Improving confidence while predicting trends in temporal disease networks

Gligorijevic, Dj., Stojanovic, J., Obradovic, Z., "Improving confidence while predicting trends in temporal disease networks", *2015 SIAM International Conference on Data Mining, Workshop on Data Mining for Medicine and Healthcare,* Vancouver, Canada, 2015

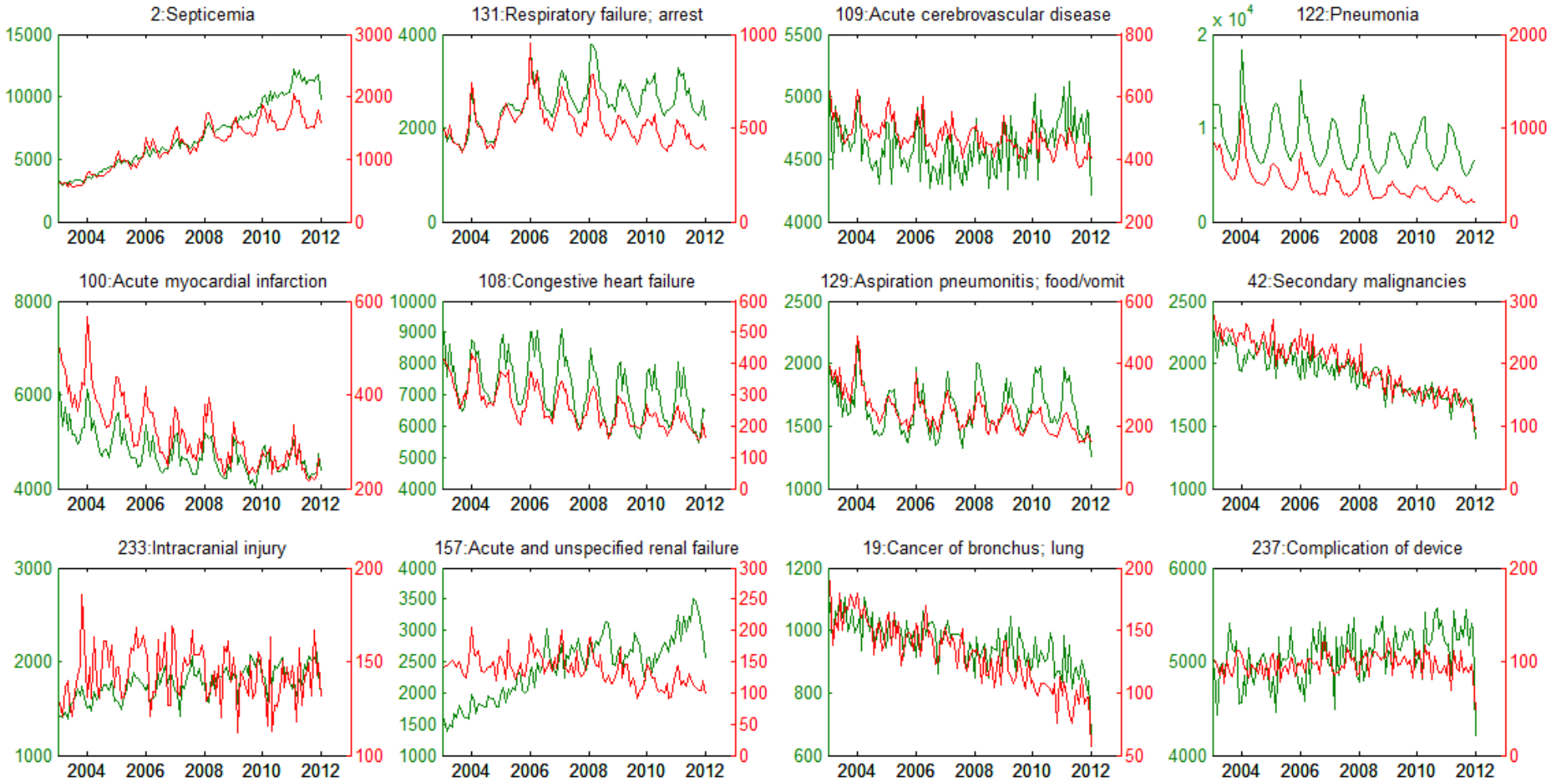
Motivation

Motivation

- Having good prediction accuracy alone is often not enough.
- Reporting uncertainty estimation of the prediction is very important, especially in domains where predictions are used for important decision-making, such as health.
- Eg., predicting admissions in a hospital as 15.026±10.000 vs. 15.000±150
- We aim to address this important topic – to improve the estimation quality of prediction uncertainty in the GCRF model
- We introduce several extensions to the Gaussian Conditional Random Fields model, which aim to provide higher quality uncertainty estimation.

