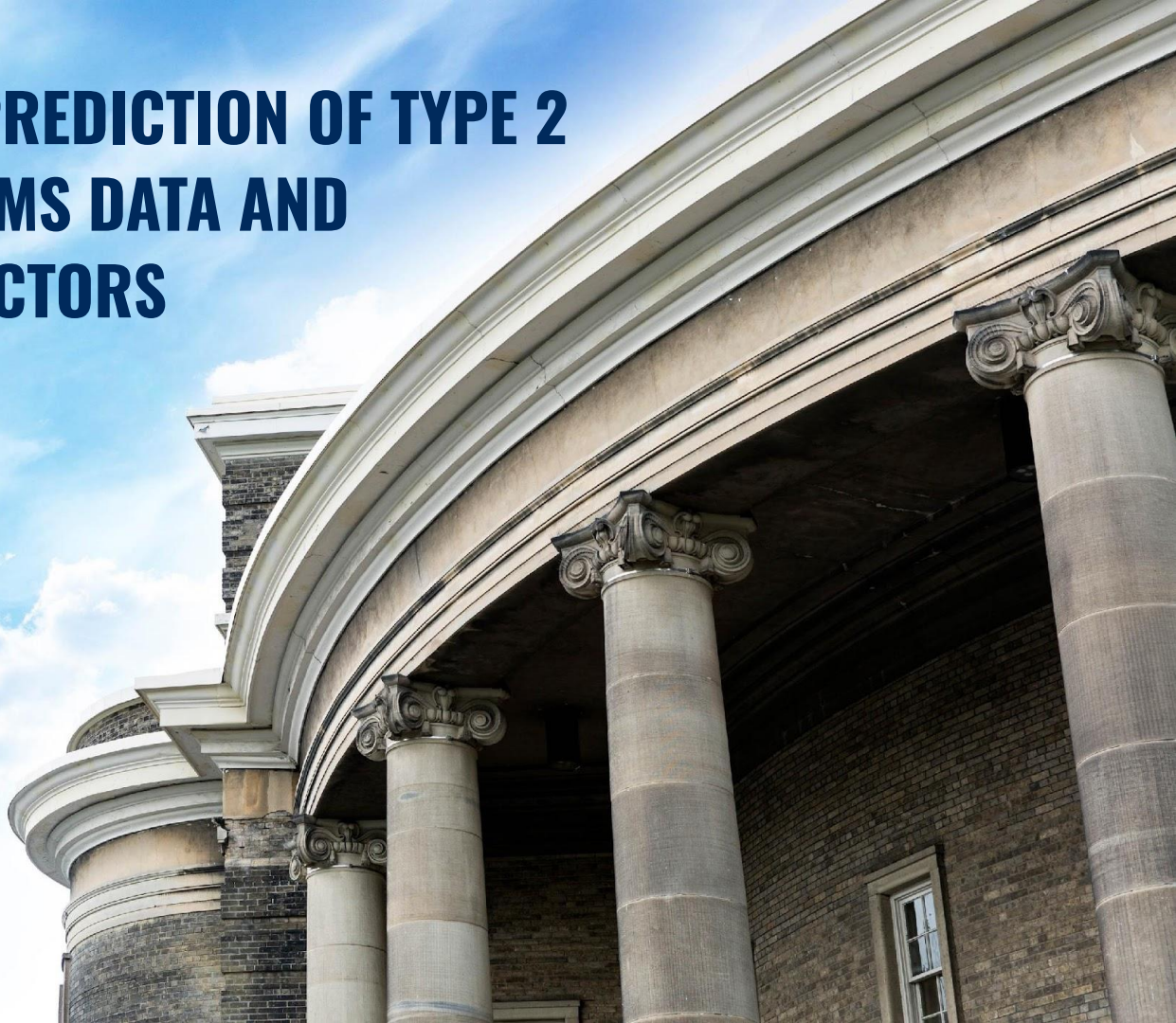
POPULATION-LEVEL PREDICTION OF TYPE 2 DIABETES FROM CLAIMS DATA AND ANALYSIS OF RISK FACTORS

KORBINIAN KOCH

OCTOBER 1ST 2021



UNIVERSITY OF TORONTO

PAPER

Big Data
Volume 3 Number 4, 2015
Mary Ann Liebert, Inc.
DOI: 10.1089/big.2015.0020

ORIGINAL ARTICLE

Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors

*and Rahul

Narges Razavian,[1] Saul Blecker,[2] Ann Marie Schmidt,[3] Aaron Smith-McLallen,[4] Somesh Nigam,[4] and David Sontag[1,*]

Abstract

We present a new approach to population health, in which data-driven predictive models are learned for outcomes such as type 2 diabetes. Our approach enables risk assessment from readily available electronic claims data on large populations, without additional screening cost. Proposed model uncovers early and late-stage risk factors. Using administrative claims, pharmacy records, healthcare utilization, and laboratory results of 4.1 million individuals between 2005 and 2009, an initial set of 42,000 variables were derived that together describe the full health status and history of every individual. Machine learning was then used to methodically enhance predictive variable set and fit models predicting onset of type 2 diabetes in 2009–2011, 2010–2012, and 2011–2013. We compared the enhanced model with a parsimonious model consisting of known diabetes risk factors in a real-world environment, where missing values are common and prevalent. Furthermore, we analyzed novel and known risk factors emerging from the model at different age groups at different stages before the onset. Parsimonious model using 21 classic diabetes risk factors resulted in area under ROC curve (AUC) of 0.75 for diabetes prediction within a 2-year window following the baseline. The enhanced model increased the AUC to 0.80, with about 900 variables selected as predictive ($p < 0.0001$ for differences between AUCs). Similar improvements were observed for models predicting diabetes onset 1–3 years and 2–4 years after baseline. The enhanced model improved positive predictive value by at least 50% and identified novel surrogate risk factors for type 2 diabetes, such as chronic liver disease (odds ratio [OR] 3.71), high alanine aminotransferase (OR 2.26), esophageal reflux (OR 1.85), and history of acute bronchitis (OR 1.45). Liver risk factors emerge later in the process of diabetes development compared with obesity-related factors such as hypertension and high hemoglobin A1c. In conclusion, population-level risk prediction for type 2 diabetes using readily available administrative data is feasible and has better prediction performance than classical diabetes risk prediction algorithms on very large populations with missing data. The new model enables intervention allocation at national scale quickly and accurately and recovers potentially novel risk factors at different stages before the disease onset.

Key words: big data analytics; data mining; machine learning; predictive analytics; risk assessment; disease prediction; longitudinal study

Introduction

The recent availability of the electronic health record and claims datasets offers an unprecedented opportu-

readmission models,[4,5] disease onset prediction,[6–13] and prediction of healthcare utilization and cost.[14]

Type 2 diabetes is a global public health challenge.

