Lecture

The Generalized Linear Model (GLM)

Objectives

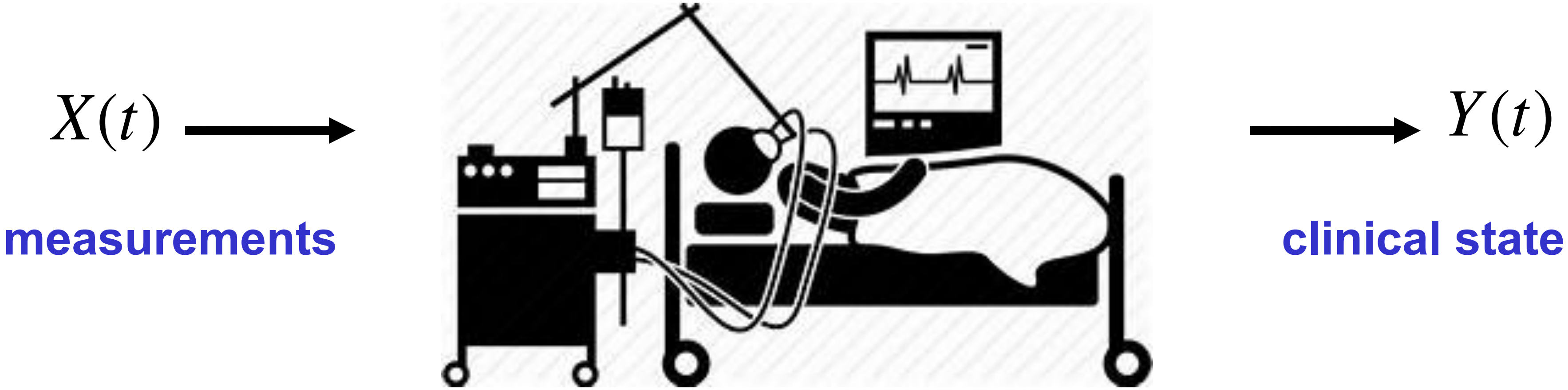
- Understand the theory of the GLM
- Understand its relation to standard linear model
- Understand how data analysis is conducted using the GLM

Outline

- Motivation for GLM use in precision care medicine
- Theory of the GLM
- Example of GLM for septic shock detection
- Summary

GLM in Precision Care Medicine

- **Measurements:** EKG, RR, SPO2 etc., electronic health record
- **Clinical States:** stable, sepsis, organ failure…