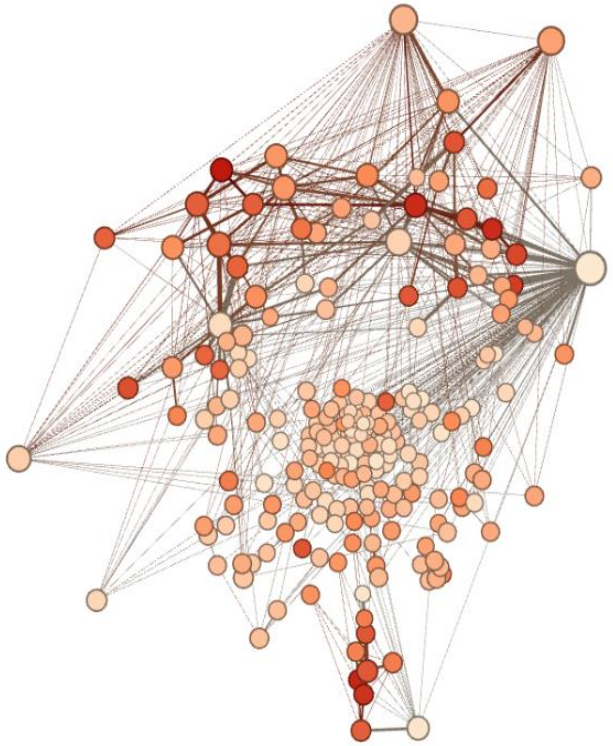
Evolution of The Top 12 Killing Diseases in SID CA

- Capture **disease trends** and estimate their value in the future



Healthcare Application: Disease Networks

- **Goal:** Predict **monthly hospital admission** for **253 disease categories** in **California** for each month of the year **2011** in order to *facilitate decision making and improve health care delivery*
- **Data:** HCUP California state inpatient database
  - Experiments conducted on 24 monthly graph snapshots (~8M inpatients)
- **Representation:** Monthly phenotype-disease graph
  - Nodes: 253 disease categories (CCS codes)
  - Links:
    - ✓ Disease comorbidities (displayed on the right)
    - ✓ **Disease similarities over the previous 3 months**



Disease comorbidity graph

Methodology and Approaches

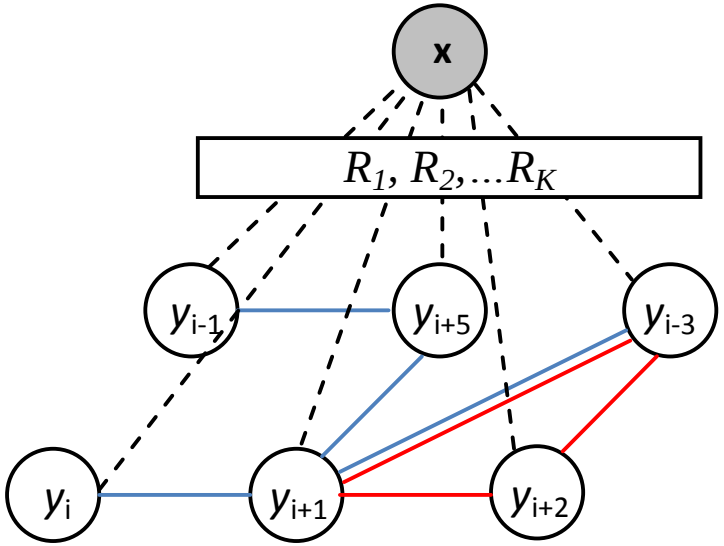
Structured Learning by Gaussian Conditional Random Fields

- Gaussian Conditional Random Field (GCRF) model:

$$P(\mathbf{y}\mid\mathbf{x})=\frac{1}{Z(\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta})}\exp\Big(\sum_{i=1}^{N}A(\boldsymbol{\alpha},y_i,\mathbf{x})+\sum_{j\sim i}I(\boldsymbol{\beta},y_i,y_j,\mathbf{x})\Big)$$

- Interpretation and modeling capabilities

$$A(\boldsymbol{\alpha},y_i,\mathbf{x})=-\sum_{k=1}^{K}\alpha_k(y_i-R_k(\mathbf{x},i))^2,\quad I(\boldsymbol{\beta},y_i,y_j,\mathbf{x})=-\sum_{l=1}^{L}\beta_l e_{ij}^{(l)}S_{ij}^{(l)}(\mathbf{x})(y_i-y_j)^2$$



- **P(y|x)** is **Gaussian distribution**
- **Learning**: finding parameters α and β is **convex optimization**
- **Inference**: Point estimate of y for given x is μ, uncertainty is Σ, where P(y|x)~N(μ, Σ)

Problem statement

- GCRF can exploit the graph structure for regression.
  - However **GCRF uncertainty estimation** is **not** taking into account:
    1. Uncertainty of unstructured predictors,
    2. Distribution of input data,

  $\Rightarrow$ **under-confident** predictions with high predictive uncertainty!

