Sepsis is a heterogeneous syndrome. Identification of distinct clinical phenotypes may allow more precise therapy and improve care.

OBJECTIVE To derive sepsis phenotypes from clinical data, determine their reproducibility and correlation with host-response biomarkers and clinical outcomes, and assess the potential causal relationship with results from randomized clinical trials (RCTs).

DESIGN, SETTINGS, AND PARTICIPANTS Retrospective analysis of data sets using statistical, machine learning, and simulation tools. Phenotypes were derived among 20,189 total patients (16,552 unique patients) who met Sepsis-3 criteria within 6 hours of hospital presentation at 12 Pennsylvania hospitals (2010-2012) using consensus k-means clustering applied to 29 variables. Reproducibility and correlation with biological parameters and clinical outcomes were assessed in a second database (2013-2014; n = 43,086 total patients and n = 31,160 unique patients), in a prospective cohort study of sepsis due to pneumonia (n = 583), and in 3 sepsis RCTs (n = 4737).

EXPOSURES All clinical and laboratory variables in the electronic health record.

MAIN OUTCOMES AND MEASURES Derived phenotype (α, β, γ, and δ) frequency, host-response biomarkers, 28-day and 365-day mortality, and RCT simulation outputs.

RESULTS The derivation cohort included 20,189 patients with sepsis (mean age, 64 [SD, 17] years; 10,022 [50%] male; mean maximum 24-hour Sequential Organ Failure Assessment [SOFA] score, 3.9 [SD, 2.4]). The validation cohort included 43,086 patients (mean age, 67 [SD, 17] years; 21,993 [51%] male; mean maximum 24-hour SOFA score, 3.6 [SD, 2.0]). Of the 4 derived phenotypes, the α phenotype was the most common (n = 6625; 33%) and included patients with the lowest administration of a vasopressor; in the β phenotype (n = 5512; 27%), patients were older and had more chronic illness and renal dysfunction; in the γ phenotype (n = 5385; 27%), patients had more inflammation and pulmonary dysfunction; and in the δ phenotype (n = 2667; 13%), patients had more liver dysfunction and septic shock. Phenotype distributions were similar in the validation cohort. There were consistent differences in biomarker patterns by phenotype. In the derivation cohort, cumulative 28-day mortality was 287 deaths of 5691 unique patients (5%) for the α phenotype; 561 of 4420 (13%) for the β phenotype; 1031 of 4318 (24%) for the γ phenotype; and 897 of 2223 (40%) for the δ phenotype. Across all cohorts and trials, 28-day and 365-day mortality were highest among the δ phenotype vs the other 3 phenotypes (P < .001). In simulation models, the proportion of RCTs reporting benefit, harm, or no effect changed considerably (eg, varying the phenotype frequencies within an RCT of early goal-directed therapy changed the results from >33% chance of benefit to >60% chance of harm).

CONCLUSIONS AND RELEVANCE In this retrospective analysis of data sets from patients with sepsis, 4 clinical phenotypes were identified that correlated with host-response patterns and clinical outcomes, and simulations suggested these phenotypes may help in understanding heterogeneity of treatment effects. Further research is needed to determine the utility of these phenotypes in clinical care and for informing trial design and interpretation.
Sepsis, defined as a dysregulated immune response to infection that leads to acute organ dysfunction, affects millions of individuals per year, and carries a high risk of death even when care is provided promptly.1,2 Although the understanding of the host immune response has advanced considerably, it has not translated into new therapies. A major barrier to progress is the overly broad definition of the syndrome, which encompasses a vast, multidimensional array of clinical and biological features. Different combinations of these features may naturally cluster into previously undescribed subtypes or phenotypes that may have different risks for a poor outcome and may respond differently to treatments. However, efforts to determine such phenotypes have remained limited and have focused primarily on patients in the intensive care unit.3–5 In addition, these phenotypes must be identifiable at or soon after hospital presentation to guide treatment.

The objectives of this investigation, the National Institutes of Health–funded Sepsis Endotyping in Emergency Care (SENECA) project, were to develop and evaluate sepsis phenotypes. The first goal was to determine whether routine clinical information available at hospital presentation could be mathematically reduced to discrete, reproducible sepsis phenotypes. The second goal was to understand whether the different clinical phenotypes were associated both with patterns among biomarkers of the host immune response and with clinical outcomes. The third goal was to explore the heterogeneity of the treatment effects and the sensitivity of clinical trial results to the frequency distributions of these phenotypes. These mathematically derived phenotypes also were compared with traditional subgrouping strategies.

Methods

The project was approved by the University of Pittsburgh institutional review board and conducted under several data use agreements (PRO15110441, PRO19030218, PRO20061050, PRO010744, PRO12110516, PRO12020657, and PRO17120315). The data for the SENECA project were obtained under a waiver of informed consent and with authorization under the Health Insurance Portability and Accountability Act. Written informed consent was obtained for clinical trial data per published trial procedures.6–8

Overview

The study approach involved several data sets and statistical approaches. For the first goal (determining phenotypes), we derived the clinical phenotypes using unsupervised clustering methods that were applied to the data available at hospital presentation in a large database of hospital encounters. We then assessed phenotype reproducibility both by comparing phenotype derivation using alternative clustering methods in the initial data set and by exploring phenotype frequency distributions in several other cohort and clinical trial data sets (figure 1 in the Supplement). For the second goal (understanding the correlation of clinical phenotypes and biological markers of the host response with clinical outcome), we first examined correlations in several data sets between the clinical phenotypes and the concurrent patterns of biomarkers, reflecting different elements of the sepsis host response. We then assessed the association of phenotypes with mortality and other clinical outcomes. For the third goal (assessing the influence of phenotypes on clinical trial results), we explored traditional analyses of heterogeneity for treatment effects on observed clinical trial data and performed simulations on 3 trial data sets to understand the potential consequences of different phenotype frequency distributions on estimation of the treatment effects.

