting to emergency departments with septic shock, defined as sepsis and refractory hypotension or elevated lactate, between 2008 and 2014 to EGDT or usual care at 81 sites (51 in ARISE and 30 in ProCESS) in Australia, New Zealand, Hong Kong, Finland, Republic of Ireland, and the United States. ProCESS also randomized patients to a third group receiving protocolized resuscitation, which we excluded because there was no equivalent in ARISE. Additional trial characteristics are in **eTable 1** (http://links.lww.com/CCM/H601). Further details on inclusion and exclusion criteria have been previously published (3, 4).

### Study Design

This study had three main goals. First, we tested the hypothesis that response to EGDT is varied by developing

and validating a model to predict individual treatment responses. Second, we identified patient characteristics that predicted individual treatment responses. Third, EGDT is a multicomponent intervention, and we determined if any HTE may be specifically due to differences in subcomponents of the EGDT protocol.

This study was approved by the University of Pittsburgh Institutional Review Board (protocol number STUDY19090326; "Supervised Clustering in Sepsis"; approved November 1, 2019). Written informed consent was obtained in both trials per published procedures in accordance with the Declaration of Helsinki.

## Primary Outcome

We selected 90-day, all-cause mortality as the primary outcome consistent with the ARISE trial (4) and an IPDMA of the three contemporary EGDT trials (12). We excluded subjects missing 90-day mortality outcome data ($n = 6$, 0.2%).

## Selection of Predictor Variables for Modeling

We used 27 predictor variables (**eTable 2**, http://links.lww.com/CCM/H601) to model individual treatment responses from six clinical domains: demographics, comorbidities, vital signs, acute severity of illness assessments, clinical laboratory values, and sites of infection. We chose these variables as they are routinely collected during clinical care and were used to elucidate HTE in our prior work (14). Details regarding missingness and sensitivity analyses with alternate approaches are available in the **Online Supplement Methods** and eTable 2 (http://links.lww.com/CCM/H601).

## Statistical Analyses

First, we constructed derivation and validation cohorts by randomly splitting study sites from both trials into two groups (41 for derivation, 40 for validation). We used this approach over others (e.g., one trial for derivation and one for validation, temporal or geographic splitting) because usual care differed across sites in ARISE and ProCESS despite harmonization efforts (12) and splitting by site allowed for better identification of HTE due to patient-level rather than site-level differences. We trained a model to predict individual treatment responses in the derivation set (cohort A) and performed formal testing of HTE using individual treatment responses estimated by the model in the validation set (cohort B), which decreased model precision as only half of the dataset is used for prediction but reduced the risk of overfitting bias. We fit a single model to the combined dataset to use the largest possible dataset to explore which patient characteristics and subcomponents of EGDT identified HTE. All statistical analyses were performed using R, Version 4.3.0.

***Model Development.*** In cohort A, we compared three models (two effect-based and one risk-based) to identify the best-performing model type. Effect-based models included causal forests (15) and R-learners based on a local-linear random forests (LLRFs) (16), which are both computationally efficient but each have different strengths (15). LLRFs provide a smooth way to model covariates that do not use step functions as in causal forests. We estimated individual absolute risk differences (iARDs) for each patient in response to EGDT representing the difference between the patient-specific covariate-adjusted risk of mortality when receiving usual care vs. when receiving EGDT (15) (akin to the difference in mortality risk between simulated "digital twins" with identical baseline characteristics but in different treatment groups). A positive iARD corresponds to lower predicted mortality with EGDT (suggesting benefit compared with usual care) and a negative iARD corresponds to a higher predicted mortality (suggesting harm). The risk-based model was a random forest predicting 90-day mortality in the usual care group (RF-risk). For all models, we set hyperparameters to their default values, and set the number of trees to 5000. All continuous variables were input into models without categorization into discrete groups. We compared model types based on three measures of discriminative performance: area under the targeting operator characteristics (AUTOCs), area under the Qini curve (AUQINI), and the adjusted AUQINI (17, 18). We also compared the two iARD models using cross-validated R-loss (for further details, **Online Data Supplement**, http://links.lww.com/CCM/H601) (16).

***Model Validation and Hypothesis Testing for Heterogeneous Effects.*** Next, we used the best-performing model, fit in cohort A, to predict iARDs, calculate measures of discriminative performance, and test for HTE in cohort B. We visualized the distribution of iARD predictions by generating histograms and

explored predicted heterogeneity for patients by plotting observed average mortality among patients with similar iARDs separately for usual care and EGDT groups. We visualized model calibration by plotting predicted iARDs against observed mortality rate differences estimated with local linear causal forests using predicted iARDs as the only covariate. We tested for HTE by performing a hypothesis test of the interaction between treatment assignment and predicted iARD in a linear regression model of transformed mortality outcome, and a *p* value of less than 0.05 for the interaction was interpreted as evidence of HTE (19). In this model, an interaction coefficient of 1 suggests iARD predictions are well-calibrated, less than 1 signifies that iARDs in the validation set are closer to 0 compared with predictions, and greater than 1 signifies that iARDs are more extreme compared with predictions (19). We then calculated the mean among the predicted iARDs in each quintile and calculated each quintile's observed absolute risk difference (ARD) and 95% CI as well as the observed relative risk reduction (RRR) and 95% CIs. We chose to split the cohorts into quintiles to assess HTE to ensure enough patients in EGDT and usual care arms per subdivision to demonstrate a difference in mortality while allowing signals of HTE at extremes to be discernable and not masked by participants with iARDs close to zero.

