Scalable Joint Modeling of Longitudinal and Point Process Data:
Disease Trajectory Prediction and Improving Management of Chronic Kidney Disease

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

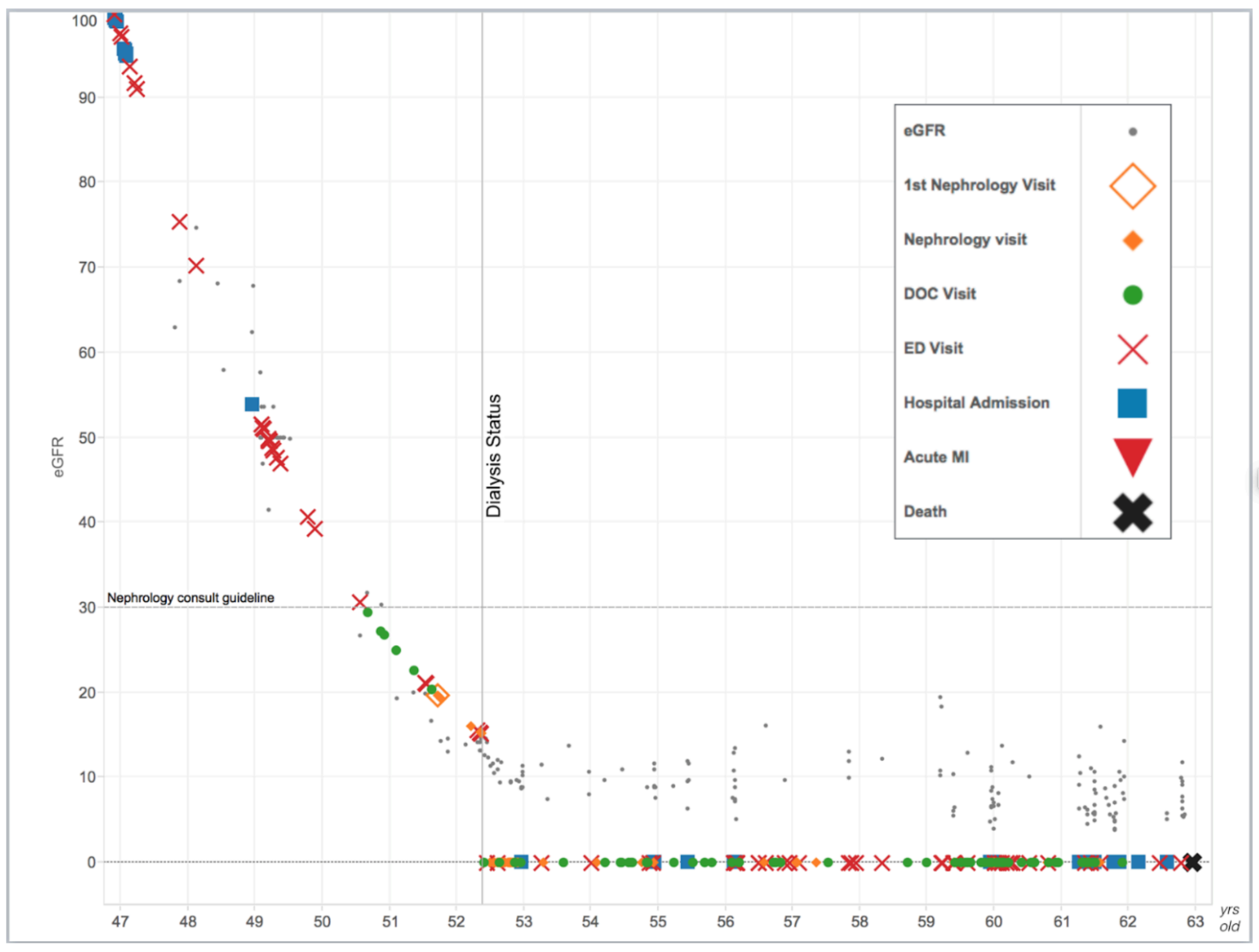

Figure 1: 15-year course of patient who experienced CKD rapid progression, other serious health events.