Data

We used data from 3 observational cohorts and 3 randomized clinical trials (RCTs)6–9 (Table 1). The first 2 cohorts (the SENECA derivation and validation cohorts) were drawn from electronic health record data on encounters at 12 community and academic hospitals within the UPMC health care system. We identified all adults (aged >18 years) who met sepsis criteria within the first 6 hours of presentation to the emergency department at the 12 hospitals during 2010 to 2012 for the derivation cohort and during 2013 to 2014 for the validation cohort.

The third cohort was the Genetic and Inflammatory Markers of Sepsis (GenIMS) study. The GenIMS study was a multicenter, prospective cohort of patients with severe community-acquired pneumonia recruited from 4 regions in the United States (western Pennsylvania, Connecticut, Tennessee, and Michigan) within 1 hour of emergency department presentation, and for whom we had rich clinical information and a variety of biomarkers for the host immune response. The GenIMS study enrolled patients hospitalized at 28 sites from 2001 to 2003.9,10

All 3 RCTs were multicenter studies that involved patients with sepsis or septic shock and had rich clinical and biomarker data. The first trial called ACCESS (A Controlled Comparison of Eritoran in Severe Sepsis) compared eritoran (a highly specific myeloid differentiation protein 2 antagonist that inhibits toll-like receptor 4) vs placebo in patients with severe sepsis at 197 sites on 6 continents from 2006 to 2010 and reported no benefit for 28-day mortality.6 The second trial called PROWESS (Activated Protein C Worldwide Evaluation in Severe Sepsis) compared activated protein C (a commonly
Table 1. Characteristics of Cohorts and Clinical Trials

<table>
<thead>
<tr>
<th>Study</th>
<th>Sepsis Endotyping in Emergency Care (SENECA)</th>
<th>Genetic and Inflammatory Markers of Sepsis (GenIMS)</th>
<th>A Controlled Comparison of Eritoran in Severe Sepsis (ACCESS)</th>
<th>Activated Protein C Worldwide Evaluation in Severe Sepsis (PROWESS)</th>
<th>Protocol-Based Care for Early Septic Shock (ProCESS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of patients</td>
<td>20,189</td>
<td>583</td>
<td>1,706</td>
<td>1,690</td>
<td>1,341</td>
</tr>
<tr>
<td>No. of sites</td>
<td>12</td>
<td>12</td>
<td>197</td>
<td>164</td>
<td>31</td>
</tr>
<tr>
<td>Location of enrollment</td>
<td>ED</td>
<td>ED</td>
<td>ED, medical ward, ICU</td>
<td>ED, medical ward, ICU</td>
<td>ED</td>
</tr>
<tr>
<td>Primary intervention</td>
<td>None</td>
<td>None</td>
<td>Eritoran (toll-like receptor 4 antagonist) vs placebo</td>
<td>EGDT as standard care vs usual care</td>
<td></td>
</tr>
<tr>
<td>Primary outcome</td>
<td>28-d mortality</td>
<td>28-d mortality</td>
<td>28-d mortality</td>
<td>28-d mortality</td>
<td>28-d mortality</td>
</tr>
<tr>
<td>Inclusion criteria</td>
<td>Sepsis-3 criteria (defined as both a body fluid culture and administration of antibiotics plus ≥2 SOFA points within first 6 h after ED arrival)</td>
<td>Community-acquired pneumonia with severe sepsis (defined by Sepsis-2 criteria and ≥3 SOFA points)</td>
<td>Severe sepsis and septic shock (defined as meeting ≥3 SIRS criteria, dysfunction in ≥1 organ, and APACHE II score of 21-37)</td>
<td>Severe sepsis and septic shock (defined as meeting ≥3 SIRS criteria and dysfunction in ≥1 organ)</td>
<td>Severe sepsis or septic shock (defined as meeting ≥2 SIRS criteria and either serum lactate level ≥4 mmol/L or refractory hypotension)</td>
</tr>
</tbody>
</table>

Abbreviations: APACHE, Acute Physiology and Chronic Health Evaluation; ED, emergency department; EGDT, early goal-directed therapy; ICU, intensive care unit; SIRS, systemic inflammatory response syndrome; SOFA, Sequential Organ Failure Assessment.

There were 16,552 unique patients among 20,189 total patients in the derivation cohort and 31,160 unique patients among 43,086 total patients in the validation cohort.

Indicates total with available data.

Measured among unique patients in each cohort.

© 2019 American Medical Association. All rights reserved.
considered reflective of the host response for sepsis and are included broadly under the domains of inflammatory, endothelial, coagulation, and vital organ function (eTable 2 in the Supplement). The primary clinical outcome was 28-day mortality in the SENECA project derivation and validation cohorts and in the GenIMS, ACCESS, and PROWESS trials. The primary clinical outcome was hospital mortality truncated at 60 days in the ProCESS trial. One-year mortality was studied in the ACCESS, PROWESS, and ProCESS trials. Other outcomes included for exploratory analyses included intensive care unit admission during hospitalization, total days of administration of a vasopressor, and total days of mechanical ventilation during hospitalization.

