Improving confidence while predicting trends in temporal disease networks

SUBTITLE

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

Title and Content Layout with List

- Methodology and Approaches
- Application
- Experiments
- Conclusion

Methodology and Approaches

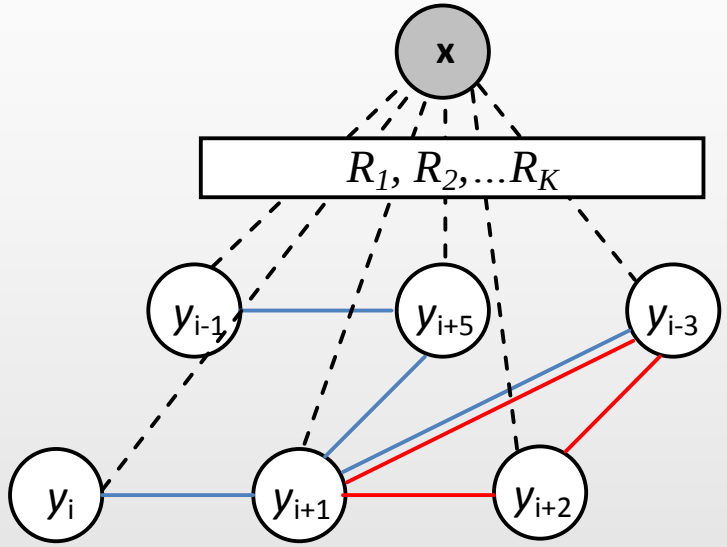
Structured Learning by Gaussian Conditional Random Fields

- Gaussian Conditional Random Field (GCRF) model:

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

- 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)

Proposed approaches for dealing with uncertainty of prediction

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

    thus often leading to underconfident predictions with high predictive uncertainty!
- **Goal**: Solve these two problems to significantly improve GCRF uncertainty estimation.
- **The idea**: Use functions instead of scalars as the GCRF parameters. Thus we compare two approaches:

1. The uGCRF approach

- ➢ Parameters of unstructured predictors, α_k, now become dependent on uncertainty estimation of unstructured predictors

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

2. The ufGCRF approach

- ➢ Parameters of unstructured predictors, α_k, now become parametrized functions of input variables X for each node in a graph

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

- ➢ Experimentally we show that this approach is better!

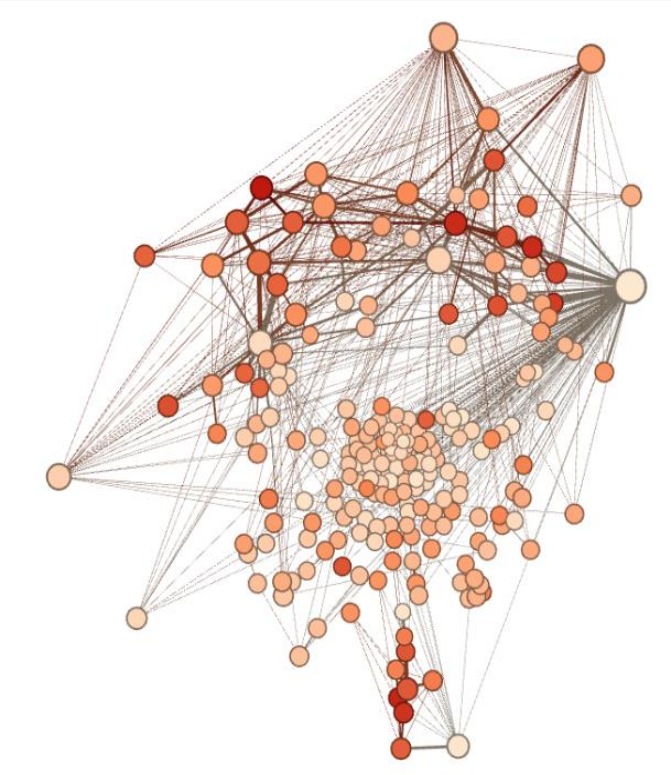
Approach setup and benchmarks

- **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.
- **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**
- **Evaluation of different models:**
  - Predictive accuracy (*Root Mean Squared Error (RMSE)*)
  - Quality of uncertainty estimate (*Negative Log Predictive Density (NLPD)*)