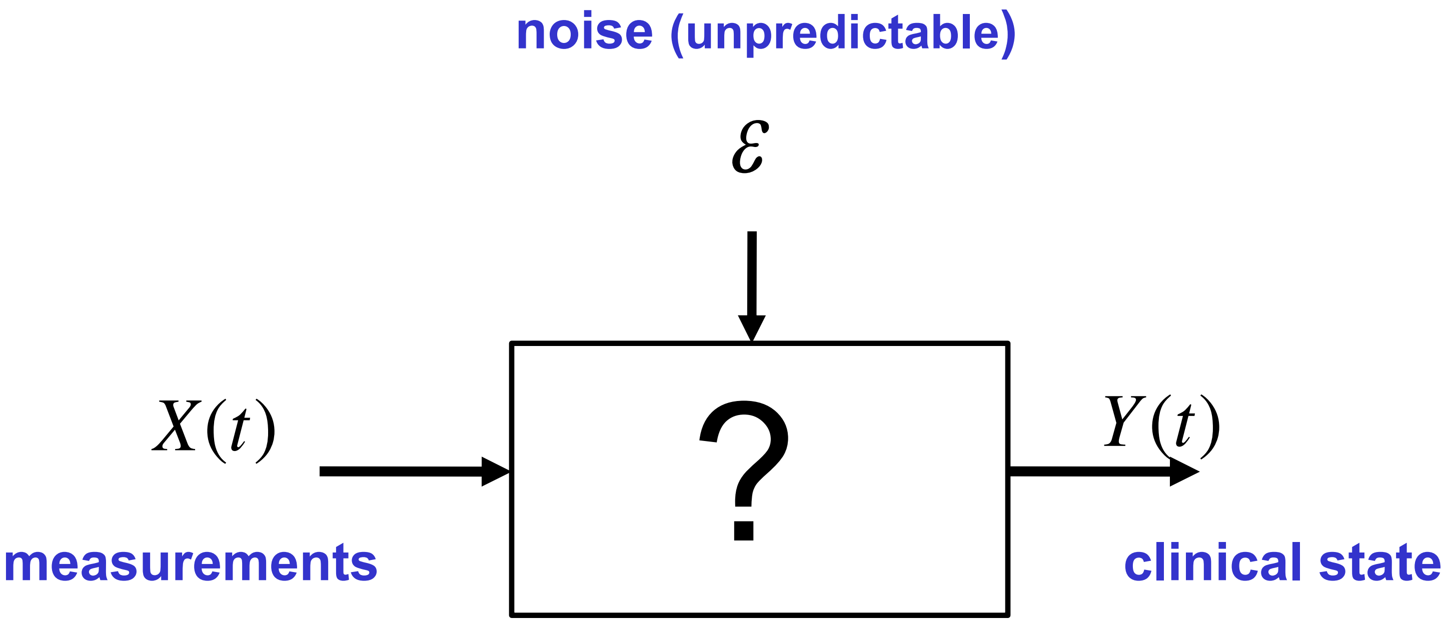
$X(t) \longrightarrow$

measurements

$\longrightarrow Y(t)$

clinical state

System to Study



noise (unpredictable)

ε

$X(t)$

measurements

?

$Y(t)$

clinical state

How can we use numerical data to model how X ‘impacts’ Y?

From Data to Model

- Data: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ $\quad X_i \in R^p \quad \forall i$

  $Y_i \in scalar \quad \forall i$

X_i's are typically non-random variables (covariates)

Y_i's are random variables

- Notation:

Constant parameter vector typically estimated from data

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad X = \begin{bmatrix} - & X_1 & - \\ - & \vdots & - \\ - & Xn & - \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}$$

From Data to Model cont…

Model is joint probability density function f:

$$f(y;X,\theta)$$

From Linear Models to GLMs

- **Linear regression models** of the form:

$$\begin{array}{c} Y = X\theta + \varepsilon \\ \varepsilon \sim N(0, \Sigma) \end{array} \quad \Leftrightarrow \quad Y \sim N(X\theta, \Sigma)$$

  are useful for relating Gaussian ***continuous valued observations*** to a set of covariates.

- Many types of data cannot be described by a Gaussian additive noise model.

- **Generalized linear models** extend a simple class of models to other data types. In particular, binary data (septic shock/non shock ~ Bernoulli)

The Linear Model: A Different Perspective

1. **Y is Gaussian which belongs to the *exponential family* of distributions:**

$$f(y\,|\,\theta) = \exp\left\{T(y)C(\theta) + H(y) + D(\theta)\right\}$$

Data and Parameters are multiplicatively separable!

2. **The likelihood function for the exponential family is:**

$$L(\theta) = f(y\,|\,\theta) = \prod_{k=1}^{K} \exp\{T(y_k)C(\theta) + H(y_k) + D(\theta)\}$$

Canonical Link function

The Linear Model: A Different Perspective

3. The likelihood for the Gaussian and its canonical link for the linear model:

$$L(\theta) = f(y \mid \theta) = \prod_{k=1}^{K} \exp\{T(y_k)C(\theta) + H(y_k) + D(\theta)\}$$

Gaussian Data

$$L(\theta) = \prod_{k=1}^{K} \left[\frac{1}{2\pi}\right]^{\frac{1}{2}} \exp\{-\frac{1}{2}(y_k - \mu_k)^2\}$$

$$= \prod_{k=1}^{K} \exp\{(y_k\mu_k - \frac{1}{2}\{y_k^2 + \mu_k^2 + \log(2\pi)\}$$