We performed several sensitivity analyses to ensure robustness of results. First, we trained a model in cohort B and tested for HTE in cohort A. Second, we performed sensitivity analyses using one trial for derivation (ARISE or ProCESS) and testing for HTE in the other trial.

***Assessment of Patient Characteristics Contributing to HTE.*** We used three approaches to determine which variables included in the model were predictive of benefit or harm in the combined dataset. First, we examined the distributions of baseline variables across quintiles of predicted iARDs. Second, we computed SHapley Additive exPlanation (SHAP) values from the model fit to calculate both traditional variable importance scores for each variable and patient-specific variable importance scores (20). The absolute value of a patient-specific SHAP is a measure of a covariate's influence on the predicted benefit from EGDT for that individual, with higher values signifying greater influence. The sign of the SHAP describes whether the patient's covariate value is accountable for making the iARD higher (driving potential benefit from EGDT) or lower (driving potential harm). Third, we used a decision tree in a "fit-the-fit" analysis of the LLRF predictions to identify variables and cutoff points for benefit or harm. Additional details are in the Online Data Supplement (http://links.lww.com/CCM/H601). Due to systematic differences in assessment of comorbidities between ARISE and ProCESS, we included the Charlson Comorbidity score as a predictor in our primary analyses and explored the role of three comorbidities that plausibly influence resuscitation in sepsis (cirrhosis, heart failure, and chronic kidney disease separately).

***Assessment of Subcomponents of EGDT in Contributing to HTE.*** Determining which components of EGDT may contribute to heterogeneity is challenging because the subcomponents (e.g., fluids and vasopressors) are deployed in both arms, and the deployment of each intervention may vary both due to usual care practices and fidelity to the EGDT protocol. Both trials reported high compliance with protocol fidelity (3, 4). Thus, we focused on whether potential HTE was influenced by variation in usual care patterns. Usual care practice could not be described at the individual-provider level for a variety of reasons but was assessed previously at the site-level in both trials using a propensity model as part of an IPDMA (12). Thus, we conducted a site-level exploratory analysis using the same model in a subset of patients from the 67 sites that enrolled at least three patients into the usual care group, which generated observed and expected values for volume of fluid administration and vasopressor usage in the usual care group (12).

We stratified sites into three groups to mimic restrictive or liberal fluid management strategies described previously (6, 8, 21): 1) low fluids and high vasopressor use, similar to restrictive fluid management strategies (ten sites); 2) high fluids and low vasopressor use, similar to liberal fluid strategies (11 sites); and 3) others (46 sites; for additional details, Online Supplement Methods, **eFig. 1**, and **eTable 3**, http://links.lww.com/CCM/H601). We fit a LLRF to the combined cohort, expanding the set of baseline covariates to include indicators for membership to one of the three management groups. We hypothesized that, if differences in fluid or vasopressor administration explain differences in treatment response, then the model predictions would only identify HTE in a subset of these three groups.

For less frequently used subcomponents of EGDT, we compared the difference in RBC transfusion and dobutamine use between EGDT and usual care arms stratified by quintiles of iARDs in the combined cohort.

## RESULTS

We included 1588 patients from ARISE (796 randomized to usual care and 792 to EGDT; 99% of enrollment) and 892 patients from ProCESS (455 randomized to usual care and 437 to EGDT; 99% of enrollment). As reported previously (12), patient characteristics were similar between the trials (**eTable 4**, http://links.lww.com/CCM/H601). After randomly splitting the 81 sites into cohorts A and B, we confirmed that patients had similar distributions of baseline characteristics overall and by treatment group (**Table 1**; and **eTable 5**, http://links.lww.com/CCM/H601).

### Model Development in the Derivation Set

In comparing the three candidate models used to rank patients in order of lowest to highest effect of EGDT within cohort A, the LLRF R-Learner had the highest measures of discriminative performance (AUTOC, AUQINI, and adjusted AUQINI), and lowest cross-validated R-loss compared with causal forest and risk-based models, indicating superior predictive performance (**eTable 6**, http://links.lww.com/CCM/H601). Similar patterns were seen in cohort B (eTable 6 and **eFig. 2**, http://links.lww.com/CCM/H601). Both causal forests and LLRF models performed well in sensitivity analyses using one trial for derivation and the other to test for HTE (**eTable 7**, http://links.lww.com/CCM/H601). Thus, we selected the LLRF model for the remainder of our analyses.