Application

Healthcare Application: Disease Networks

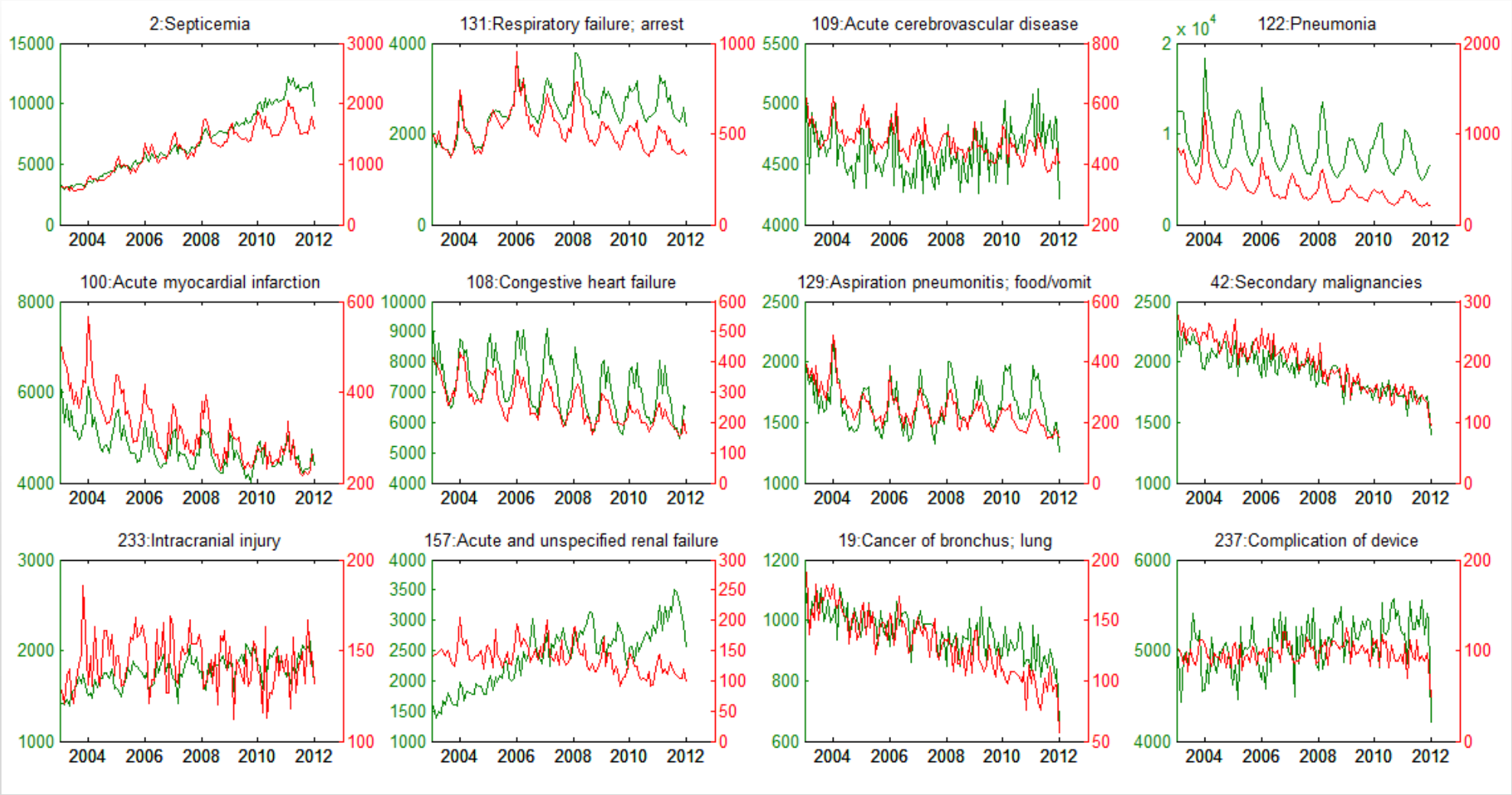
- **Goal:** Predict monthly hospital admission for **253 disease categories** in **California** for each month of the year **2011** in order to *facilitate decision making and improve health care delivery*
- **Hypothesis:** Exploiting structural relationships among diseases will improve prediction quality
- **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**
- **Data:** Experiments are conducted on 24 monthly graph snapshots (~8M inpatients) built using HCUP
  - California state inpatient database



