SUPPLEMENTARY DIGITAL CONTENT BACKGROUND

Principal component analysis (PCA), first proposed by Karl Pearson in 1901, is a classic method for extracting and distilling the inter-relationships among a large number of correlated variables.[1] PCA is often regarded as the first true 'multivariate' statistical method. PCA's essential purpose is to understand the best fit plane through a system of points in multidimensional space, thereby distilling complex interactions among variables to essential unifying relationships represented by this lower dimensional plane (technically known as a 'hyperplane'). In this sense, PCA can be thought of as a multivariable form of the Pearson correlation, where the best-fit line has been replaced by a best fit, hyperplane. PCA is closely related to the technique of exploratory factor analysis (EFA) developed by Charles Spearman in 1904 for developing psychometric scales, and the two approaches are often used interchangeably.[2] Both PCA and EFA have been used extensively in human test and measurement theory, and paved the way for most modern neuropsychological testing batteries and standardized educational tests.[3] PCA and related data-driven pattern-detection methods have also been used extensively in other fields that require integration of numerous variables such as meteorology, economics, physics, , physiology, and molecular biology.[4-7]

However, PCA has not been exploited to its full potential in medicine for the purpose of patient classification and disease taxonomy. Doing so requires a close collaboration among applied statisticians, biologists and physicians. The present paper represents the fruit of one such collaboration. The goal of the paper is not to provide a definitive review of PCA, as numerous other resources are available: as of 1 January 2013, over 21,000 Pubmed-indexed PCA-related

papers exist in biomedicine alone (for an in-depth primer, see Vyas *et al*.[8]). Rather, the goal of the present paper is to understand and categorize coagulopathy as a multivariate pattern derived directly from clotting factor level data. The rationale for taking this approach is that it provides the first step toward a data-defined (i.e. human decision-free) syndromic definition of what it means to be 'coagulopathic'. For a theoretical overview of a data-intensive syndromic approach in traumatic disorders, we refer interested readers to prior work from members of our team..[9] The findings reported in the main body of the present text set the stage for ongoing work to develop robust diagnostic tests to improve biological mechanism-level understanding as well as outcome-level prediction in management of critically injured patients. In this supplement we provide a brief primer on the historical development of PCA, additional methodical considerations, and provide further details of the results that could not be included in the main text due to space limitations.

SUPPLEMENTARY DIGITAL CONTENT METHODS

Mathematics of the approach

Linear PCA is achieved through the linear algebraic approach of singular value decomposition (SVD) applied to the cross-correlation matrix of all variables;[10] (see Supplementary Figure 1, **Supplemental Digital Content 2, http://links.lww.com/TA/A241).** When SVD is applied to a symmetrical square matrix such as the correlation matrix it is known as eigenvalue decomposition; PCA is therefore also referred to as eigenvector decomposition. In the context of the current paper, the variables analyzed consisted of numerous clotting factors measurements from trauma patients. The inter-relationships among the clotting factors were first distilled to

their cross correlations in a set of the 163 patients in the dataset. PCA then shuffled the correlation matrix such that variables that are highly correlated (both positively and negatively) were clustered together. The function determining this clustering was constrained to maximize the variance explained within the original correlation matrix, yielding the first solution: this is labeled as principal component 1 (PC1). This procedure was then repeated a second time with the additional constraint that this second solution be uncorrelated with (i.e., orthogonal to) the first, yielding PC2. This procedure was then repeated iteratively and each successive solution labeled sequentially as PC3, PC4, *et cetera* until 100% of the variance in the dataset was accounted for. In non-linear variants of PCA, the variance-maximization is coupled to an optimal scaling transformation through an alternating least squares method that simultaneously linearizes the variables and maximizes the variance explained in the PCs.[11]