### Model Validation and Hypothesis Testing for Heterogeneous Effects

In our validation analysis in cohort B, we observed a significant interaction between iARD prediction and treatment (coefficient, 1.8; 95% CI, 0.76–2.8; $p < 0.001$), supportive of HTE. In cohort B patients, the overall absolute risk difference comparing EGDT with usual care was 2% (95% CI, –5% to 4%) and the RRR was 1.00 (95% CI, 0.8–1.2). When using the model derived from cohort A to rank order patients in cohort B from lowest to highest likelihood of benefit, both the predicted and observed treatment effects ranged widely. The mean predicted iARD in quintile 1 was –6% and the observed 90-day mortality was 31% in the EGDT and 23% in the usual care groups, representing an observed ARD of –8% (95% CI, –19% to 4%) and RRR of 1.34 (95% CI, 0.89–2.01) suggesting increased mortality with EGDT. In contrast, in quintile 5, the mean predicted iARD was 7% with observed mortality of 21% in the EGDT and 33% in the usual care group thereby representing an observed ARD of 12% (95% CI, 1–23%) and a RRR of 0.64 (95% CI, 0.42–0.96) suggesting decreased mortality with EGDT (**Fig. 1**). Plots of mortality rates by treatment group for patients with similar iARD predictions showed that patients with iARD near 0 had relatively low risk in each treatment group, and patients with more extreme iARDs in both directions had increasingly high baseline mortality risk (**Fig. 2**). Similar results were observed in sensitivity analyses reversing derivation and validation cohorts (**eFigs. 3** and **4**, http://links.lww.com/CCM/H601), with consistent findings of significant treatment by predicted iARD interaction (coefficient, 1.3; 95% CI, 0.4–2.1; $p = 0.001$). Notably, the full spectrum of iARDs was broad ranging from –13.7% to 13.9% when cohort B was used as validation and from –28.5% to 16.5% when cohort A was used for validation. Findings were also consistent in sensitivity analyses using alternative approaches for handling missing data values, although not all analyses reached statistical significance (**eFigs. 5–7**, http://links.lww.com/CCM/H601). Treatment by predicted iARD interaction remained significant in sensitivity analyses using ProCESS for derivation and ARISE for validation ($p = 0.003$) and vice versa ($p = 0.003$; **eFig. 8**, http://links.lww.com/CCM/H601). For the model derived in cohort A, metrics of discriminative performance calculated in cohort B were excellent; metrics calculated in cohort A were similarly strong for the model derived in cohort B (**eTable 8**, http://links.lww.com/CCM/H601).

### Assessment of Patient Characteristics Contributing to HTE

Circulating albumin levels were most predictive of treatment response in SHAP (**eFig. 9**, http://links.lww.com/CCM/H601), decision tree analyses (**eFig. 10**, http://links.lww.com/CCM/H601), and when baseline characteristics across iARD quintiles were

## TABLE 1.
## Baseline Patient Characteristics in Cohorts A and B

| Variable | Cohort A ($n = 1148$) | Cohort B ($n = 1332$) |
|---|---|---|
| Trial membership | | |
| Australasian Resuscitation of Sepsis Evaluation | 742 (64.6) | 846 (63.5) |
| Protocolized Care for Early Septic Shock | 406 (35.4) | 486 (36.5) |
| Demographics | | |
| Age, yr | 65 (52–76) | 62 (50–73) |
| Male | 668 (58.2) | 774 (58.1) |
| Vital signs | | |
| Temperature (°C) | 37 (36–38) | 37 (37–38) |
| Respiratory rate, bpm | 22 (18–28) | 22 (18–28) |
| Heart rate, bpm | 100 (90–120) | 110 (92–120) |
| Mean arterial pressure, mm Hg | 67 (59–79) | 67 (59–78) |
| Systolic blood pressure, mm Hg | 95 (83–110) | 95 (83–110) |
| Severity of illness | | |
| Acute Physiology and Chronic Health Evaluation II | 15 (11–20) | 15 (11–20) |
| Total Sequential Organ Failure Assessment | 4 (2–6) | 4 (2–6) |
| Glasgow Coma Scale | 15 (14–15) | 15 (14–15) |
| Charlson Comorbidity Index | 1 (0–3) | 1 (0–2) |
| Clinical laboratories | | |
| Albumin, g/dL | 3.2 (2.7–3.7) | 3.2 (2.7–3.7) |
| Hemoglobin, g/dL | 12 (11–14) | 12 (11–14) |
| $Pao_2$, mm Hg | 87 (64–140) | 89 (67–130) |
| Bilirubin, mg/dL | 0.9 (0.6–1.6) | 0.93 (0.6–1.5) |
| Blood urea nitrogen, mg/dL | 27 (18–43) | 26 (18–41) |
| Creatinine, mg/dL | 1.6 (1.1–2.4) | 1.5 (1–2.4) |
| Glucose, mg/dL | 130 (100–170) | 130 (100–180) |
| Lactate, mmol/L | 4.2 (2.3–5.8) | 4.1 (2.2–5.7) |
| Platelets, $10^9$/L | 200 (140–280) | 200 (130–260) |
| Oxygen saturation, % | 98 (95–100) | 97 (95–99) |
| WBC, $10^9$/L | 13 (7.5–19) | 13 (7.3–19) |
| Infection site | | |
| Lung | 408 (36.2) | 432 (33.2) |
| Urinary tract | 219 (19.4) | 282 (21.7) |
| Abdominal | 114 (10.1) | 130 (10.0) |
| Organ support | | |
| Mechanical ventilation | 143 (12.5) | 135 (10.1) |
| Vasopressors | 244 (21.3) | 253 (19.0) |

bpm = beats/min.
Cells show median (interquartile range) for continuous variables, and frequency (percent) for binary variables.

