Scalable Modeling of Multivariate Longitudinal Data:
Prediction of Chronic Kidney Disease Progression

Joseph Futoma[1], Mark Sendak[2,3], C. Blake Cameron MD[3,4], Katherine Heller[1]

[1]Dept. of Statistical Science, [2] Institute for Health Innovation, [3] School of Medicine, [4] Division of Nephrology
Duke University

Figure 1: 15-year course of patient who experienced CKD rapid progression, other serious health events.

Motivation

- Chronic Kidney Disease (CKD): progressive loss of kidney function; high morbidity.
- Diagnosed using eGFR: extremely noisy estimate of kidney function.
- Systematically underdiagnosed; progression can be slowed/halted if detected early.
  - $< 10\%$ **with moderate CKD,** $< 50\%$ **with advanced CKD aware of illness.**
- Kidney function trajectory more important than eGFR value / CKD stage.
- Goal: flexible model for multivariate longitudinal clinical data.

Electronic Health Records (EHRs)

- Stores information captured about patients during encounters with health system.
- ICD-9 diagnosis codes: structured, hierarchical, primarily for billing, subjective.
- **Laboratory test results**: objective clinical data.
- Many inherent limitations and problems to working with live EHR data.

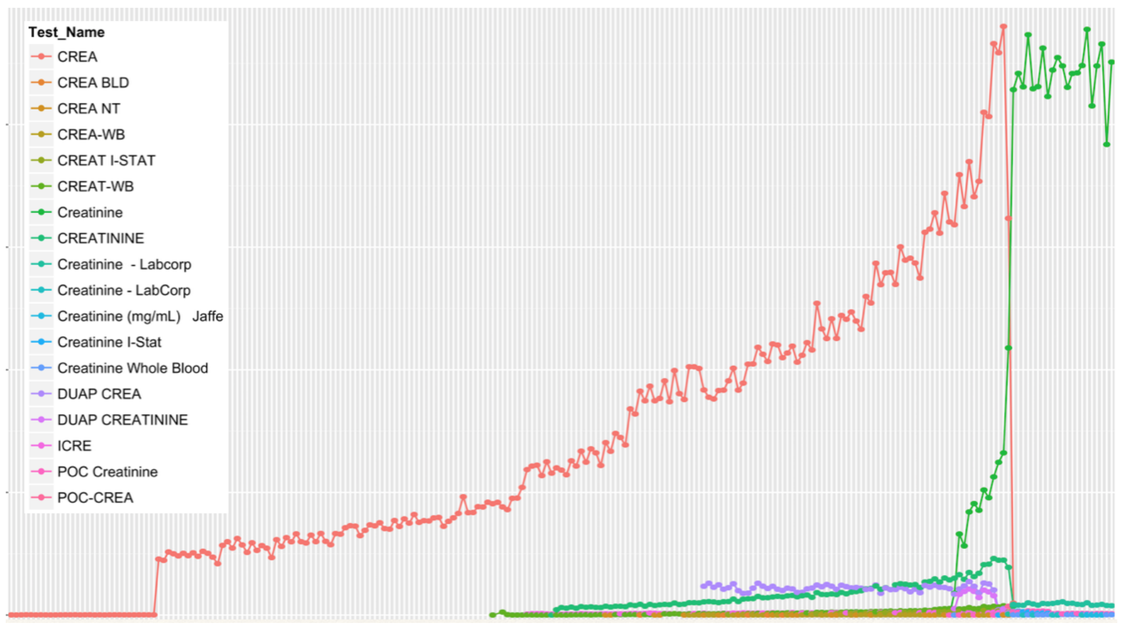

Figure 2: Monthly counts for different lab names for Serum Creatinine, 1996-2015.