**Statistical Methods**
To derive the phenotypes, we first assessed the candidate variable distributions, missingness, and correlation (eTable 3 in the Supplement). Multiple imputation with chained equations was used to account for missing data (eTable 4 and eMethods in the Supplement) and log transformation was used for nonnormal data. After evaluating correlation, we excluded highly correlated variables using rank-order statistics in the sensitivity analyses (eFigure 2 in the Supplement). Ordering points to identify the clustering structure (OPTICS) plots were used to determine the optimal clustering strategy. Based on these plots, we applied consensus $k$ means clustering to 29 variables using a partitioning approach. To determine the optimal number of phenotypes with consensus $k$ means clustering, we evaluated a combination of phenotype size, clear separation of the consensus matrix heatmaps, characteristics of the consensus cumulative distribution function plots, and adequate pairwise-consensus values between cluster members ($>0.8$). Once optimal phenotype number was determined, patterns of clinical variables were visualized in 3 ways: (1) $t$-distributed stochastic neighbor embedding plots (which show multidimensional data in 2 dimensions), (2) alluvial plots (which show the proportional distribution of phenotype members across specific variables), (3) chord diagrams (which show how phenotypes differ by major variable groups; eMethods in the Supplement), and (4) ranked plots of variables by the mean standardized difference between the phenotype pairs.

To assess the reproducibility of the phenotypes, we first used a latent class analysis to derive the groups (eMethods in the Supplement). In the latent class analysis, the optimal phenotype number was confirmed using a combination of Bayesian information criteria, adequate size, high median probabilities of group membership within each phenotype, maximum entropy (a measure between 0 and 1 indicating better classification), and clinical features of potential groups. We also determined the proportion of patients with a probability of phenotype assignment on the margin, which was defined as between 45% and 55%. We assessed how robust the phenotypes were to sensitivity analyses of the derivation method, including (1) excluding variables with high missingness (eg, erythrocyte sedimentation rate, C-reactive protein, premature neutrophil count [bands]); (2) excluding both highly missing and highly correlated variables (sodium, hemoglobin, blood urea nitrogen, and alanine aminotransferase); and (3) using a 12-hour window of electronic health record data after hospital presentation (eMethods in the Supplement).

To determine the reproducibility in the external data, we used the SENECA validation cohort and rederived groups using consensus $k$ means clustering. Then, in the GenIMS study and in the 3 RCTs, we predicted phenotype based on the clinical characteristics of typical cluster members in the SENECA derivation cohort. Predictions arose from the Euclidean distance from each patient to the centroid of each SENECA phenotype (eMethods in the Supplement). We studied the frequency and clinical characteristics of the predicted phenotype groups in the GenIMS study and in the 3 RCTs.

We determined the correlation of the phenotypes with 27 biomarkers of the host immune response and compared the mean (SD), the median (interquartile range [IQR]), and the ratio of biomarker distributions across phenotypes as appropriate. The $\chi^2$ test was used to compare in-hospital, 28-day, and 365-day mortality. The cumulative mortality was illustrated using probability plots and the differences were tested using the log-rank test.

To understand the implications of the phenotypes on the RCT estimates of the treatment effects, we conducted Monte Carlo simulations (10 000 iterations per simulation) in which the only variable modified was the proportion of phenotypes enrolled in the existing trial data set using random sampling with replacement. Six scenarios were created for each of the 3 trials (eMethods in the Supplement), in which the range of phenotypes was varied. The frequency for the range of phenotypes was informed in simulated trials using upper and lower bounds up to twice that observed across the hospitals in the SENECA derivation and validation cohorts. We also tested logistic regression models for 28-day and 365-day mortality using phenotype, treatment assignment, and their interaction as covariates (eMethods in the Supplement).

Several analyses were conducted to ensure the phenotypes were not simply recapitulations of more traditional clinical groups. First, we tested whether the phenotypes were explained by traditional measures of illness severity, such as the SOFA score or the Acute Physiology and Chronic Health Evaluation (APACHE) score. For the SENECA derivation cohort, alluvial plots were used to inspect whether the phenotypes overlapped with the SOFA score. We also determined the overlap of the phenotypes with the quartiles of APACHE and SOFA scores in the 3 RCTs. We further inspected the biomarker profiles and mortality by APACHE quartile in the ProCESS trial. Simulations in the ProCESS trial were also repeated, varying the proportions of the 4 severity-of-illness quartiles instead of the phenotypes, and comparing the potential causal relationship with the estimates of treatment benefit or harm (eMethods in the Supplement). Second, we explored whether the phenotypes were explained by the site of the infection. In the ACCESS trial, which includes independent adjudication of the source of infection, we generated alluvial plots and the proportions for infection sites across phenotypes. We measured the frequency of the phenotypes in a subset of patients with sepsis from a single source (bacteremia) among patients in the SENECA derivation cohort.

Data are presented as mean (SD) or median (IQR). For comparisons, we used analysis of variance and the Kruskal-Wallis test for continuous data and the $\chi^2$ test for categorical data. The threshold for statistical significance was less than .05 for
2-sided tests. There was no adjustment for the type I error rate due to multiple comparisons; therefore, the findings from these analyses should be considered exploratory. Analyses were performed with Stata version 14.2 (StataCorp) and R versions 3.4.1 and 3.5.0 (R Foundation for Statistical Computing).