| Quintile | Deaths EGDT | Total N EGDT | Deaths Usual Care | Total N Usual Care | Mean iARD prediction | % ARD (95% CI) | Relative Risk (95% CI) |
|---|---|---|---|---|---|---|---|
| Q1 | 47 | 151 | 27 | 116 | -5.9 | -7.8 (-19.3,3.6) | 1.34 (0.89,2.01) |
| Q2 | 30 | 137 | 23 | 129 | -1.7 | -4.1 (-14.4,6.3) | 1.23 (0.75,2) |
| Q3 | 27 | 124 | 24 | 142 | 0.6 | -4.9 (-15.2,5.4) | 1.29 (0.79,2.11) |
| Q4 | 22 | 136 | 28 | 130 | 2.8 | 5.4 (-4.8,15.5) | 0.75 (0.45,1.24) |
| Q5 | 28 | 132 | 45 | 135 | 6.6 | 12.1 (0.8,23.4) | 0.64 (0.42,0.96) |

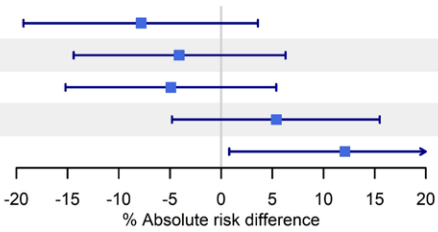


**Figure 1.** Observed absolute risk difference (ARD) with 95% CIs from early goal-directed therapy (EGDT) compared with usual care in quintiles of patients stratified by predicted individual ARDs (iARDs). Quintiles represent patients in cohort B stratified by predicted treatment response from the local-linear random forest R-learner model derived in cohort A. Negative ARD represents higher mortality on EGDT compared with usual care (harm from EGDT) and positive ARD represents lower mortality on EGDT compared with usual care (benefit from EGDT). Quintile 1 (Q1) represents the patients with the lowest predicted iARDs (predicted to benefit the least from EGDT) and quintile 5 (Q5) represents the patients with the highest predicted iARDs (predicted to benefit the most from EGDT). Relative risk of 90-d mortality in EGDT group compared with usual care is also provided with 95% CI. Q2 = quintile 2, Q3 = quintile 3, Q4 = quintile 4.

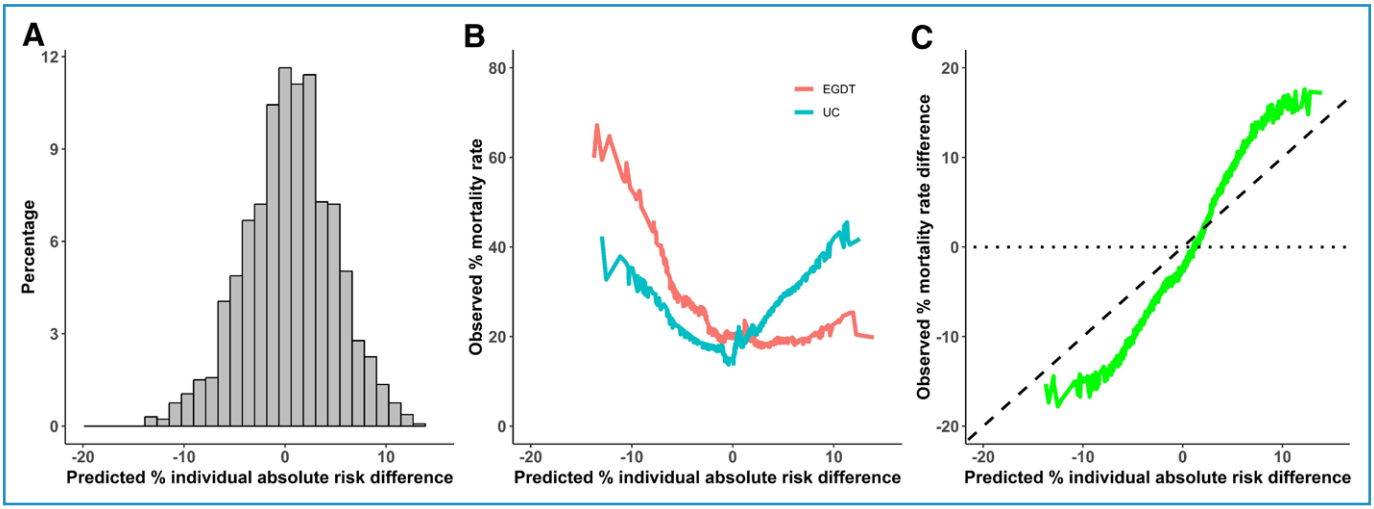


**Figure 2.** Examination of predicted individual absolute risk differences (iARDs) in cohort B. Predicted iARDs in cohort B determined based on a local linear random forest R-learner model derived in cohort A. Histogram of iARD predictions is provided in (**A**). Plot of predicted iARDs against smoothed observed mortality rates in each treatment group is provided in (**B**). Plot of iARD predictions against observed mortality rate differences is provided in (**C**). Curves in (**B**) and (**C**) were generated using random forest-based smoothing functions using cohort B patient predictions and 90-d mortality status. Negative iARD predictions represent a higher risk of mortality on early goal-directed therapy (EGDT) compared with usual care (UC) (probable harm) and positive predictions represent a lower risk of mortality on EGDT. The *dashed line* in (**C**) represents perfect calibration intercept and slope of the iARD predictions in cohort B; the *dotted line* represents the calibration of randomly generated predictions.