Proposed Model

- Hierarchical latent variable model: capture dependencies between lab trajectories.
- $\vec{y_{ip}}$: observed values for patient i, lab/biomarker $p = 1, \ldots, P$. $\vec{y_i}$: all labs.
- Assume independence in conditional likelihood (z_i, b_i, c_i latents for person i):

$$p(\vec{y_i}|z_i, b_i, c_i; x_i) = \prod_{p=1}^{P} p(\vec{y_{ip}}|z_i, b_i, c_i; x_i)$$

- **Model for a Single Trajectory**:
  - *Population component*: Fixed intercept from baseline covariates.
  - *Subpopulation component*: Latent subpopulation $z_i \in \{1, \ldots, G\}$, unique B-spline trajectory.
  - *Individual component*: Random intercept, slope (long-term individual deviations).
  - *Structured noise component*: Transient trends in trajectory, GP with OU kernel.

$$y_{ip}(t) \sim N(\mu_{ip}(t), \sigma_p^2) \quad (1)$$

$$\mu_{ip}(t) \sim \mathcal{GP}(\Lambda^{(p)} x_i + \Phi_z(t)^\top \beta_{z_{ip}}^{(p)} + \Phi_l(t)^\top b_{ip}, K_p) \quad (2)$$

$$K_p(t, t') = a_p^2 \exp\{-l_p^{-1}|t - t'|\} \quad (3)$$

- **Inducing Dependence**:
  - Induce dependence among mean functions $\mu_{ip}(t)$ in two ways:
    1. Long-term deviations (random intercepts, slopes) b_{ip} correlated via joint multivariate normal:

$$\vec{b_i} = (b_{i1}, \ldots, b_{iP})^\top \sim N(0, \Sigma_b). \quad (4)$$

    2. Subtypes/clusters per lab z_{ip} correlated via mixture of multinomials:

$$z_{ip}|c_i \sim \text{Multinomial}(\Psi_{c_i}^{(p)}) \quad (5)$$

$$c_i \sim \text{Multinomial}(\pi_i),\ \pi_{ig} = \frac{e^{w_g^\top x_i}}{\sum_{g'=1}^{G} e^{w_{g'}^\top x_i}} \quad (6)$$

- **Inference**: Fit joint model with stochastic variational inference.
  - Mean field variational distribution, for now.
  - Lower bound has closed form; automatic differentiation for gradients.

Results

- 44,519 patients with at least moderate stage CKD, 5+ eGFR values. Other labs:
  - Serum Albumin, Bicarbonate, Calcium, Phosphorus; Urine Albumin/Creatinine Ratio.
- Processing: mean in monthly bins; $t = 0$: first eGFR < 60.
- Given trained model, predict future labs given labs up to t for each test patient.
- Evaluation: Mean Absolute Error on held-out lab values.
- Baseline: (Schulam & Saria, NIPS 2015) for a single trajectory.

Predictions with data up to...		$t=1$				$t=2$			$t=4$	
Lab	Model	$(1,2]$	$(2,4]$	$(4,8]$	$(8,19]$	$(2,4]$	$(4,8]$	$(8,19]$	$(4,8]$	$(8,19]$
eGFR	Schulam	**8.86***	**10.43***	**12.05**	**13.69**	**8.84*****	**11.08****	**13.23***	**9.39*****	**12.29***
	Proposed	9.12	10.67	12.28	14.21	9.26	11.73	13.99	10.12	13.07
Serum Alb.	Schulam	0.59	0.79	1.09	1.53	0.60	0.88	1.28	0.63	0.96
	Proposed	**0.34*****	**0.39*****	**0.47*****	**0.63*****	**0.35*****	**0.45*****	**0.63*****	**0.40*****	**0.58*****
Serum Bicarb.	Schulam	1.92	2.06	2.13	2.31	1.93	2.06	2.21	1.89	**2.14**
	Proposed	1.87	1.97	2.04	2.31	1.89	1.99	2.31	1.87	2.24
Serum Calc.	Schulam	0.74	1.02	1.62	2.89	0.72	1.26	2.27	0.85	1.53
	Proposed	**0.37*****	**0.44*****	**0.58*****	**0.80*****	**0.39*****	**0.54*****	**0.80*****	**0.46*****	**0.73*****
Serum Phos.	Schulam	1.02	1.35	1.46	1.44	1.17	1.36	1.34	1.13	1.15
	Proposed	**0.57*****	**0.68*****	**0.88*****	**1.23**	**0.65*****	**0.88*****	1.25	**0.82*****	1.23
Urine ACR	Schulam	1.17	1.30	1.44	1.64	1.14	1.30	1.53	1.11	1.41
	Proposed	**0.92*****	**1.02*****	**1.17*****	**1.44**	**0.96*****	**1.13***	1.45	**1.02**	1.42