## Results

### Patients

Among 1,309,025 patient encounters in the SENECA derivation cohort (eFigure 3 in the Supplement), 87,844 patients (6.7%) had suspected infection within 6 hours of hospital presentation and 20,189 met Sepsis-3 criteria (eTable 5 in the Supplement). The mean SOFA score was 3.9 (SD, 2.4) and the mean serum lactate level was 3.2 mmol/L (SD, 3.2 mmol/L). Among 111,938 encounters in the SENECA validation cohort, more patients had suspected infection (n = 103,259; 9.2%) and met Sepsis-3 criteria within 6 hours (n = 43,086); however, the demographic characteristics and SOFA scores were similar (eTables 5 and 6 in the Supplement). The SENECA validation cohort included 43,086 patients (mean age, 67 [SD, 17] years; 21,993 [51%] male; mean maximum 24-hour SOFA score, 3.6 [SD, 2.0]). Patients in the GenIMS cohort (total of 2,320 enrolled, including 583 patients with sepsis) had more comorbidities and respiratory symptoms (eg, elevated respiratory rate, lower oxygen saturation). Across the SENECA derivation and validation cohorts and the GenIMS cohort, the in-hospital mortality ranged from 6% to 14% and from 16% to 23% among patients who required intensive care. In the 3 RCTs (eTable 7 in the Supplement), a total of 4,737 patients (1,706 in the trial on eritoran, 1,690 in the trial on activated protein C, and 1,341 in the trial on early goal-directed therapy) participated at 392 sites and short-term mortality (ie, at 28 days in 2 trials and at 60 days in 1 trial) ranged from 19% to 28%.

### Derivation of Clinical Sepsis Phenotypes

In the SENECA derivation cohort, the consensus k means clustering models found that a 4-class model was the optimal fit with the 4 phenotypes of α, β, γ, and δ (eFigures 4 and 5 and eTable 8 in the Supplement). Consensus matrix plots and the relative change under the cumulative distribution function curve implied little statistical gain by increasing to a 5- or 6-class model. The size and characteristics of the phenotypes in the 4-class model appear in Table 2 and Figure 1. Phenotypes ranged in size (from 13% to 33% of the cohort) and differed broadly in clinical characteristics and organ dysfunction patterns. When ranking continuous variables by the standardized mean difference between phenotypes (Figure 2), patients with the α phenotype had fewer abnormal laboratory values and less organ dysfunction; those with the β phenotype were older, had greater chronic illness, and were more likely to present with renal dysfunction; those with the γ phenotype were more likely to have elevated measures of inflammation (eg, white blood cell count, premature neutrophil count [bands], erythrocyte sedimentation rate, or C-reactive protein), lower albumin level, and higher temperature; and those with the δ phenotype had elevated serum lactate levels, elevated levels of transaminases, and hypotension (eFigure 6-8 in the Supplement).

Variables such as sex, sodium level, glucose level, and white blood cell count contributed least to phenotype differences (eFigure 9 in the Supplement). Phenotypes also varied across the 12 SENECA hospitals as follows: α phenotype ranged from 24% to 42%; β phenotype ranged from 19% to 30%; γ phenotype ranged from 23% to 50%; and δ phenotype ranged from 5% to 23% (eFigure 10 in the Supplement). There was no difference across phenotypes in the rate of peripheral blood culture as the first body fluid culture after hospital presentation, whereas the rate of intravenous antibiotics (vs other routes of administration) ranged from 76% to 93% (eTable 9 and eFigure 11 in the Supplement).

### Reproducibility

Latent class analysis confirmed the statistical fit of the 4-class model (Figure 2 and eFigure 12 and eTable 10 in the Supplement). Bayesian information criteria decreased as class numbers increased from 2 to 4 while entropy was preserved (>0.8). The clinical characteristics of the phenotypes were similar when derived using this method as well as by visualization with t-distributed stochastic neighbor embedding plots (eFigure 13 and eTable 11 in the Supplement). There was strong separation in the likelihood of membership for patients assigned to a given phenotype compared with those assigned to other phenotypes (eFigure 14 in the Supplement).

Phenotypes also were derived in the SENECA validation cohort and showed similar optimal phenotype numbers, frequency of phenotypes, and clinical characteristics as observed in the primary analysis (Figure 2; eFigures 15 and 16 and eTable 12 in the Supplement). No substantial changes were evident after excluding variables with high missingness (eTable 13 in the Supplement), after excluding variables with both high missingness and correlation (eTable 14 in the Supplement), and when the window for capturing data was expanded to 12 hours after hospital presentation (eTable 15 in the Supplement).

In the GenIMS cohort in which patients had sepsis due only to pneumonia (eMethods and eFigure 17 in the Supplement), all 4 phenotypes were present, albeit with slightly different frequencies compared with the SENECA derivation cohort. The clinical characteristics of the phenotypes were largely the same (eTable 16 in the Supplement). When the phenotypes were predicted in the 3 RCTs (eFigure 18-21 in the Supplement), the frequency distributions and clinical characteristics were also similar to the SENECA derivation cohort (eTable 17 for ACCESS, eTable 18 for PROWESS, and eTable 19 for ProCESS in the Supplement).

### Correlation of Phenotypes With Biomarker Profiles

Broad differences were observed in the distributions of the host-response biomarkers across phenotypes (Figure 3). Of the 27 biomarkers measured in 4 studies, 23 were significantly different across phenotypes in at least 1 study (P < .05). In general, there was an increase in the markers of inflammation and in abnormal coagulation in both the γ and δ phenotypes compared with the α or β phenotypes (Figure 4). For example, in the GenIMS study, the ratio of the δ phenotype to the α phenotype for the

