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 a 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

---

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

- 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)

\[
CI = \frac{1}{|\mathcal{V}|} \sum_{T_i} \sum_{T_j > T_i} 1 \cdot \frac{f(x_j)}{f(x_i)}
\]
Simultaneously achieve two important metrics:

- Accurate prediction of event times, and
- Good ranking of the relative risks of two patients

\[ \alpha \mathcal{L}_{\text{obs}}(t_i, f(x_i|\Theta)) + (1 - \alpha) \mathcal{L}_{\text{cen}}(E_{ij}, f(x_i|\Theta), f(x_j|\Theta)) + g(\Theta) \]
Multi-task RankSvx

- Capture association between different diabetes complications

\[
\sum_{k=1}^{M} \alpha \sum_{i \forall c_{ki} = 1} \mathcal{L}_{\text{obs}}(t_{ki}, w_k^\top x_i) - (1 - \alpha) \sum_{kij} \sigma [w_k^\top (x_j - x_i)] + \text{tr} \left( \frac{\lambda_1}{2} W^\top W + \frac{\lambda_2}{2} \Omega_0 \right) \Omega^{-1} + \frac{\lambda_3}{2} \log |\Omega| + \eta \sum_{k=1}^{M} ||w_k||^2
\]

\[
W \sim \mathcal{M}\mathcal{N}(0, \Gamma, \Omega)
\]

Risk Association Matrix \( \Omega \)
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.

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

Table 3: Data statistics and patient characteristics.

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

RankSvx vs traditional survival models and regression model

MTL-RankSvx vs STL-RankSvx
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!