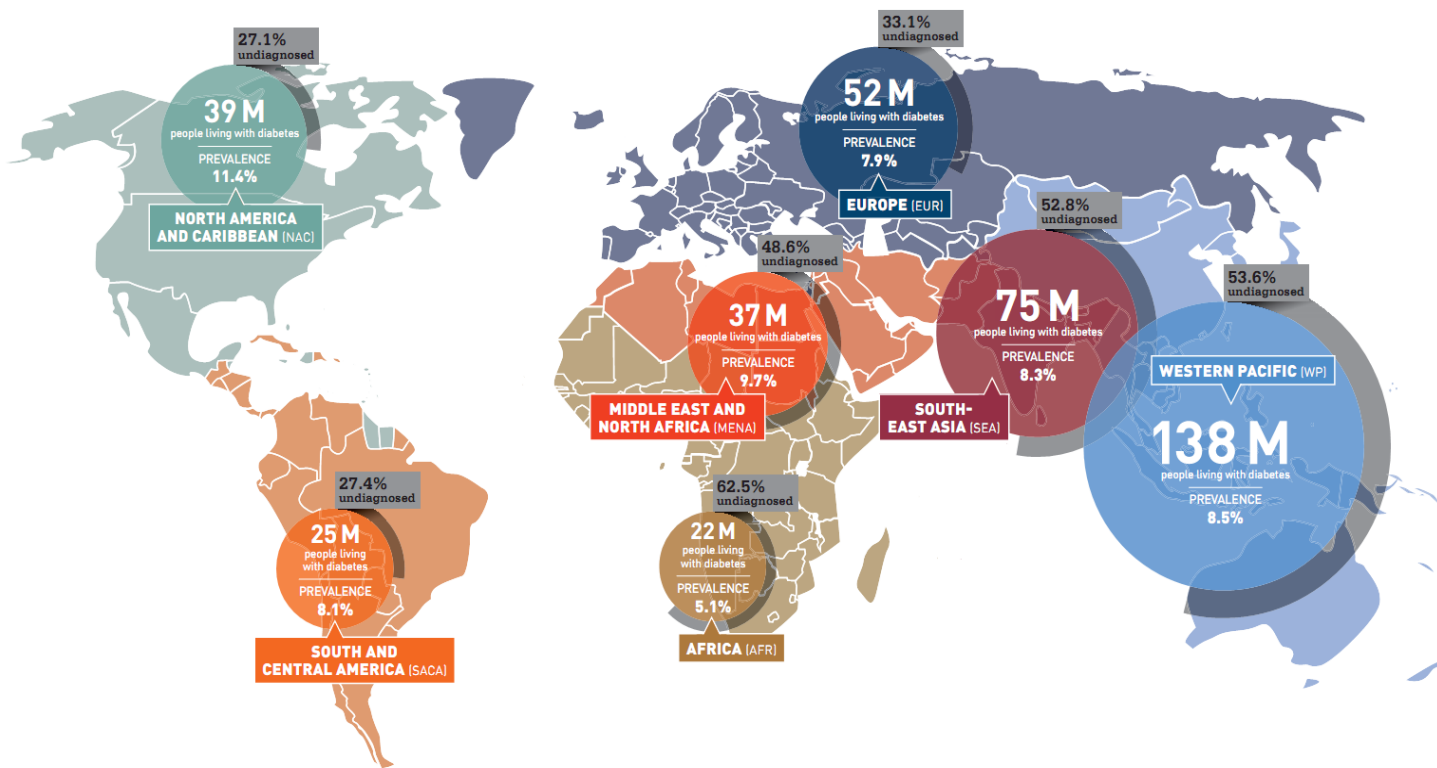


Predicting Chronic Comorbid Conditions of Type 2 Diabetes in Newly-Diagnosed Diabetic Patients