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.**
- Most CKD patients die from heart disease before kidney failure.
- Goal: jointly model risk of future loss of kidney function, cardiac complications.

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.

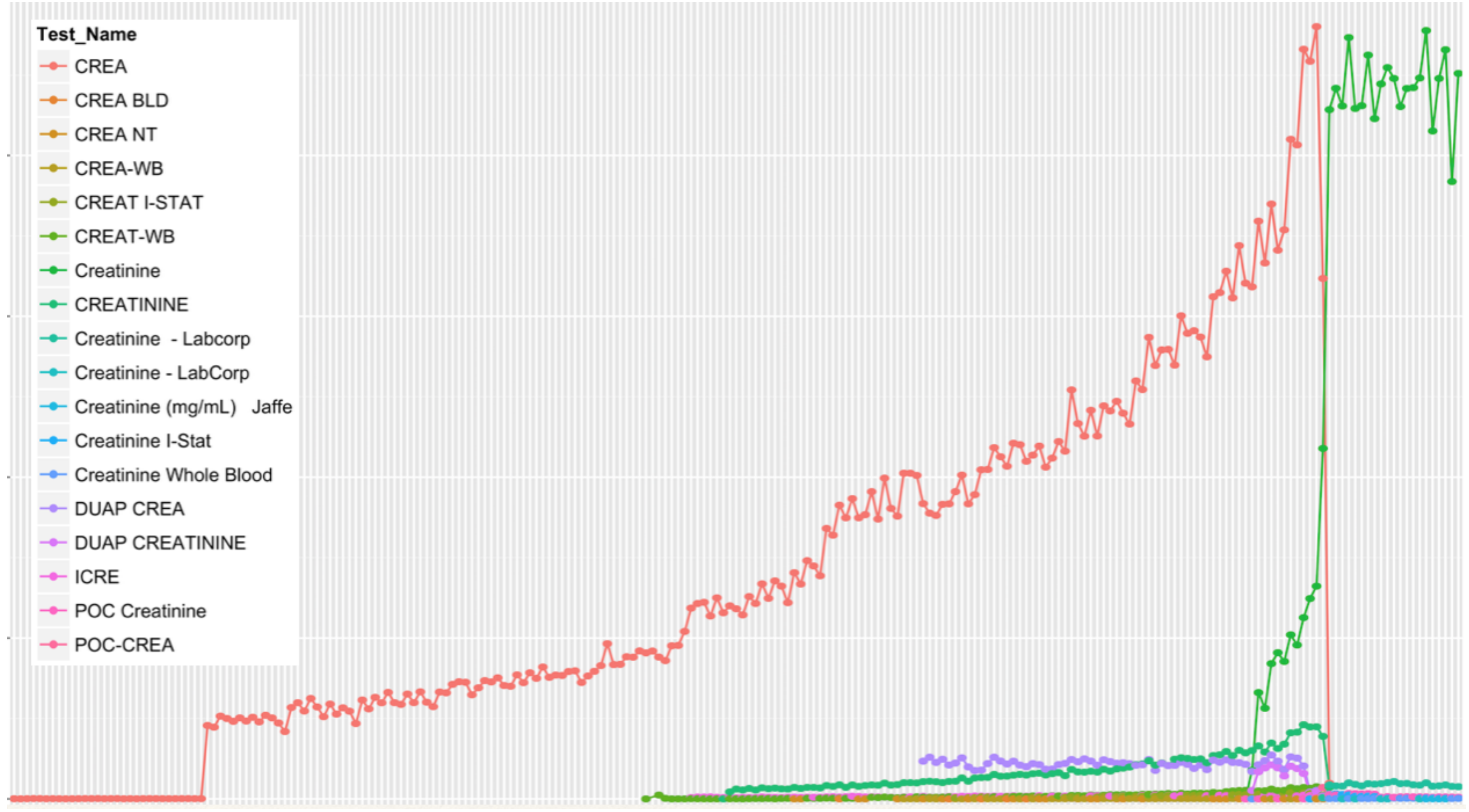

Figure 2: Monthly counts for different lab names for Serum Creatinine, 1996-2015.