(jama.com)
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Total</th>
<th>Phenotype</th>
<th>α</th>
<th>β</th>
<th>γ</th>
<th>δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of patients (%)</td>
<td>20 189 (100)</td>
<td>6625 (33)</td>
<td>5512 (27)</td>
<td>5385 (27)</td>
<td>2667 (13)</td>
<td></td>
</tr>
<tr>
<td>Age, mean (SD), y</td>
<td>64 (17)</td>
<td>60 (18)</td>
<td>71 (15)</td>
<td>65 (16)</td>
<td>63 (17)</td>
<td></td>
</tr>
<tr>
<td>Sex, No. (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>10 022 (50)</td>
<td>3372 (51)</td>
<td>2624 (48)</td>
<td>2559 (48)</td>
<td>1467 (55)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>10 167 (50)</td>
<td>3253 (49)</td>
<td>2888 (52)</td>
<td>2826 (52)</td>
<td>1200 (45)</td>
<td></td>
</tr>
<tr>
<td>Race, No. (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>15 640 (77)</td>
<td>5165 (78)</td>
<td>4221 (77)</td>
<td>4269 (79)</td>
<td>1985 (74)</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>2428 (12)</td>
<td>805 (12)</td>
<td>797 (14)</td>
<td>539 (10)</td>
<td>287 (11)</td>
<td></td>
</tr>
<tr>
<td>Otherb</td>
<td>2121 (11)</td>
<td>655 (10)</td>
<td>494 (9)</td>
<td>577 (11)</td>
<td>395 (15)</td>
<td></td>
</tr>
<tr>
<td>Organ Dysfunction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elixhauser comorbidities, mean (SD)</td>
<td>1.8 (1.2)</td>
<td>1.5 (1.1)</td>
<td>2.4 (1.2)</td>
<td>1.7 (1.1)</td>
<td>1.7 (1.1)</td>
<td></td>
</tr>
<tr>
<td>Surgery, No. (%)</td>
<td>2727 (14)</td>
<td>696 (11)</td>
<td>786 (14)</td>
<td>825 (15)</td>
<td>420 (16)</td>
<td></td>
</tr>
<tr>
<td>Reached maximum within 24 h, mean (SD)</td>
<td>3.7 (1.0)</td>
<td>3.7 (1.0)</td>
<td>3.7 (1.0)</td>
<td>3.7 (1.0)</td>
<td>3.7 (1.3)</td>
<td></td>
</tr>
<tr>
<td>SIRS criteria</td>
<td>1.8 (1.2)</td>
<td>1.4 (1.1)</td>
<td>1.2 (1.0)</td>
<td>2.3 (1.0)</td>
<td>2.5 (1.1)</td>
<td></td>
</tr>
<tr>
<td>SOFA score</td>
<td>3.9 (2.4)</td>
<td>3.0 (1.4)</td>
<td>3.5 (1.7)</td>
<td>4.0 (2.3)</td>
<td>6.6 (3.7)</td>
<td></td>
</tr>
</tbody>
</table>

(continued)
Table 2. Characteristics of the 4 Phenotypes (continued)

<table>
<thead>
<tr>
<th>Characteristic*</th>
<th>Total</th>
<th>α</th>
<th>β</th>
<th>γ</th>
<th>δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcomes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mechanical ventilation, median (IQR), d&lt;sup&gt;d&lt;/sup&gt;</td>
<td>5 (2-10)</td>
<td>4 (2-9)</td>
<td>4 (2-9)</td>
<td>6 (3-13)</td>
<td>4 (2-9)</td>
</tr>
<tr>
<td>Administration of a vasopressor, median (IQR), d&lt;sup&gt;d&lt;/sup&gt;</td>
<td>3 (2-5)</td>
<td>2 (2-4)</td>
<td>3 (2-4)</td>
<td>3 (2-5)</td>
<td>3 (2-5)</td>
</tr>
<tr>
<td>Admitted to intensive care unit, No. (%)&lt;sup&gt;d&lt;/sup&gt;</td>
<td>9063 (45)</td>
<td>1644 (25)</td>
<td>1778 (32)</td>
<td>3381 (63)</td>
<td>2260 (85)</td>
</tr>
<tr>
<td>In-hospital mortality, No. (%)</td>
<td>2082 (10)</td>
<td>126 (2)</td>
<td>286 (5)</td>
<td>818 (15)</td>
<td>852 (32)</td>
</tr>
</tbody>
</table>

Abbreviations: IQR, interquartile range; SIRS, systemic inflammatory response syndrome; SOFA, Sequential Organ Failure Assessment.

<sup>d</sup> Includes Chinese, Filipino, Hawaiian, American Indian/Alaskan Native, Asian, Hawaiian/others Pacific Islander, Middle Eastern, Native American, not specified, or Pacific Islander.

* Correlates to minimum or maximum value (as appropriate) within 6 hours of hospital presentation. The variables in this table were log transformed for modeling (eTable 3 in the Supplement). Comparisons across all 4 phenotypes were performed using the Kruskal-Wallis test, analysis of variance, or the x² test (P < .01 for all comparisons).

The median level of IL-6 was 5.0 (IQR, 1.6-13.2); in the ACCESS trial, it was 7.7 (IQR, 1.4-16.6); in the PROWESS trial, it was 3.0 (IQR, 0.7-24.6); and in the ProCESS trial, it was 8.3 (IQR, 1.4-67.7). Similar findings comparing the δ phenotype vs the α phenotype were present for IL-10 level (ranges of levels across the studies for median level of IL-10, 1.3-6.2), but were less prominent for tumour necrosis factor (range of levels across the studies for tumour necrosis factor, 1.0-4.6; Figure 3).

Coagulation markers such as thrombin-antithrombin complex, plasminogen activator inhibitor 1, and D-dimer were significantly greater in the δ phenotype compared with the other phenotypes (P < .001; Figure 4 and eTables 20-23 in the Supplement). The levels of some markers of endothelial dysfunction (eg, intercellular adhesion molecule 1, E-selectin) were highest in the γ phenotype (P < .01), other markers were higher in the δ phenotype (eg, vascular cell adhesion molecule 1), and other markers were not different across groups (eg, P-selectin, P = .37). Markers of renal injury (eg, insulin-like growth factor-binding protein 7, collagen type 4, tissue inhibitor of metalloproteinase 2) were highest in both the β and δ phenotypes (P < .01).

