Demonstrating the Advantages of Applying Data Mining Techniques on Time-Dependent Electronic Medical Records

Uri Kartoun, PhD[1,2], Vishesh Kumar, MD[1,2], Su-Chun Cheng, ScD[3], Sheng Yu, PhD[3], Katherine Liao, MD, MPH[2,4], Elizabeth Karlson, MD[2,4], Ashwin Ananthakrishnan, MBBS, MPH[2,5], Zongqi Xia, MD, PhD[2,5], Vivian Gainer, MS[6], Andrew Cagan, BSc[6], Guergana Savova, PhD[2,7], Pei Chen, MS[2,7], Shawn Murphy, MD, PhD[6], Susanne Churchill, PhD[5], Isaac Kohane, MD, PhD[2,5], Peter Szolovits, PhD[8], Tianxi Cai, ScD[3], Stanley Y. Shaw, MD, PhD[1,2]

1. Center for Systems Biology, Massachusetts General Hospital (MGH), Boston, MA; 2. Harvard Medical School, Boston, MA; 3. Department of Biostatistics, Harvard School of Public Health, Boston, MA; 4. Division of Rheumatology, Brigham and Women's Hospital (BWH), Boston, MA; 5. i2b2 National Center for Biomedical Computing, BWH, Boston, MA; 6. Research Computing, MGH, Boston, MA; 7. Clinical Natural Language Processing Program, Boston Children's Hospital, Boston, MA; 8. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Summary: We demonstrate several advantages of applying data mining techniques on time-dependent Electronic Medical Records (EMR), specifically: 1) combining structured and unstructured variables improves the accuracy of a type-2 diabetes (T2D) classification algorithm, 2) conducting a quantitative survey of multiple comorbidities is important in T2D especially cardiovascular complications with hazard ratios, 3) analyzing time dependent variables can clarify time dependent contributions to variety of comorbidities, and specifically of the "obesity paradox", and 4) demonstrating that an unbiased examination of physician treatment patterns reveals changes over time consistent with clinical trials.

Background: Cohorts assembled from EMR present a potentially powerful resource to study T2D and cardiovascular complications at population scale. Recent reports have demonstrated the utility of EMR analysis to discover genotype-phenotype correlations, sub-categories of disease, and adverse drug events.

Methods: We developed a classification algorithm to identify T2D patients based on characteristics including clinical notes, diagnosis and procedure codes, medications, and laboratory tests. We analyzed an EMR database at MGH and BWH considering patients who received care between 1990 - 2013. We applied logistic regression with the adaptive LASSO using different combinations of variables such as structured variables only, unstructured variables only, and combination of all variables. To determine the level of association between clinical and demographic variables with mortality we developed baseline and lagged-time varying Cox regression models that included an adjustment to ethnicity and time varying covariates. To assess how therapeutic choices change over time, we calculated sparse covariance matrices for heart failure related concepts extracted from clinical notes.

Results: Our classification algorithm identified 65,099 T2D patients with a specificity of 97% and PPV of 96% based on "gold standard" physician chart review. 56,691 patients (87.1%) had two and 38,449 patients (59.1%) had four or more chronic conditions, demonstrating the complexity of the cohort. Cox regression models indicated statistically significant HRs > 1 for CHF, CAD, and CVD, and HRs < 1 for PCI and CABG. Increasing BMI was associated with lower mortality as compared to the reference BMI (< 25 kg/m^2). Further stratifying the results into 1, 3 and 5 years analysis, this "obesity paradox" is strikingly obvious at short-term follow-up of 1 year, suggesting that patients with low BMI were suffering from chronic medical conditions (*e.g.*, malignancy or inflammatory conditions) increasing their 1 year mortality. However, at 3 and 5 years follow-up, we do see increase in mortality with increasing BMI levels likely related to increase in the burden of cardiovascular events.

Discussion: We implemented classification, prediction, and natural language processing techniques in multiple scenarios to create and to analyze a highly complex and large cohort. This cohort recapitulates many findings from traditionally ascertained cohorts while enabling additional analyses (e.g., utilizing physician notes or richer temporal data), illustrating its utility for a variety of discovery efforts.

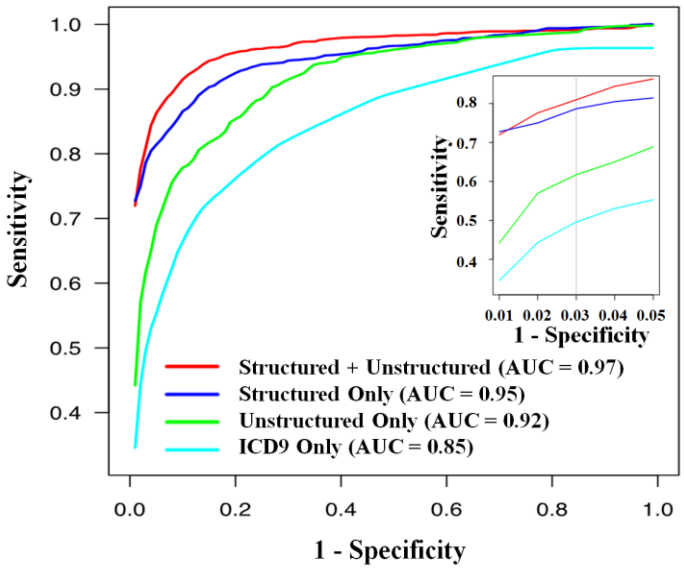

Figure 1. Receiver operating characteristic curves and area under the curve (AUC) values of the algorithms using varying combinations of structured and unstructured data

Table 1. 65,099-patient cohort characteristics considering diagnosis codes

(a) Most common comorbidities and combinations of comorbidities (for no comorbidities: n = 1,117):

		Exactly 1 comorbidity (n = 5,422)	Exactly 2 comorbidities (n = 7,152)	Exactly 3 comorbidities (n = 11,090)	Exactly 4 comorbidities (n = 10,643)	Exactly 5 comorbidities (n = 8,664)
Most Common	Diabetes	78.0%	85.6%	92.0%	94.1%	94.8%
	Hypertension	6.0%	48.8%	81.2%	88.2%	91.2%
	Hyperlipidemia	3.4%	27.2%	65.6%	73.9%	76.5%
Most Common Combination		Diabetes (78.0%)	Diabetes & Hypertension (2,868 patients)	Diabetes & Hypertension & Hyperlipidemia (5,795 patients)	Diabetes & Hypertension & Hyperlipidemia & Arthritis (2,421 patients)	Diabetes & Hypertension & Hyperlipidemia & Arthritis & Depression (571 patients)

(b) Most common comorbidities considering a minimum number of comorbidities threshold per patient:

	≥ 0 Comorbidities (n = 63,230)	≥ 1 Comorbidities (n = 62,113)	≥ 2 Comorbidities (n = 56,691)	≥ 3 Comorbidities (n = 49,539)	≥ 4 Comorbidities (n = 38,449)	≥ 5 Comorbidities (n = 27,806)
Diabetes	90.6%	92.2%	93.6%	94.8%	95.6%	96.1%
Hypertension	76.7%	78.1%	85.0%	90.2%	92.8%	94.6%
Hyperlipidemia	64.0%	65.2%	71.1%	77.4%	80.8%	83.5%

(c) Prevelance of additional common combinations of comorbidities:

Combination	Patients associated with the combination (%)
Diabetes & Hypertension	69.7%
Diabetes & Hyperlipidemia	58.5%
Hypertension & Hyperlipidemia	55.7%
Ischemic Heart Disease & Hypertension	18.6%
Ischemic Heart Disease & Hyperlipidemia	17.1%