Online Data Supplement for:

Phenotypic clusters within sepsis-associated multiple organ dysfunction syndrome

Authors:
Daniel B. Knox, MD[1,2](danknoxmd@gmail.com), Michael J. Lanspa MD[1,2](michael.lanspa@imail.org), Kathryn G. Kuttler, PhD[3](Kathryn.Kuttler@imail.org), Simon C. Brewer, PhD[4](simon.brewer@geog.utah.edu), Samuel M. Brown, MD MS[1,2](Samuel.Brown@imail.org)

[1]Pulmonary and Critical Care, Intermountain Medical Center
[2]Pulmonary and Critical Care, University of Utah School of Medicine
[3]Homer Warner Center for Informatics Research, Intermountain Healthcare
[4]Geography Department, University of Utah

This online data supplement contains: Appendix, eTable 1, eTable 2, eTable 3, and eTable 4.

Appendix: A technical description of self-organizing maps

The self-organizing map (SOM), introduced by Kohonen,(1, 2) is a widely used neural network method for the classification and visualization of high-dimensional data sets.(3, 4) SOMs are similar in some respects to multi-dimensional scaling and ordination,(5) in that they allow reduction of multi-dimensional data onto a simpler set of dimensions. SOMs are also parallel in some respects to cluster analysis in that they classify multivariate data by grouping together similar observations. However, the groups that are identified in a SOM are ordered/mapped into a two-dimensional parameter space, while preserving the topological characteristics of the input data.(6) This results in an organization of the data, where similar observations are placed in the same group, and the position of each group in the final map depends on the similarity among groups.

SOMs are a form of neural network and fall in the category of machine learning techniques. As such they require training against a set of *n* observations. Each observation represents an *input vector*, here a patient, of length *m*, where *m* is the number of variables recorded for each observation, in the case of the present study the SOFA subscores. Variables are standardized to avoid bias where parameters have different magnitudes of variation. The map or grid of the SOM is made up of a two-dimensional lattice of nodes or neurons, with each node linked to either four or six neighbors.(2) Higher dimensions are sometimes used, but these present additional difficulties for visualization and post-processing. Each node has an associated vector of *m* weights, with each weight corresponding to the number of variables in the input dataset.

The basic SOM is an unsupervised classification method, i.e. no target data is used to constrain the final groups. A supervised version does exist,(7) but requires a target classification. As we employed SOM as an exploratory technique to identify novel

phenotypes, we employed the unsupervised version of SOM.

The SOM method proceeds as follows. The input data is used to train the grid of nodes, using a competitive learning technique.(2) The grid is initialized by setting the weights of each node to a random value. Then, at each training iteration (usually 100 to 1000 iterations; we employed 500 in the current study):

1. An input vector is selected at random and compared to all the nodes on the map.
2. The most similar node or best matching unit (BMU) is identified using dissimilarity metric.
3. The *m* weights of the BMU are then adjusted by a small amount (α) towards the values of the input vector.
4. In addition, the weights of the nodes within a certain radius of the BMU are also adjusted toward the input vector, although this adjustment diminishes with distance.
5. Once all adjustments are complete, another input vector is chosen, and steps 2 through 4 are repeated

As the iterations continue and the node weights are continuously updated, the observations are progressively shuffled between nodes. The neighborhood effect forces closer nodes to be more similar, and distant nodes to be more dissimilar, with the net effect that "the low-dimensional lattice of node vectors begins to replicate major topological structures existing in the n-dimensional space."(3) The final map represents a two-dimensional parameter space defined by co-variation in the parameters, and the final set of weights for each node represents a *prototype* input vector for that node. For our data, the node weights would therefore represent an "idealized" patient associated with that node. These weights may be visualized and used to display how that parameter varies across the map, and by comparison between multiple parameters, it is possible to visualize the areas of the parameter space where different parameters show co-variation (positive or negative).

While there is no single objective function to optimize, the fit of the SOM to the data can be estimated by calculating the average distance between each input vector and its BMU. As the number of iterations increases, this value decreases and ultimately stabilizes, at which point the organization is optimal, and little reassignment of input vectors occurs with further training iterations. (In our data, stability was observed after approximately 250 iterations, so we used 500 iterations.)