Relationship With Mortality and Organ Support

Phenotypes were associated with short- and long-term outcomes (eTables 24 and 25 in the Supplement). In the SENeca derivation cohort, the fewest in-hospital deaths occurred in the α phenotype (n = 126; 2%) compared with the β phenotype (n = 286; 5%), the γ phenotype (n = 818; 15%), and the δ phenotype (n = 852; 32%) (P < .001). Across all cohorts and trials, the 28-day mortality (Figure 5) and the 365-day mortality (eFigure 22 in the Supplement) were highest in the δ phenotype compared with the other phenotypes (P < .001). In the SENeca derivation cohort (n = 16,552 unique patients), cumulative 28-day mortality was 287 of 5691 (5%) for the α phenotype, 561 of 4420 (13%) for the β phenotype, 1031 of 4318 (24%) for the γ phenotype, and 897 of 2223 (40%) for the δ phenotype. In the SENeca validation cohort (n = 31,600 unique patients), cumulative 28-day mortality was 837 (9%) for the α phenotype, 923 (11%) for the β phenotype, 854 (9%) for the γ phenotype, and 1278 (29%) for the δ phenotype. Intensive care unit admission rates were higher in the δ phenotype compared with the other phenotypes (P < .01), whereas days of mechanical ventilation and administration of a vasopressor were variable across studies.

Differential Estimated Treatment Effects by Phenotype and Sensitivity of the Clinical Trial Results to Changes in Phenotype Distributions in the Trial Simulations

The estimated treatment effects by phenotype were variable in the observed data in the ACCESS, PROWESS, and ProCESS trials (eFigures 23-28 in the Supplement). Standard treatment × phenotype interactions were only significant in the ProCESS trial, but not for the other 2 trials based on the P < .05 criteria. The primary findings of the trial simulations appear in Figure 6 (more detailed examples appear in eFigures 29-31 in the Supplement). In general, the trials had similar baseline characteristics between simulation scenarios and original trial populations. For example, a doubling of the δ phenotype did not change the demographics and increased the mean baseline SOFA score from 7.2 (SD, 3.6) points to only 8.6 (SD, 3.6) points in the ProCESS trial (eTables 26-28 in the Supplement). The mortality rates for the control group were also stable across the simulations and were within the typically reported ranges (eTable 29 and eFigure 32 in the Supplement). For example, a doubling of the highly morbid δ phenotype was only associated with an increase in the mortality rate for usual care from 26% to 31% in the ACCESS trial, from 31% to 39% in the PROWESS trial, and from 19% to 26% in the ProCESS trial.

The trial conclusions about the treatment effects were relatively robust to large changes in the proportion of patients with the β and γ phenotypes. Despite modest changes to the baseline characteristics in the trial populations, the changes to the distributions for the α and δ phenotypes had substantial effects (Figure 6). For example, in the ProCESS trial, which under
In A, the ribbons connect from an individual phenotype to an organ system if the group mean is greater or lesser than the overall mean for the entire cohort. For example, the δ phenotype (light blue) is more likely to have members with abnormal cardiovascular and hepatic dysfunction (ribbons connect with these portions of the circle) vs β phenotype members (light purple) who are more likely to have kidney dysfunction and other abnormal variables (eg, increased age, comorbidity). In B-E, each phenotype is highlighted separately and the ribbons connect to the different patterns of clinical variables and organ system dysfunctions on the top of the circle.
Figure 2. Comparison of Variables That Contribute to Clinical Phenotypes in the SENECA Derivation Cohort (n = 20 189)

In all panels, the variables are standardized such that all means are scaled to 0 and SDs to 1. A value of 1 for the standardized variable value (x-axis) signifies that the mean value for the phenotype was 1 SD higher than the mean value for both phenotypes shown in the graph as a whole. ALT indicates alanine transaminase; AST, aspartate transaminase; Bands, also known as premature neutrophil count; BUN, blood urea nitrogen; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; GCS, Glasgow Coma Scale; INR, international normalized ratio; PaO₂, partial pressure of oxygen; SENECA, Sepsis Endotyping in Emergency Care; SBP, systolic blood pressure; WBC, white blood cell.
the baseline phenotype distribution had a 0% chance of finding benefit with early goal-directed therapy for 60-day inpatient mortality (and an 85% and 15% chance of finding no difference or harm, respectively), the chance of finding benefit increased to 35% when the α phenotype represented the majority of the population (eFigure 29 in the Supplement).

In contrast, when the δ phenotype was increased to 50% of the ProCESS trial population, there was a greater than 60% chance of finding that early goal-directed therapy was harmful. In the ACCESS trial (eFigure 30 in the Supplement), which under the baseline phenotype distribution had a 0% chance of finding benefit, a 91% chance of finding no difference, and a 9% chance of finding harm for 28-day mortality, an increase in the δ phenotype from 14% to 44% of the trial population resulted in 29% of simulated trials concluding etorfan caused harm. In the PROWESS trial, which had an 82% chance of finding a positive effect with the baseline phenotype distribution, 50% of the simulated trials showed no difference when the frequency of the α phenotype was increased to represent the majority of the trial population (eFigure 31 in the Supplement).

Comparison With Traditional Subgroups of Patients With Sepsis

The 4 phenotypes could not be described by severity of illness or site of infection alone. In the SENECA derivation cohort, all 4 phenotypes included both patients with and without organ dysfunction in all SOFA categories (Figure 1). The mean SOFA scores at hospital presentation were lower in patients with the α phenotype (3.0 [SD, 1.4]) and higher in patients with the δ phenotype (6.6 [SD, 3.7]), but overlapped in patients with the β phenotype (3.5 [SD, 1.7]) and in those with the γ phenotype (4.0 [SD, 2.3]) (Table 2 and eFigures 33 and 34 in the Supplement). In the ACCESS trial, although the δ phenotype had a greater proportion of patients with intraabdominal infections, there was a broad distribution for site of infection in each phenotype (eFigure 35 and eTable 30 in the Supplement). There was a similarly broad distribution for phenotypes among patients with sepsis due to bacteremia alone (n = 1714; eFigure 36 in the Supplement).

