A Non-parametric Learning Method for Confidently Estimating Patient's Clinical State and Dynamics

William Hoiles† and Mihaela van der Schaar†,*, †UCLA, * University of Oxford

Problem and Objectives

• **In the US, every year:**
▷ **200,000** hospitalized patients experience cardiopulmonary arrests.
▷ **75**% of those patients die.
▷ **50**% of those patients could have been saved by early transfer to ICU (Hershey 1982).

• **Our goal:**
▷ Develop an algorithm for estimating a patient's clinical state using the offline EHR data to allow for **timely ICU admission**.

Our Algorithm

• **Our clinical state algorithm has the following features:**
▷ **[Non-parametric]:** number of clinical states is learned from the EHR data.
▷ **[Bias-immune]:** the bias created by therapeutic intervention censoring is removed.
▷ **[Confidence guarantees].**
• **Three steps for learning:**
1- Physiological model learning.
2- Model refinement.
3- Domain knowledge incorporation.

The Physiological Model

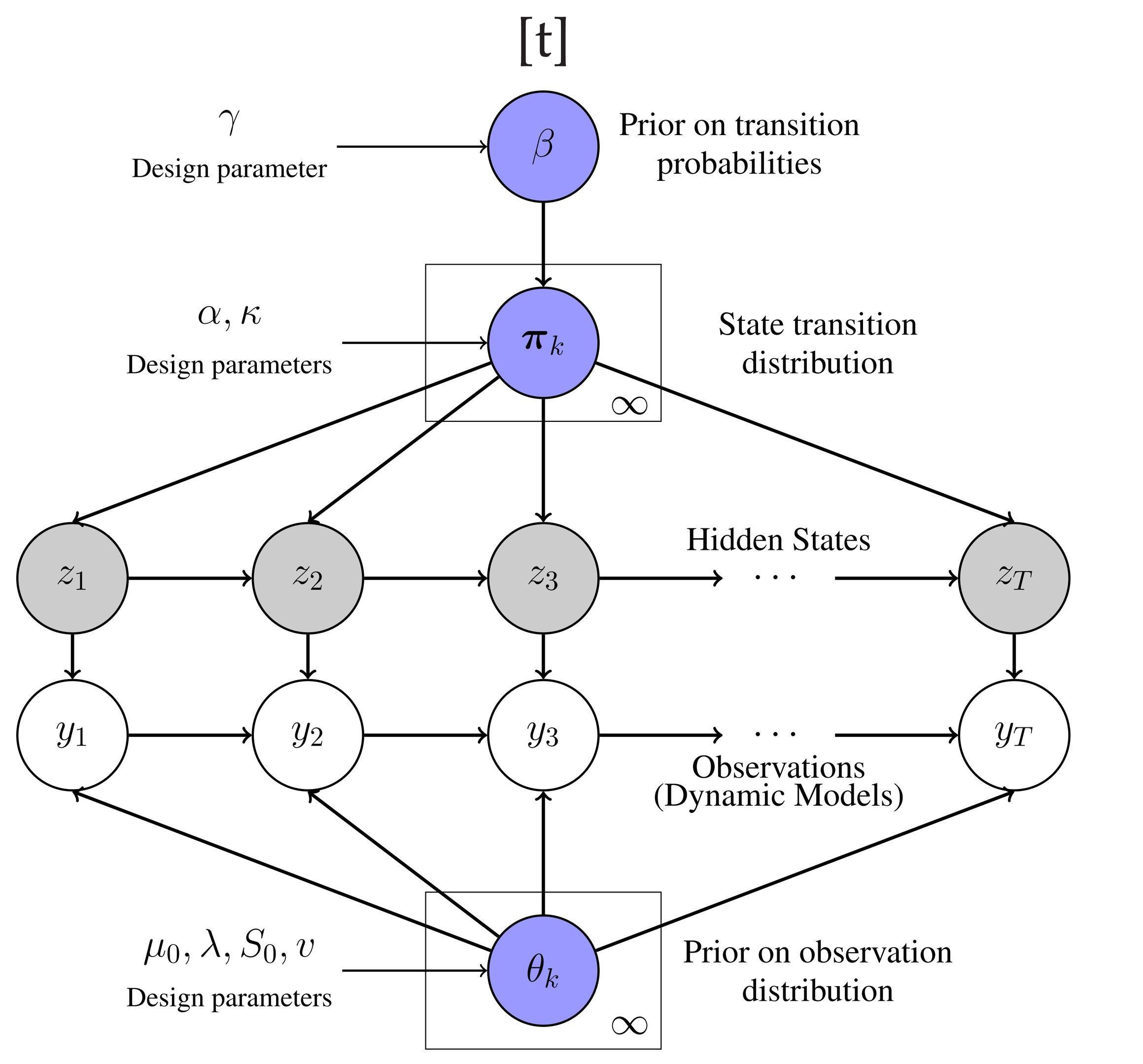
• **Latent variable model:** clinical states $\{z_t\}_{t\in\mathbb{N}_+}$ are hidden and manifest through lab tests and vital signs $\{y_t\}_{t\in\mathbb{N}_+}$.
Model for time series data → HMM
Non-parametric inference → HDP prior.
• **Sticky HDP-HMM with Gaussian emissions:**
▷ Conditional on the clinical states, the physiological variables are Gaussian.