Disease comorbidity graph

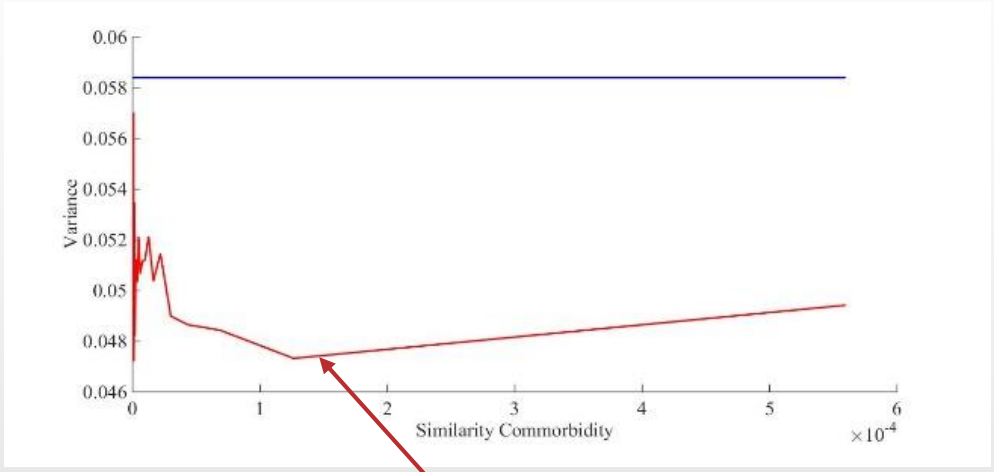
Healthcare Application: Evolution of The Top 12 Killing Diseases in SID CA

- We are able to capture disease trends and estimate their value in the future!

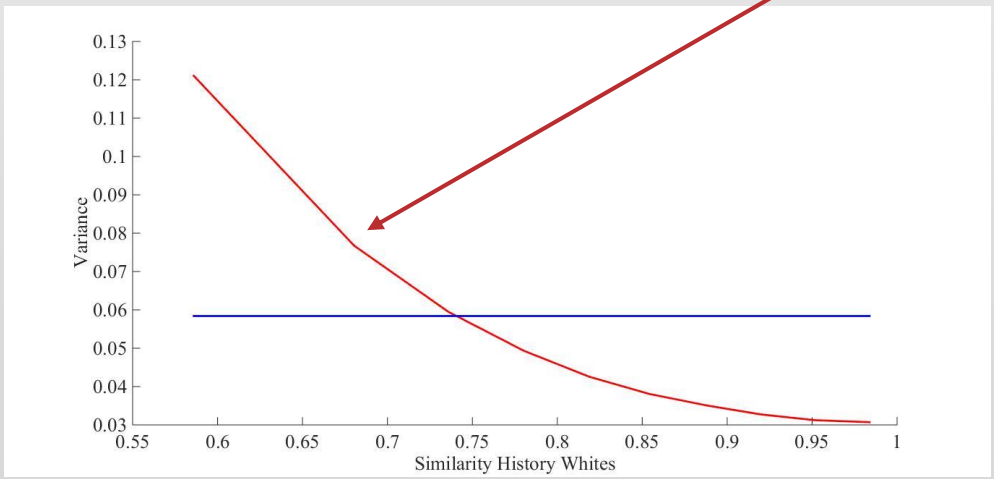


Healthcare Application: Utilization of disease graph structure

- For admission count of diseases we normalize values and predict with linear and non-linear predictors with different values of lag
- For structure information we are considering several graphs:
  - Comorbidity graph
  - Jenson-Shannon graph
  - Common history graph
- Using variogram technique (smoother drop is desirable) we discover that using Common history graph is most beneficial for our problem.