Proposed Joint Model

- Hierarchical latent variable model: capture dependencies between trajectory, events.
- $\vec{y_i}$: observed eGFRs at times $\vec{t_i}$; $\vec{u_i}$: times of cardiac events (may be none).
- Assume independence in conditional likelihood (z_i, b_i, f_i, v_i latents for person i):

$$p(\vec{y_i}, \vec{u_i}|z_i, b_i, f_i, v_i; x_i) = p(\vec{y_i}|z_i, b_i, f_i; x_i)p(\vec{u_i}|z_i, b_i, f_i, v_i; x_i).$$

- **Longitudinal Submodel**:
  - Likelihood further factorizes: $p(\vec{y_i}|z_i, b_i, f_i) = \prod_{j=1}^{N_i} p(y_i(t_{ij})|z_i, b_i, f_i)$.

$$y_i(t) = m_i(t) + \epsilon_i(t),\ \ \epsilon_i(t) \overset{iid}{\sim} N(0, \sigma_\epsilon^2) \tag{1}$$
$$m_i(t) = \Phi_p(t)^\top \Lambda x_{ip} + \Phi_z(t)^\top \beta_{z_i} + \Phi_l(t)^\top b_i + f_i(t). \tag{2}$$

  - *Population component*: Fixed intercept, slope from baseline covariates.
  - *Subpopulation component*: Latent subpopulation $z_i \in \{1, \ldots, G\}$, unique B-spline trajectory.
  - *Individual component*: Random intercept, slope.
  - *Structured noise component*: Transient trends in trajectory, GP with OU kernel.

- **Point Process Submodel**:
  - Poisson process model, with Cox proportional hazards rate function:

$$p(\vec{u_i}|z_i, b_i, f_i, v_i) = \prod_{k=1}^{K_i} r_i(u_{ik}) \exp\{-\int_{T_i^-}^{T_i^+} r_i(t)dt\} \tag{3}$$
$$r_i(t) = r_0(t) \exp\{\gamma^\top x_{ir} + \alpha m_i(t) + \delta m_i'(t) + v_i\} \tag{4}$$

  - γ, α, δ: association between risk for events and baseline covariates, $m_i(t), m_i'(t)$ in (2).
- **Inference**: Fit joint model with stochastic variational inference.
- Mean field variational distribution; sparse GPs for f_i (pseudo-inputs $\vec{t_i}$).
- Lower bound has closed form; automatic differentiation for gradients.

Results

- 23,450 patients with at least moderate stage CKD, 10+ eGFR readings.
- Fit joint models to eGFR trajectory and heart attack (AMI), stroke (CVA) events.
- Longitudinal submodel evaluation: MSE/MAE on held-out eGFR values.
- Point Process submodel evaluation: AUROC, AUPR predicting future events.

Longitudinal Submodels	MSE	MAE
Joint Model (CVA)	**147.31**	**9.01**
Joint Model (AMI)	152.78	9.15
[Schulam, 2015]	155.80	9.27

		1 yr.	2 yr.	3 yr.	4 yr.	5 yr.
CVA: AUROCs	Joint Model	**0.786**	**0.746**	**0.727**	**0.742**	**0.740**
	Cox	0.574	0.597	0.602	0.606	0.587
	Time-varying Cox	0.576	0.557	0.563	0.593	0.566
AMI: AUROCs	Joint Model	**0.755**	**0.704**	**0.737**	0.654	**0.663**
	Cox	0.704	0.676	0.617	0.599	0.640
	Time-varying Cox	0.640	0.652	0.647	**0.663**	0.655
CVA: AUPRs	Joint Model	**0.423**	**0.389**	**0.370**	**0.405**	**0.400**
	Cox	0.065	0.101	0.123	0.137	0.134
	Time-varying Cox	0.062	0.086	0.114	0.157	0.130
AMI: AUPRs	Joint Model	**0.163**	**0.128**	**0.172**	**0.166**	**0.119**
	Cox	0.052	0.059	0.051	0.065	0.083
	Time-varying Cox	0.048	0.057	0.067	0.088	0.103

Table 1: Top: Longitudinal submodel results. Bottom: Point Process submodel results.

