Early Prediction of Diabetes Complications from Electronic Health Records: A Multi-task Survival Analysis Approach

Bin Liu, Ying Li, Zhaonan Sun, Soumya Ghosh, Kenney Ng

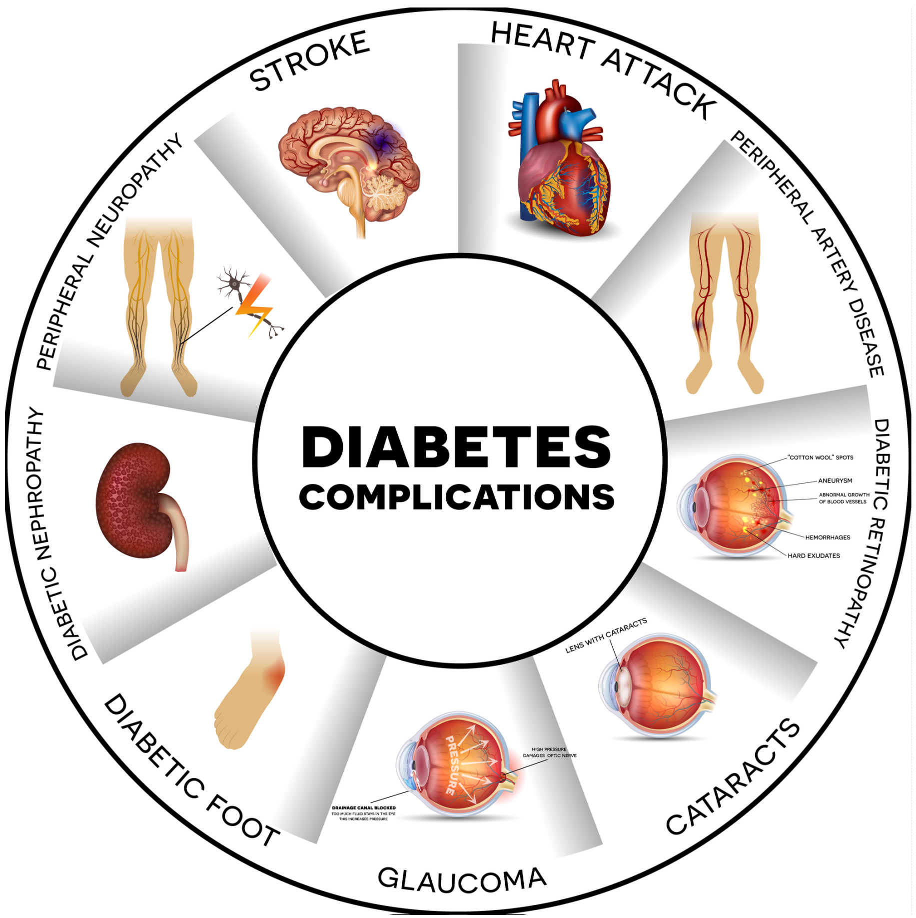
Center for Computational Health
IBM Thomas J. Watson Research Center

The Burden of Type 2 Diabetes

- Type 2 diabetes is chronic disease with a long term metabolic disorder characterized by
  - Either resists the effects of insulin, or cannot produce enough insulin
  - High blood sugar (Hyperglycemia)
- United States (2017)
  - 30.3 million people have diabetes (**9.4%** of the U.S. population)
    - 23.1 million diagnosed
    - 7.2 million undiagnosed
  - 90% to 95% of them are type 2 diabetes
  - Cost hundreds of billions of dollars per year

Centers for Disease Control and Prevention. National diabetes statistics report: estimates of diabetes and its burden in the united states. National Diabetes Statistics Report, 2017.

Type 2 Diabetes Complications



- Blindness
- Skin conditions
- Foot damage
- Nerve damage
- Kidney failure
- Stroke
- Heart attack
- Even death

Early Prediction of Complications from EHRs

- Electronic health records (EHRs) are readily available

- Research questions:
  - *When* will a patient develop complications after the initial T2DM diagnosis?
  - Given the EHR records of two group of patients, which group is more likely to develop complications?