The *canonical link function* is then

$$C(\theta) = \mu_k = E(Y_k) \overset{\text{linear model}}{=} [X\theta]_k = \sum_{j=1}^{p} \theta_j x_{k,j}$$

The Generalized Linear Model

1. Y belongs to the *exponential family of distributions*

$$L(\theta) = f(y\,|\,\theta) = \prod_{k=1}^{K} \exp\{T(y_k)C(\theta) + H(y_k) + D(\theta)\}$$

2. The canonical link function is a *linear function of the parameters*

$$C(\theta) = \theta_0 + \sum_{j=1}^{J} \theta_j x_j$$

All the probability models we have studied, Bernoulli, binomial, Poisson, Gaussian, gamma, exponential, inverse Gaussian, beta belong to the exponential family!

The Exponential Family

$$L(\theta) = f(y \mid \theta) = \prod_{k=1}^{K} \exp\{T(y_k)C(\theta) + H(y_k) + D(\theta)\}$$

Poisson Data: number of arrivals in 1 time unit; $y_k \sim Poisson(\lambda_k)$

$$L(\theta) = \prod_{k=1}^{K} \frac{\lambda_k{}^{y_k} \exp\{-\lambda_k\}}{y_k\,!}$$

$$= \prod_{k=1}^{K} \exp\{y_k \;\log(\lambda_k) - \log(y_k\,!) - \lambda_k\}$$

The canonical link function is

$$C(\theta) = \log(\lambda_k) = \sum_{j=1}^{J} \theta_j x_{kj}$$

The Exponential Family

$$L(\theta) = f(y \mid \theta) = \prod_{k=1}^{K} \exp\{T(y_k)C(\theta) + H(y_k) + D(\theta)\}$$

Bernoulli Data: success (1) or failure (0); $y_k \sim Bernoulli(p_k)$

$$L(\theta) = \prod_{k=1}^{K} p_k^{\,y_k}(1 - p_k)^{1-y_k}$$

$$= \prod_{k=1}^{K} \exp\{y_k \log(p_k) + (1 - y_k)\log(1 - p_k)\}$$

$$= \prod_{k=1}^{K} \exp\{y_k \log(\frac{p_k}{1 - p_k}) + \log(1 - p_k)\}$$

The canonical link function

$$C(\theta) = \log(\frac{p_k}{1 - p_k}) = \sum_{j=1}^{J} \theta_j x_{kj}$$

Summary of Generalized Linear Models

Model	Link Equation
Gaussian	$\mu_k = \sum_{j=1}^{J} \theta_j x_{kj}$
Poisson	$\log(\lambda_k) = \sum_{j=1}^{J} \theta_j x_{kj}$
Bernoulli	$\log(\frac{p_k}{1-p_k}) = \sum_{j=1}^{J} \theta_j x_{kj}$

Model Goodness-of-Fit and Analysis

A. Deviance (Analog of the Residual Sum of Squares):

$$-2\log f(\mathrm{y} \mid \theta)$$

where in the Gaussian case $-2\log f(\mathrm{y} \mid \theta) = -2\log(\hat{\sigma}^2)$