compared (**Table 2**). For example, younger patients with lower albumin levels were more likely to be harmed by EGDT, whereas older patients with normal albumin levels were more likely to benefit. SHAP analysis suggested that relationships between covariates and predicted treatment response were complex. For instance, albumin level showed a steep increasing linear relationship with treatment response for albumin values less than 3 g/dL followed by constant (flat) contributions for values greater than or equal to 3 g/dL, whereas age showed a linear relationship (**eFig. 11**, http://links.lww.com/CCM/H601). Temperature, heart rate, and circulating glucose values were also important predictors of treatment response, with either a U- or S-shaped relationship between these variables and treatment response. The remaining variables also contributed to treatment response, although individual contributions varied.

Notably, two patients may have had similar iARD predictions but due to different covariate patterns. For instance, two patients in the combined cohort had received iARD predictions less than –14%, suggesting

## TABLE 2.

**Comparison of Baseline Characteristics Across Quintiles of Predicted Individual Absolute Risk Difference for All Patients (Cohorts A and B), From a Model Derived in the Combined Sample**

| Variable | Quintile 1 (*n* = 496) | Quintile 2 (*n* = 496) | Quintile 3 (*n* = 496) | Quintile 4 (*n* = 496) | Quintile 5 (*n* = 496) |
|---|---|---|---|---|---|
| Demographics | | | | | |
| Age, yr | 55 (43–66) | 54 (42–67) | 63 (52–74) | 68 (58–76) | 72 (65–80) |
| Male | 296 (59.7) | 287 (57.9) | 294 (59.3) | 286 (57.7) | 279 (56.2) |
| Vital signs | | | | | |
| Temperature (°C) | 37 (36–38) | 38 (37–38) | 38 (37–38) | 37 (37–38) | 37 (36–38) |
| Respiratory rate, bpm | 23 (18–28) | 23 (19–28) | 22 (18–29) | 22 (18–26) | 22 (18–27) |
| Heart rate, bpm | 120 (100–140) | 120 (100–130) | 100 (94–120) | 100 (88–110) | 94 (82–110) |
| Mean arterial pressure, mm Hg | 65 (58–77) | 68 (60–81) | 68 (61–78) | 68 (60–80) | 65 (58–77) |
| Systolic blood pressure, mm Hg | 92 (81–110) | 96 (85–110) | 97 (85–110) | 95 (84–110) | 93 (82–110) |
| Severity of illness | | | | | |
| Acute Physiology and Chronic Health Evaluation II | 17 (12–22) | 14 (9.8–19) | 13 (9–17) | 14 (11–18) | 19 (14–23) |
| Total Sequential Organ Failure Assessment | 5 (3–8) | 4 (2–5) | 3 (2–5) | 3 (2–5) | 5 (3–6) |
| Glasgow Coma Score | 15 (14–15) | 15 (15–15) | 15 (14–15) | 15 (15–15) | 15 (13–15) |
| Charlson | 2 (0–4) | 1 (0–2) | 0 (0–2) | 1 (0–2) | 2 (0–3) |
| Clinical laboratories | | | | | |
| Albumin, g/dL | 2.2 (1.9–2.6) | 3.3 (2.8–3.8) | 3.4 (3–3.8) | 3.4 (3–3.8) | 3.3 (3–3.7) |
| Hemoglobin, g/dL | 11 (9.1–13) | 13 (11–15) | 13 (12–14) | 13 (11–14) | 12 (10–14) |
| $Pao_2$, mm Hg | 88 (66–150) | 84 (61–130) | 86 (67–130) | 90 (68–140) | 89 (65–140) |
| Bilirubin, mg/dL | 1.2 (0.6–2.3) | 1 (0.64–1.7) | 0.99 (0.6–1.4) | 0.87 (0.58–1.4) | 0.75 (0.5–1.2) |
| Blood urea nitrogen, mg/dL | 27 (17–46) | 23 (15–33) | 23 (17–33) | 26 (18–36) | 40 (25–64) |
| Creatinine, mg/dL | 1.5 (1.1–2.7) | 1.4 (1–2.2) | 1.3 (1–1.9) | 1.4 (1–2) | 2.1 (1.3–3.3) |
| Glucose, mg/dL | 110 (85–140) | 120 (96–160) | 130 (110–170) | 140 (110–180) | 150 (120–230) |
| Lactate, mmol/L | 4.6 (2.6–6.7) | 4.3 (2.6–5.9) | 3.9 (2.2–5.4) | 3.7 (2–5.1) | 4.2 (2.2–5.2) |
| Platelets, $10^9$/L | 170 (91–260) | 190 (130–250) | 190 (140–250) | 210 (150–280) | 230 (170–320) |
| Oxygen saturation, % | 98 (95–100) | 98 (95–99) | 98 (95–100) | 97 (95–100) | 97 (94–100) |
| WBC, $10^9$/L | 12 (5.9–18) | 12 (6–18) | 12 (7.8–18) | 13 (8–19) | 15 (10–21) |
| Infection site | | | | | |
| Lung | 164 (34.3) | 160 (33) | 179 (36.8) | 178 (36.3) | 159 (32.6) |
| Urinary tract | 76 (15.9) | 94 (19.4) | 101 (20.8) | 119 (24.2) | 111 (22.8) |
| Abdominal | 60 (12.6) | 47 (9.7) | 45 (9.3) | 37 (7.5) | 55 (11.3) |
| Organ support | | | | | |
| Mechanical ventilation | 84 (16.9) | 33 (6.7) | 38 (7.7) | 42 (8.5) | 81 (16.3) |
| Vasopressors | 149 (30) | 80 (16.1) | 67 (13.5) | 91 (18.3) | 110 (22.2) |

bpm = beats/min.