UNIVERSITY OF TORONTO

DISCLAIMER

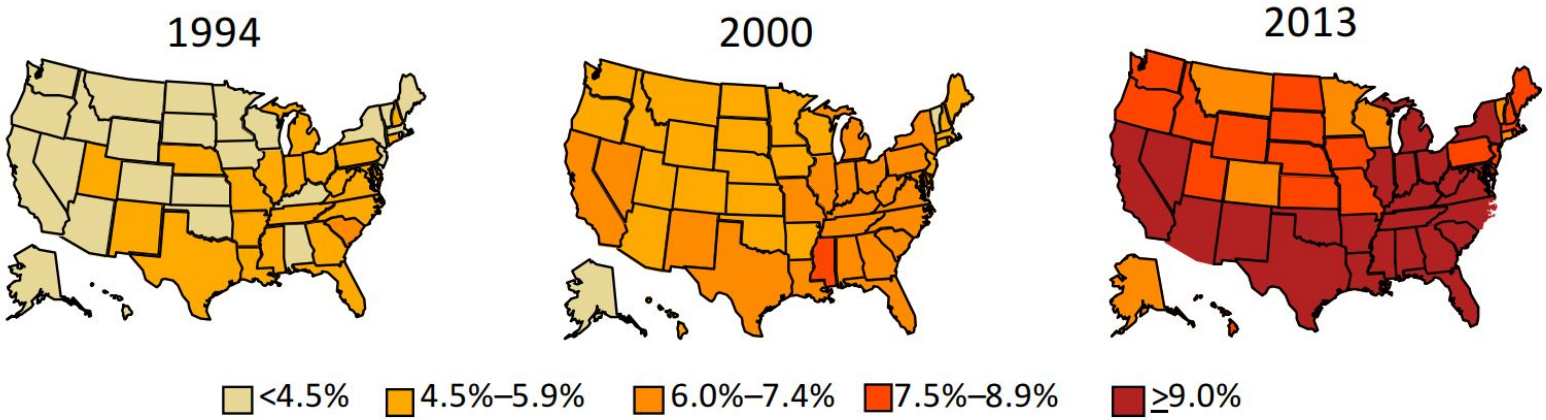
Some of the slides are inspired by **David Sontag**'s MIT lecture 'Machine Learning for Healthcare'. Content from his lecture will be indicated with:

Sontag, 2019

David is a great lecturer! Check out his course at:

https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-s897-machine-learning-for-healthcare-spring-2019/lecture-notes/index.htm

DIABETES PREVALENCE



Sontag, 2019

COST OF DIABETES

1 in 7 health care dollars is spent in treating diabetes and its complications

American Diabetes Association, 2020

⇒ detect at-risk population early and intervene

https://www.diabetes.org/resources/statistics/cost-diabetes

TRADITIONAL RISK ASSESSMENT FORM

Finnish Diabetes Association

TYPE 2 DIABETES RISK ASSESSMENT FORM

Circle the right alternative and add up your points.

1. Age

0 p. Under 45 years
2 p. 45–54 years
3 p. 55–64 years
4 p. Over 64 years

2. Body-mass index
(See reverse of form)

0 p. Lower than 25 kg/m^2
1 p. 25–30 kg/m^2
3 p. Higher than 30 kg/m^2

3. Waist circumference measured below the ribs (usually at the level of the navel)

	MEN	WOMEN
0 p.	Less than 94 cm	Less than 80 cm
3 p.	94–102 cm	80–88 cm
4 p.	More than 102 cm	More than 88 cm

6. Have you ever taken medication for high blood pressure on regular basis?

0 p. No
2 p. Yes

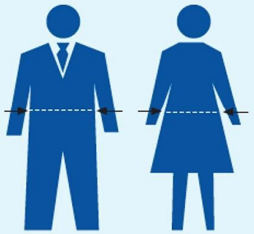
7. Have you ever been found to have high blood glucose (eg in a health examination, during an illness, during pregnancy)?

0 p. No
5 p. Yes

8. Have any of the members of your immediate family or other relatives been diagnosed with diabetes (type 1 or type 2)?

0 p. No
3 p. Yes: grandparent, aunt, uncle or first

4 p. More than 102 cm More than 88 cm



4. Do you usually have daily at least 30 minutes of physical activity at work and/or during leisure time (including normal daily activity)?

0 p. Yes
2 p. No

5. How often do you eat vegetables, fruit or berries?

0 p. Every day
1 p. Not every day

0 p. No
3 p. Yes: grandparent, aunt, uncle or first cousin (but no own parent, brother, sister or child)
5 p. Yes: parent, brother, sister or own child

Total Risk Score

☐ The risk of developing type 2 diabetes within 10 years is

Lower than 7	Low: estimated 1 in 100 will develop disease
7–11	Slightly elevated: estimated 1 in 25 will develop disease
12–14	Moderate: estimated 1 in 6 will develop disease
15–20	High: estimated 1 in 3 will develop disease
Higher than 20	Very high: estimated 1 in 2 will develop disease

Please turn over

Test designed by Professor Jaakko Tuomilehto, Department of Public Health, University of Helsinki, and Jaana Lindström, MFS, National Public Health Institute.

Finnish Diabetes Association, https://www.diabetes.fi/files/502/eRiskitestilomake.pdf

DIABETES 2 RISK PREDICTION MODELS

- ARIC
- KORA
- FRAMINGHAM
- AUSDRISC
- FINDRISC
- San Antonio Model

REPLACING QUESTIONNAIRES WITH CLAIMS DATA

- Claims data = data that the insurance company has (invoices, tests, ...)
- Readily available
- No time and cost intensive screening at doctors office necessary
- Immediately available and always up-to-date (effortless 're-taking')
- But: new dangers introduced with ML approach (to be discussed later)