PC extraction rules

In essence, PCs can be conceptualized as synthetic multivariables that capture the majority of the meaningful variance in the original variables—in our case, variance in coagulation factors and their inter-relationships. Note from Supplementary Figure 2 (**Supplementary Digital Content 3, http://links.lww.com/TA/A242)** that each successive PC accounts for an ever-diminishing additional percentage of the variance from the original correlation matrix. At a certain point, the value-added by additional PCs diminishes below a level that provides meaningful understanding of the original variables and their inter-relationships. Based on this point of asymptotic diminishing returns, a number of prominent 20th-century statisticians proposed rules for PC extraction, retention, and interpretation. The first rule, proposed by Kaiser,[12] is based on the

concept that eigenvalues > 1.0 indicate that a multivariate solution provides more information (variance explained) than considering each individual variable as a solitary, unrelated unit. This "Kaiser rule" is embedded as a default in many popular statistical software packages. In our dataset, the Kaiser rule suggests retention of the first 3 PCs (see Figure 2A, **Supplemental Digital Content Figure 3, http://links.lww.com/TA/A242**). A second rule, proposed by Cattell,[13] involves plotting the PCs in rank order by eigenvalue and retaining those above the elbow of this so-called 'scree plot'; examination of the Scree plot for this dataset suggests retention of the first 2-3 PCs (see Supplementary Figure 2B, **Supplemental Digital Content Figure 3, http://links.lww.com/TA/A242**). Monte Carlo studies have identified the Kaiser rule as overly liberal and scree plot-based selection as overly conservative as approaches for determining the number of PCs to retain for interpretation and use in subsequent analyses.[14] In this case, a plot of the total variance explained by the PCs indicates that the top 3 PCs together account for 67% of the variance in the dataset (see Supplementary Figure 2C, **Supplemental Digital Content Figure 3, http://links.lww.com/TA/A242**). Based on these results we opted to retain PCs 1-3 for further interpretation, naming, and inclusion in predictive modeling (see main body Results section). More recent literature has introduced additional proposed selection rules not utilized here, such as component saturation, iterative cross-validation, and shrinkage methods.[10,15]

PC interpretation and naming

Once PCs are mathematically determined, researchers can explore their conceptual meaning by examining *PC loadings*—which are equivalent to Pearson correlations between each original

variable and the PC multivariable (see main text, **Table 1**). In addition, the *PC scores* can be calculated for each patient by multiplying the raw variable values by the PC loadings and then summing these weighted variables. In the context of the current paper, PC scores for each patient can be thought of their position along composite metrics reflecting the different types of coagulopathy (global clotting factor depletion vs. fibrinolytic vs. consumptive coagulopathy). These scores can be leveraged in prediction models (e.g, **Tables 2-5** and **Figure 1** in main body of the paper).

SUPPLEMENTARY DIGITAL CONTENT REFERENCES

1. Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophy Magazine*. 1901;2(6):559-72.
2. Spearman C. "General intelligence " objectively determined and measured. *Am J Psychol*. 1904;15:201-92.
3. Harman HH. Modern factor analysis. 3d ed. Chicago; London: University of Chicago Press; 1976. xx, 487 p.p.
4. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*. 2000;97(18):10101-6.
5. Berret B, Bonnetblanc F, Papaxanthis C, Pozzo T. Modular control of pointing beyond arm's length. *J Neurosci*. 2009;29(1):191-205.
6. Monahan AH. Nonlinear principal component analysis by neural networks: Theory and application to the Lorenz system. *J Climate*. 2000;13(4):821-35.

7. Janes KA, Yaffe MB. Data-driven modelling of signal-transduction networks. *Nat Rev Mol Cell Biol*. 2006;7(11):820-8.

8. Vyas S, Kumaranayake L. Constructing socio-economic status indices: how to use principal components analysis. *Health Policy Plan*. 2006;21(6):459-68.

9. Ferguson AR, Stuck ED, Nielson JL. Syndromics: A Bioinformatics Approach for Neurotrauma Research. *Transl Stroke Res*. 2011;2(4):438-54.

10. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York, NY: Springer; 2009. xxii, 745 p.p.

11. Linting M, Meulman JJ, Groenen PJ, van der Koojj AJ. Nonlinear principal components analysis: introduction and application. *Psychol Methods*. 2007;12(3):336-58.

12. Kaiser HF. The Application of Electronic-Computers to Factor-Analysis. *Educ Psychol Meas.* 1960;20(1):141-51.

13. Cattell RB. Scree Test for Number of Factors. *Multivar Behav Res*. 1966;1(2):245-76.

14. Guadagnoli E, Velicer WF. Relation of Sample-Size to the Stability of Component Patterns. *Psychol Bull*. 1988;103(2):265-75.

15. Stevens JP. Applied multivariate statistics for the social sciences (5th ed.). 5th ed. New York: Routledge; 2009.