  - For example prediction of Sepsis admission: ~ 9,059 ± 15,867

- **Goal**: Solve these two problems to significantly improve GCRF uncertainty estimation.

- **The idea**: Use **functions** instead of **scalars** as the GCRF parameters.

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \exp\left(-\sum_{i=1}^{N} \boldsymbol{\alpha}_k (y_i - R_k(x, i))^2 - \sum_{l=1} \boldsymbol{\beta}_l e_{ij}^{(l)} S_{ij}^{(l)}(x)(y_i - y_j)^2\right)$$

1. The uGCRF approach

- Parameters of unstructured predictors, α_k, now become dependent on uncertainty estimation of unstructured predictors $\sigma_{k,p}^2$ (uncertainty of predictor k in time step p)

$$\alpha_{k,p} = \frac{e^{u_{k,p}}}{\sigma_{k,p}^2} \qquad \beta_l = e^{v_l}$$

- It captures the uncertainty of unstructured predictors thus providing "healthier" degree of belief towards it,
- Cons: not able to *adapt to the errors,* model is making, while predicting.

2. The ufGCRF approach

- Parameters of unstructured predictors, α_k, now become parametrized functions (feed-forward Neural Networks) of input variables X for each node in a graph

$$\alpha_k(\theta_k, x) = e^{u_k(x,\theta_k)} = e^{\sum \theta_l x_l} \qquad \beta_l = e^{v_l}$$

- It better adapts to errors the unstructured model is making, as it uses previous time-steps as inputs
- Experimentally we demonstrate that this approach is better!

Experiments

Experimental set-up

- Prediction of monthly hospital admission for **253 disease categories (nodes)** in **California** for 12 months in **2011**
- We normalize values and predict with linear and non-linear predictors with different values of lag
- **Modelling setup:**
  - Use all models in an autoregressive fashion and predict one-step-ahead
  - Move 12 month training window and obtain next month's prediction, repeat for 1 year.
- **Evaluation:**
  - Predictive accuracy (*Root Mean Squared Error (RMSE)*)
  - Quality of uncertainty estimate (*Negative Log Predictive Density (NLPD)*)

Benchmarks

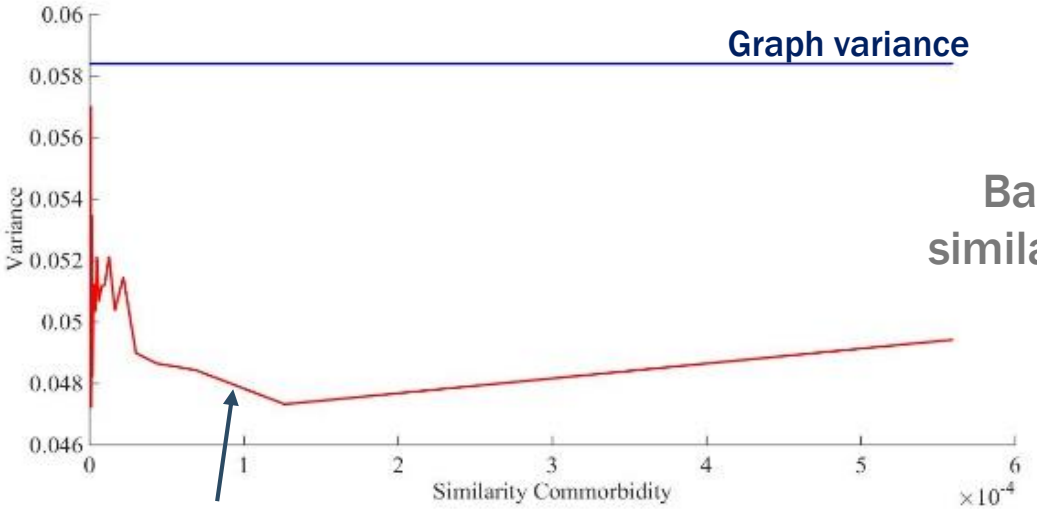
- Linear and non-linear unstructured models are trained with up to 3 previous time steps used as inputs (lag1, lag2, lag3):

  - Linear Regression (lag1, lag2, lag3)
  - Gaussian Processes Regression (lag1, lag2, lag3)
  - GCRF
  - uGCRF (GCRF with parameters sensitive to uncertainty of unstructured predictors)
  - ufGCRF (GCRF with parameters modeled as feed-forward NN)