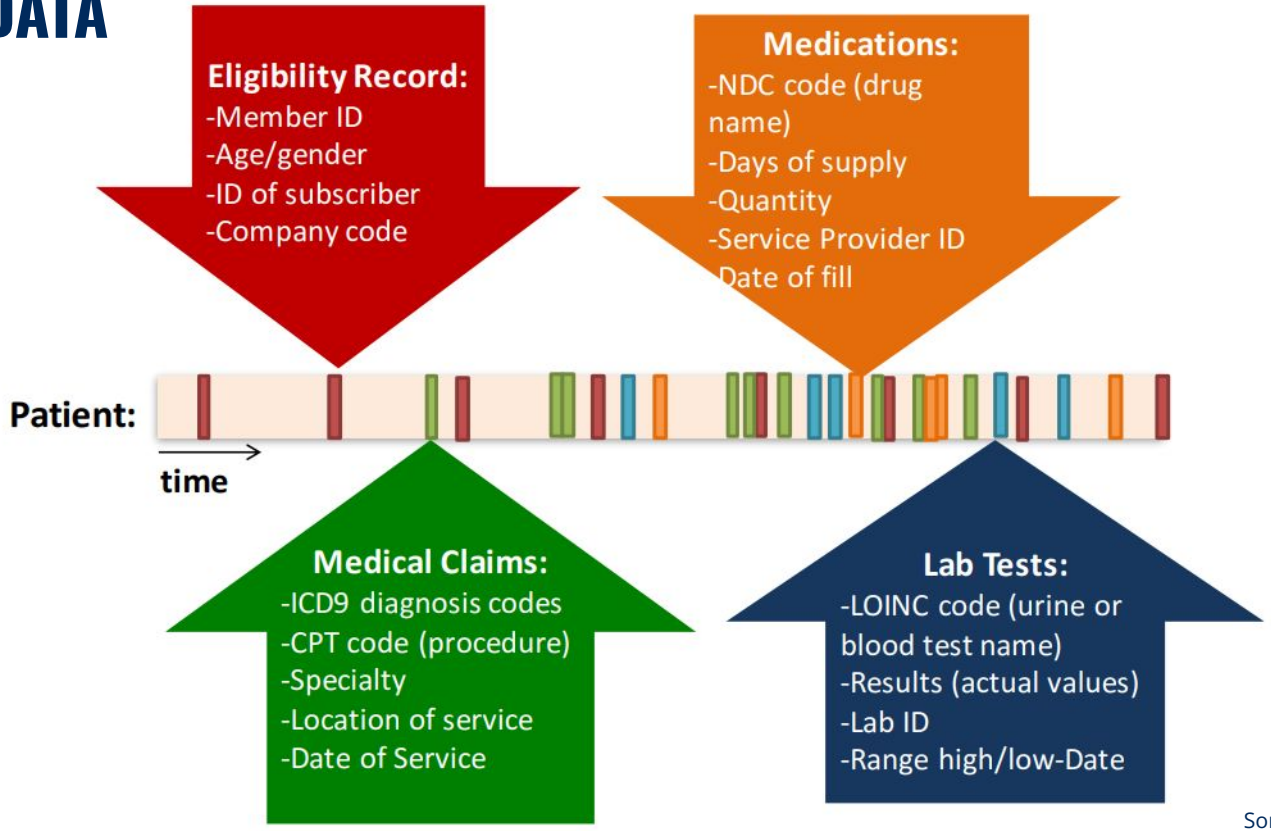
adapted from Sontag, 2019

REPLACING QUESTIONNAIRES WITH ML MODELS

Questionnaire		*Claims Data*
1. Age	→	**Age**
6. Have you ever taken medication for high blood pressure on regular basis?	→	**e.g. Lisinopril prescription**
5. How often do you eat vegetables, fruit or berries?	→	**¯_(ツ)_/¯**

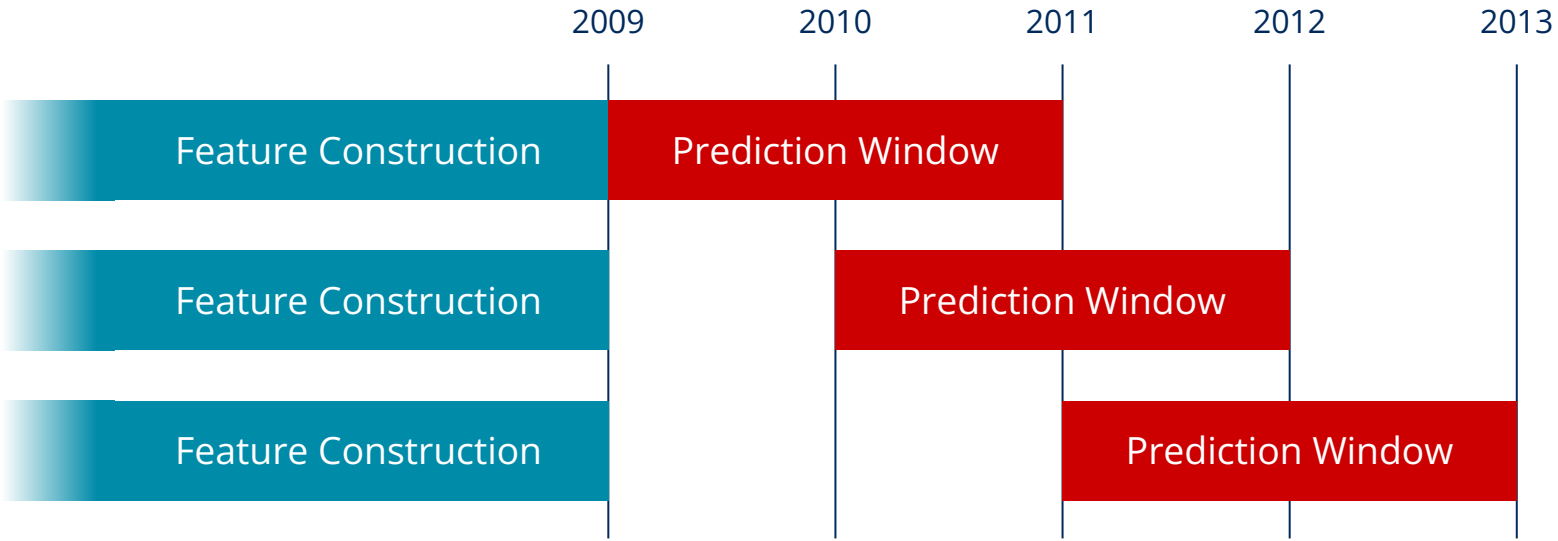
⇒ ML must find surrogates for missing features

