A new representation of disease conditions and treatment pathways accurately predicts mortality and chronic diseases

Che Ngufor, PhD[1], Pedro Caraballo, M.D.[1] , Thomas J. O Byrne[1], David Chen, Ph.D.[1], Nilay D. Shah, Ph.D[1], Michael Steinbach, Ph.D[2], Gyorgy Simon, Ph.D[2]
[1]Mayo Clinic, Rochester; [2]University of Minnesota, Minneapolis

Introduction

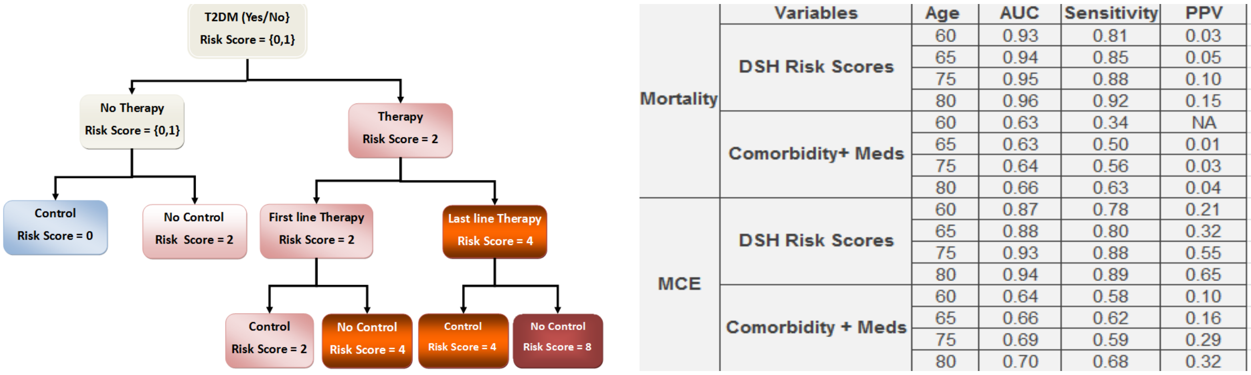
The number of individuals in the U.S. suffering from multiple chronic conditions (MCC) has increased substantially and continues to rise.[1] Patients with MCCs require significantly more health care resources, incur high cost of care, and face greater mortality, functional decline, and worse quality of life.[2] At the basic level, the distinct MCCs act as competing risks, and co-occur either by chance, through causal relationships, or by common underlying risk factors. Consequently, interventions that address one comorbidity without reducing the severity of the others will offer limited benefit in terms of quality of life and mortality. Because slowing the progression of one comorbidity may hinder the progression of others, understanding the interactions among MCCs and treatment pathways offers the opportunity to strengthen intervention strategies. However, identifying these relationships or ascertaining the likelihood of adverse health outcomes have been hindered by the marked complexity and heterogeneity of patients with MCC.

Predictive analytics based on longitudinal data can provide crucial information for making better clinical decision about MCC patients. However, it also raises an important question about the appropriate representation of MCCs for predicting health outcomes. Numerous comorbidity measures have been developed for research using electronic health records (EHR) and administrative claims data such as the Charlson and Elixhauser Comorbidity Index (CI) and the Agency for Healthcare Research and Quality's (AHRQ) Clinical Classification Software(CCS).[3] Current Comparative Effectiveness Research and Evidence Based Practices (EBP) tend to focus on inefficient comorbidity representations such as simple summaries (e.g. counts, sum, freqeuncies, any or last oberved) of individual conditions or aggregate measures of the CI and CCS within a specific observation window. However, these representations do not reflect an individual's comorbid disease history and severity, and do not account for interactions with other risk factors or the different treatment pathways that a patient may take. This thus leaves a large gap in our knowledge about how to optimally manage individuals with complex MCCs. Different subpopulations of patients can exhibit different relationships between patient characteristics and outcomes. Thus information about the interactions and subgroups that maximally capture heterogeneity in MCCs and treatment pathways can facilitate diagnosis, enhance preventive strategies, improve quality of life, and help create smart EBP guidelines.

To address this critical knowledge gap, in this study, we introduce a novel representation of patient data called Disease Severity Hierarchy (DSH) that explores specific diseases and their known treatment pathways in a nested fashion to create subpopulations in a clinically meaningful way. As the DSH tree is traversed from the root towards the leaves, we encounter subpopulations that share increasing richer amounts of clinical details such as similar disease severity, illness trajectories, and time to event that are discriminative, and suitable for learning risk stratification models.