$$y_t|z_t \sim \mathcal{N}(\mu(\mathbf{z_t}), \boldsymbol{\Sigma}(\mathbf{z_t})).$$

▷ **To control the rate of self-transitions**, we use a **sticky HDP** with a stick-breaking construction as a prior for the transition probabilities π_k.

$$v_k \sim \text{Beta}(1,\gamma),\ \beta_k = v_k \prod_{l=1}^{k-1}(1-v_l),$$

$$\pi_k \sim \text{DP}\left(\alpha+\kappa, \frac{\alpha\beta+\kappa\delta_k}{\alpha+\kappa}\right),\ k=1,2,\ldots$$



Model Learning

• **Non-parametric Bayesian Inference:**
▷ Compute the posterior probability on the model parameters $(\pi, \mu, \boldsymbol{\Sigma})$ given an offline EHR dataset $\mathcal{D}$.
▷ Prior on the transition parameters

$$\beta \sim \text{Dir}(\gamma/L, \ldots, \gamma/L)$$

$$\pi_k \sim \text{Dir}(\alpha\beta_1, \ldots, \alpha\beta_k + \kappa, \ldots, \alpha\beta_k)$$

▷ Conjugate priors on the Gaussian emissions: Normal-Inverse-Wishart distribution. (Normal-inverse-gamma distribution is 1-D equivalent.)
▷ Output of this phase: a segmentation of the physiological streams and parameter estimates.