Utilization of disease graph structure

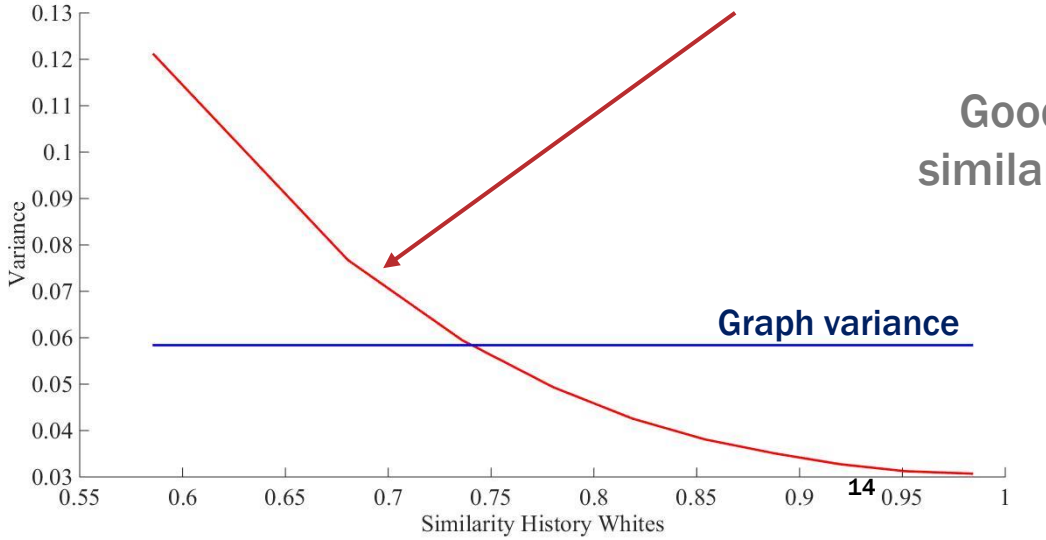
- We are considering several **graphs**:
  - ***Disease comorbidity graph***
  - *Jenson-Shannon graph* (based on distribution of admitted whites in history)
  - *Common history graph* (based on distribution of admitted whites in previous 3 months)

- Using variogram technique (***smoother drop is desirable***) we discover that using Common history graph is most beneficial for our problem.



Bad similarity

Comorbidity graph (above) vs **Common history graph** (below)

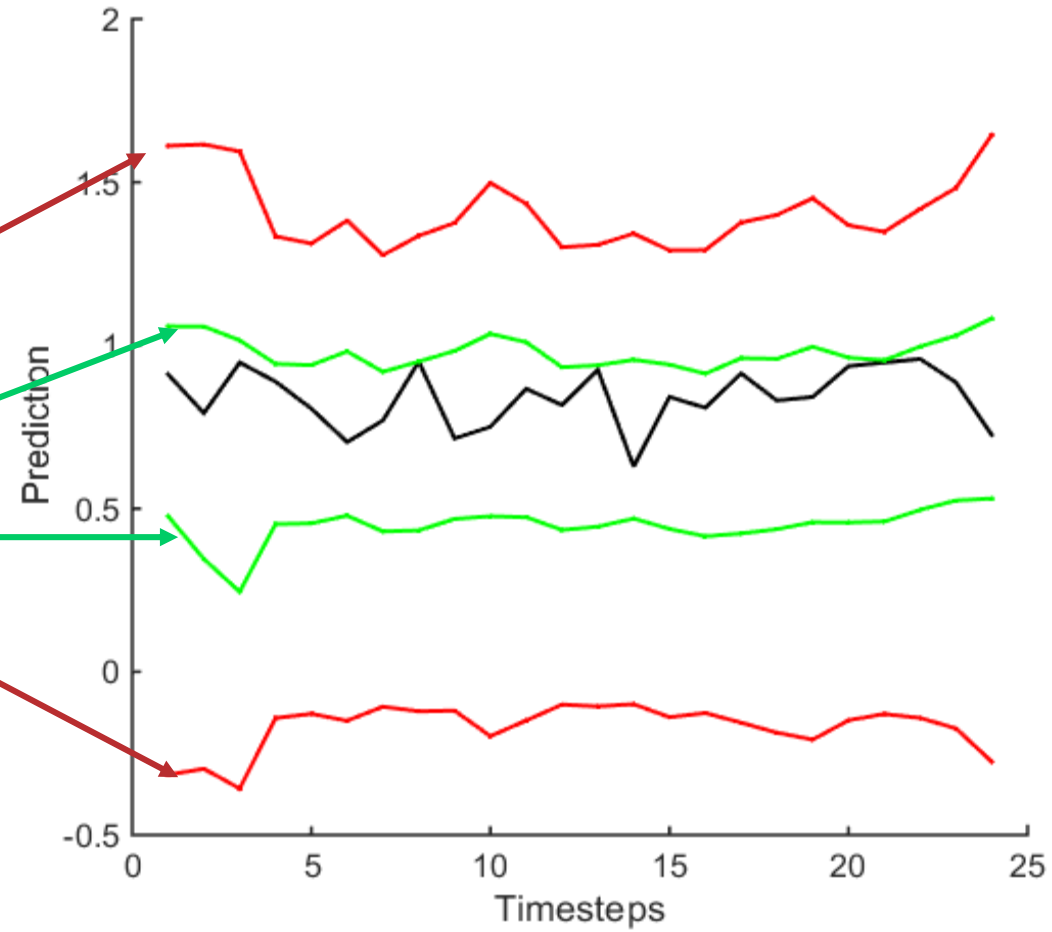


Good similarity

Experiment 1: Disease admission for each of 12 months – Hepatitis admission prediction

- Confidence estimation ($\mu \pm 1.96 * \sigma$, where μ is mean and σ is standard deviation) of predicted admission for 12 months using ufGCRF was much better than when using GCRF.