Narges Razavian^{PhD, NYU}, Saul Blecker^{MD, NYU}, Ann Marie Schmidt^{MD, NYU}, Aaron Smith-McLallen^{PhD, IBC}, Somesh Nigam^{PhD, IBC}, David Sontag^{PhD, NYU}

Background¹



Comorbid Complications²:

- Cardiovascular:** Myocardial infarction, Ventricular fibrillation, cardiac arrest, Ischemic heart disease, Heart failure
- Cerebrovascular:** Subarachnoid, Intracerebral, Nontraumatic extradural, Subdural hemorrhage, Occlusion of cerebral arteries, stroke, Transient cerebral ischemia, Occlusion and stenosis of precerebral arteries, Other and ill-defined cerebrovascular disease
- Renal:** Hemodialysis, Peritoneal dialysis, Acute or chronic kidney failure, kidney transplant
- Eye:** Background retinopathy, Macular edema, Proliferative retinopathy, retinal detachment or Vitreous hemorrhage, Operations of vitreous

Methods

Our prediction task:

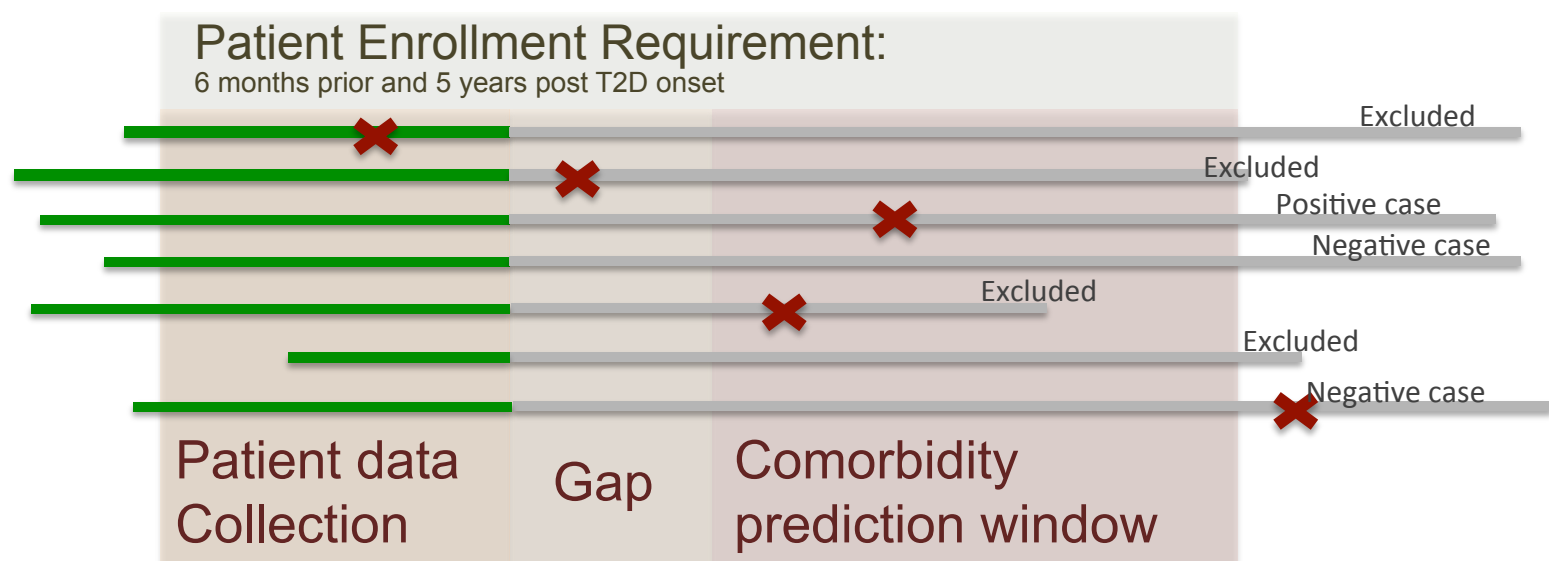
- Given Patient Variables up to T2D, can we predict the risk of 4 comorbidities within 1 to 5 years after the onset?
- What variables are predictive?
- What variables differ between genders?