Cells show median (interquartile range) for continuous variables, and frequency (percent) for binary variables. Quintile 1 for each trial includes the patients with the lowest predicted individual absolute risk differences (iARDs) representing the patients with the least predicted benefit (or greatest harm) from early goal-directed therapy (EGDT) compared with usual care, and quintile 1 includes the patients with the highest iARDs representing the patients with the highest predicted benefit from EGDT. iARDs are estimated from a local linear random forest R-learner model trained in the combined cohort.

harm with EGDT. SHAP analysis revealed that the model generated the prediction of harm almost entirely based on low albumin levels for one patient (**eFig. 12*A***, http://links.lww.com/CCM/H601), but the prediction of harm for another patient was based primarily on contributions from low albumin, age, and glucose (**eFig. 12*B***, http://links.lww.com/CCM/H601). Some differences in variable importance were noted in sensitivity analyses using one trial for derivation and the other for validation, but albumin remained the most important variable, and the top ten variables were mostly consistent across analyses (**eFigs. 13** and **14**, http://links.lww.com/CCM/H601).

In exploratory analyses investigating differences in individual comorbidities across iARD quintiles, we determined that cirrhosis was most prevalent in the quintile of patients predicted to be harmed by EGDT and heart failure was most prevalent in the quintile of patients predicted to benefit the most from EGDT (**eTable 9**, http://links.lww.com/CCM/H601).

### Assessment of Subcomponents of EGDT Contributing to HTE

Exploratory analyses on whether the way interventions were deployed influenced HTE were inconclusive. For fluid and vasopressor use, as expected, the sites in the restrictive fluid management group had higher differences in fluid administration between EGDT and usual care groups (**eFig. 15**, http://links.lww.com/CCM/H601), while those in the liberal group sites had higher differences in vasopressor use (**eFig. 16**, http://links.lww.com/CCM/H601). However, the magnitude of HTE did not vary across these sites (**eTable 10**, http://links.lww.com/CCM/H601). Similarly, for RBC transfusion and dobutamine use, no differences were observed between usual care and EGDT arms (**eTable 11**, http://links.lww.com/CCM/H601).

## DISCUSSION

In a secondary analysis of two multicenter, international trials testing resuscitation strategies, we found large variation in ITEs of EGDT compared with usual care. Although neither trial demonstrated a statistically significant ATE, the range of iARDs within these trials included benefit of a magnitude similar to that of the initial EGDT trial but also included iARDs of an equal magnitude of harm. Circulating albumin levels were most predictive of treatment response with younger adults with lower albumin values predicted to have the highest harm from EGDT. Exploratory analyses suggest EGDT may harm patients with cirrhosis and benefit patients with heart failure, but, due to differences comorbidity assessment between trials, these findings are hypothesis generating and highlight a need to consider comorbidities in future studies investigating iARDs to resuscitation.

When the ARISE (4), ProCESS (3), and PROMISe (5) trials of EGDT did not replicate the benefits demonstrated in the original Rivers trial (2), diverse opinions emerged for these discrepant findings, including shifts in usual care practices over time and lack of efficacy (12). Our findings suggest that EGDT may indeed be efficacious but only for a subset of patients, and the discrepant results from trials may have reflected differences in enrollment of patients in whom EGDT was beneficial and harmful.

Importantly, similar HTE likely exists in other trials of sepsis resuscitation. Our results therefore have implications for the design of future resuscitation trials, which should be much larger compared with contemporary trials to allow randomization of patients to several intervention arms across multiple groups to account for potential HTE for each resuscitation component. Several studies investigating this concept identified subtypes using baseline covariates and demonstrated differential treatment effects in subtypes (14, 22, 23). Consistent with recent studies of ITEs (10, 24, 25), our current approach is different and has strengths. First, identifying ITEs may be advantageous over subtyping when variables determining subtype membership differ from those determining treatment response or when individual treatment responses vary within subtypes. Second, the iARD values in our study provide an estimate of benefit or harm at an individual level and may allow personalizing treatments. Third, modern model explanation tools such as SHAP allow investigators to understand the contributions of predictor variables in treatment response, which may be different for individual patients.