- It is critical for designing personalized treatment plans

Main Challenges

- Data censoring in time-to-event modeling
  - Limited duration of the study period
  - Losing track of patients during the observation period

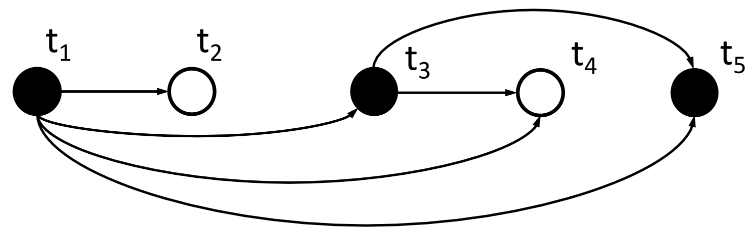
- Capture the correlations between multiple T2DM complications
  - Different complications are manifestations of a common underlying condition — high blood sugar
  - Modeling complications as independent of one another leads to suboptimal models

Survival Analysis

IBM Research

- Cox model: maximizes a partial likelihood objective
  – Does not directly model event probability
  – Depends only on the relative ordering of event times (not actual times)
- Parametric survival models
  – Assume baseline hazard function follows some distribution
  – Not flexible enough to capture the complex event patterns
- Concordance index (CI)

$$\mathrm{CI} = \frac{1}{|\mathcal{V}|} \sum_{\substack{T_i \\ \forall c_i = 1}} \sum_{T_j > T_i} \mathbf{1}_{f(x_j) > f(x_i)}$$

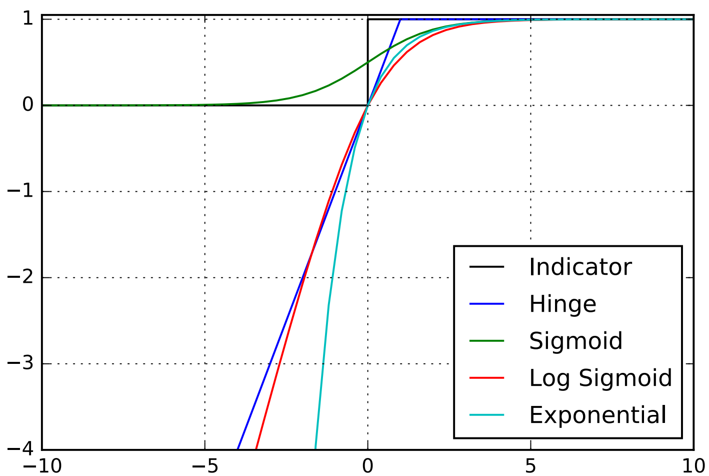


Event order graph G with edge $\mathcal{E}_{i,j}$

RankSvx: Data-driven Survival Model

- Simultaneously achieve two important metrics
  - Accurate prediction of event times, and
  - Good ranking of the relative risks of two patients