In the analyses to further explore whether the derived phenotypes were proxies for severity of illness, the pattern of baseline clinical variables and host-response biomarkers differed across the APACHE quartiles from the pattern for the 4 phenotypes (eTables 31 and 32 and eFigures 37 and 38 in the Supplement). The range of short-term mortality rates across the APACHE III quartiles was similar to the range across the 4 phenotypes (eFigure 38 in the Supplement). However, enrichment of the ProCESS trial using APACHE III quartiles was associated with smaller changes in the trial conclusions compared with phenotype enrichment (eFigure 39 in the Supplement).

Discussion

In this retrospective analysis of data sets from patients with sepsis, 4 clinical phenotypes of sepsis were derived using...
Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis

The phenotypes were identified using clustering methods and were derived from data collected at the time of hospital presentation. The phenotypes were multidimensional, differed in their demographics, laboratory abnormalities, patterns of organ dysfunction, and were not homologous with traditional patient groupings such as by site of infection, organ dysfunction patterns, or severity of illness. The frequency and characteristics of the phenotypes were reproducible in additional cohorts and using different machine learning methods. The 4 sepsis phenotypes were strongly correlated with patterns of the host immune response, mortality, and other clinical outcomes. In simulations of 3 large, multicenter trials, conclusions about the estimated treatment benefit or harm were sensitive to phenotype distributions, especially the α and δ phenotypes.

These sepsis phenotypes can be identified at the time of patient presentation to the emergency department, and thus could be useful with regard to early treatment and enrollment in clinical trials. Only routinely available data were used in the clustering models, and the phenotypes were derived from a large observational cohort to ensure generalizability. Phenotype frequency distributions and characteristics were similar in studies with different definitions for sepsis. For example, the SENECA derivation and validation cohorts used electronic health record criteria for Sepsis-3,2 the GenIMS study used Sepsis-2,10 the ProCESS trial enrolled patients with early septic shock and used broad sepsis criteria, and both the ACCESS and PROWESS trials enrolled patients later in their clinical course and patients with more organ failure.

Of the 4 phenotypes identified, the δ phenotype was most strongly correlated with abnormal values of host-response biomarkers as well as clinical features of cardiovascular and liver dysfunction. These characteristics are similar to previously reported subclasses, including the hyperinflammatory subphenotype reported in acute respiratory distress syndrome, a condition most commonly caused by sepsis.18 The δ phenotype also resembles sepsis endotypes derived using transcriptomic analyses of circulating immune cells (such as the inflammatropic cluster, sepsis response signature 1, or the Molecular Diagnosis and Risk Stratification of Sepsis [MARS] 2 cluster) described in patients with sepsis in the intensive care unit.3–5 In contrast, the α phenotype had fewer laboratory abnormalities and less septic shock, which resembles the MARS 3 and sepsis response signature 2 endotypes reported in the same series, and which were found to have predominant expression of adaptive immune and B-cell development pathways.3–5 This concordance between clinical phenotypes and more computationally intensive transcriptomic endotypes could help identify subsets of patients most likely to benefit from particular immunomodulation strategies.

Figure 4. Ratio of Additional Biomarkers in Heatmap

Heatmap shows the ratio of the median biomarker value for various markers of the sepsis host response grouped by those reflecting coagulation, endothelium, inflammation, and renal injury. Orange represents a greater median biomarker value for that phenotype compared with the median for the entire study, whereas colors in the tan to brown range represent lower median biomarker values compared with the median for the entire study. Empty cells are those for which the biomarker was not measured. The factor V, factor X, plasminogen, protein C, and protein S biomarkers were reversed on the scale to coordinate the color map. The IL-1β and IL-12 biomarkers are not shown due to having less than 0.5-fold changes. ACCESS indicates A Controlled Comparison of Eritoran in Severe Sepsis; COL-4, collagen type 4; GenIMS, Genetic and Inflammatory Markers of Sepsis; ICAM, intercellular adhesion molecule 1; IGFBP-7, insulin-like growth factor–binding protein 7; KIM-1, kidney injury molecule 1; PAI-1, plasminogen activator inhibitor 1; ProCESS, Protocol-Based Care for Early Septic Shock; PROWESS, Activated Protein C Worldwide Evaluation in Severe Sepsis; TAT, thrombin–antithrombin; TIMP-2, tissue inhibitor of metalloproteinase 2; TNF, tumor necrosis factor; VCAM, vascular cell adhesion molecule.

2013
Figure 5. Short-term Mortality by Phenotype

A SENECA derivation cohort (n = 16 652) 

B SENECA validation cohort (n = 31 160) 

C GenIMS cohort (n = 583) 

D ACCESS trial (n = 1706) (eritoran vs placebo) 

E PROWESS trial (n = 1690) (drotrecogin alfa vs placebo) 

F ProCESS trial (n = 1341) (EGDT vs protocolized standard care vs usual care) 

All panels show significant differences in mortality by phenotype (log-rank P < .001). In the SENECA derivation and validation cohorts, in the GenIMS cohort, and in the 3 randomized clinical trials, clinical phenotypes are associated with short-term mortality. This suggests that phenotypes are generalizable and prognostic across data sets with different severity, temporality, and definitions of sepsis and septic shock. ACCESS indicates A Controlled Comparison of Eritoran in Severe Sepsis; GenIMS, Genetic and Inflammatory Markers of Sepsis; ProCESS, Protocol-Based Care for Early Septic Shock; PROWESS, Activated Protein C Worldwide Evaluation in Severe Sepsis; SENECA, Sepsis Endotyping in Emergency Care.