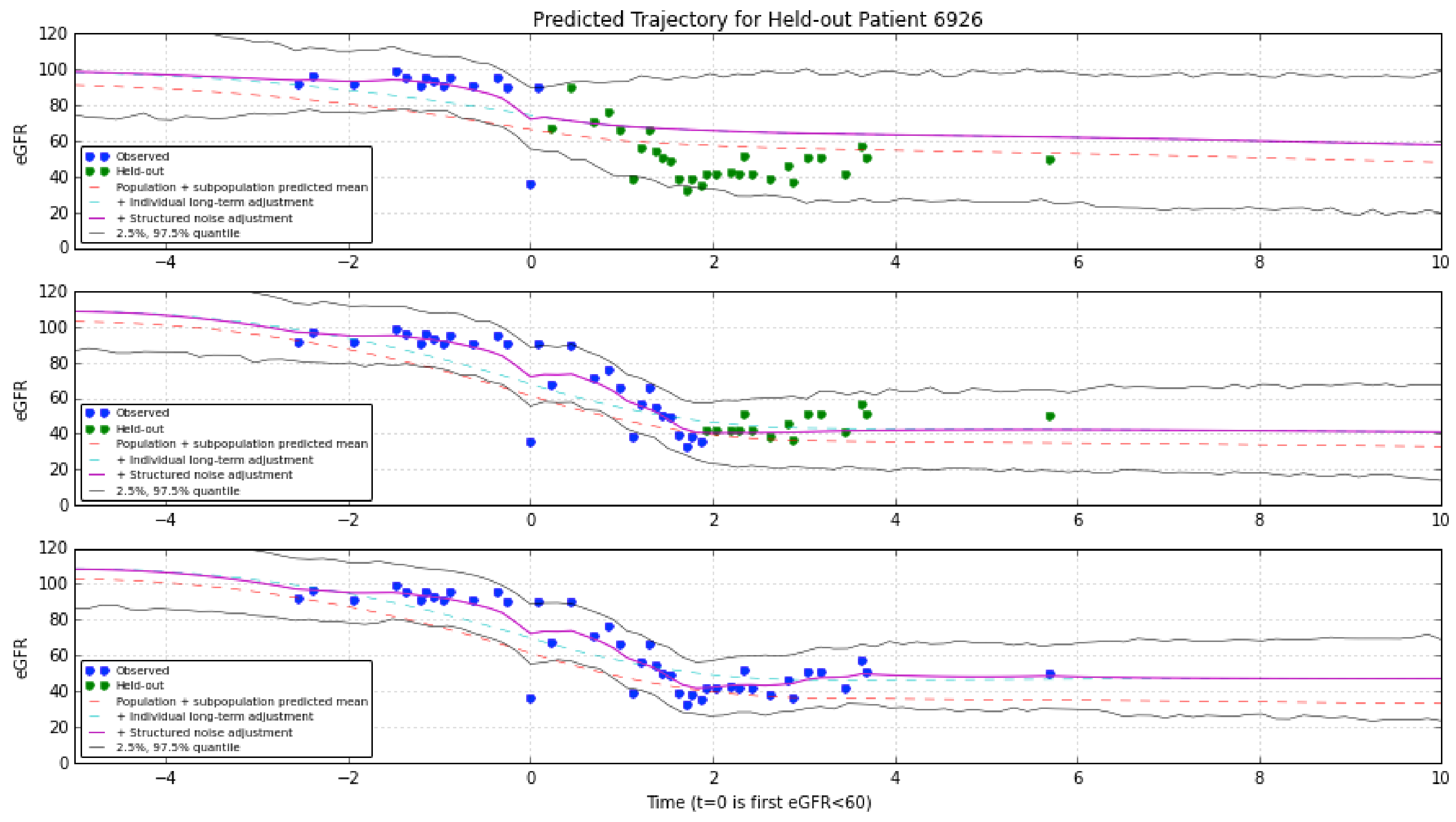

Figure 3: Dynamic predictions of disease trajectory from longitudinal submodel.

