Online Supplement to accompany:

Sequence Patterns in the Resolution of Clinical Instabilities in Community-Acquired Pneumonia and Association with Outcomes

Gavin W. Hougham, PhD[1,2], Sandra A. Ham, MS[2], Gregory W. Ruhnke, MD, MS, MPH[1], Elizabeth Schulwolf, MD[3], Andrew D. Auerbach, MD[4], Jeffrey L. Schnipper, MD[5], Peter J. Kaboli, MD[6], Tosha B. Wetterneck, MD, MS[7], David Gonzalez, MD[8], Vineet M Arora, MD, MAPP[9], David O. Meltzer, MD, PhD[1,2]

[1]Department of Medicine/Section of Hospital Medicine, University of Chicago, Chicago, IL, USA; [2]Center for Health and the Social Sciences, University of Chicago, Chicago, IL, USA; [3]Loyola University Medical Center, Chicago, IL, USA; [4]University of California, San Francisco, CA, USA; [5]Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA; [6]University of Iowa, Iowa City, IA, USA; [7]University of Wisconsin School of Medicine and Public Health, Madison, WI, USA; [8]University of New Mexico, Rio Rancho, NM, USA; [9]Department of Medicine/Section of General Internal Medicine, University of Chicago, Chicago, IL, USA.

This paper tests whether a set of techniques to analyze categorically coded data in sequences, as developed in the fields of genomics, computational linguistics, and quantitative sociology, can be used to help generate and answer questions in clinical medicine. We provide here additional information about the analysis in text, tabular, and graphic form.

Supplementary Table S1 compares key features distinguishing typical, multi-channel, and hybrid approaches to the analysis of categorically coded data in sequences. Traditional methods of displaying and analyzing series of numbers (as in, e.g., time series analysis) are ill-suited to analyzing strings of events or states over time, while sequence analysis methods have been developed for just this purpose. Key features of a "simple" sequence analysis are shown in the first column of Supplementary Table S1, where items on the X-axis represent linear time (as in, e.g., a trajectory of events or states, such as a series of occupational roles in a career trajectory) or linear space (as in, e.g., a string of nucleotides comprising a section of DNA). Each position in the trajectory or string corresponding to a unique event, state, or nucleotide can be coded with a unique symbol ("A" "B" "C" or "B" "C" "A"…) and the whole string thereby comprising a unit of analysis for comparison with other such strings. If the analyst wishes to track the ordered occurrences of more than one phenomenon at a time (say, one's lifetime marital history or residential locale history in parallel with one's job history), then multi-channel sequence analysis may be appropriate. In multi-channel analysis, each phenomenon is tracked and coded separately, and conceptualizing this is not difficult, but the analytic methods to do so are not well developed.[1] The third column in the table describes the approach we used, which was based on determining whether and when each clinical indicator for each patient resolved.

Supplementary Figure S1 is similar to the proportional distributions in Figure 2 of the main paper showing the rank order of indicator stabilization for the whole sample, but stratified by participating site. This figure shows the proportional distributions of the rank order of indicator stabilization as stacked bars, with different shades of color assigned to each order of stabilization. By inspection, these plots show similar distribution patterns across sites, suggesting that no one site was disproportionately influential in later stages of analysis.

Supplementary Figure S2 shows sequence index plots of the eight clusters with representative sequences (N = 1,326). The indicator stabilization patterns for all 1,326 patients are shown here, stratified by derived cluster (one panel for each). Each panel shows all patient sequences for that cluster using sequence index plots in the "Sequences" part of the panel, which show every sequence observed. A smaller set of two representative sequences from each cluster are in the "Representatives" part of each panel. The representativeness statistic denotes the percent of all sequences in each cluster that are similar to the illustrated sequences within a moderate radius of optimal matching dissimilarity scores around the representative sequences. These representative sequences are defined as being within a few changes in ranks from all sequences in that cluster[2] and reflect the homogeneity/heterogeneity of sequence patterns within each cluster.