Our cohort :

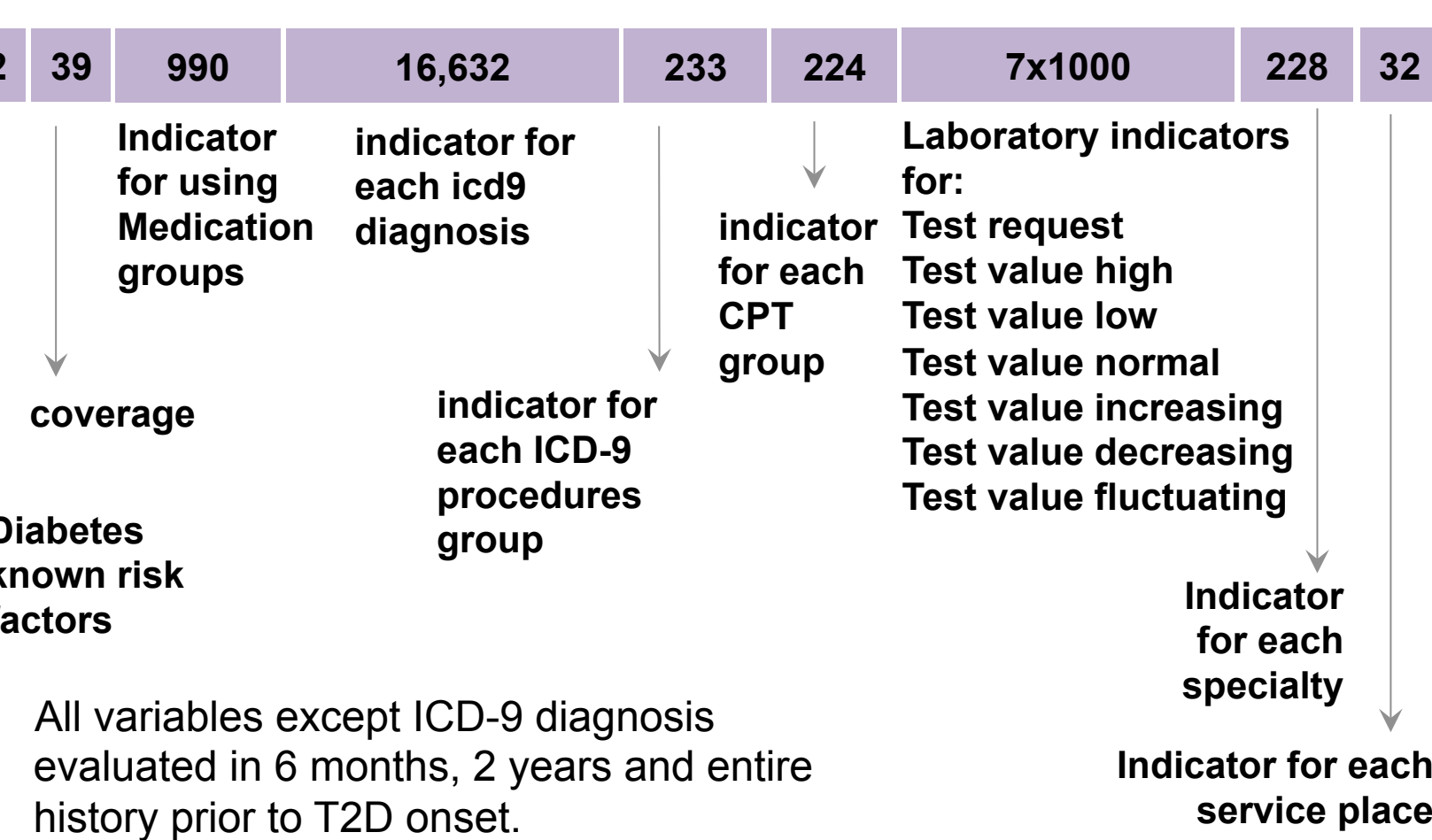
- 4.1 million insurance beneficiaries of age ≥ 18 years between 2005 and 2013. We only included patients newly-diagnosed with T2D,
- Continuous enrollment for 6 months pre- and 5 years post-onset,
- Without the comorbid condition diagnosis before T2D.