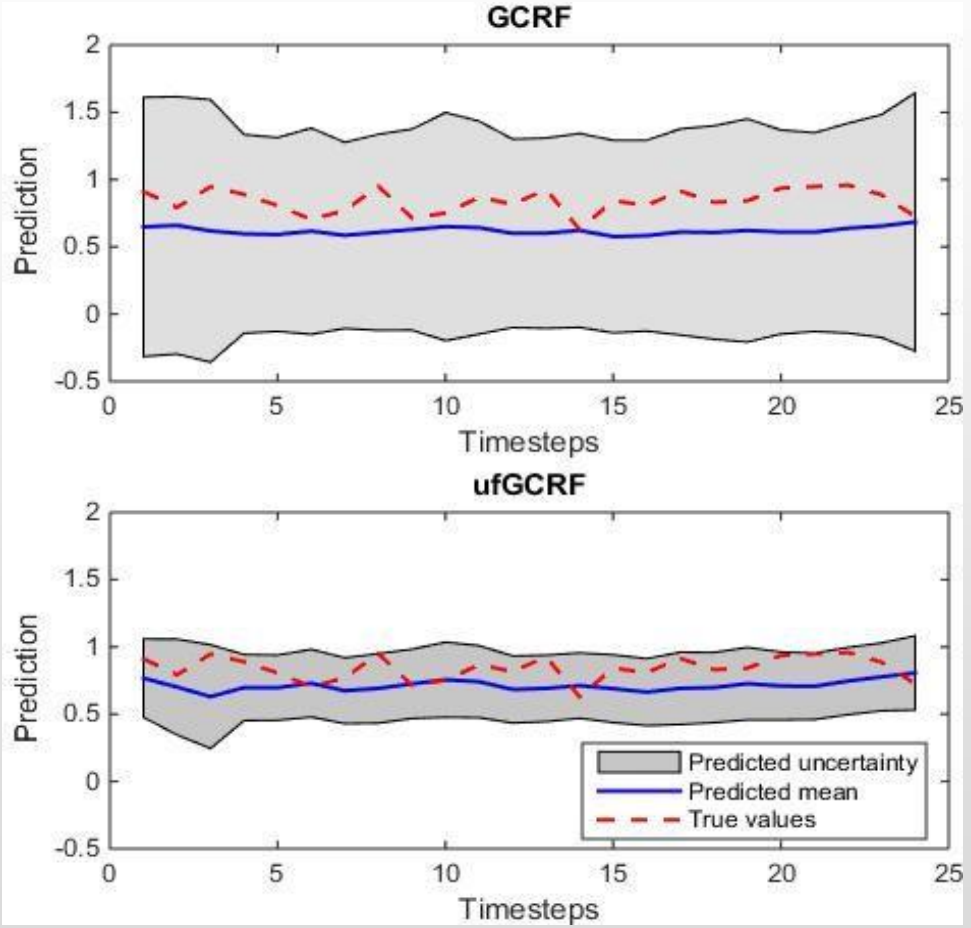
Disease comorbidity graph (above) vs Common history graph (below)