Supplementary Figure S3 shows plots of sequence clusters by cluster means of outcome measures. Panels A-C plot mean 30-day mortality, mean LOS, and mean total hospitalization costs by time to maximum stabilization for each sequence cluster (N=1,326). For example, 30-day mortality is lowest for patients in the Tachycardic and Febrile cluster, intermediate for those in the Tachycardic/Hypoxic, Stabilized Fast, and Stabilized Slow clusters; and highest in the Hypoxic, Mental/Oral/Hypoxic, and Inpatient Mortality clusters.

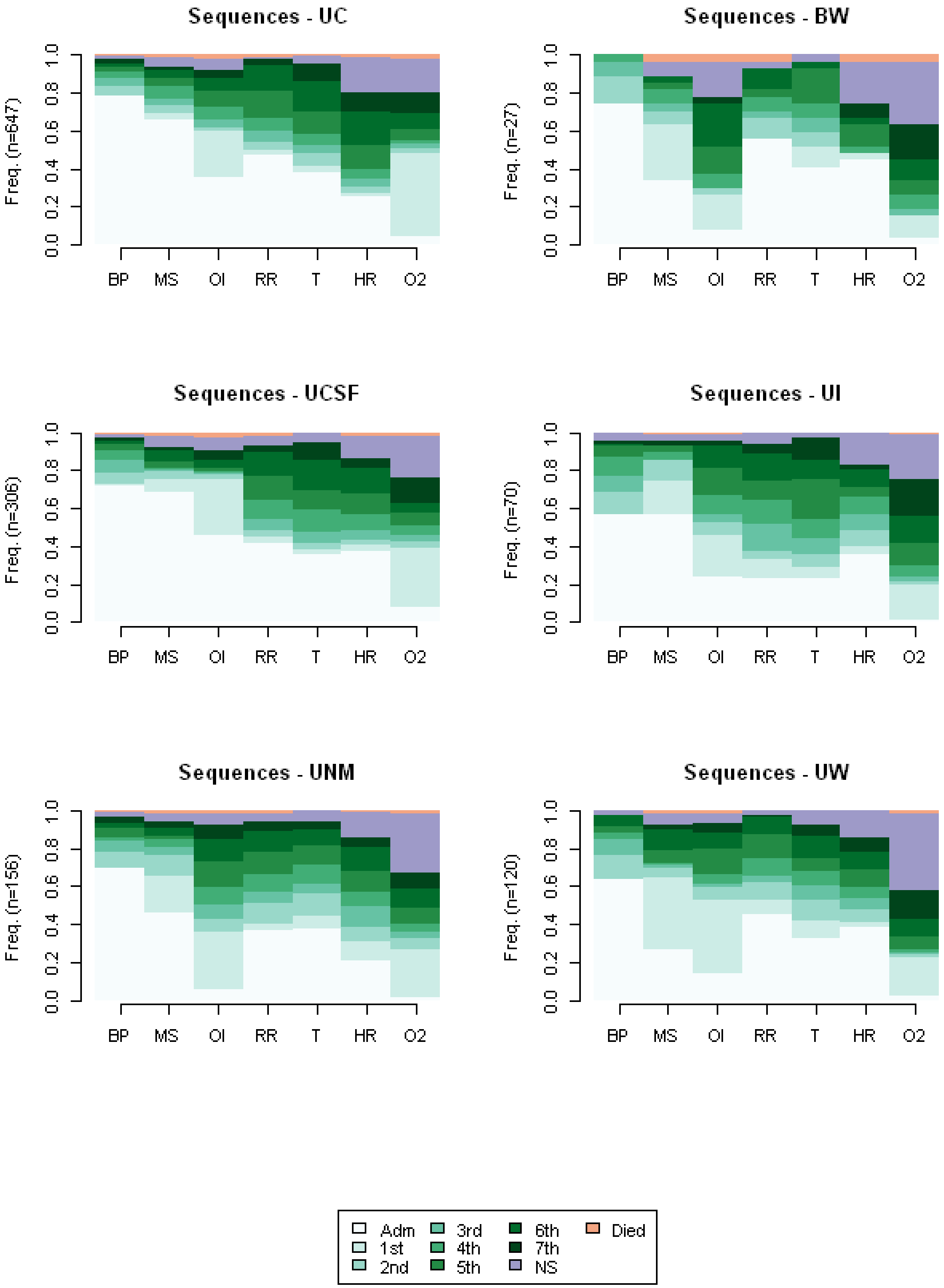
REFERENCES