B. Akaike's Information Criterion:

$$-2\log f(\mathrm{y} \mid \hat{\theta}_{ML}) + 2p$$

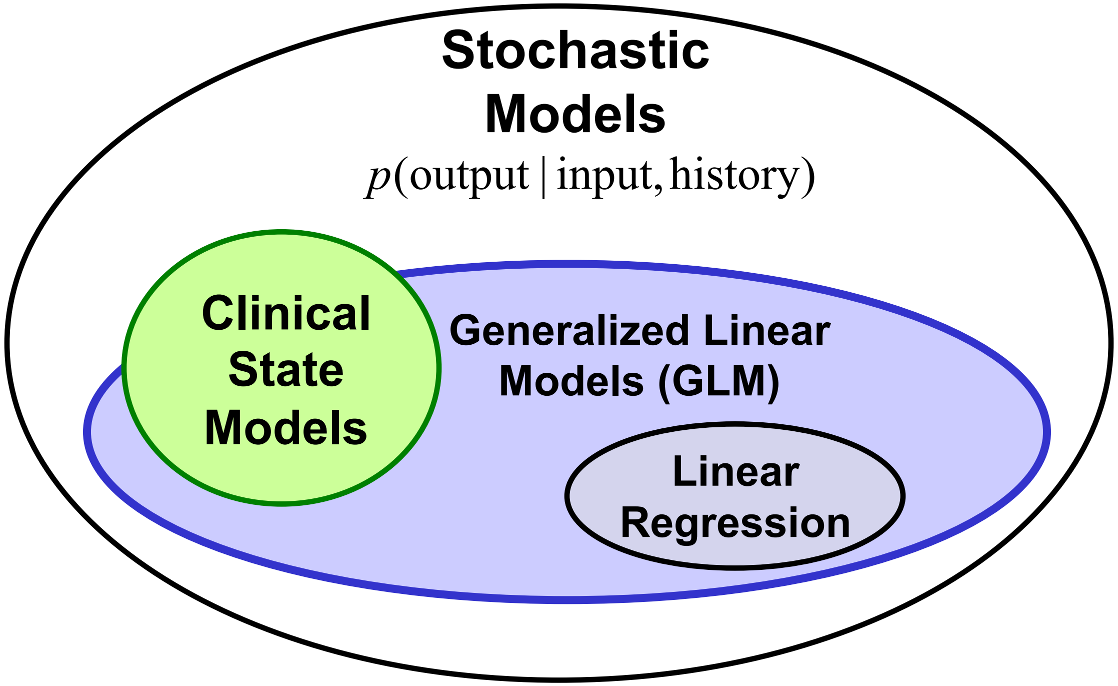
For maximum likelihood estimates it measures the trade-off between maximizing the likelihood (minimizing $-2\log f(\mathrm{y} \mid \hat{\theta}_{ML})$) and the numbers of parameters p in the model.

C. Standard Errors of the Coefficients and t-tests

t-statistic = Coefficient Estimate/SE

Properties of the GLM

- Convex likelihood surface
- Estimators asymptotically have minimum MSE
- All model estimation is efficient: ***iterative reweighted least squares***



GLM Clinical State Models

$$\log\left(\frac{p_k}{1-p_k}\right) = \theta_0 + \sum_{i=1}^{I} \alpha_i f_i(\text{Physiological Covariates(k)})$$

$$+ \sum_{j=1}^{J} \beta_j g_j(\text{Demographics})$$

$$+ \sum_{k=1}^{K} \gamma_k h_k(\text{EHR})$$

- By selecting an appropriate set of basis functions we can capture arbitrary functional relations.
- Analysis of relative contributions of components to clinical state

Truccolo W, Eden UT, Fellows MR, Donoghue JP, Brown EN. (2004) *J. Neurophys* 93:1074-1089

Summary of GLM Theory

- Generalization of the Gaussian Linear Model (McCulloch and Nelder)

- Can be used for any probability models in the exponential family.

- Is a maximum likelihood analysis and all its optimality properties.

- An efficient computational framework using iteratively reweighted least squares.

- GLM is available as a toolbox in all major statistics packages and Matlab.

Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU

Liu, R, Greenstein, JL, Granite, SJ, Fackler, JC, Bembea, MM, Sarma, SV, Winslow, RL

Scientific Reports, 2019

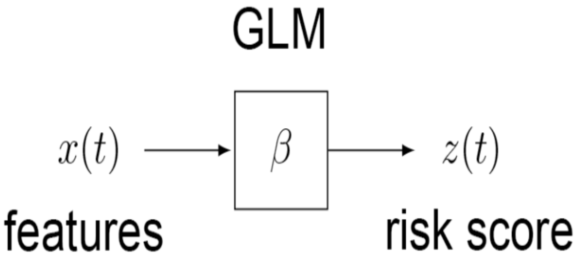
Background

- Sepsis is life-threatening organ dysfunction caused by dysregulated host response to infection[1]
- Septic shock is a subset of sepsis with profound circulatory, cellular, and metabolic abnormalities[1]
- Sepsis and septic shock are the leading causes of in-hospital mortality -the most costly medical conditions[2,3]
- Septic shock patients treated within the first hour have 80% survival rate, but each hour of delayed treatment increases mortality by 8%[4]
- An automated system able to identify patients with sepsis who are likely to develop septic shock has the potential to improve patient outcome by providing a time window for intervention

Methods

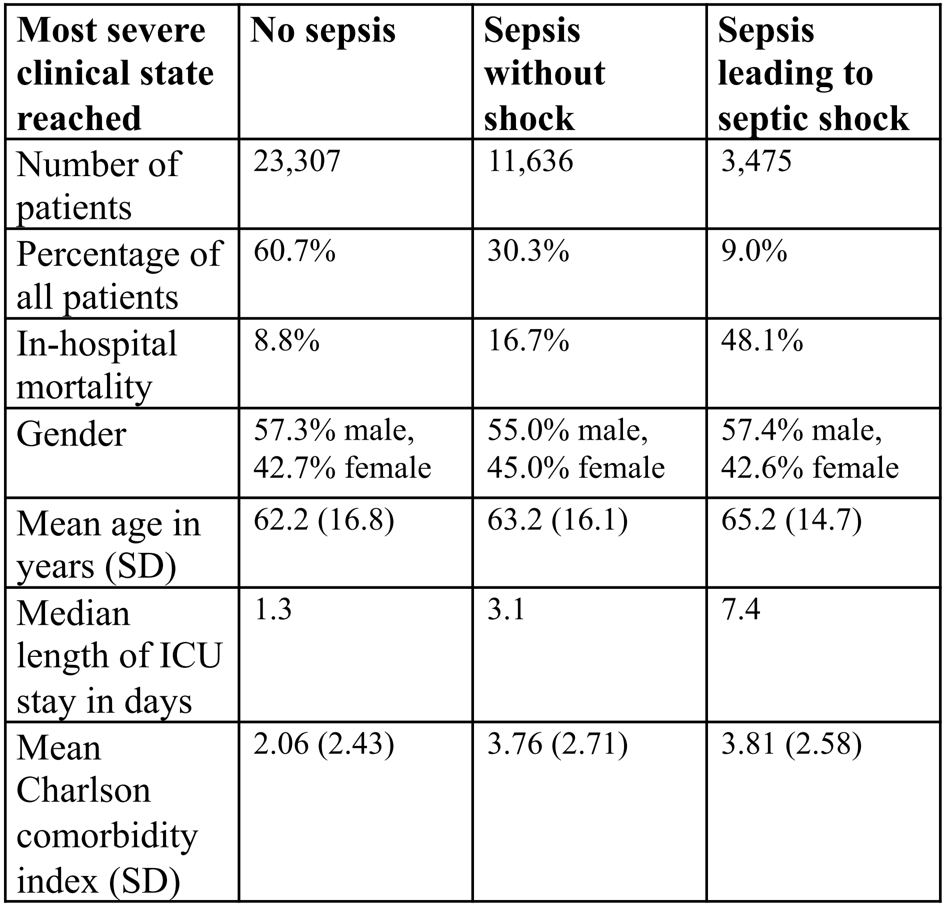
- We hypothesize the existence of a physiologically distinct state of sepsis; patients with sepsis who enter this state are highly likely to develop septic shock