We acknowledge some caveats in the interpretation of findings. We examined a small set of baseline covariates available in both ARISE and ProCESS to predict treatment response, although a richer set of covariates, such as detailed measures of organ support and intervention delivery, may have provided greater insights. Several physiologic variables that may guide sepsis resuscitation such as vasopressor dose and central venous oxygen saturation were not universally available. As the variables that contribute to response

to resuscitation remain unclear, future HTE studies should strive to strike a balance between including as predictors both variables that already have a plausible causal relationship to treatment response and other variables where a causal association is not as clear but may be possible. We acknowledge some minor differences in variable importance rank order when models were derived using only half of the available data; however, five variables consistently emerged as the most important, and our approach of using the entire cohort for variable importance would provide the most precise estimates for a patient enrolled in a future prospective trial. Missingness in our predictor variables may also affect precision of ITE estimates but our results were robust across multiple sensitivity analyses. Prospective studies are needed to advance precision medicine approaches and refine ITE-based approaches, but retrospective analyses such as ours serve to provide insight into variables and model designs, and help avoid randomizing patients to treatment arms where they may be harmed.

In conclusion, our results suggest that individual responses to resuscitation varied in patients enrolled in EGDT trials. Future studies should seek to understand underlying mechanisms and future trial designs should incorporate approaches that learn HTE and tailor resuscitation strategies.

1 *Division of Pulmonary, Allergy, and Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA.*

2 *Veterans Affairs Pittsburgh Healthcare System, Pittsburgh, PA.*

3 *Department of Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA.*

4 *Division of General Internal Medicine, University of Pittsburgh, Pittsburgh, PA.*

5 *Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA.*

6 *Department of Mathematics and Applied Mathematics, University of Crete, Heraklion, Crete, Greece.*

7 *Australian and New Zealand Intensive Care Research Centre, Monash University, Melbourne, VIC, Australia.*

8 *Intensive Care National Audit & Research Centre, London, United Kingdom.*

9 *Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA.*

10 *Department of Critical Care, University of Melbourne, Melbourne, VIC, Australia.*

11 *Department of Intensive Care, Austin Hospital, Melbourne, VIC, Australia.*

12 *Department of Intensive Care, Royal Melbourne Hospital, Melbourne, VIC, Australia.*

13 *Data Analytics Research and Evaluation, Austin Hospital, Melbourne, VIC, Australia.*

14 *Department of Emergency Medicine, University of Pittsburgh, Pittsburgh, PA.*

*Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccmjournal).*

*Drs. Shah, Talisa, Chang, Seymour, Angus, and Yende were involved in concept and design. Dr. Talisa, Dr. Chang, and Mr. Kennedy were involved in statistical analysis. All authors were involved in analysis and interpretation of data. Drs. Shah, Talisa, Angus, and Yende were involved in drafting of article. All authors were involved in critical revision of article. Drs. Shah, Talisa, Angus, and Yende were involved in obtained funding. Drs. Angus and Yende were involved in supervision. All authors give their approval to the submission of this article.*

*Study investigators were supported in this project by grants from the National Institutes of Health through awards GM141081 (to Drs. Angus, Yende, Tang, Chang, Shah, and Talisa), R35GM119519 (to Dr. Seymour and Mr. Kennedy), R21GM144851 (to Dr. Seymour and Mr. Kennedy), K23GM132688 (to Dr. Mayr), K23GM122069 (to Dr. Shah), R21HL168070 (to Dr. Shah), R01HL164835 (to Dr. Seymour, Dr. Cooper, Dr. Triantafyllou, and Mr. Kennedy), and from the National Health and Medical Research Council through award GNT2008447 (to Dr. Higgins). The project described was additionally supported by the National Institutes of Health through Grant Number UL1TR001857. Finally, the University of Pittsburgh Center for Research Computing supported this study through resources provided for high throughput computing under award S10OD028483 from the National Institutes of Health.*

*Drs. Shah's, Talisa's, Chang's, Tang's, Cooper's, Angus's, and Yende's institutions received funding from the National Institutes of Health (NIH; GM141081). Dr. Shah's institution received funding from the NIH (K23GM122069, R21HL168070). Dr. Shah, Dr. Talisa, Dr. Chang, Dr. Triantafyllou, Dr. Tang, Mr. Kennedy, Dr. Cooper, Dr. Yealy, Dr. Seymour, Dr. Angus, and Dr. Yende received support for article research from the NIH. Dr. Talisa disclosed the study was supported by the NIH (UL1TR001857; S10OD028483). Dr. Triantafyllou, Mr. Kennedy, Dr. Cooper, and Dr. Seymour's institutions received funding from the NIH (R01HL164835). Dr. Mayr's institution received funding from the NIH (K23GM132688); he disclosed government work; and he received personal fees from Baxter for serving on a racial disparities advisory board outside the submitted work. Dr. Higgins' institution received funding from the National Health and Medical Research Council (NHMRC) (GNT2008447); she received support for article research from the NHMRC. Mr. Mouncey's institution received funding from the National Institute for Health and Care Research Health Technology Assessment program; he received support for article research from the National Institute for Health and Care Research. Mr. Kennedy's and Dr. Seymour's institutions received funding from the NIH (R21GM144851). Mr. Kennedy's, Dr. Cooper's, and Dr. Seymour's institution received funding from the National Institute for General Medical Sciences (R35GM119519). Dr. Yealy's institution received funding from*

*the NIH. Dr. Seymour received funding from Beckman Coulter and Octapharma. Dr. Angus received funding from AM-Pharma and Abionyx. The remaining authors have disclosed that they do not have any potential conflicts of interest.*

*Drs. Shah and Talisa contributed equally.*

*For information regarding this article, E-mail: yendes@upmc.edu*

*The contents of this article do not represent the views of the U.S. Department of Veterans Affairs or the U.S. government.*