1. **Gauthier JA, Widmer ED, Bucher P, Notredame C**. Multichannel Sequence Analysis Applied to Social Science Data. Sociol Methodol. 2010;40:1-38. doi:DOI 10.1111/j.1467-9531.2010.01227.x
2. **Gabadinho A, Ritschard G, Studer M, Muller NS**. Mining sequence data in R with the TraMineR package: A user's guide. University of Geneva, 2011. (http://mephisto.unige.ch/traminer/) Accessed July 10, 2013.

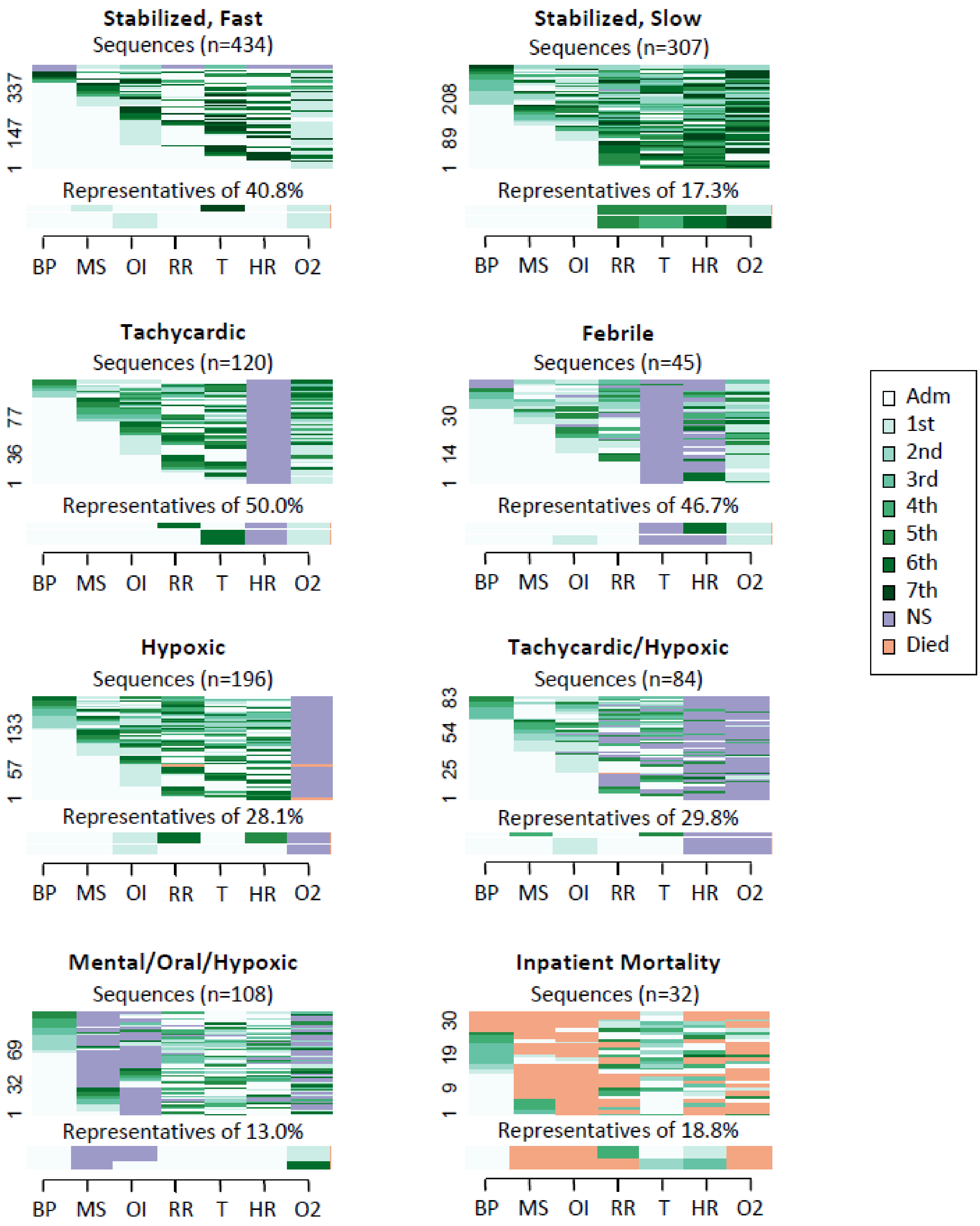
Supplementary Table S1. Comparison of key features distinguishing typical, multi-channel, and hybrid approaches to the analysis of categorically coded data in sequences.