- We apply three different machine learning methods: generalized linear models (GLM), XGBoost, and recurrent neural networks (RNN), to characterize the pre-shock state and model risk of impending transition from sepsis into septic shock



Data Summary

- MIMIC-III contains 38,418 adult patients with sufficient data to evaluate Sepsis-3



Most severe clinical state reached	**No sepsis**	**Sepsis without shock**	**Sepsis leading to septic shock**
Number of patients	23,307	11,636	3,475
Percentage of all patients	60.7%	30.3%	9.0%
In-hospital mortality	8.8%	16.7%	48.1%
Gender	57.3% male, 42.7% female	55.0% male, 45.0% female	57.4% male, 42.6% female
Mean age in years (SD)	62.2 (16.8)	63.2 (16.1)	65.2 (14.7)
Median length of ICU stay in days	1.3	3.1	7.4
Mean Charlson comorbidity index (SD)	2.06 (2.43)	3.76 (2.71)	3.81 (2.58)

Table S4: Statistics and demographic information on MIMIC-III clinical database

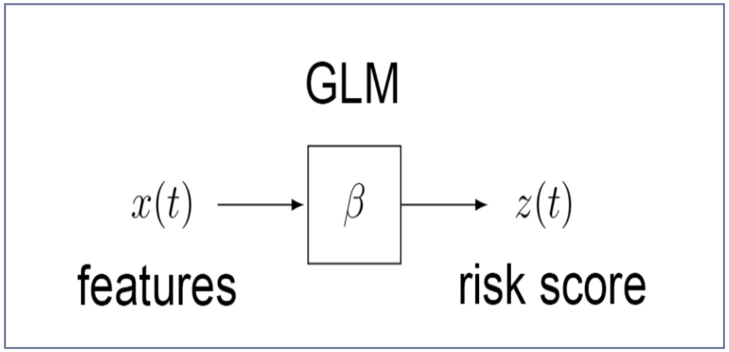
Data Labels

- Third International Consensus Definitions for Sepsis and Septic Shock are applied to generate clinical labels[7]

- Sepsis is defined as suspected infection (as determined by concomitant orders for blood cultures and antibiotics) and a increase in SOFA score of 2 points or more

- Septic shock is defined as sepsis with persisting hypotension requiring vasopressors to maintain mean arterial pressure (MAP) >= 65mmHg, and a serum lactate concentration >2 mM (18 mg/dL) despite adequate fluid resuscitation