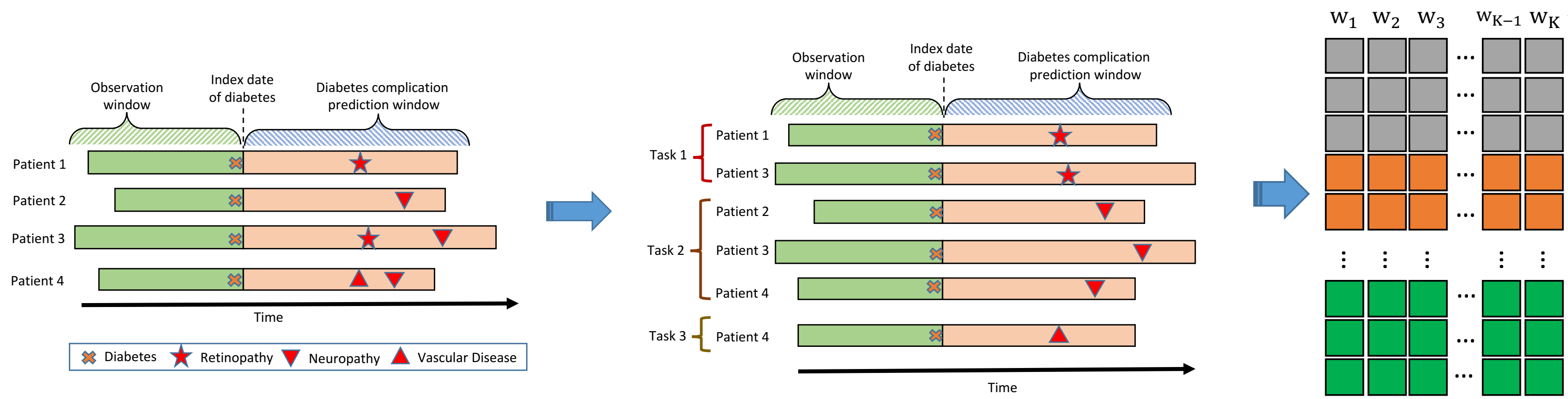
$$\alpha \overbrace{\mathcal{L}_{\text{obs}}\big(t_i, f(\mathbf{x}_i|\Theta)\big)}^{\text{Observed events}} + (1-\alpha)\overbrace{\mathcal{L}_{\text{cen}}\big(\mathcal{E}_{ij}, f(\mathbf{x}_i|\Theta), f(\mathbf{x}_j|\Theta)\big)}^{\text{Censored events}} + g(\Theta)$$



Approximate concordance index

Multi-task RankSvx

- Capture association between different diabetes complications



$$\mathbf{W} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Gamma}, \mathbf{\Omega})$$

$$\sum_{k=1}^{M} \left[\alpha \sum_{\substack{i \\ \forall c_{ki}=1}} \mathcal{L}_{\text{obs}}\left(t_{ki}, \mathbf{w}_k^\top \mathbf{x}_i\right) - (1-\alpha) \sum_{\mathcal{E}_{ij}^k} \log \sigma \left[\mathbf{w}_k^\top (\mathbf{x}_j - \mathbf{x}_i) \right] \right]$$

$$+ \operatorname{tr}\left[\left(\frac{\lambda_1}{2} \mathbf{W}^\top \mathbf{W} + \frac{\lambda_2}{2} \mathbf{\Omega}_0 \right) \mathbf{\Omega}^{-1} \right] + \frac{\lambda_3}{2} \log |\mathbf{\Omega}| + \frac{\eta}{2} \sum_{k=1}^{M} ||\mathbf{w}_k||^2$$

Risk Association Matrix Ω

Experimental Setup and Data

- De-identified patients between the years 2011 and 2015 from a large electronic medical claims database
- T2DM patient cohort
  - I. The frequency ratio between Type 2 diabetes visits to Type 1 diabetes visits is larger than 0.5; AND
  - II-a. The patient have two (2) or more Type 2 diabetes records on different days; OR
  - II-b. The patient received insulin and/or antidiabetic medication
- Prediction variables:
  - Patient demographics: age, gender and weight index.
  - ICD codes: 359 ICD features

Experimental Setup and Data

- T2DM patient cohort from a large electronic medical claims database

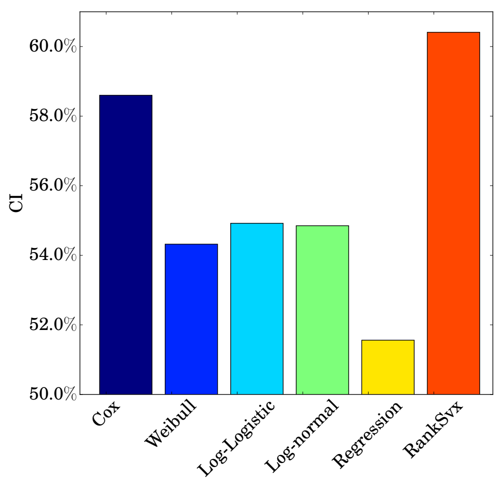
Table 2: List of the five T2DM complications in this study.