Methods

Study population: We used data for 15,391 adults, age $45 - 85$ years, included in the Rochester Epidemiology Project (REP) database who received primary care at Mayo Clinic in 2004-2015. Subjects entered the study at their age in Jan 1 2004 and were still alive on Dec 31 2010. Patients were then followed until Dec 31 2015. Primary outcomes considered include all-cause mortality and major cardiovascular event (MCE) in 2011 - 2015. ***DSH construction:*** We focus on four common diseases: type 2 diabetes mellitus (T2DM), hypertension, hyperlipidemia and obesity and constructed their DSH for comparison with traditional representations. We included all primary care patients with or without an indication of these conditions at any time during the study period. We developed models based on features taken in 2004-2010 to predict mortality and MCE in 2011 - 2015. Specifically, for each patient, we collected demographic data, time-stamped diagnosis of the four disease conditions, medications, and laboratory results measured in 2004-2010 and use these to construct the DSH trees for each disease. For most disease conditions, associated laboratory tests exist, and medications indicated for these conditions are also known. However, most existing representation techniques ignore these relationships, thus eroding the interpretability and clinical applicability of the results. DSH is designed to account for known relationships in EHR by encoding disease severity. At each time point (hospital visit), DSH considered the full clinical context of a disease at several nested levels of details. Starting at the population level (root) we determined if a patient had a disease condition or not (e.g T2DM). If the patient is diseased, we looked for any information regarding whether the condition was treated (e.g. prescription of Metfomin); next we consider the



	Variables	Age	AUC	Sensitivity	PPV
Mortality	DSH Risk Scores	60	0.93	0.81	0.03
		65	0.94	0.85	0.05
		75	0.95	0.88	0.10
		80	0.96	0.92	0.15
	Comorbidity+ Meds	60	0.63	0.34	NA
		65	0.63	0.50	0.01
		75	0.64	0.56	0.03
		80	0.66	0.63	0.04
MCE	DSH Risk Scores	60	0.87	0.78	0.21
		65	0.88	0.80	0.32
		75	0.93	0.88	0.55
		80	0.94	0.89	0.65
	Comorbidity + Meds	60	0.64	0.58	0.10
		65	0.66	0.62	0.16
		75	0.69	0.59	0.29
		80	0.70	0.68	0.32

(a) DSH tree for T2DM

(b) Performance of DSH risk scores compared to traditional method

Figure 1: DSH tree and Performace Results

aggressiveness of the therapy (first-line / last-line drug); and finally whether the patient was under control or not. A patient is under control, if the lab result or vital sign associated with the condition was within its predefined normal range. This nested or hierarchical representation of information helped us to assess the severity of the disease. Figure 1 (a) shows the DSH for T2DM, where the leaves indicate if the patient is in control or not based on hemoglobin A1c. Notice that the structure of DSH is a binary tree, where each node has at most two children. Further, any branch on a right child leads to intensification of the disease severity. We used this property to quantitively represent the information embedded in DSH. Specifically, we assigned "risk" scores to the nodes, where the score of the right child doubles that of its parent. The risk score of the root node is either 1 (diseased) or 0 (disease free). ***Training and validation:*** We developed age specific survival models,[4] where instead of the traditional time-on-study, we used the subject's age as the time scale to predict the risk of death and MCE at age 60, 65, 75, and 80 years based on the DSH, while adjusting for sex and race. The use of age time scale provides an expressive and flexible way to control the effect of age especially for older adults. It also provides a relatively meaningful basis on which to examine how risk varies over time.[4] We trained the random survival forest (RSF) model through a 5-fold cross-validation procedure.

Results

The median age of the study population in 2004 was 47.6 years, with 57.7% female and 88.5% white. 3.3% of the patients died during the follow up period (2011-2015), while 25.7% had MCE. Figure 1 (b) presents the AUC, Sensitivity, and PPV of the RSF model trained using DSH risk scores and standard representation of the four comorbid conditions. Specifically, we considered three formats for the standard representation: a diagnosed condition persisted throughout 2004-2010, most frequent value (yes/no) of the condition, and the last observed value of the condition in 2004-2010. We also included indicators of medication use in 2004-2010. The figure clearly shows that the performance of RSF based on DSH risk scores significantly outperformed the traditional representation of comorbidities.

Conclusion

The proposed DSH risk scores effectively and accurately predict the age at which a patient maybe at risk of dying or developing MCE significantly better than traditional representation of disease conditions. DSH utilizes known relationships among various entities in EHR data to capture disease severity in a natural way and has the additional benefit of being expressive and interpretable. This novel patient representation can help support critical decision making, development of smart EBP guidelines, and enhance healthcare care and disease management by helping to identify and reduce disease burden among high-risk patients.

Acknowledgement: This study was supported by NSF grants IIS-1602394 and IIS-1602198

References

1. Ward BW. Multiple chronic conditions among US adults: a 2012 update. Preventing chronic disease. 2014;11.
2. Newman AB, Boudreau RM, Naydeck BL, Fried LF, Harris TB. A physiologic index of comorbidity: relationship to mortality and disability. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences. 2008 Jun 1;63(6):603-9.
3. Chu, Yu-Tseng, Yee-Yung Ng, and Shiao-Chi Wu. "Comparison of different comorbidity measures for use with administrative data in predicting short-and long-term mortality." BMC health services research 10.1 (2010): 140.
4. Griffin, Beth Ann, et al. "Use of alternative time scales in Cox proportional hazard models: implications for time-varying environmental exposures." Statistics in medicine 31.27 (2012): 3320-3327.