	Typical sequences	Multi-channel sequences	Hybrid sequences used in this study
Orienting dimension	X-axis represents time	X-axis represents time	X-axis represents categorical events
Time signature	Time is usually in regular intervals	Time is usually in regular intervals	Time is often in irregular intervals
What are the sequences representing?	Sequences are categorical states or categorical events	Sequences are categorical states or categorical events	Sequences are categories of relative temporal order
How to show state sequences	States possess duration, shown using index plots	States possess duration, shown using symbolic sequence enumeration	Event orders are treated as states, shown using sequence index plots
Analytical approach for event sequences	Events analyzed for transitions from one state to another	Events analyzed for transitions from one state to another	Events (ranks, for indicator stabilization) analyzed for sequential order
How many sequence trajectories are involved?	Single state occurs at one time	Multiple dimensions co-occur, treated as separate trajectories	Multiple events occur at one time, treated as a single trajectory
Example sequences	A-B-C-D-E-F-G	B-B-B-B-B-B-B-B-B G-G-H-H-H-H-H-H-H K-K-L-L-L-L-L-L-L-L N-N-N-N-O-O-O-O-O R-R-R-R-R-R-R-R-S U-U-U-U-U-V-V-V-V X-X-X-Y-Y-Y-Y-Y-Y	1st-2nd-2nd-5th-7th-6th-4th

This table compares three analytical approaches to sequence analysis. The first and most widely used typical approach organizes time along the X-axis, where a series of categorical states or events are arrayed in a single linear trajectory (In the context of computational genomics, this type of sequence coding represents not time but physical location in linear space). Examples of this approach might be looking at the order and spacing history of vaccinations during childhood, or a work history in adulthood. This conception may be extended to a multi-channel or multi-dimensional approach by adding other categorical state or event trajectories in parallel. An example of this approach might be looking at work history over time in parallel with marital history or locale of residence over time. The hybrid sequence analysis used in this study aligns categorical events along the X-axis, where sequences capture relative temporal order of the categorical events. In this study, the seven clinical instabilities are assigned ranks according to when they attained stabilization. This coding scheme allows for the possibility of instabilities to attain stabilization one-by-one during the course of a hospitalization or in tandem with one or more instabilities. Analysis then proceeds on the ranks as indicated in the example sequence 1st-2nd-2nd-5th-7th-6th-4th.