## REFERENCES

1. Tseng CH, Chen TT, Wu MY, et al: Resuscitation fluid types in sepsis, surgical, and trauma patients: A systematic review and sequential network meta-analyses. *Crit Care* 2020; 24:693
2. Rivers E, Nguyen B, Havstad S, et al; Early Goal-Directed Therapy Collaborative Group: Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N Engl J Med* 2001; 345:1368–1377
3. Pro CI, Yealy DM, Kellum JA, et al: A randomized trial of protocol-based care for early septic shock. *N Engl J Med* 2014; 370:1683–1693
4. Peake SL, Delaney A, Bailey M, et al; ARISE Investigators; ANZICS Clinical Trials Group: Goal-directed resuscitation for patients with early septic shock. *N Engl J Med* 2014; 371:1496–1506
5. Mouncey PR, Osborn TM, Power GS, et al; ProMISe Trial Investigators: Trial of early, goal-directed resuscitation for septic shock. *N Engl J Med* 2015; 372:1301–1311
6. Self WH, Semler MW, Bellomo R, et al; CLOVERS Protocol Committee and NHLBI Prevention and Early Treatment of Acute Lung Injury (PETAL) Network Investigators: Liberal versus restrictive intravenous fluid therapy for early septic shock: Rationale for a randomized trial. *Ann Emerg Med* 2018; 72:457–466
7. Meyhoff TS, Sivapalan P, Perner A: Restriction of intravenous fluid in ICU patients with septic shock. Reply. *N Engl J Med* 2022; 387:857
8. Wiedemann HP, Wheeler AP, Bernard GR, et al; National Heart, Lung, and Blood Institute Acute Respiratory Distress Syndrome (ARDS) Clinical Trials Network: Comparison of two fluid-management strategies in acute lung injury. *N Engl J Med* 2006; 354:2564–2575
9. Zampieri FG, Damiani LP, Bagshaw SM, et al; BRICNet: Conditional treatment effect analysis of two infusion rates for fluid challenges in critically ill patients: A secondary analysis of balanced solution versus saline in intensive care study (BaSICS) trial. *Ann Am Thorac Soc* 2023; 20:872–879
10. Seitz KP, Spicer AB, Casey JD, et al: Individualized treatment effects of Bougie versus stylet for tracheal intubation in critical illness. *Am J Respir Crit Care Med* 2023; 207:1602–1611
11. Goligher EC, Lawler PR, Jensen TP, et al; REMAP-CAP, ATTACC, and ACTIV-4a Investigators: Heterogeneous treatment effects of therapeutic-dose heparin in patients hospitalized for COVID-19. *JAMA* 2023; 329:1066–1077
12. Rowan KM, Angus DC, Bailey M, et al; PRISM Investigators: Early, goal-directed therapy for septic shock—a patient-level meta-analysis. *N Engl J Med* 2017; 376:2223–2234
13. Shah FA, Talisa V, Angus DC, et al: A novel ensemble learning approach to understand heterogeneity of treatment effect in critical care trials. *Am J Respir Crit Care Med* 2020; 201:A1644
14. Seymour CW, Kennedy JN, Wang S, et al: Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA* 2019; 321:2003–2017
15. Athey S, Tibshirani J, Wager S: Generalized random forests. *Ann Statist* 2019; 47:1148–1178
16. Nie X, Wager S: Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 2021; 108:299–319
17. Belbahri M, Murua A, Gandouet O, et al: Qini-based uplift regression. *Ann Appl Stat* 2021; 15:1247–1272
18. Yadlowsky S, Fleming S, Shah N, et al: Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv* Preprint posted online November 15, 2021. doi: 10.48550/arXiv.2111.07966
19. Chernozhukov V, Demirer M, Duflo E, et al: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India. National Bureau of Economic Research Working Paper Series 2018; No. 24678. Available at: https://www.nber.org/papers/w24678. Accessed December 15, 2019
20. Lundberg SM, Erion G, Chen H, et al: From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020; 2:56–67
21. Shapiro NI, Douglas IS, Brower RG, et al; National Heart, Lung, and Blood Institute Prevention and Early Treatment of Acute Lung Injury Clinical Trials Network: Early restrictive or liberal fluid management for sepsis-induced hypotension. *N Engl J Med* 2023; 388:499–510
22. Antcliffe DB, Burnham KL, Al-Beidh F, et al: Transcriptomic signatures in sepsis and a differential response to steroids. From the VANISH randomized trial. *Am J Respir Crit Care Med* 2019; 199:980–986
23. Wong HR, Hart KW, Lindsell CJ, et al: External corroboration that corticosteroids may be harmful to septic shock endotype A patients. *Crit Care Med* 2021; 49:e98–e101
24. Pirracchio R, Hubbard A, Sprung CL, et al; Rapid Recognition of Corticosteroid Resistant or Sensitive Sepsis (RECORDS) Collaborators: Assessment of machine learning to estimate the individual treatment effect of corticosteroids in septic shock. *JAMA Netw Open* 2020; 3:e2029050
25. Buell KG, Spicer AB, Casey JD, et al: Individualized treatment effects of oxygen targets in mechanically ventilated critically ill adults. *JAMA* 2024; 331:1195–1204