Experiments

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

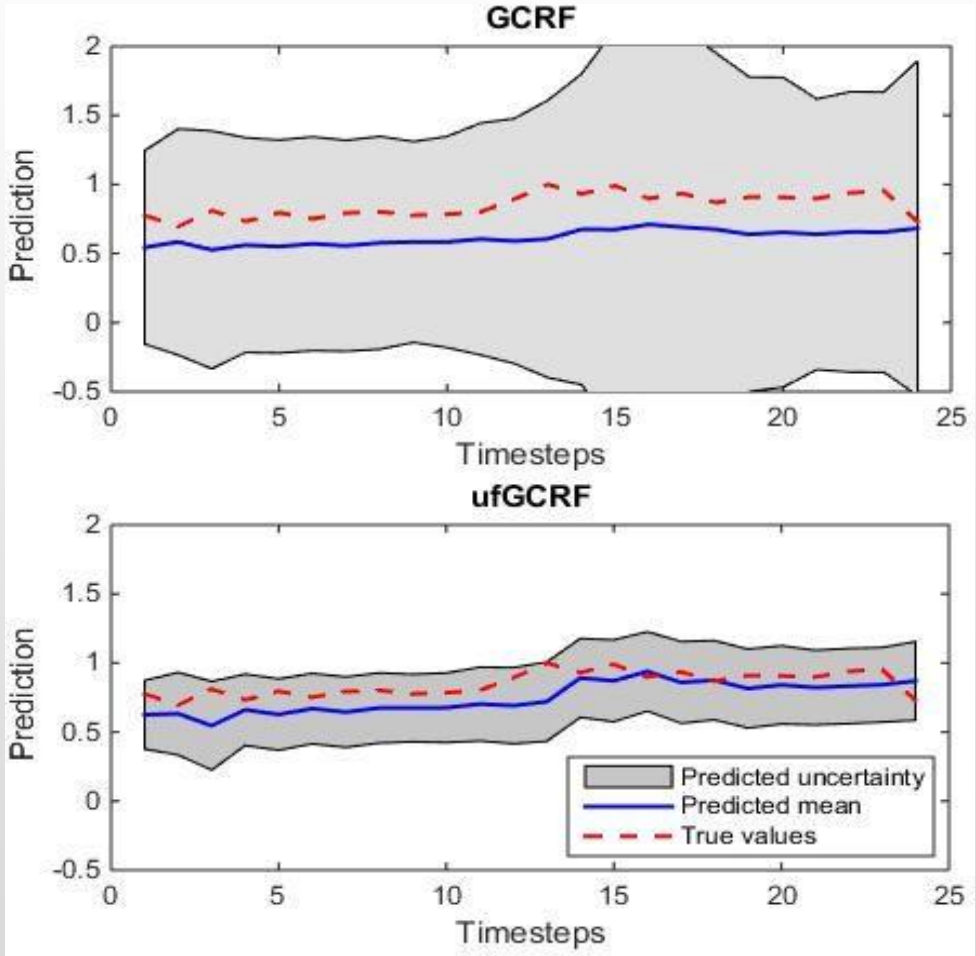
- Confidence estimation (μ±1.96*σ, where μ is mean and σ is standard deviation) of predicted admission for 12 months using ufGCRF was much better than when using GCRF.
  - ✓ **Admissions predicted by GCRF: ~ 442 ± 544**
  - ✓ **ufGCRF prediction: ~ 527 ± 289**
  - ✓ **True admissions: : ~ 478 ± 167**



Prediction mean (blue line) and uncertainty (gray area) with true values (red dashed line) of admission for Sepsis in test period of 12 months of 2011

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

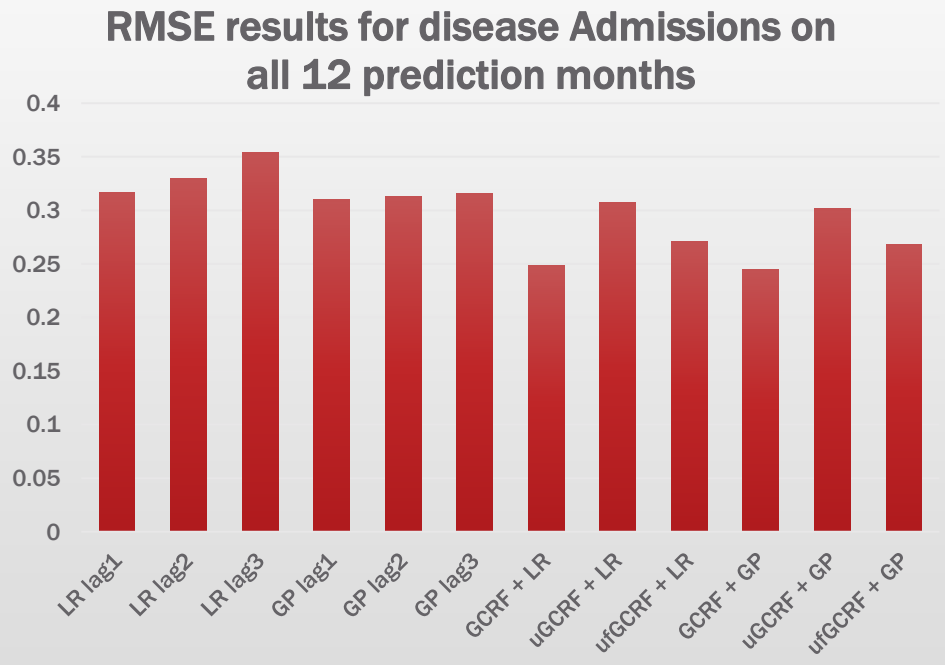
- Confidence estimation ($\mu \pm 1.96 * \sigma$, where μ is mean and σ is standard deviation) of predicted admission for 12 months using ufGCRF was much better than when using GCRF.
  - ✓ **Admissions predicted by GCRF: ~ 9,059 ± 15,867**
  - ✓ **ufGCRF prediction: ~ 10,791 ± 3,539**
  - ✓ **True admissions: : ~ 11,400 ± 4,128**