There are several choices in setting up the SOM, including the value of α and the neighborhood radius, the dissimilarity metric and the size of the SOM grid. Generally, both α and the radius are set to relatively high values at the outset of the training process to allow for quick but crude organization of the data.(2) As training continues, these values can be reduced. This adaptive learning allows finer or more local tuning of the map, so that a match between an input vector and the BMU may only update the immediate neighbors or even just the BMU itself.

By default, multivariate Euclidean distances are used as the dissimilarity metric, where dissimilarity between observation i and node j across the set of variables *m* is given by:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

This metric is used as standard in nearly all applications of SOMs and was used in the current study.

The choice of the number of nodes is important. Smaller SOMs result in a classification of data very similar to k-means clustering. With higher numbers of nodes, the SOM is a topological representation of the original data, and termed a high-resolution SOM.(8) Higher resolution SOMs can identify smaller-scale features in the data, and may be used as base maps to display other aspects of the data, including time-dependent trajectories

for longitudinal observations(9, 10) For the present study, we employed a high resolution 24X24 map.

The final map is generally post-processed to identify larger features, particularly when the number of nodes is high. This is usually done by hierarchical or *k*-means clustering of the node weights, which helps identify continuous, internally homogenous parts of the map.(8) It is worth noting that the resulting *k* clusters can be quite different from those obtained by clustering of the original dataset, as clustering is performed on the organized results. By comparing the mapped weights within and between each of the clusters, it is possible to identify the characteristics of each group, as well as the in-group variability, which is usually ignored in classical cluster analysis. In our study, we defined these clusters as phenotypes of the multiple organ dysfunction syndrome, and the individual nodes represent the internal variability within a given phenotype (Fig. 3).

The classification generated by the SOM may be further used with parameters that were not used in generating the organization, by mapping them onto the SOM nodes. This mapping is carried out for each node by averaging the value of the variable for all observations belonging to that node. This provides a simple method to investigate the relationship between different nodes, or node groups, and the value of the external variables.

eTable 1 Simple multivariate logistic regression of 30-day mortality, after backwards elimination.

	Univariate OR (95% confidence interval)	Univariate P*	Multivariate OR (95% confidence interval)	Multivariate P*
APACHE II (per point)	1.09 (1.07-1.10)	<0.0001	1.07 (1.06, 1.10)	<0.0001
SOFA (per point)	1.19 (1.16-1.23)	<0.0001	1.05 (0.99, 1.11)	0.09
Age (per year)	1.03 (1.03-1.04)	<0.0001	1.03 (1.03, 1.04)	<0.0001
Lactate (per mg/dL)	1.15 (1.11-1.19)	<0.0001	1.08 (1.05, 1.12)	<0.0001
Urinary source	0.58 (0.44-0.76)	<0.0001	0.58 (0.42, 0.79)	0.0005
Cluster membership	NA	<0.0001	NA	0.0006
Cluster 1	0.55 (0.40-0.73)	<0.0001	NA (referent)	NA
Cluster 2	0.56 (0.45-0.70)	<0.0001	1.67 (1.14, 2.46)	NA
Cluster 3	2.58 (2.06-3.23)	<0.0001	1.20 (0.80, 1.82)	NA
Cluster 4	1.39 (1.03-1.86)	0.03	2.25 (1.46, 3.48)	NA

*P values are from the Likelihood Ratio test. Overall odds ratios are not defined for multinomial predictors, so odds ratios are not reported for overall cluster membership, and multivariate odds ratios for clusters are reported in comparison to cluster 1. P values are not defined for the multivariate comparisons for individual clusters.

APACHE II: Acute Physiology and Chronic Health Evaluation score, 2nd version. SOFA: Sequential Organ Failure Assessment; OR: odds ratio.

eTable 2 Simple multivariate linear regression of ICU-free days

	β	95% confidence interval	P
APACHE II (per point)	-0.30	(-0.37, -0.24)	<0.001
SOFA	-0.43	(-0.62, -0.25)	<0.001
Age (per year)	-0.04	(-0.06, -0.02)	<0.001
Lactate (per mg/dL)	-0.31	(-0.43, -0.18)	<0.001
Urinary source	2.18	(1.32, 3.04)	<0.001
Cluster 2 (versus Cluster 1)	-2.16	(-3.23, -1.10)	<0.001
Cluster 3 (versus Cluster 1)	-2.50	(-3.75, -1.25)	<0.001
Cluster 4 (versus Cluster 1)	-2.87	(-4.15, -1.58)	<0.001

eTable 3 Stratified logistic regression of 30-day mortality after backwards elimination.