* The cumulative mortality data are only for unique patients in the SENECA derivation cohort (16 652 of 20 189 total patients) and in the SENECA validation cohort (31 160 of 43 086 total patients).
In RCT simulations, variations in the phenotypes had small changes in the distribution of average baseline characteristics, yet resulted in unstable trial conclusions. For example, the ACCESS trial found no benefit from eritoran on 28-day mortality. Yet, when the δ phenotype (the phenotype with the greatest proportion of intraabdominal infections) was increased shown in the top set of graphs by varying α (panel B) and δ (panel C). ACCESS indicates A Controlled Comparison of Eritoran in Severe Sepsis; EGDT, early goal-directed therapy; HBN, harm, benefit, or neutral; ProCESS, Protocol-Based Care for Early Septic Shock; PROWESS, Activated Protein C Worldwide Evaluation in Severe Sepsis; SENECA, Sepsis Endotyping in Emergency Care.

In RCT simulations, variations in the phenotypes had small changes in the distribution of average baseline characteristics, yet resulted in unstable trial conclusions. For example, the ACCESS trial found no benefit from eritoran on 28-day mortality. Yet, when the δ phenotype (the phenotype with the greatest proportion of intraabdominal infections) was increased
to nearly half of the trial population, more than one-third of the simulated trials suggested harm from eritoran. This finding is consistent with animal models that suggest toll-like receptor 4 signaling aids bacterial clearance from the peritoneum in patients with intraabdominal sepsis.21 The high proportion with activated protein C and no benefit when the proportion of patients with the α phenotype was increased in the simulated PROWESS trial raises the possibility that such patients were also more common in the subsequent negative trials of activated protein C.22,23

The largest changes were seen in the ProCESS trial, which found no benefit from early goal-directed therapy compared with usual care. In simulations, when the δ phenotype was increased, early goal-directed therapy was harmful in more than half of the trials. This finding supports data from 2 RCTs conducted in low- to middle-income countries that found harm from early goal-directed therapy in select populations.24,25 Increases in the α phenotype suggested benefit from early goal-directed therapy, similar to the initial report by Rivers et al.26 These data highlight the importance of characterizing the heterogeneity of sepsis when comparing across trials with different conclusions.

These findings have additional implications. First, completed trials may have unrecognized heterogeneity in the treatment effects by clinical phenotype that were not apparent when analyzing (I) the entire cohort, (2) subgroups based on individual variables, or (3) stratification based on risk of death.27 However, a secondary analysis of treatment × phenotype interactions may be limited by small sample sizes. Second, these proof-of-concept clinical phenotypes could be incorporated prospectively in future study designs that test new biologically active therapeutics. Novel designs could enrich for a priori phenotypes as well as confirm the boundaries around predictive phenotypes during the trial.28

**Limitations**

This study has several limitations. First, only routinely available clinical data in the electronic health record were used to identify phenotypes, and the inclusion of other data such as clinicians’ impression, protein biomarkers, immune cell gene expression, or pathogen variables during derivation could change phenotype assignments. However, there appears to be some similarity between the clinical phenotypes derived in this study and those described in other series using such data.

Second, the statistical approach involved a variety of supervised decisions such as (1) the time window for capturing data at hospital presentation was 6 hours, (2) the selection of candidate variables, and (3) the handling of variable distributions. Changes to the initial assumptions, the time window for data capture, or the choice of optimal cluster number could alter the results. The findings were consistent when the electronic health record window of 12 hours was used.

Third, because missing data were common for some variables included in the clustering models, multiple imputation was used in the primary analysis. However, variables with high missingness were excluded from the sensitivity analyses and similar results were still found.

Fourth, differences in short- and long-term prognosis were present across phenotypes, perhaps due to different features of the validation cohorts, such as the definition of sepsis, demographics, or burden of organ dysfunction.

Fifth, characteristics of clinical phenotypes were derived initially from a single integrated health system in the United States at a single moment in clinical care. Although phenotypes were found to be generalizable in the other data sets examined, further exploration is necessary, especially using data from low- and middle-income countries, more recent clinical trials, and longitudinal cohorts.

**Conclusions**

In this retrospective analysis of data sets from patients with sepsis, 4 clinical phenotypes were identified that correlated with host-response patterns and clinical outcomes, and simulations suggested these phenotypes may help in understanding heterogeneity of treatment effects. Further research is needed to determine the utility of these phenotypes in clinical care and for informing trial design and interpretation.

**Author Contributions:** Dr Seymour had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. **Concept and design:** Seymour, Kennedy, Chang, Clermont, Cooper, Gomez, Opal, van der Poll, Vodovotz, Yealy, Angus. **Acquisition, analysis, or interpretation of data:** Seymour, Kennedy, Wang, Chang, Elliott, Xu, Berry, Huang, Kellum, Talisa, Visweswaran, Vodovotz, Weiss, Yende. **Drafting of the manuscript:** Seymour, Kennedy, Wang, Elliott, Berry, van der Poll, Vodovotz, Yealy, Angus. **Critical revision of the manuscript for important intellectual content:** Seymour, Kennedy, Chang, Elliott, Xu, Berry, Clermont, Cooper, Gomez, Huang, Kellum, Mi, Talisa, Visweswaran, Vodovotz, Weiss, Yende, Angus. **Statistical analysis:** Seymour, Kennedy, Wang, Chang, Elliott, Xu, Berry, Mi, Talisa, Weiss, Angus.
REFERENCES