Use of Clinical Phenotypes and Non-negative Tensor Factorization for Heart Failure Prediction

Yu Deng, BS1; Al’ona Furmanchuk, PhD1; Robert Chen, BS2; Faraz S. Ahmad, MD1; Jimeng Sun, PhD2; Abel Kho, MD1

1Feinberg School of Medicine, Northwestern University, Chicago, IL
2College of Computing, Georgia Institute of Technology, Atlanta, GA

Introduction
Heart failure (HF) is a diverse syndrome associated with multiple risk factors and diseases. Heart failure affects millions of adults in the United States annually, and for some subtypes of patients with HF, 5-year mortality is higher than 50%. In order to intervene earlier to reduce morbidity and mortality with targeted therapies it is important to identify patients at high risk for HF. Electronic health records (EHR) provide extensive information on patients that one can use as features for outcome prediction. Many machine learning (ML) methods are able to deal with large feature sets and complete the prediction task successfully. However, there are few efficient algorithms that also yield results that are easily interpretable. Furthermore, when applied to EHR data, such methods (e.g. logistic regression) will struggle from over-fitting due to low number of events per variable. Recent applications of non-negative tensor factorization (NTF) to EHR. Recent applications of non-negative tensor factorization (NTF) to EHR records1 show potential compromise between sparseness of EHR data and ease of comprehension of the model. Here, we assessed the performance of a novel ML method for detecting patients at risk of HF. In the first step, the NTF algorithm generates phenotypes of patients specific to patient diagnosis and pharmacological profiles prior to HF. Each patient is assigned to its own “fingerprint” label according to his membership to the different phenotypes. The fingerprints were then used as a feature set to predict HF. Ease of comprehension is achieved through transforming the high dimensional space (raw features) to a reduced clinical phenotype space (“fingerprints”).

Methods
We extracted medication (RxNorm codes) and diagnosis (ICD9 codes) from 1/1/2005 to 11/30/2016 on patients with possible HF from the Northwestern Medicine Enterprise Data Warehouse (EDW). We mapped diagnoses codes into a higher level of hierarchy based on PheWAS2 and Anatomical Therapeutical Chemical Classification System3. We further reduced the study population to HF cases and controls only. We identified a HF case if he/she has (i) at least 2 HF diagnoses from outpatient encounters; or (ii) at least 2 medications prescribed for a HF diagnosis; or (iii) at least 1 HF diagnosis from an outpatient encounter and at least 1 medication prescribed for a HF diagnosis. This definition was previously validated by Geisinger Clinic4. For each case, we matched control patients on gender, year of birth and similar encounter information. We analyzed records of each patient within a 2-year time window prior to HF onset date to ensure that each patient had the same length of prediction window. Only cases and controls having a full 2-year period of records and at least 1 visit during these 2 years were kept in the final sample. Once a HF patient was identified, the first appearance of HF related medication or diagnosis was assigned as the onset date. A tensor was constructed with 3 modes including a patient mode, medication mode and diagnosis mode. Each tensor element represents the co-occurrence between medication and diagnosis for a certain patient during the available time window. NTF was applied on the tensor to generate patient phenotypes. The cutoff parameters (alpha: 1.0; gamma: 0.001, 0.08, 0.07) for NTF were set to reduce the noise of data and keep the phenotypes concise. A series of dummy variables were created for each patient based on phenotype membership. Those variables were then used as predictors for logistic regression (LR) and random forest (RF) prediction. The predictive performance of this method was compared to the performance of algorithm based on raw features transformed lower dimension space of principal components (PCA). The 3 main principal components were used as predictors to RF algorithm. The area under receiver operator characteristic curves (AUC) were used to evaluate model discrimination ability (Fig 1).

Results and Discussion
We identified 518 subjects as qualified HF cases in the EHR. Patient onset age ranges from 20 to 89 years old with majority of HF onset age at 45-89 years old. For both RF and LR prediction, “fingerprint” labels outperformed raw features and principal components when the number of clinical phenotypes is greater than 20. The AUC score for raw feature/PCA based method is 0.628±0.145 (green solid and dashed lines in Fig. 1), while corresponding values for “fingerprint” –based methods are 0.733±0.05 (LR: red dots and lines in Fig. 1) and 0.760±0.136 (RF: blue crosses and lines in Fig.1). Based on how NTF handles noise of the data, number of cases and controls at the final
step (orange line in Fig.1) will be smaller than original number of patients selected for the study. Therefore, for small samples, accuracy of the final prediction will be significantly affected by the reduction of overall number of the patients. In our study, the case:control ratio is approximately 1:3. Even after adding weights to controls for imbalance adjustment, analysis will still suffer from reduction of HF case patients during NTF. We used the number of HF cases remained after NTF together with AUC value as criteria for selection of optimal number of phenotypes. It appears that 90-phenotypes is the optimal number (both classification performance and number of patients reach a plateau). We sorted phenotypes importance based on use of features in RF prediction. The 5 phenotypes out of top 12 were selected as the most significant risk phenotypes (table 1) for developing HF event. 66.73% of future HF case patients are in those five phenotypes, while contribution of these phenotypes to controls is only 16.24%.

Figure 1. AUC plot to show performance of models. Error bars are plotted to represent standard deviation of AUCs.

Table 1. Contribution to HF cases and controls for each phenotype

<table>
<thead>
<tr>
<th>Phenotype label and description</th>
<th>Phenotype contribution to cases (case # in a phenotype/# of all cases)</th>
<th>Phenotype contribution to controls, %,(control # in each phenotype/# of all controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Atrial fibrillation/flutter &amp; edema</td>
<td>17.67</td>
<td>5.71</td>
</tr>
<tr>
<td>2. Lipoid disorders &amp; hypertension</td>
<td>16.63</td>
<td>6</td>
</tr>
<tr>
<td>3. Hypertension, ischemic heart disease</td>
<td>12.68</td>
<td>1.9</td>
</tr>
<tr>
<td>4. Type 2 diabetes mellitus</td>
<td>12.27</td>
<td>1.9</td>
</tr>
<tr>
<td>5. Hypertension, lipoid disorder, ischemic heart disease, ill-defined complications of HF</td>
<td>7.48</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Physician evaluation shows clinical meaningfulness of the 5 phenotypes that contributed the most to the risk of subsequent HF. The prominent comorbidities in the five phenotypes in decreasing order of contribution were atrial fibrillation/flutter, lipoid disorders, hypertension, and type 2 diabetes mellitus, with the fifth being an amalgam of lipoid metabolism, hypertension and ischemic heart disease. Clinically, this order may suggest that some acute conditions (e.g. atrial fibrillation) may contribute more to subsequent heart failure risk in the short term (2-year window) than chronic conditions which may take years to lead to subsequent cardiovascular injury.

Conclusions
Our preliminary results show better performance of NTF membership linked classification over other popular dimension reduction method and suggested a plausible hierarchy of HF risk based on associated comorbidities. NTF also shows potential in capturing novel HF prediction phenotypes (i.e. features) with distinct clinical characteristics. Such phenotypes are likely to improve HF prediction and help discover mechanisms of HF pathogenesis.

Acknowledgements
This research is supported by the National Science Foundation, under Grant IIS-1417819.

References
3. Organization WH. ATC/DDD Index 2017