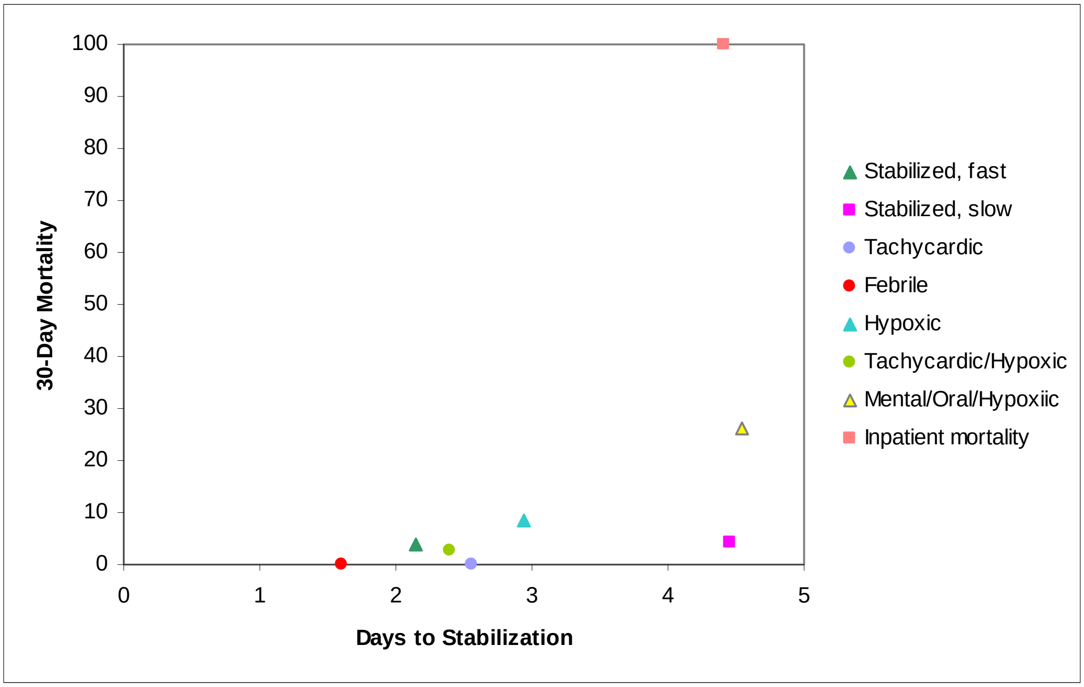

Sequences - UC
Freq. (n=647)
Sequences - BW
Freq. (n=27)
Sequences - UCSF
Freq. (n=306)
Sequences - UI
Freq. (n=70)
Sequences - UNM
Freq. (n=156)
Sequences - UW
Freq. (n=120)
0.0
0.2
0.4
0.6
0.8
1.0
BP
MS
OI
RR
T
HR
O2
Adm
1st
2nd
3rd
4th
5th
6th
7th
NS
Died

Supplementary Figure S1. Frequency distributions of the rank order of indicator stabilizations by site (N=1,326). Similar to Figure 2 in the main paper, but stratified by participating site, this figure shows the proportional distributions of the rank orders of indicator stabilization as stacked bars, with different shades of color assigned to each order of stabilization. UC=University of Chicago; BWH=Brigham and Women's Hospital; UCSF=University of California at San Francisco; UI=University of Iowa; UNM=University of New Mexico; UW=University of Wisconsin. White=indicator stable at admission; Varying shades of green=stabilized nth or tied for nth (i.e., 1st through 7th); Purple=Indicator not stabilized before discharge; Orange=Patient died in-hospital before indicator stabilized. BP=blood pressure; MS=return to baseline mental status; OI=ability to feed by oral intake; RR=respiratory rate; T=temperature; HR=heart rate; O2=Blood oxygen saturation.