Prediction Framework

Patients are only included if they are continuously enrolled throughout the patient enrollment requirement window, and do not have comorbidity onset before the comorbidity prediction window. We use 67% of the cohort to train our system, and evaluate the quality on 33% remaining patients data not used during training.



Patient Variables



- All variables except ICD-9 diagnosis evaluated in 6 months, 2 years and entire history prior to T2D onset.

Machine Learning Method

Prediction model: L1-regularized Logistic Regression

- Total Variable Size: 42,000 with diagnosis variables, and 27,000 without.
- Total Patients Per Complications: Between 25,000 to 33,000
- Total training time < 5 minutes
- Evaluation for new patient: < 0.1 seconds
- Total Memory requirement for training < 64GB RAM
- Total Memory requirement for evaluating new patient with existing model < 1GB
- Python runtime environment.

Prediction models can be developed elsewhere and plugged into administrative or EHR infrastructure.

For software package including patient visualization visit:

<https://github.com/nyuvis/patient-viz>

Software for patient variable extraction on public CMS³ data visit:

https://github.com/nyuvis/patient-viz/tree/master/feature_extraction

For additional information visit <http://clinicalml.org/> or contact

razavian@cs.nyu.edu or dsontag@cs.nyu.edu

Results

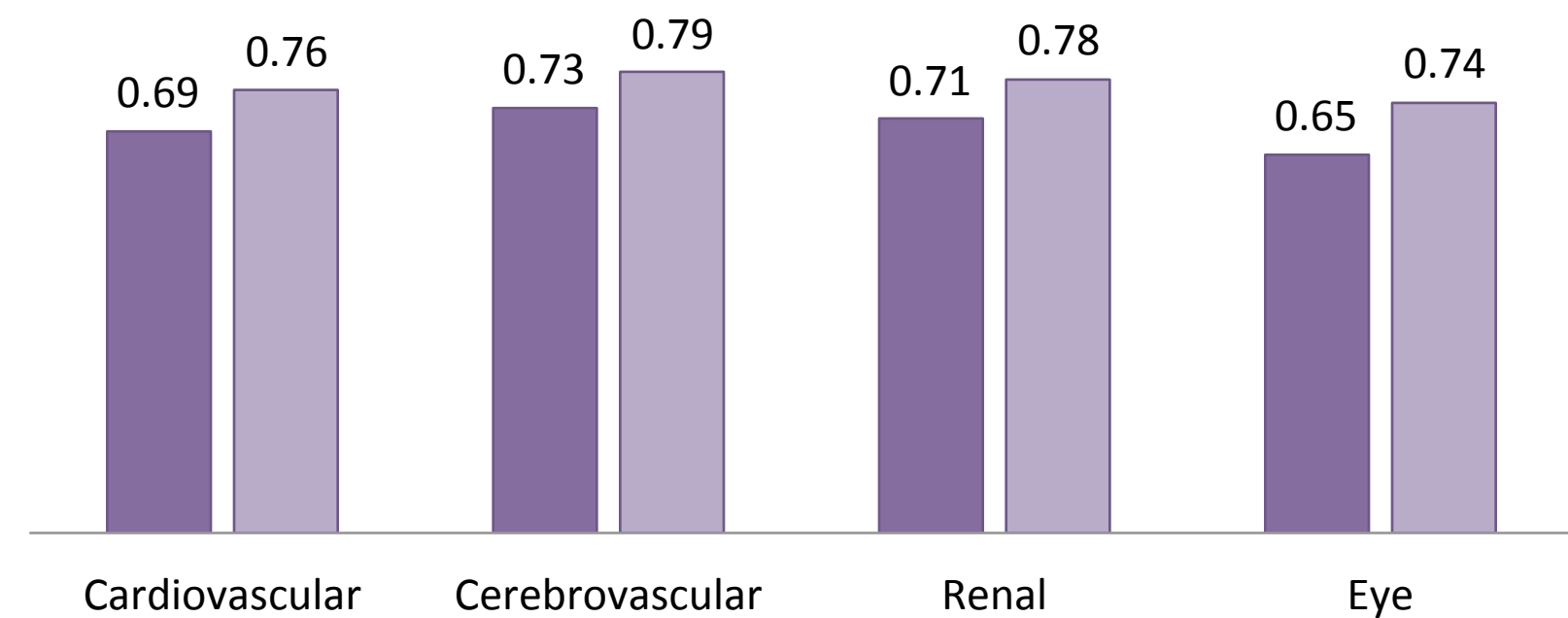
T2D Complications Cohort in Independence Blue Cross Cohort:

Comorbidity	Total included in training	Total Positive cases during training
Cardiovascular	25,421	4,896
Cerebrovascular	32,962	2,678
Renal	32,496	3,658
Eye	33,775	2,699

Prediction Quality and Risk Factors

Prediction Quality for Comorbid Complications, for a newly diagnosed T2D Patient

■ AUC with every variable except ICD9s ■ AUC with ICD9s included as variables



Cardiovascular Events Predictive Variables (Total 317 variables)* Lists are sorted by p-values

Top Lab Factors	Top Diagnosis Factors
Platelets-low in the past 2 years (loinc-777-3) (OR 1.87 [1.46 2.38])	History of cardiovascular disease (ICD9 429.2) (OR 7.75[6.57 9.15])
Albumin-high in the past 2 years (loinc-1751-7) (OR 1.13 [1.06 1.21])	Atrial fibrillation (ICD9 427.31) (OR 10.87[9.35 12.65])
Cholesterol.total/Cholesterol.in HDL -high in the past 2 years (loinc-9830-1) (OR 1.13 [1.06 1.21])	Atherosclerosis of native arteries of the extremities (ICD9 440.20) (OR 13.46[10.30 17.61])
Alanine aminotransferase -high in the past 6 months (loinc-1742-6) (OR 0.77 [0.67 0.88])	Abdominal aneurysm without mention of rupture (ICD9 441.4) (OR 9.79 [7.15 13.40])
Erythrocyte mean corpuscular volume -low in the past 6 months (loinc-787-2) (OR 0.73 [0.59 0.91])	Sinoatrial node dysfunction (ICD9 427.81) (OR 4.83 [3.97 5.88])

Cerebrovascular Events Predictive Variables (Total 216 variables)

Top Lab Factors	Top Diagnosis Factors
Alanine aminotransferase -high in the past 6 months (loinc-1742-6) (OR 0.58 [0.47 0.71])	Epistaxis (ICD9 784.7) (OR 2.34 [1.98 2.77])
Creatinine -high in the past 24 months (loinc-2160-0) (OR 2.34 [2.10 2.62])	Peripheral vascular disease (ICD9 443.9) (OR 4.06 [3.73 4.42])
Erythrocyte mean corpuscular volume -high in the entire past history (loinc-787-2)(OR 2.05[1.60 2.63])	Obstructive chronic bronchitis with (acute) exacerbation (ICD9 491.21) (OR 3.19 [2.79 3.64])
Coagulation surface induced -request for test (loinc-3173-2) (OR 1.99 [1.58 2.51])	Hypertension (ICD9 401.9) (OR 2.66 [2.46 2.86])
Erythrocyte distribution width -high (loinc-788-0) (OR 1.46 [1.30 1.62])	Convulsions (ICD9 780.39) (OR 3.46 [2.94 4.09])
Lymphocytes/100 leukocytes -decreasing (loinc-736-9) (OR 1.30 [1.13 1.50])	Syncope and collapse (ICD9 780.2) (OR 3.51 [3.21 3.84])