  - ✓ **GCRF prediction : ~ 442 ± 544**
  - ✓ **ufGCRF prediction: ~ 527 ± 289**
  - ✓ **True admissions: : ~ 478 ± 167**

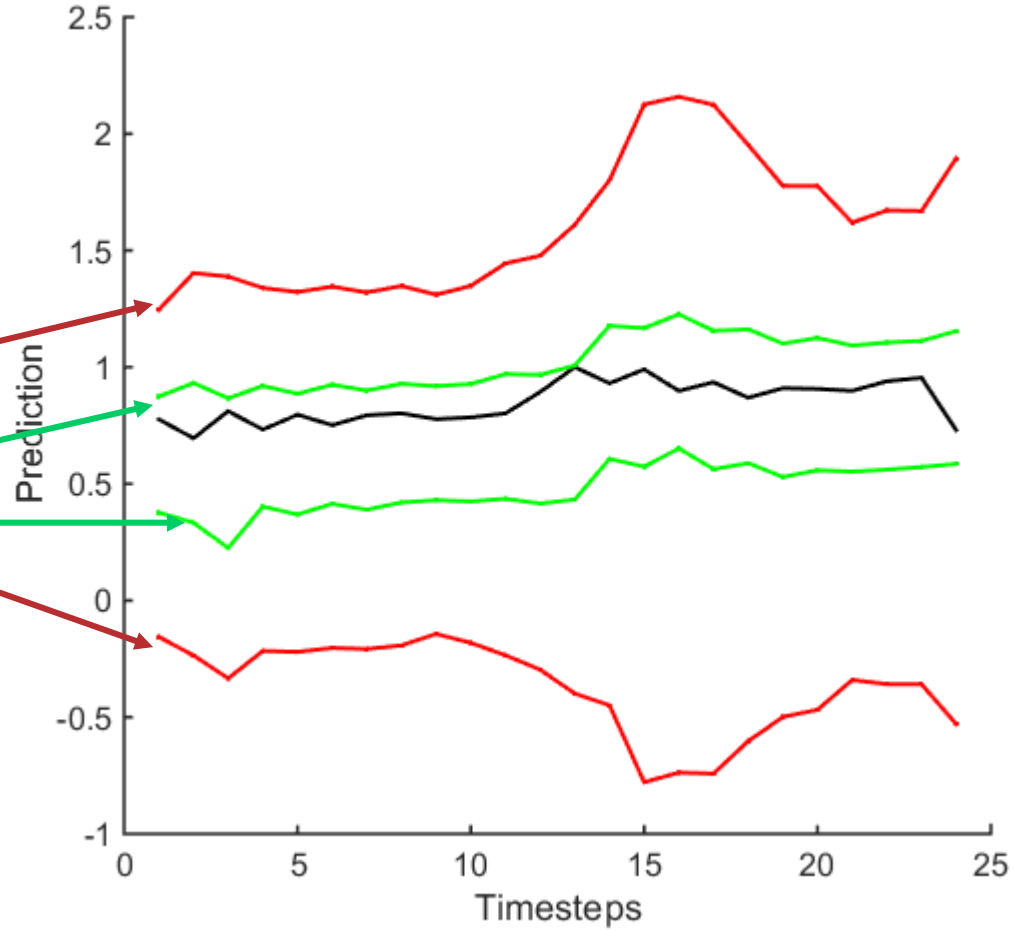


GCRF uncertainty region (red) vs ufGCRF uncertainty region (blue). *Values are normalized*.

Experiment 1: Disease admission for each of 12 months – Sepsis admission prediction