Table 1: Mean Absolute Errors across all labs from 10 fold cross validation. Bold indicates p-value from one-sided, paired t-test comparing methods was $< .05$. *,**,*** indicate $p < .01$, $< .001$, $< .0001$, respectively.

CKD Rounding Application

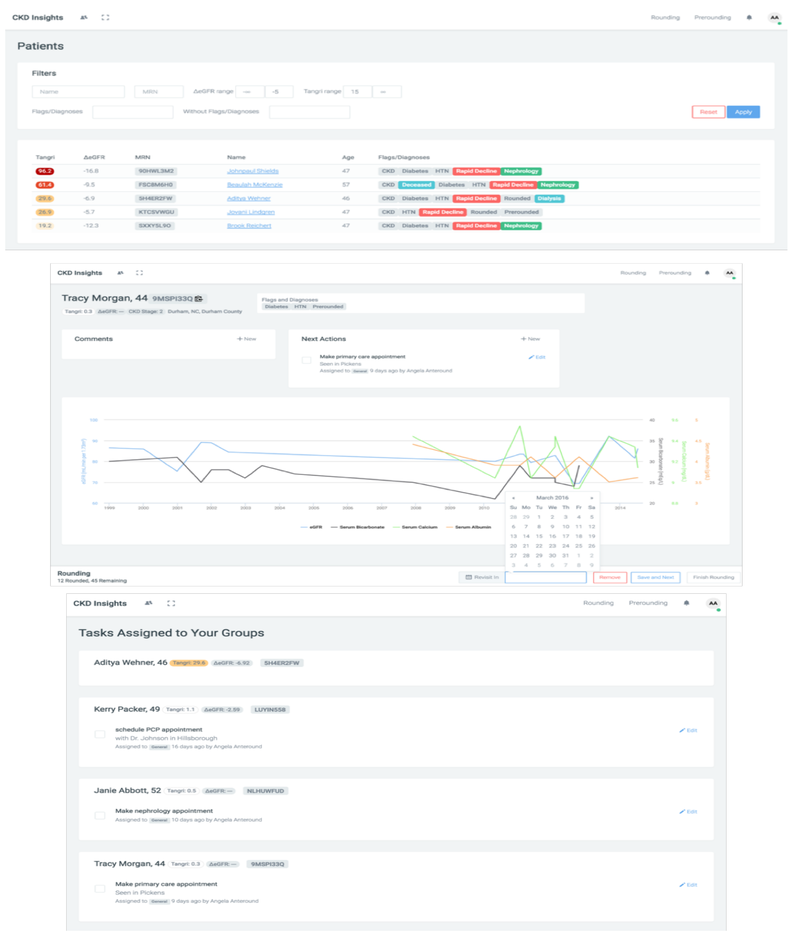

Figure 3: Snapshots from our CKD rounding application (with synthetic data). **Top**: pre-rounding table of patients, with risk scores, flags. **Middle**: patient data, other relevant info to possibly make an intervention. **Bottom**: list of tasks for each group at rounds.

Conclusion

- Novel model for multivariate longitudinal clinical data; scales well.
- Adds value as a clinical decision support tool, supplement clinical "gestalt"
- Future work: more flexible models; joint model with events (e.g. admissions).