Prediction mean (blue line) and uncertainty (gray area) with true values (red dashed line) of admission for Sepsis in test period of 12 months of 2011

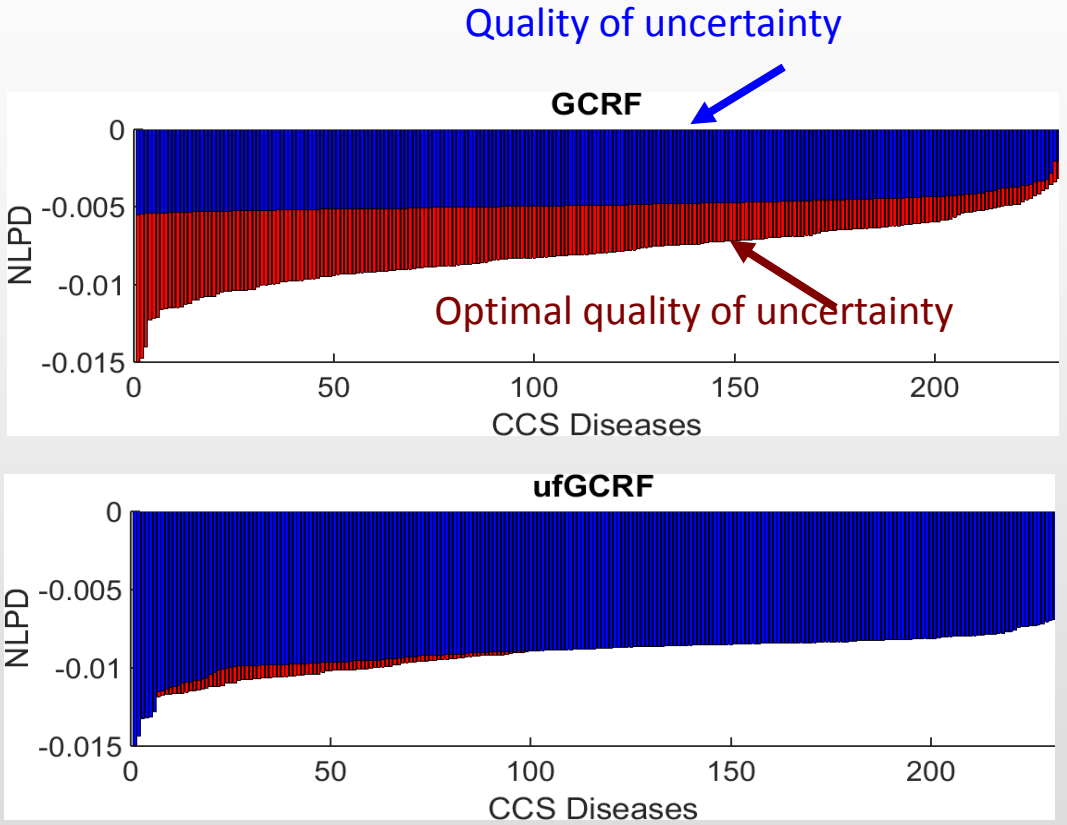
Experiment 2: ufGCRF compared to 10 alternative methods on all diseases

- ufGCRF provided the best balance between predictive accuracy and quality of uncertainty estimation!
- Predictive Accuracy (all diseases for 12 months)
- RMSE results (smaller is better) are evaluated on the normalized admission count (admission rate).
  - Graph structure improves predictive accuracy
  - Errors of unstructured predictors (~31% - ~36%), while error of GCRF modes is ~24%.
  - Two extensions uGCRF (error ~30%) and ufGCRF(error ~27%) introduce additional error to predictive accuracy.