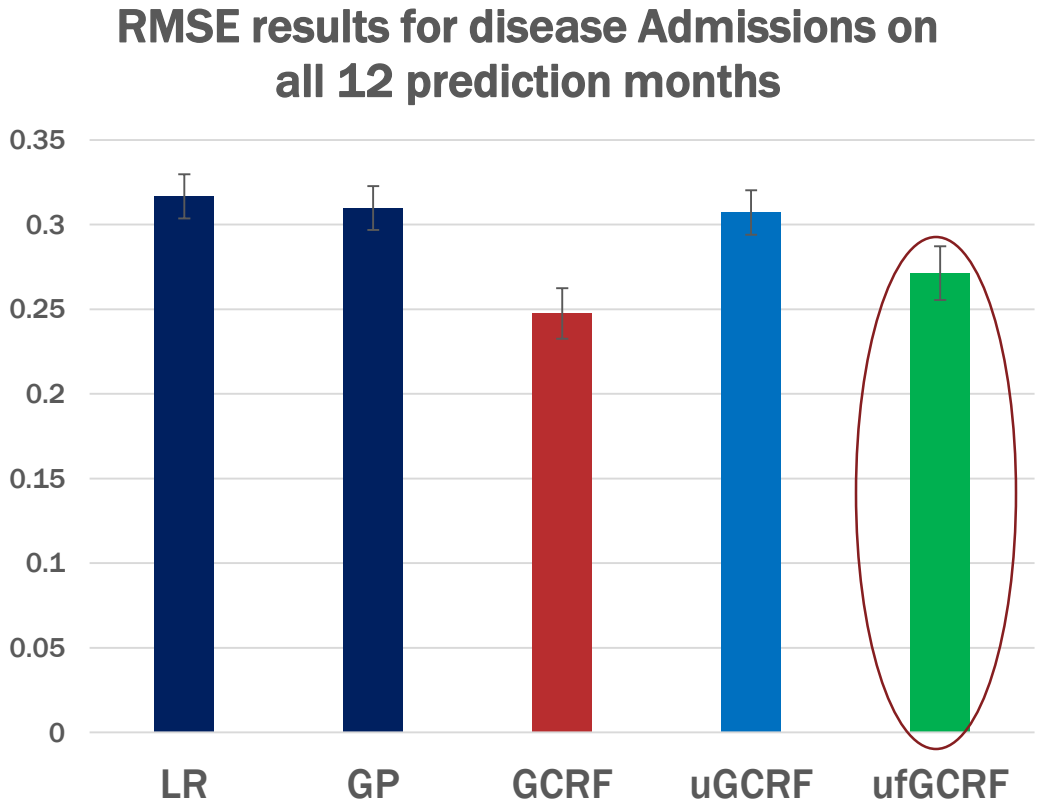
- Confidence estimation ($\mu \pm 1.96 * \sigma$, where μ is mean and σ is standard deviation) of predicted admission for 12 months using ufGCRF was much better than when using GCRF.

  - ✓ **GCRF prediction : ~ 9,059 ± 15,867**
  - ✓ **ufGCRF prediction: ~ 10,791 ± 3,539**
  - ✓ **True admissions: : ~ 11,400 ± 4,128**