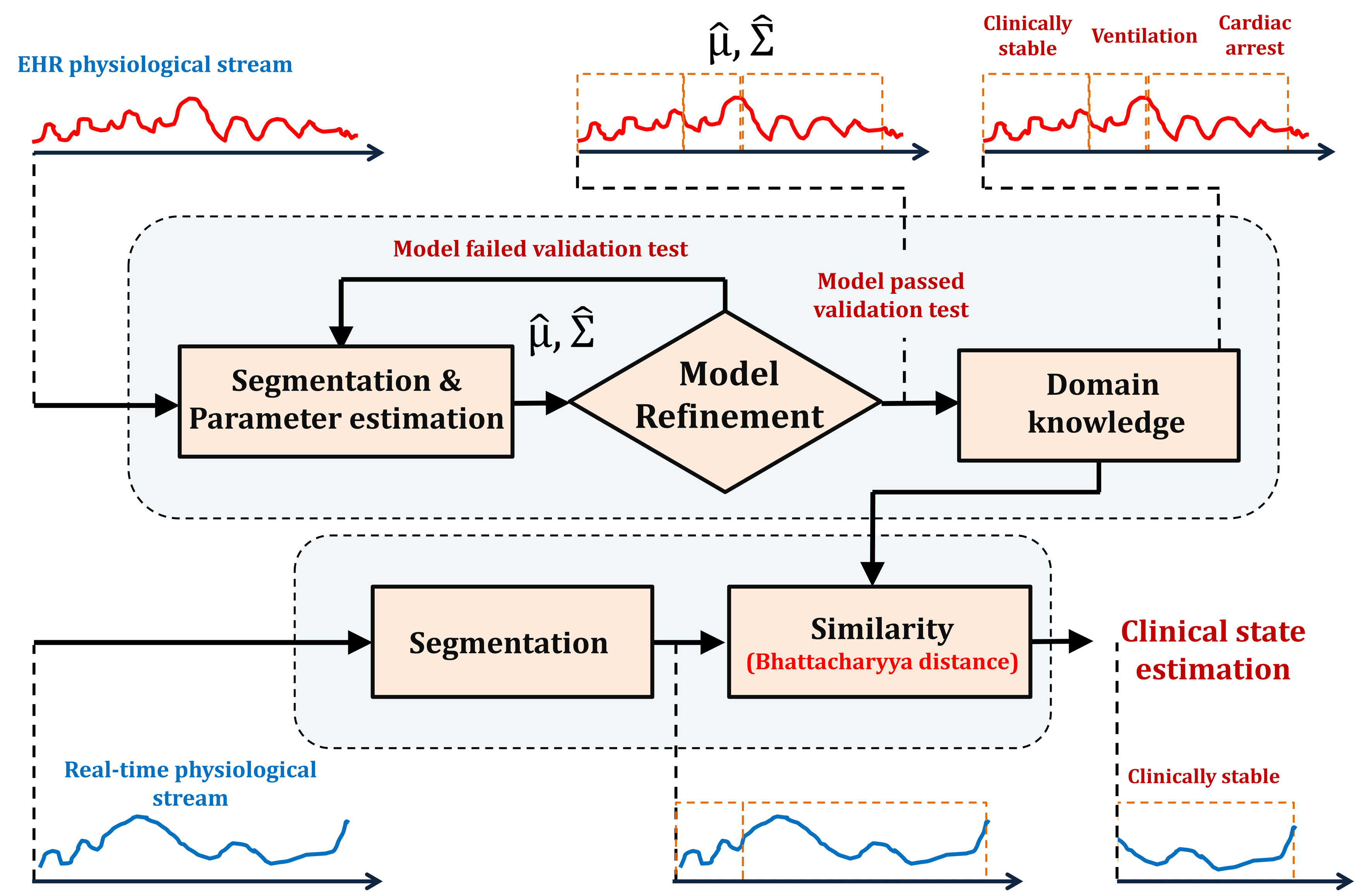
Model Refinement

The learned model is validated in 3 steps:
• **Goodness-of-fit:**
▷ Test the validity of the Gaussian distribution using an **improved Bonferroni method**.
• **Sample complexity:**
▷ Check that the sample size in each segment is sufficient using a multidimensional **empirical Bernstein bound**.
• **Check state distinctness:**
Use a **permutation test** to ensure that the discovered clinical state are distinct (have different μ and $\boldsymbol{\Sigma}$).
If the learned model fails the above tests, hyper-parameters are re-adjusted.

Clinician Domain Knowledge Incorporation

To assess the patients' clinical states, attach clinical interpretation to the learned states:
▷ Clinicians provide labels in the EHR dataset by marking specific segments of the physiological streams with clinical assessments/conditions. We use the **Bhattacharyya distance** to associate the discovered states with the "domain-knowledge-based" states labeled by clinicians.

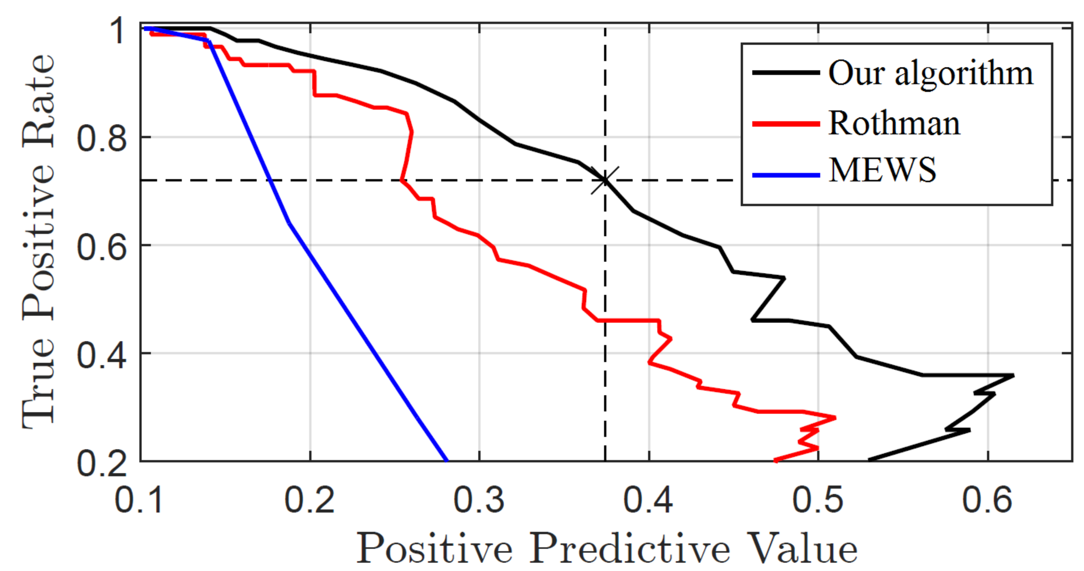
This removes the bias in the learned parameters created by censoring due to interventions.



Real-World Clinical State Estimation

• **We applied our algorithm to a heterogeneous cohort of 6,094 patients:** admissions to Ronald Reagan UCLA medical center (March 2013-February 2016).

• **Our model anticipates clinical deterioration many hours before clinicians. The model outperforms the Rothman index (currently deployed in more than 70 US hospitals).**



Method	Our Algorithm	Rothman	MEWS	Logistic Reg.	RF	SVM
TPR/PPV (%)	71.9/37.4	53.9/34.5	28.1/26.3	55.7/30.7	44.5/31.1	32.2/29.9

• Clinical impact: many cardiac arrests prevented!