	OR	95% confidence interval	P
Phenotype 1 (AUC 0.81)			
Age	1.06	1.03-1.08	<0.001
APACHE II	1.10	1.05-1.16	<0.001
Phenotype 2 (AUC 0.77)			
Age	1.04	1.02-1.05	<0.001
APACHE II	1.07	1.03-1.10	<0.001
Lactate	1.12	1.04-1.21	0.003
Pneumonia	0.62	0.38-0.99	0.05
Urinary infection	0.27	0.15-0.51	<0.001
Soft tissue infection	0.23	0.09-0.61	0.003
Phenotype 3 (AUC 0.74)			
Age	1.04	1.02-1.05	<0.001
APACHE II	1.07	1.04-1.10	<0.001
Lactate	1.11	1.05-1.17	<0.001
Phenotype 4 (AUC 0.73)			
APACHE II	1.09	1.04-1.14	0.001
Elixhauser comorbidity index	1.26	1.08-1.46	0.003

APACHE II: Acute Physiology and Chronic Health Evaluation score. SOFA: Sequential Organ Failure Assessment.

eTable 4 Stratified multivariate linear regression of ICU-free days

	β	95% confidence interval	P
Overall population (adjusted R^2 0.26)			
Phenotype 1 (adjusted R^2 0.14)			
Age	-0.05	(-0.09,0.01)	0.01
APACHE II	-0.26	(-0.39,-0.14)	<0.001
Phenotype 2 (adjusted R^2 0.14)			
APACHE II	-0.26	(-0.41, -0.21)	0.001
Lactate	-0.40	(-0.63, -0.16)	0.001
Urinary infection	3.5	(2.01, 5.00)	<0.001
Soft tissue infection	2.8	(0.94, 4.63)	0.003
Phenotype 3 (adjusted R^2 0.15),			
Age	-0.9	(-0.14, -0.03)	0.002
APACHE II	0.29	(-0.43, -0.15)	<0.001
Lactate	-0.39	(-0.62, -0.16)	0.001
Urinary infection	3.3	(0.18, 6.34)	0.04
Phenotype 4 (adjusted R^2 0.25)			
SOFA score	-0.98	(-1.45, -0.50)	<0.001
APACHE II	-0.29	(-0.48, -0.09)	0.005
Elixhauser Comorbidity Index	-0.57	(-1.10, -0.05)	0.03

References for Online Data Supplement

1. Kohonen T. Self-organized formation of topologically correct feature maps. *Biological cybernetics* 1982; 43: 59-69.
2. Kohonen T. Self-Organizing Maps, ser. *Information Sciences Berlin: Springer* 2001; 30.
3. Skupin A, Esperbé A. Towards High-Resolution Self-Organizing Maps of Geographic Features. *Geographic Visualization: Concepts, Tools and Applications* 2008: 159-181.
4. Boyack KW, Newman D, Duhon RJ, Klavans R, Patek M, Biberstine JR, Schijvenaars B, Skupin A, Ma N, Börner K. Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS One* 2011; 6: e18029.
5. Giraudel J, Lek S. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. *Ecological Modelling* 2001; 146: 329-339.
6. Kiviluoto K. Topology preservation in self-organizing maps. IEEE International Conference on Neural Networks; 1996. p. 294-299.
7. Melssen W, Wehrens R, Buydens L. Supervised Kohonen networks for classification problems. *Chemometrics and Intelligent Laboratory Systems* 2006; 83: 99-113.
8. Skupin A. The world of geography: Visualizing a knowledge domain with cartographic means. *Proceedings of the National Academy of Sciences* 2004; 101: 5274-5278.
9. Skupin A, Hagelman R. Visualizing demographic trajectories with self-organizing maps. *GeoInformatica* 2005; 9: 159-179.
10. Delmelle E, Thill J-C, Furuseth O, Ludden T. Trajectories of multidimensional neighbourhood quality of life change. *Urban Studies* 2013; 50: 923-941.