Renal Events Predictive Variables (Total 272 variables)

Top Lab Factors	Top Diagnosis Factors
Creatinine -high in the entire past history (loinc-2160-0) (OR 5.83 [5.29 6.42])	Congestive heart failure (ICD9 428.0) (OR 4.35 [4.03 4.70])
Creatinine -high in the past 6 months (loinc-2160-0) (OR 6.53 [5.78 7.38])	Hyper-potassemia (ICD9 276.7) (OR 5.98 [5.13 6.97])
Urea nitrogen -high in the entire past history (loinc-3094-0) (OR 3.56 [3.22 3.93])	Proteinuria (ICD9 791.0) (OR 4.45 [3.89 5.10])
Glomerular filtration rate/1.73 sq Predicted -low in the past 2 years (loinc-33914-3) (OR 2.98 [2.66 3.34])	Anemia of other chronic disease (ICD9 285.29) (OR 5.43 [4.60 6.42])
-Hematocrit -low in the entire past history (loinc-4544-3) (OR 1.99 [1.79 2.21])	Gout (ICD9 274.9) (OR 2.41 [2.18 2.67])
Albumin/Creatinine -high in the entire past history (loinc-9318-7) (OR 2.86 [2.41 3.39])	Benign renovascular hypertension (ICD9 405.11) (OR 9.68 [7.17 13.06])

Eye Complication Predictive Variables (Total 226 variables)

Top Lab variables	Top Diagnosis variables
Cholesterol -high in the past 2 years (loinc-2093-3) (OR 0.78 [0.73 0.84])	Nonexudative senile macular degeneration (ICD9 362.51) (OR 4.01 [3.51 4.58])
Urea nitrogen -high in the past entire history (loinc-3094-0) (OR 1.43 [1.26 1.61])	Hypertensive retinopathy (ICD9 362.11) (OR 8.00 [6.56 9.75])
Chloride -increasing in the past 6 months (loinc-2075-0) (OR 0.77 [0.69 0.86])	Long-term use of insulin (ICD9 V5867) (OR 3.10 [2.70 3.56])
Thyrotropin -high in the past 2 years (loinc-3016-3) (OR 0.60 [0.46 0.76])	Primary open angle glaucoma (ICD9 365.11) (OR 2.30 [2.06 2.58])
Bilirubin.non-glucuronidated -decreasing in the entire history (loinc-1971-1) (OR 0.57 [0.43 0.77])	Visual loss (ICD9 369.9) (OR 15.19 [9.93 23.22])
Eosinophils -decreasing in the past entire history (loinc-711-2) (OR 0.79 [0.67 0.92])	Ingrowing nail (ICD9 703.0) (OR 1.57 [1.40 1.76])

Conclusions

- Type 2 Diabetes and its comorbidities increasingly strain healthcare sectors and societies worldwide, and population-level early intervention to avoid severe complications is necessary.
- Using longitudinal patient data, data driven approaches enable automatic risk factor identification and risk assessment, without the need for unaffordable large scale screening tests.
- Using our method we can predict risk of each comorbid complication for newly diagnosed T2D patients, and discover potentially novel risk factors, significantly differing in different subpopulations.
- Next step includes intervention allocation based on the risk prediction.

References

1- Diabetes Atlas 6th Edition, International Diabetes Federation, 2014
 2- Newton KM, Wagner EH, Ramsey SD, McCulloch D, Evans R, Sandhu N, Davis C. *The use of automated data to identify complications and comorbidities of diabetes: a validation study.* J Clin Epidemiol 1999;52:199-207
 3- CMS Medicare claims synthetic public use files 2008-2010. <http://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/>