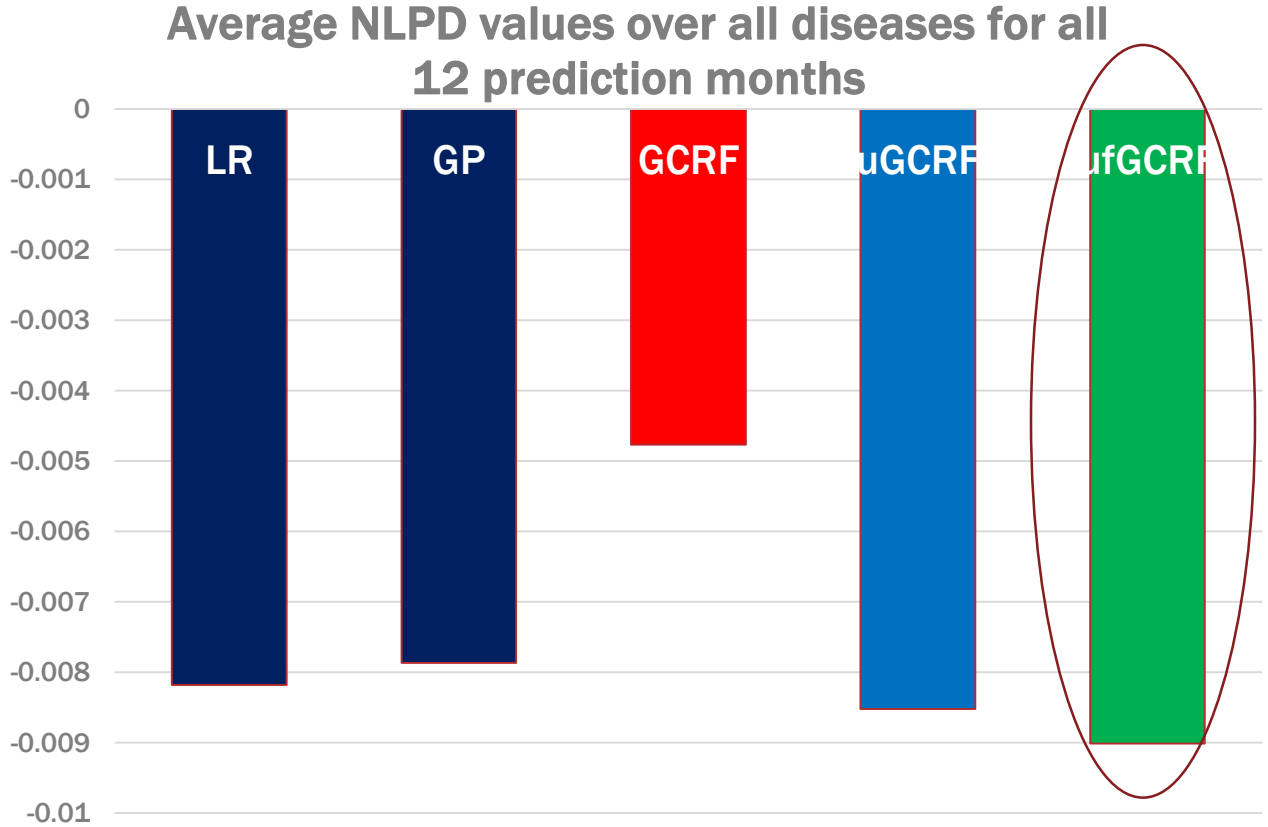


GCRF uncertainty region (red) vs ufGCRF uncertainty region (blue). *Values are normalized*.

Experiment 2: ufGCRF compared to alternative methods on all diseases – RMSE and NLPD results



- Graph structure improves predictive accuracy
- Two extensions uGCRF and ufGCRF introduce small additional error to predictive accuracy.
- ufGCRF provided the best balance between predictive accuracy and quality of uncertainty estimation



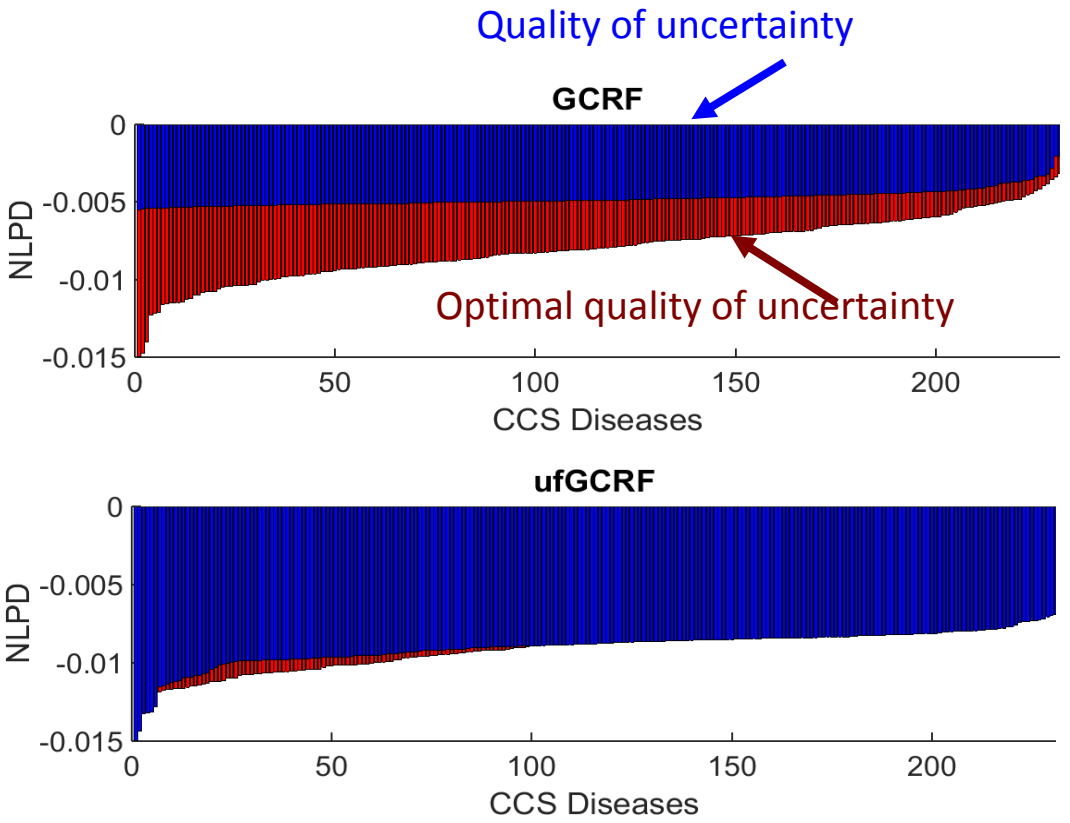
- GCRF provides lower quality of uncertainty estimation in this dataset.
- The two extensions significantly improve predictive accuracy outperforming all of the unstructured predictors.

Experiment 3: Quality of uncertainty estimate for all diseases admission for each of 12 months

- The uncertainty estimation is evaluated by the NLPD metric (lower values are better).
  - *Red bars* - optimal uncertainty for achieved predictions of models
  - *Blue bars* - achieved uncertainty quality for each disease.