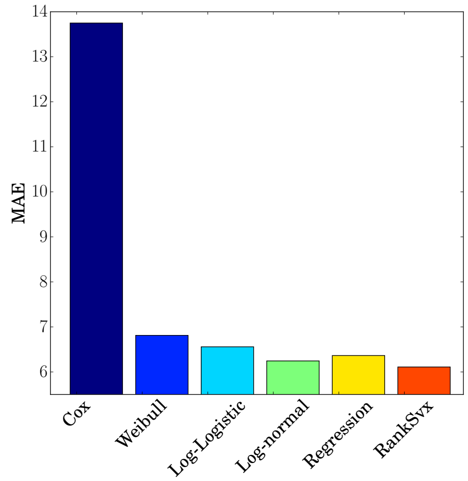
T2DM Complication (Abbreviation)	Description	Example ICD codes
Retinopathy (RET)	eye disease caused by damage to the blood vessels in the tissue at the back of the eye (retina)	25050, 25052, 24950, 24951, 36201-36207, E08311-E0839
Neuropathy (NEU)	nerve damage most often affecting the legs and feet	25060, 25062, 24960, 24961
Nephrology (NEP)	kidney disease characterized by hardening of the glomerulus	25040, 25042, 24940, 24941
Vascular Disease (VAS)	vascular diseases including peripheral vascular disease, cardiovascular disease, and cerebrovascular diseases	25070, 25072, 24970, 24971, E0851, E08621-E08622, E0859
Hyperosmolar (HYPER)	serious condition caused by high blood sugar levels	25020, 25022, 24920, 24921, E0800, E0900, E1100, E1300

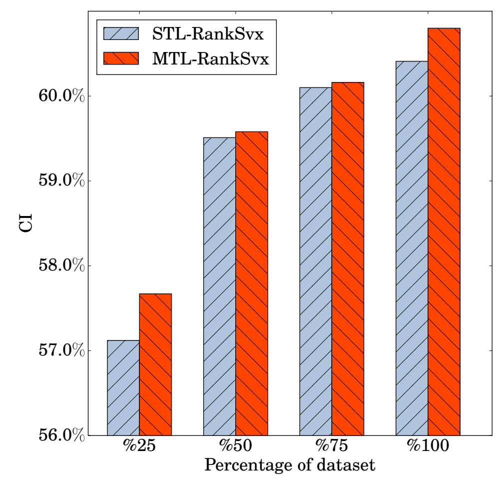
Table 3: Data statistics and patient characteristics.

Complication	# instances	# observations	Female ratio	Average age (SD)	19–44 pct.	45–54 pct.	55–64 pct.
RET	5604	1868	35.03%	52.50 (8.58)	17.02%	33.21%	49.50%
NEU	11874	3958	36.97%	52.53 (8.59)	16.97%	33.01%	49.82%
NEP	4074	1358	37.02%	52.52 (8.91)	17.53%	31.44%	50.86%
VAS	2517	839	39.85%	53.17 (8.31)	15.06%	31.55%	53.12%
HYPER	651	217	36.41%	52.00 (8.90)	19.35%	32.72%	47.93%

Result Comparisons







RankSvx vs traditional survival models and regression model

MTL-RankSvx vs STL-RankSvx

Conclusion

- To the best of our knowledge, this paper presents the first study to investigate the early prediction of T2DM complications from EHRs
- A novel data-driven survival analysis approach for time-to-event modeling
- Developed a multi-task version of the survival model
- Extensive experiments validated the performance of our model

Future Work

- Incorporating more features or new feature representations can potentially improve prediction performance

- Analyze and identify the important associated risk factors by feature selection

- Adapting our models to other chronic diseases and other types of electronic health record data

Thanks!