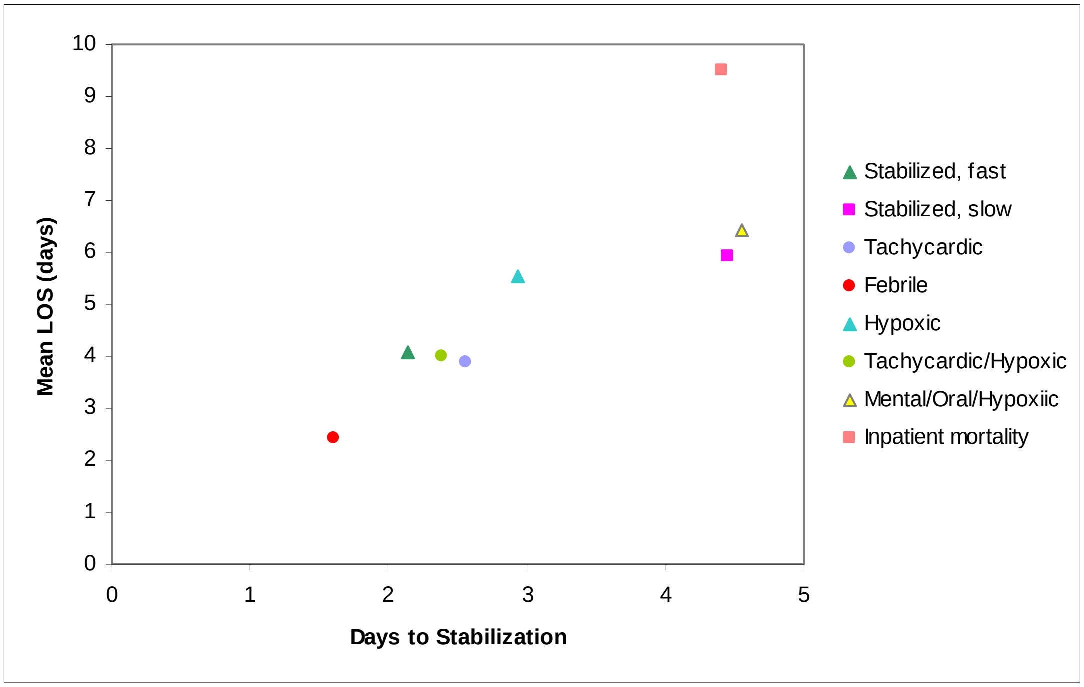

Stabilized, Fast
Sequences (n=434)
1
147
337
Representatives of 40.8%
BP MS OI RR T HR O2
Stabilized, Slow
Sequences (n=307)
1
89
208
Representatives of 17.3%
BP MS OI RR T HR O2
Tachycardic
Sequences (n=120)
1
36
77
Representatives of 50.0%
BP MS OI RR T HR O2
Febrile
Sequences (n=45)
1
14
30
Representatives of 46.7%
BP MS OI RR T HR O2
Hypoxic
Sequences (n=196)
1
57
133
Representatives of 28.1%
BP MS OI RR T HR O2
Tachycardic/Hypoxic
Sequences (n=84)
1
25
54
83
Representatives of 29.8%
BP MS OI RR T HR O2
Mental/Oral/Hypoxic
Sequences (n=108)
1
32
69
Representatives of 13.0%
BP MS OI RR T HR O2
Inpatient Mortality
Sequences (n=32)
1
9
19
30
Representatives of 18.8%
BP MS OI RR T HR O2
□ Adm
□ 1st
□ 2nd
□ 3rd
□ 4th
□ 5th
□ 6th
□ 7th
□ NS
□ Died

Supplementary Figure S2. Sequence index plots of the eight clusters with representative sequences (*N* = 1,326). The indicator stabilization patterns for all 1,326 patients are shown here, stratified by cluster. Each cluster shows all patient sequences for that cluster in the top position of the panel, and a smaller set of representative sequences from the cluster are in the bottom position of each panel. The representativeness statistic denotes the percent of all sequences in a cluster that are similar to the illustrated sequence or sequences within a moderate radius of optimal matching dissimilarity scores around the representative sequence(s). White=indicator stable at admission; Varying shades of green=stabilized nth or tied for nth (i.e., 1st through 7th); Purple=Indicator not stabilized before discharge; Orange=Patient died in-hospital before indicator stabilized. BP=blood pressure; MS=return to baseline mental status; OI=ability to feed by oral intake; RR=respiratory rate; T=temperature; HR=heart rate; O2=Blood oxygen saturation.