- $NLPD = \frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - y_{i*})^2}{2\sigma_{i*}^2} + \log(\sigma_{i*}^2)$

- ufGCRF outperformed GCRF's uncertainty estimation for each disease (uncertainty estimates are near optimal ones for obtained prediction quality)

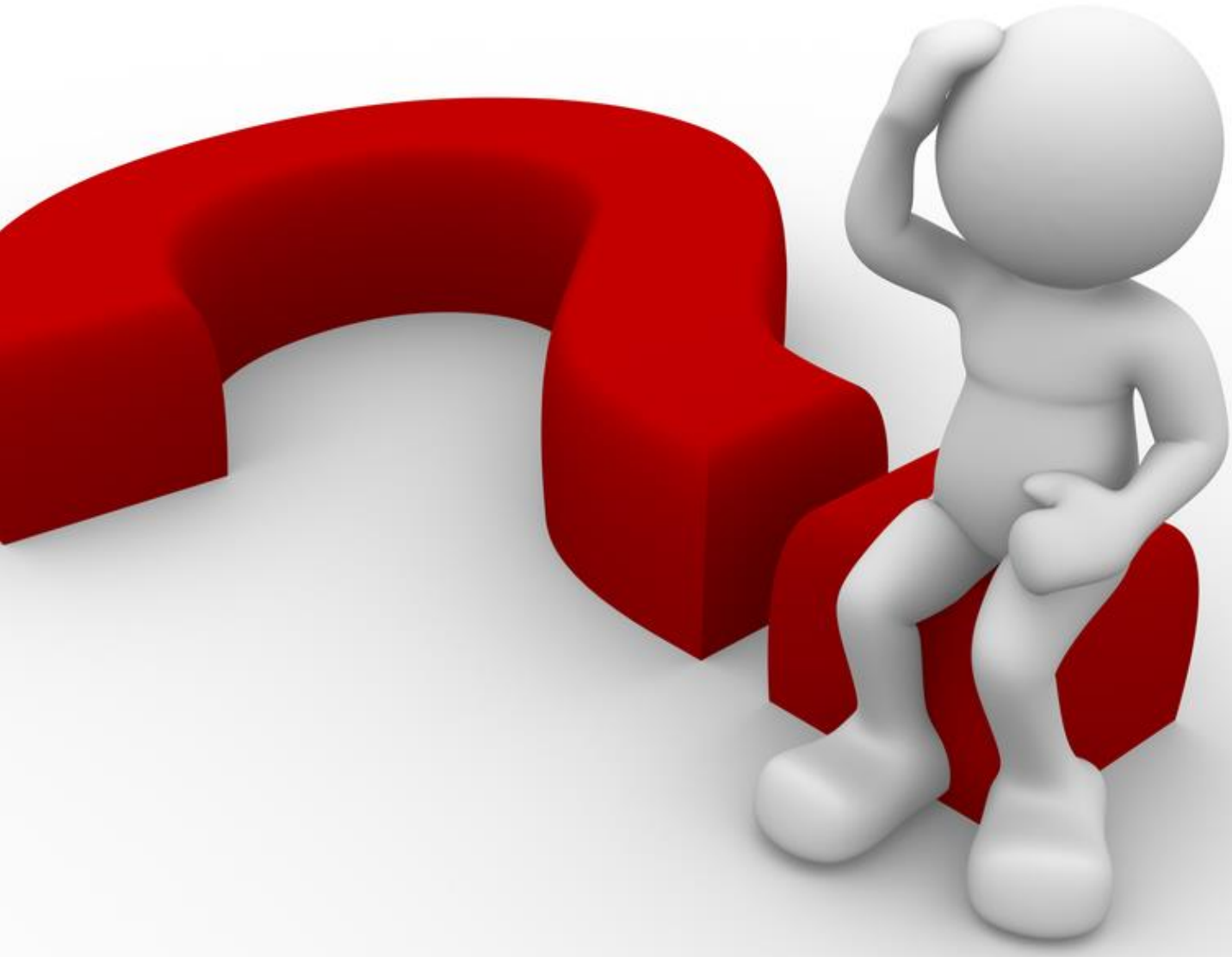


Optimal (red) vs. achieved (blue) uncertainty quality when using GCRF (top) and ufGCRF (bottom)

Conclusions

Conclusions

- In this study, the GCRF model is applied to a challenging problem of admission rate prediction, based on a temporal graph built from HCUP (SID) data
- In the experiments we characterize:
  - several unstructured (Linear Regression and Gaussian Processes with lag 1, lag 2 and lag 3) and
  - structured predictors (original GCRF, uGCRF and ufGCRF) for their predictive error and quality of uncertainty estimation.
- All three structured models outperformed unstructured ones in terms of predictive error, showing that structure brings useful information to this prediction task
- Even though the original GCRF model showed the best performance in predictions, it had the lowest quality of uncertainty estimation. Introducing small predictive error, uGCRF and ufGCRF models gained large improvements in uncertainty estimation, especially the ufGCRF model that had the better performance in prediction of these two GCRF model extensions.



Thank you for your attention!

Questions?