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

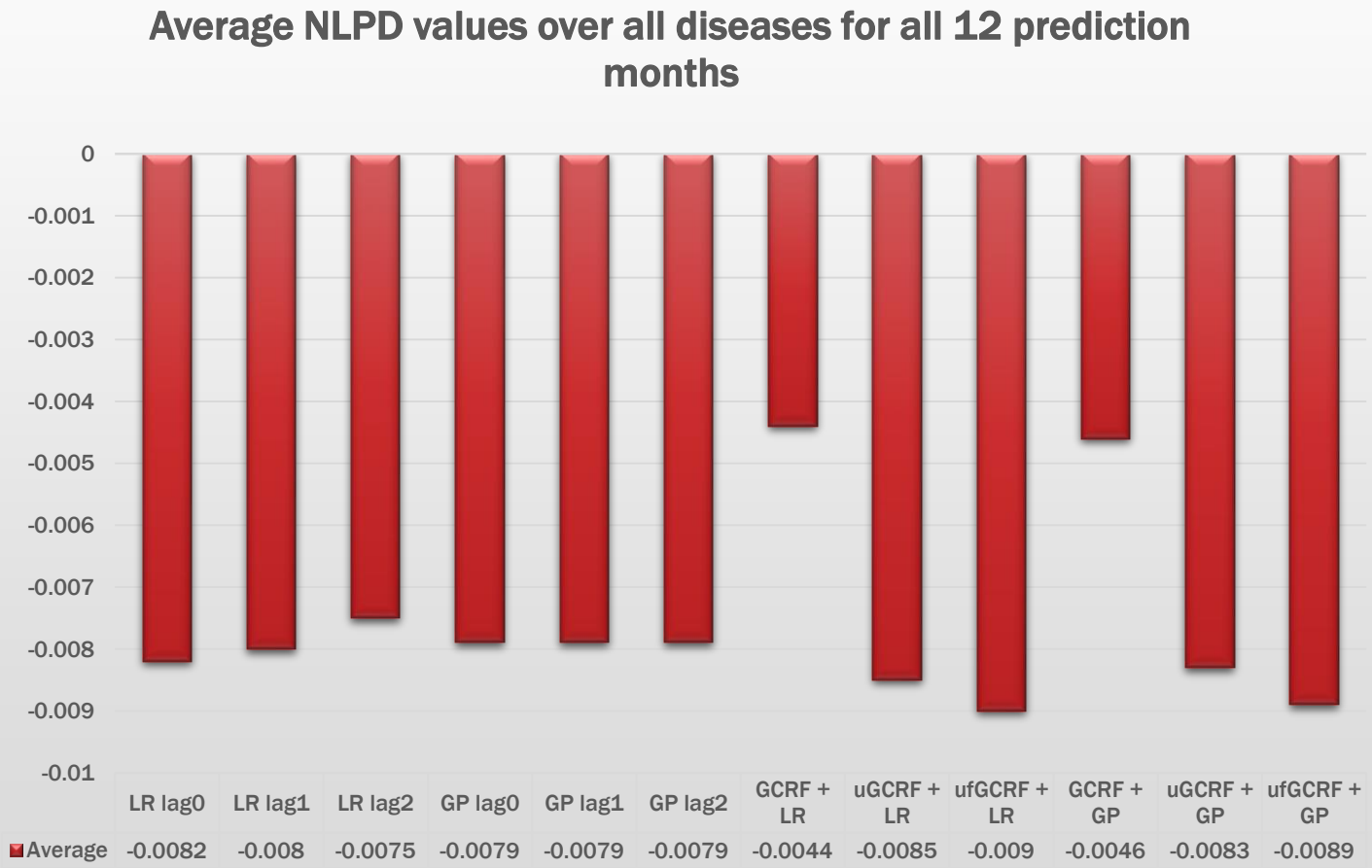
- The uncertainty estimation is estimated by the NLPD metric (lower values are better).
  - *Red bars* on plots represent optimal uncertainty for achieved predictions of models and blue lines represent achieved uncertainty quality for each disease.
- ufGCRF outperformed GCRF's uncertainty estimation for each disease (uncertainty estimates are near optimal ones for obtained prediction quality)
- On the right we see achieved uncertainty quality for three GCRF models:
  - GCRF model where parameters are **scalars** on top
  - GCRF model where parameters are **neural networks**



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

Experiment 4: Quality of uncertainty estimate (all diseases for 12 months)

- Uncertainty estimation is evaluated with NLPD metric (smaller is better) which takes into account predictive accuracy and how close is estimated variance to true variance of the data.
- GCRF provides lower quality of uncertainty estimation in this dataset. The two extensions significantly improve predictive accuracy outperforming all of the unstructured predictors.