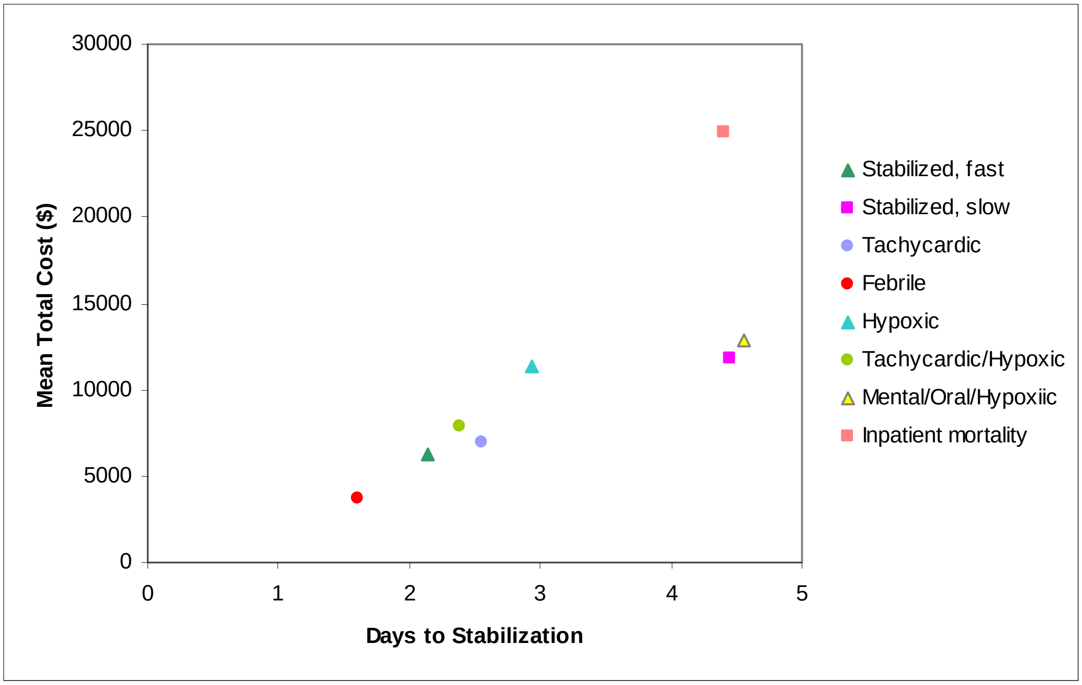
A.


100
90
80
70
60
50
40
30
20
10
0
30-Day Mortality
0
1
2
3
4
5
Days to Stabilization
▲ Stabilized, fast
■ Stabilized, slow
● Tachycardic
● Febrile
▲ Hypoxic
● Tachycardic/Hypoxic
△ Mental/Oral/Hypoxiic
■ Inpatient mortality

B.


10
9
8
7
6
5
4
3
2
1
0
Mean LOS (days)
0
1
2
3
4
5
Days to Stabilization
▲ Stabilized, fast
■ Stabilized, slow
● Tachycardic
● Febrile
▲ Hypoxic
● Tachycardic/Hypoxic
△ Mental/Oral/Hypoxiic
■ Inpatient mortality

C.



Supplementary Figure S3. Plots of sequence clusters by cluster means of outcome measures. Panels A-C plot mean 30-day mortality, mean LOS, and mean total hospitalization costs by time to maximum stabilization for each sequence cluster (*N* = 1,326). For example, 30-day mortality is lowest for patients in the Tachycardic and Febrile cluster, intermediate for those in the Tachycardic/Hypoxic, Stabilized Fast, and Stabilized Slow clusters; and highest in the Hypoxic, Mental/Oral/Hypoxic, and Inpatient Mortality clusters.

Corresponding Author: *Gavin W. Hougham, PhD; Department of Medicine/Section of Hospital Medicine, University of Chicago, 5841 S. Maryland Avenue (MC 5000), Chicago, IL, USA (e-mail: ghougham@bsd.uchicago.edu).*