UTILIZED DATA



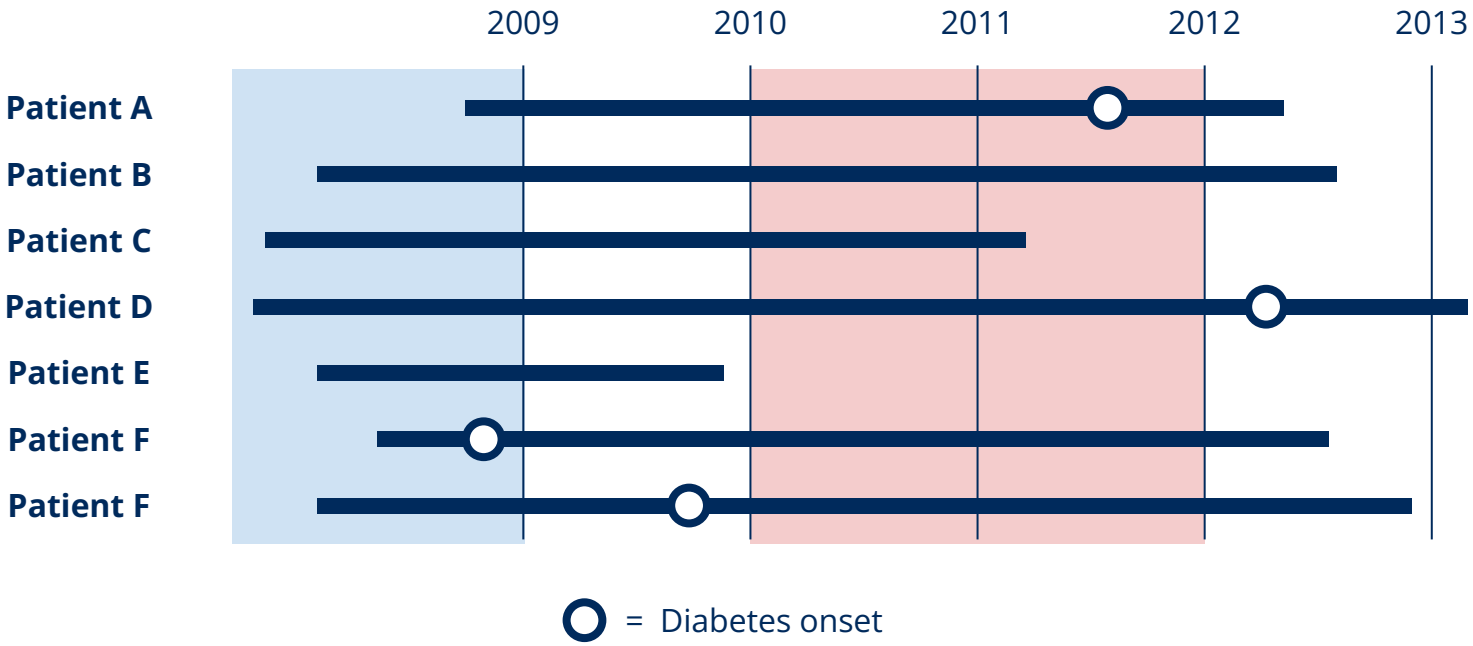
Sontag, 2019

PREDICTION TIMEFRAMES



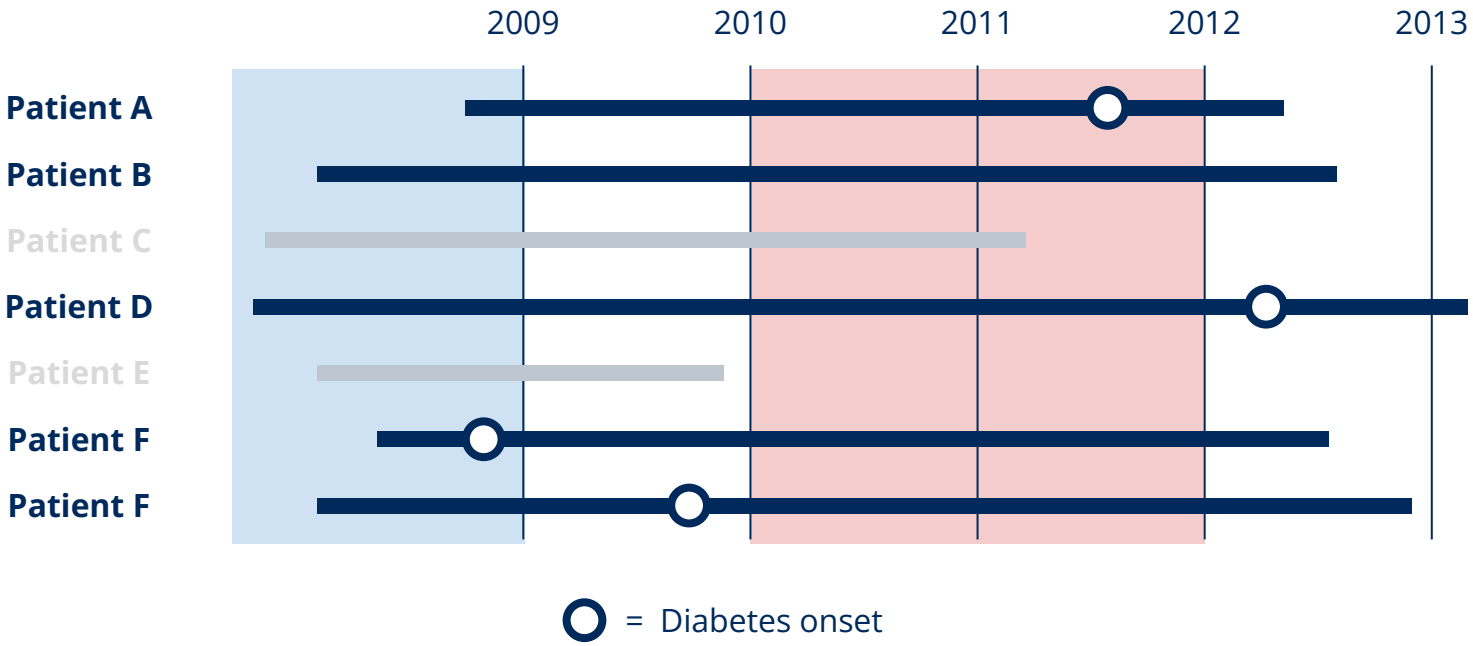
Sontag, 2019

LABELING



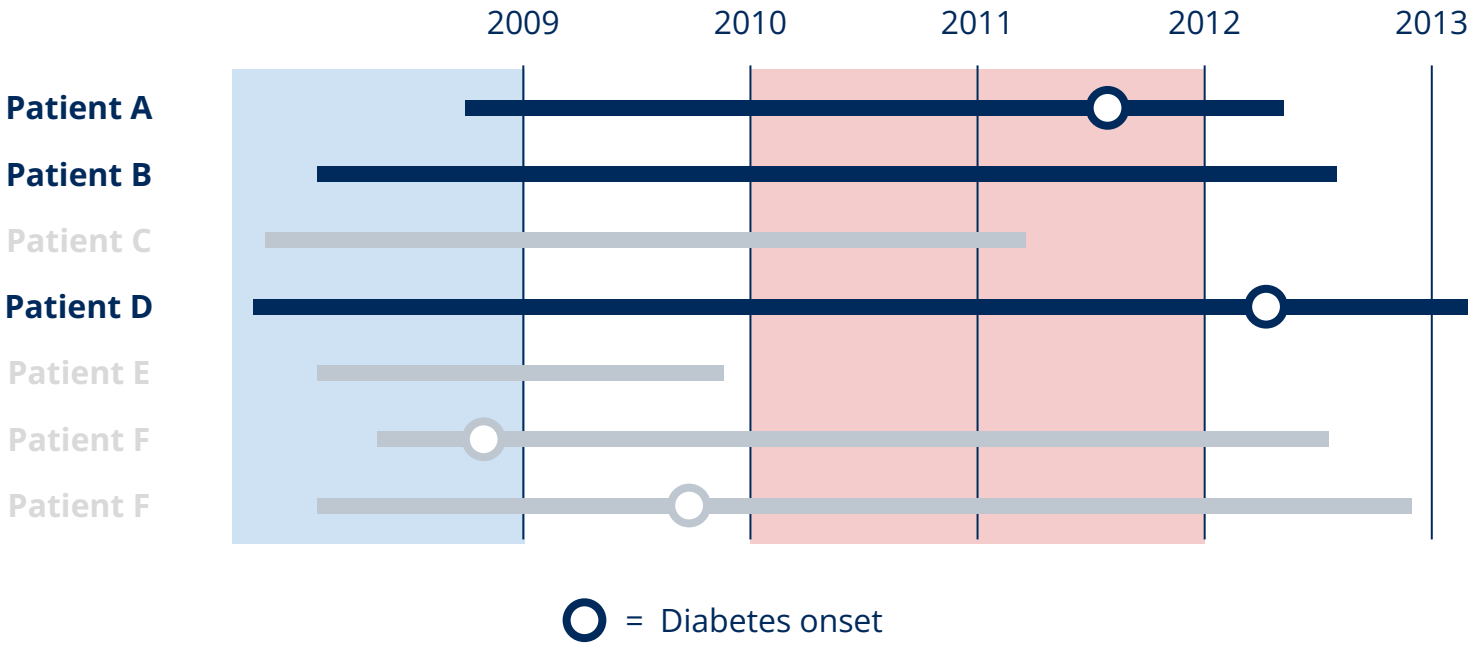
adapted from Razavian et al., 2015

LABELING



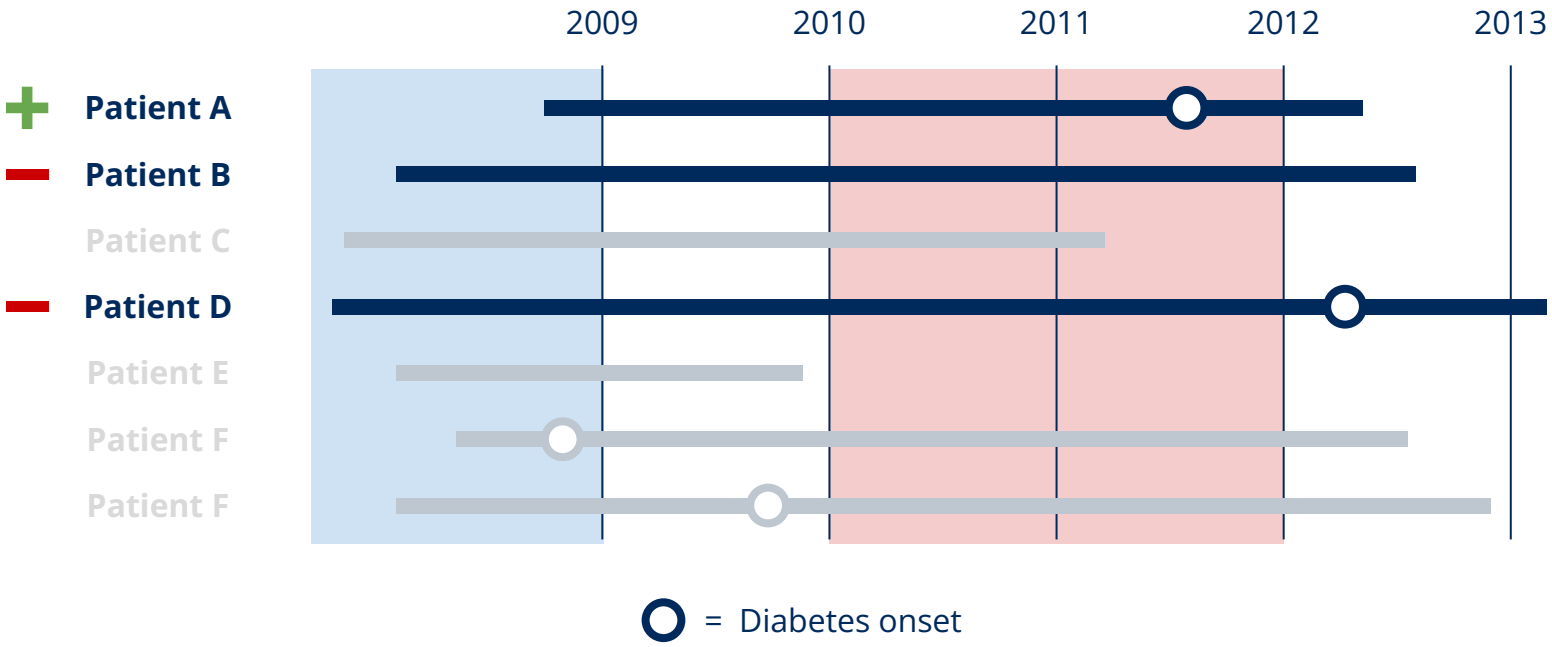