GLM

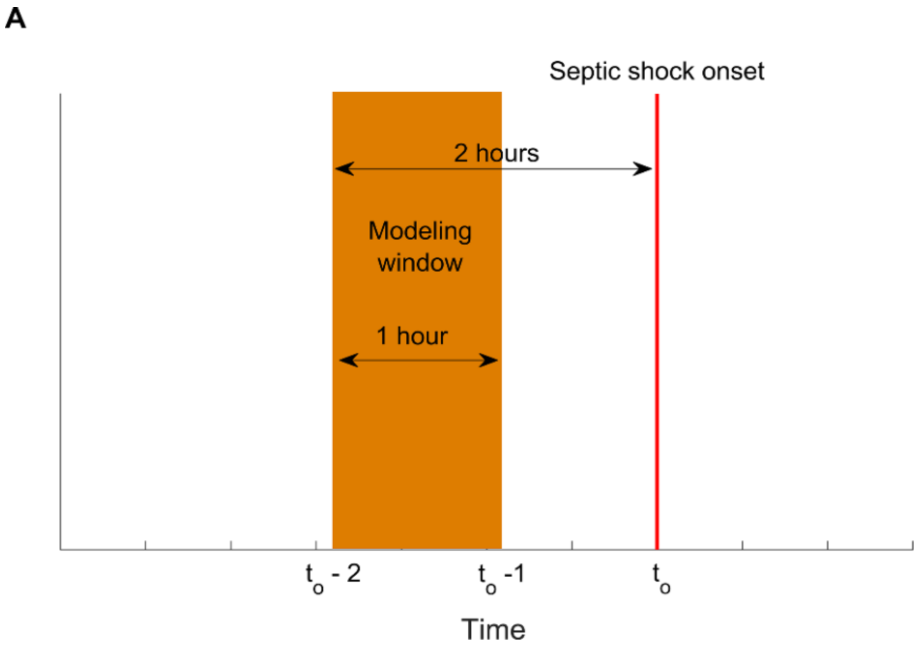


$$z\left(x(t)\right)=\frac{e^{\beta_0+\beta^T x(t)}}{1+e^{\beta_0+\beta^T x(t)}}$$

$$\beta=\underset{\beta\in\mathbb{R}^{p+1}}{\operatorname{argmin}}\left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\frac{e^{\beta_0+\beta^T x_i}}{1+e^{\beta_0+\beta^T x_i}}\right)+\lambda\|\beta\|_1\right)$$

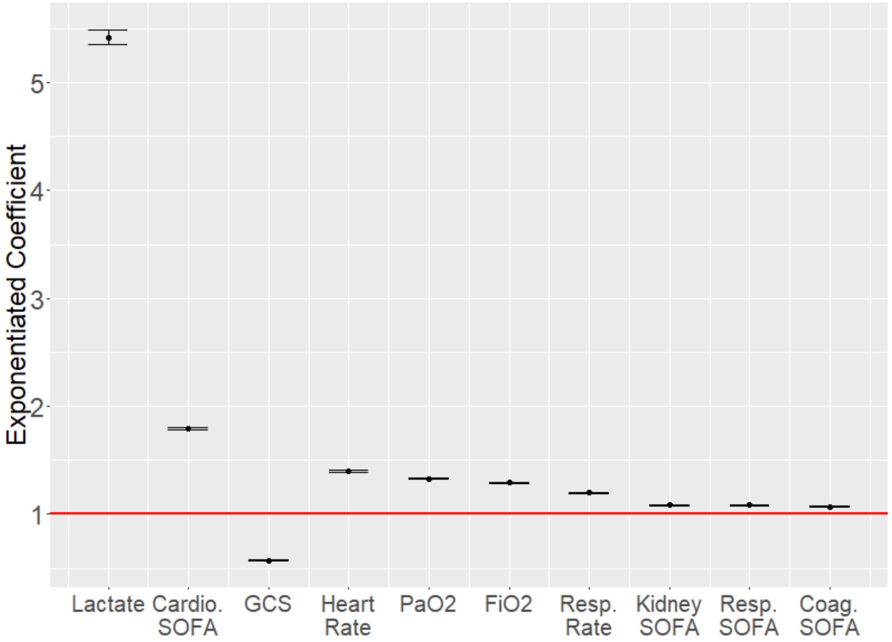
- Data from sepsis patients who never go into shock have all time points labeled 0

- Data from patients who transition have a 1-hr long window labeled 1 (pre-shock)