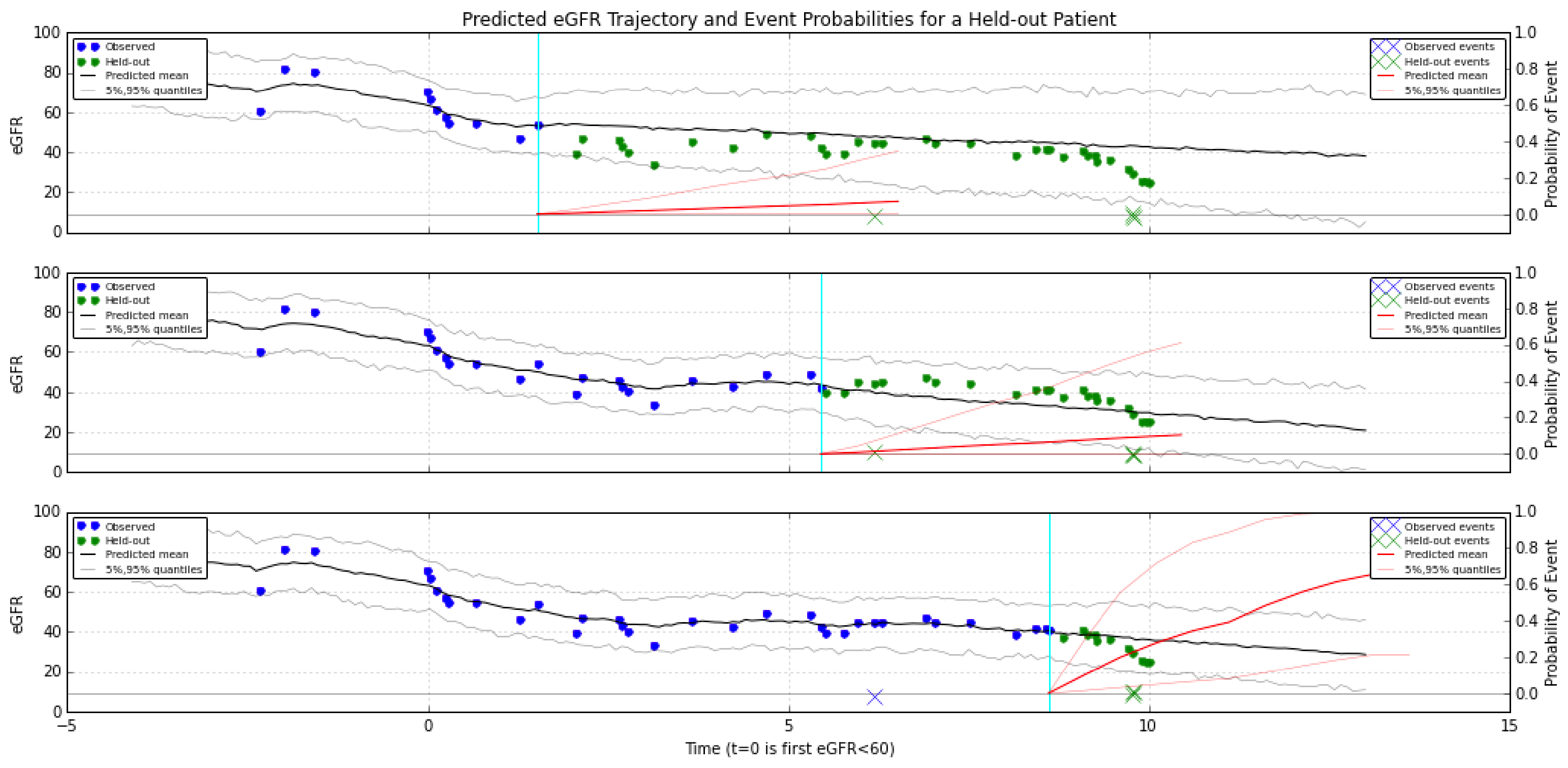

Figure 4: Dynamic predictions of disease trajectory, risk of CVA event from joint model.

Conclusion

- Novel joint model for longitudinal, point process data.
- First scalable stochastic variational inference algorithm for this model class.
- Future work: multivariate in longitudinal, point process data; more flexible models.