adapted from Razavian et al., 2015

LABELING



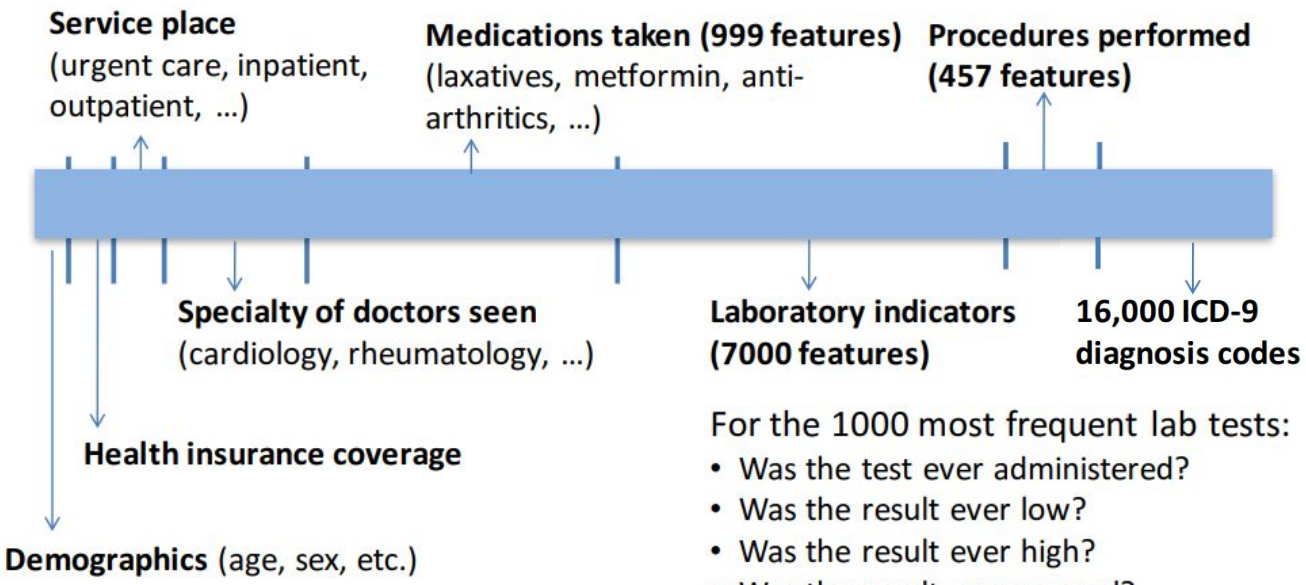
adapted from Razavian et al., 2015

LABELING



adapted from Razavian et al., 2015

DATA TRANSFORMATION

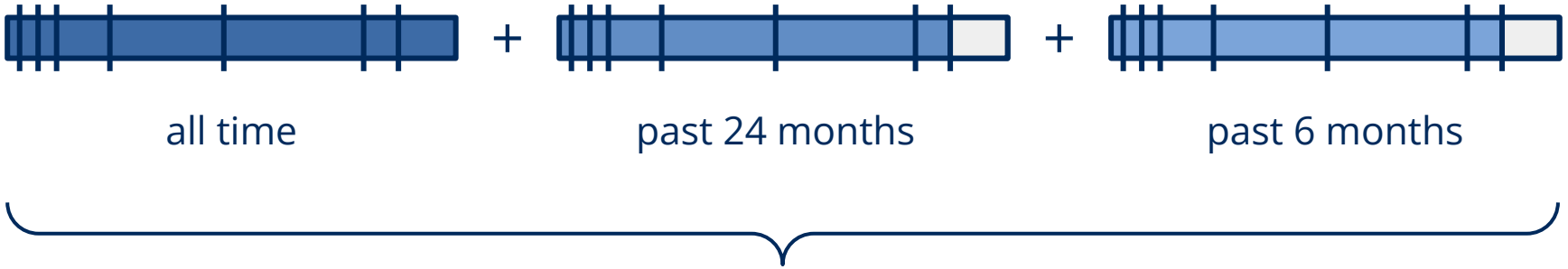


For the 1000 most frequent lab tests:

- Was the test ever administered?
- Was the result ever low?
- Was the result ever high?
- Was the result ever normal?
- Is the value increasing?
- Is the value decreasing?
- Is the value fluctuating?

adapted from Sontag, 2019

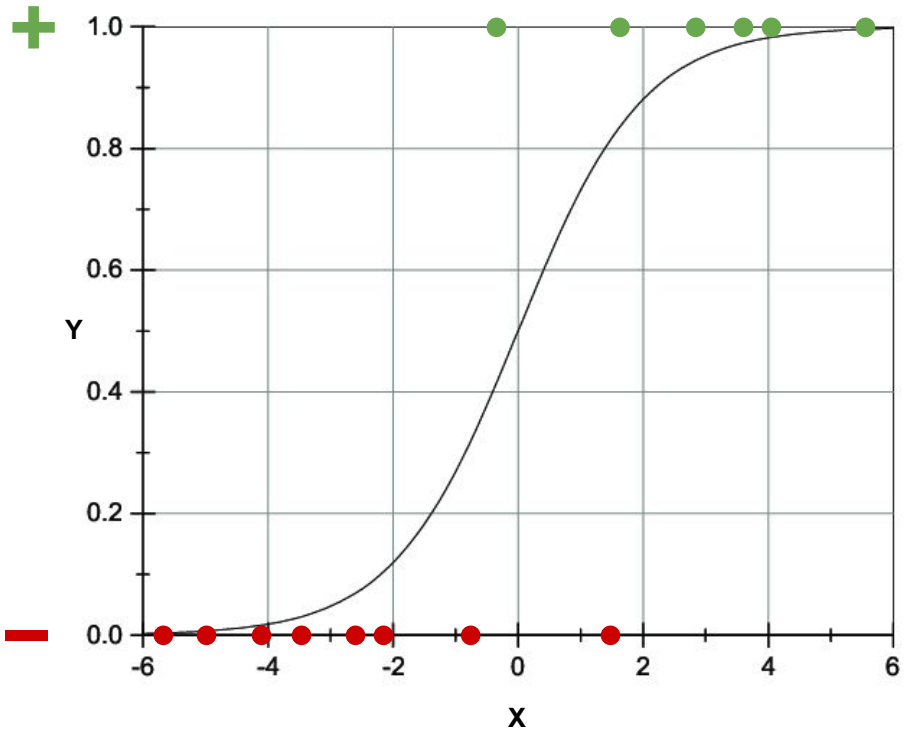
TIME BUCKETING



42.000 features

Sontag, 2019

LOGISTIC REGRESSION



MODEL

$$L(w) = \sum_{i=1}^{n} l(x_i, y_i; w) + \lambda \|w\|$$

L1-regularized logistic regression

built-in feature selection
sparse solution

optimize predictive performance

BASELINE: TRADITIONAL RISK FACTORS

- ARIC
- KORA
- FRAMINGHAM
- AUSDRISC
- FINDRISC
- San Antonio Model

age

hypertension

gender

obesity

HDL-Cholesterin

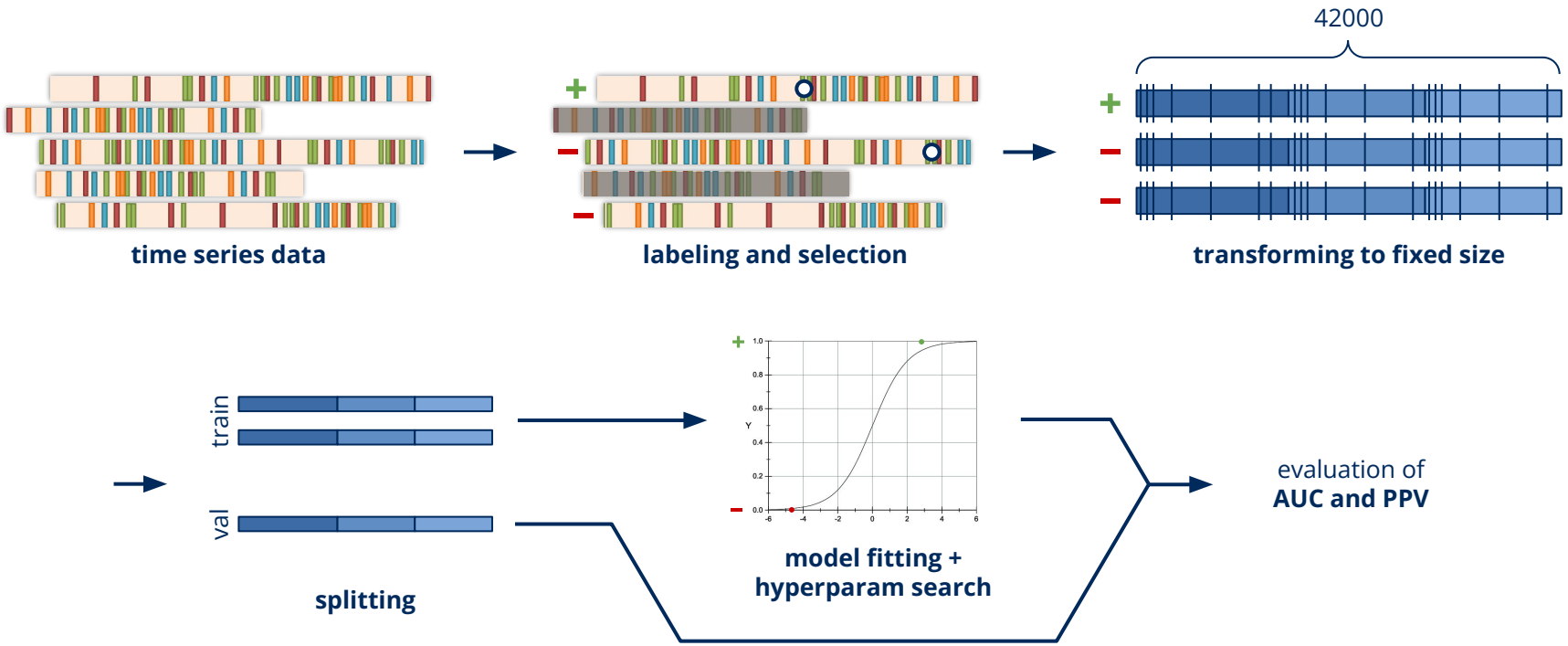
...