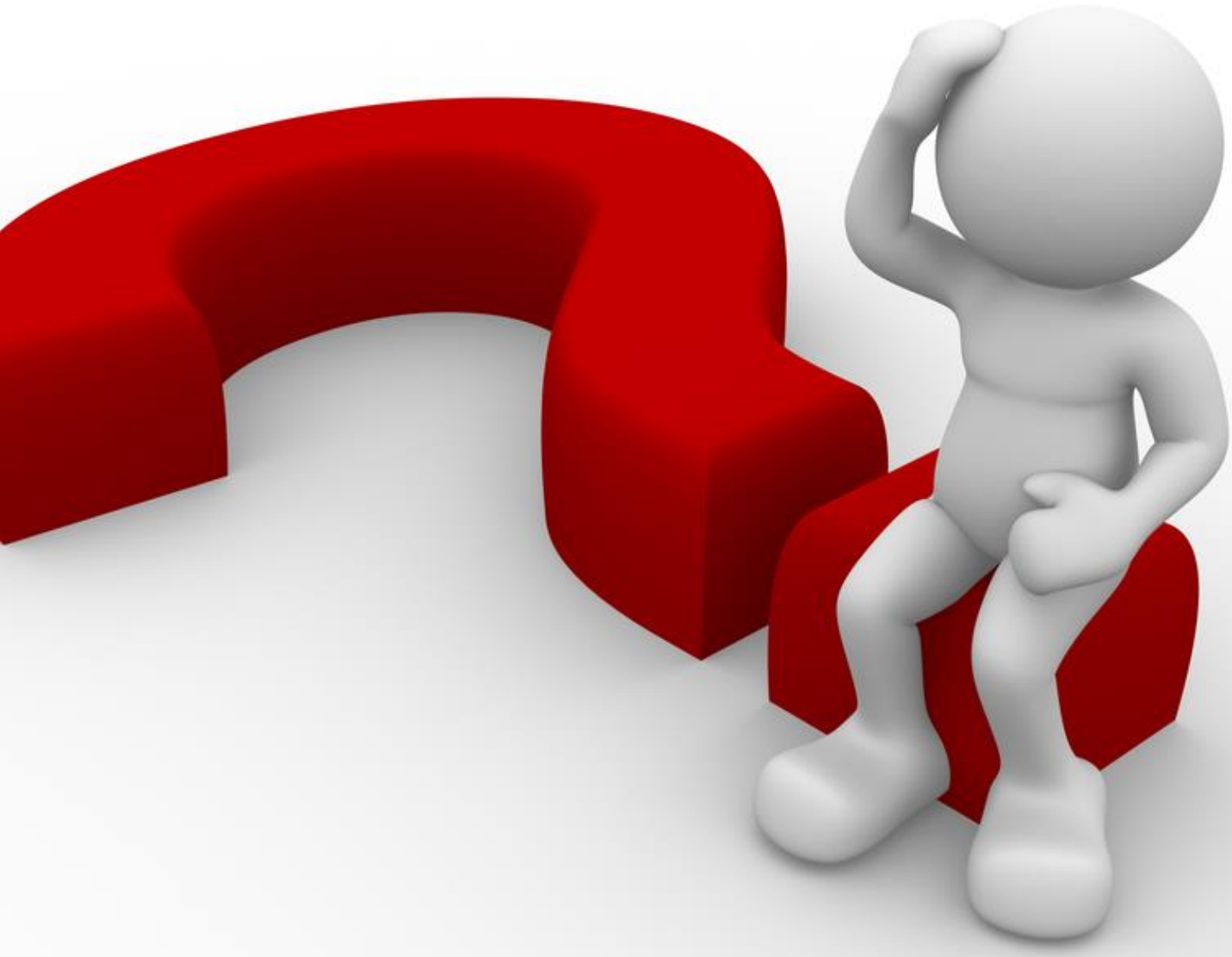
Conclusions

Conclusions

Add more conclusions.....

- In our experiments ufGCRF provides the best balance between predictive accuracy and uncertainty estimation quality

Maybe remarks here...



Thank you for your attention!

Questions?