GLM Results

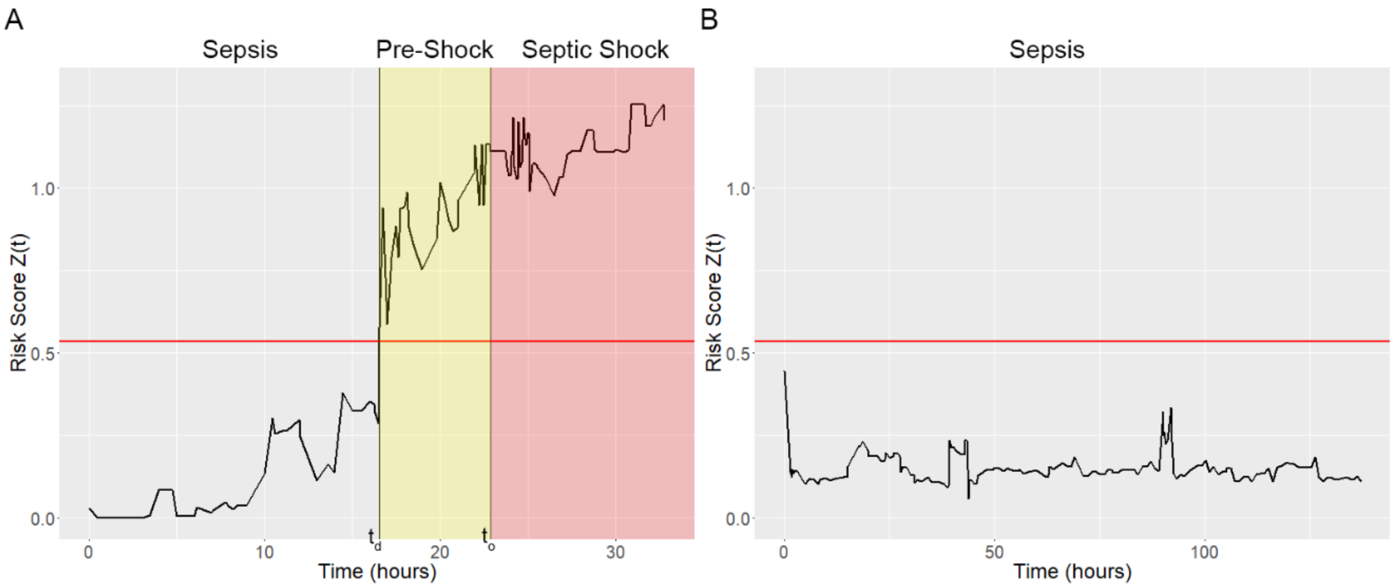
- Each variable is normalized to have a population mean of 0, and a standard dev of 1
- Magnitude of each coefficient is the relative importance of the corresponding feature to the risk of developing septic shock
- Exponentiated coefficients are equivalent to odds-ratios: e.g. a lactate 1 standard deviation above the mean means that a patient is ~5x as likely to develop septic shock



Abbreviations: CVP – Central Venous Pressure; PaO2: Arterial partial pressure of oxygen; Cardio. SOFA – Cardiovascular SOFA Score; SBP – Systolic Blood Pressure; GCS – Glasgow Coma Scale; BUN – Blood Urea Nitrogen; WBC – White Blood Cell Count; Resp. SOFA – Respiratory SOFA Score; Resp. Rate – Respiratory Rate.

Early Warning Time (EWT) Results

- We achieve early prediction of impending transition to septic shock with **0.93 AUC, 88% sensitivity, 84% specificity, 7 hour median EWT**



References

1. Rhodes, A. *et al.* Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016. *Intensive Care Medicine* **43**, 304-377, doi:10.1007/s00134-017-4683-6 (2017).
2. Liu, V. *et al.* Hospital Deaths in Patients With Sepsis From 2 Independent Cohorts. *Jama* **312**, 90, doi:10.1001/jama.2014.5804 (2014).
3. Torio, C. M. & Moore, B. J. in *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs* (2006).
4. Kumar, A. *et al.* Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock*. *Critical Care Medicine* **34**, 1589-1596, doi:10.1097/01.ccm.0000217961.75225.e9 (2006).
5. Joon, L. *et al.* Open-access MIMIC-II database for intensive care research. 8315-8318, doi:10.1109/iembs.2011.6092050 (2011).
6. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Scientific Data* **3**, 160035, doi:10.1038/sdata.2016.35 (2016).
7. Singer, M. *et al.* The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *Jama* **315**, 801, doi:10.1001/jama.2016.0287 (2016).

Acknowledgments

Lecture materials adapted from Woods Hole
"Neuroinformatics" Course taught by
Uri Eden, Sri Sarma, and Emery Brown

Reference

McCullagh P, Nelder JA. Generalized Linear Model, 2nd Edition. Chapman and Hall, 1989.