14 established features

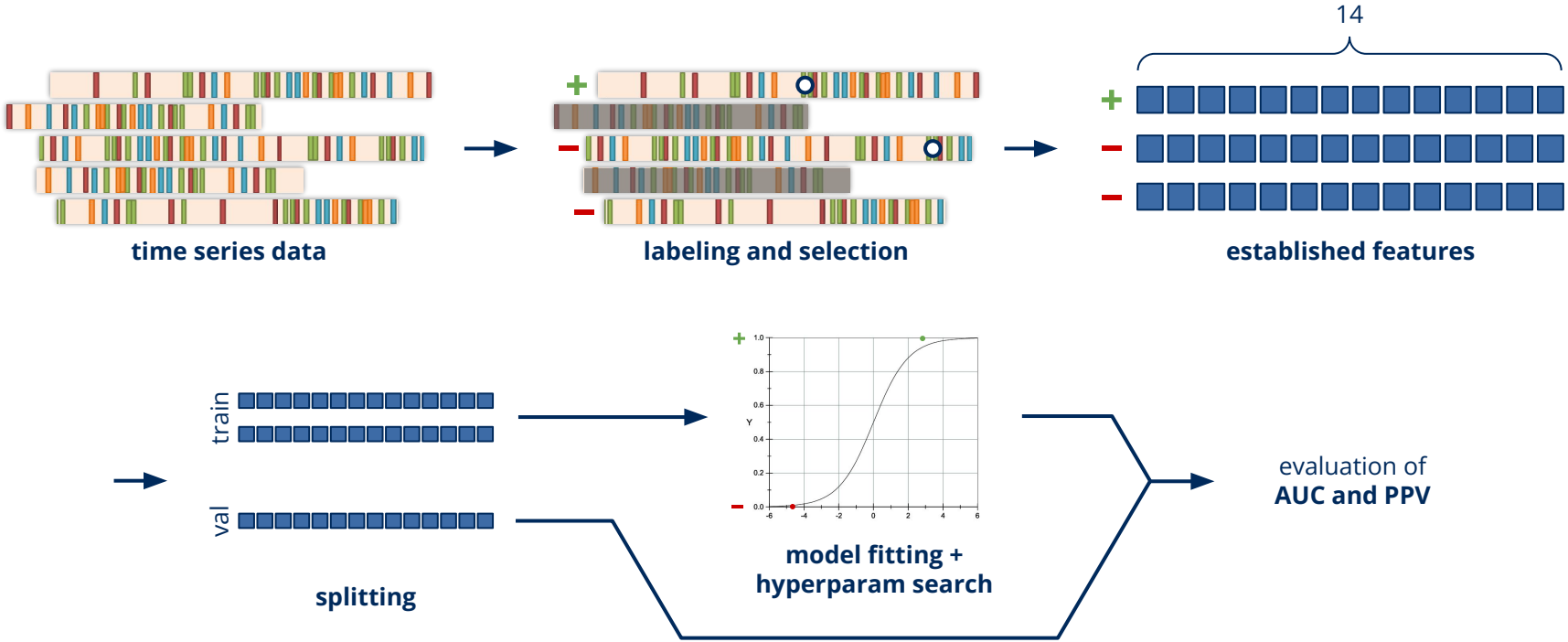
METHOD

- 67% training set
- 33% validation set
- hyperparameter search on training set with 5-fold cross-validation
- λ = [0.0001, 0.001, **0.1**, 1, 10]
- Reported: **AUC**, **Positive Predictive Value** (PPV)

PIPELINE: FULL MODEL

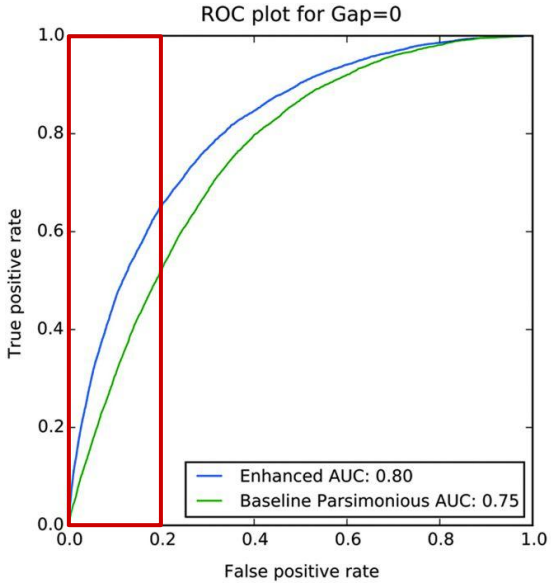


PIPELINE: BASELINE MODEL

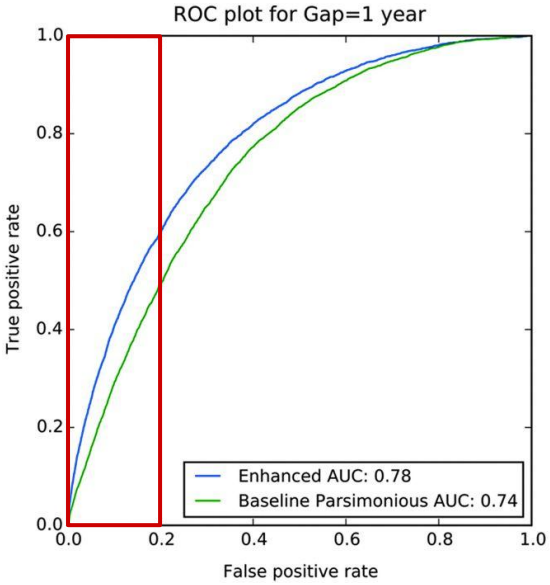


RESULTS: ROC/AUC

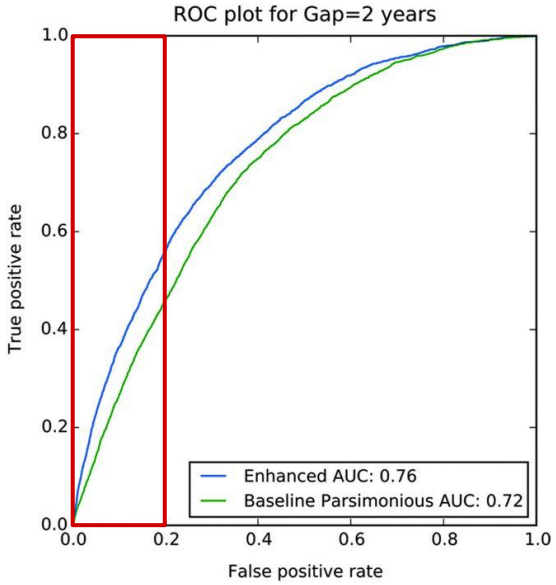
967 non-zero weight features



769 non-zero weight features



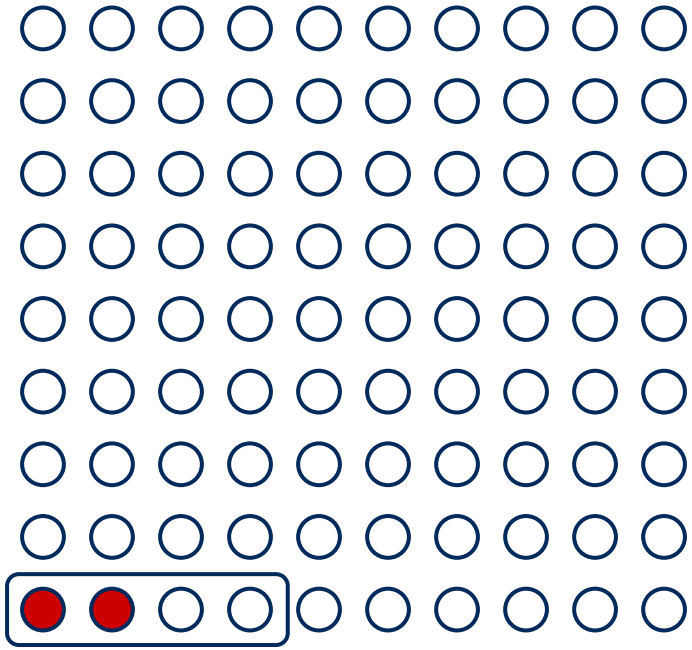
538 non-zero weight features



Razavian et al., 2015

POSITIVE PREDICTIVE VALUE (PPV) VS. SENSITIVITY (TPR)

$$\text{PPV} = \frac{\text{True Positives}}{\text{All Positives}}$$



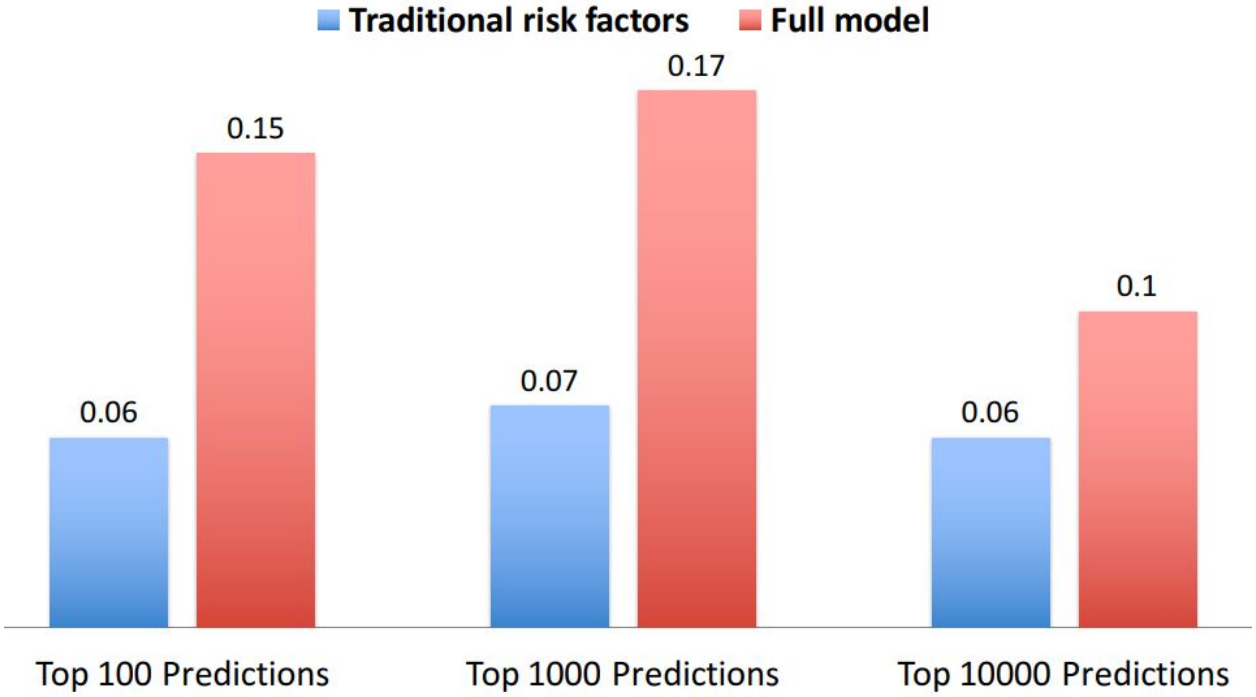
100 % **sensitivity (TPR)**

97.9 % **specificity (TNR)**

50 % **PPV**

RESULTS: PPV

1-year gap



Sontag, 2019

ODDS RATIO (OR)

	Diabetic	Non-diabetic
X	A	B
not X	C	D

$$\textbf{Odds ratio} = \frac{\textbf{Odds} \text{ that diabetic person has X } \textbf{(A/C)}}{\textbf{Odds} \text{ that non-diabetic person has X } \textbf{(B/D)}}$$

RESULTS: RISK FACTORS AND OR

1-year gap

Variable type	*Variable evaluation period*[a]	*Variable description*	*OR (95% CI)*
ICD9 history	Entire history	Impaired fasting glucose (ICD9-790.21)	4.17 (3.87 4.49)
	Entire history	Abnormal glucose NEC (ICD9-790.29)	4.07 (3.76 4.41)
	Entire history	Hypertension (ICD9-401)	3.28 (3.17 3.39)
	Entire history	Obstructive sleep apnea (ICD9-327.23)	2.98 (2.78 3.20)
	Entire history	Obesity (ICD9 278)	2.88 (2.75 3.02)
	Entire history	Abnormal blood chemistry (ICD9-790.6)	2.49 (2.36 2.62)
	Entire history	Hyperlipidemia (ICD9 272.4)	2.45 (2.37 2.53)
	Entire history	Shortness of breath (ICD9-786.05)	2.09 (1.99 2.19)
	Entire history	Esophageal reflux (ICD9-530.81)	1.85 (1.78 1.93)
	Entire history	Acute bronchitis (ICD9-466.0)	1.44 (1.37 1.50)

Razavian et al., 2015

RESULTS: RISK FACTORS AND OR

1-year gap

Variable type	*Variable evaluation period*[a]	*Variable description*	*OR (95% CI)*
Laboratory test	Entire history	Hemoglobin A1c/hemoglobin.total—high (LOINC-4548-4)	5.75 (5.42 6.10)
	Past 2 years	Glucose—high (LOINC-2345-7)	4.05 (3.89 4.21)
	Past 2 years	Hemoglobin A1c/hemoglobin.total—request for test	3.42 (3.27 3.58)
	Entire history	Hemoglobin A1c/hemoglobin.total—request for test	3.13 (3.00 3.26)
	Entire history	Cholesterol.In HDL—low (LOINC-2085-9)	2.78 (2.66 2.92)
	Entire history	Cholesterol.total/cholesterol.In HDL—high (LOINC-9830-1)	2.29 (2.19 2.40)
	Entire history	Cholesterol.In VLDL—request for test (LOINC-13458-5)	2.23 (2.13 2.33)
	Entire history	Carbon dioxide—request for test (LOINC-2028-9)	1.58 (1.53 1.64)
	Past 2 years	Glomerular filtration rate/1.73 Sq. M.Predicted.Black—request for test (LOINC-48643-1)	1.58 (1.52 1.64)

Razavian et al., 2015

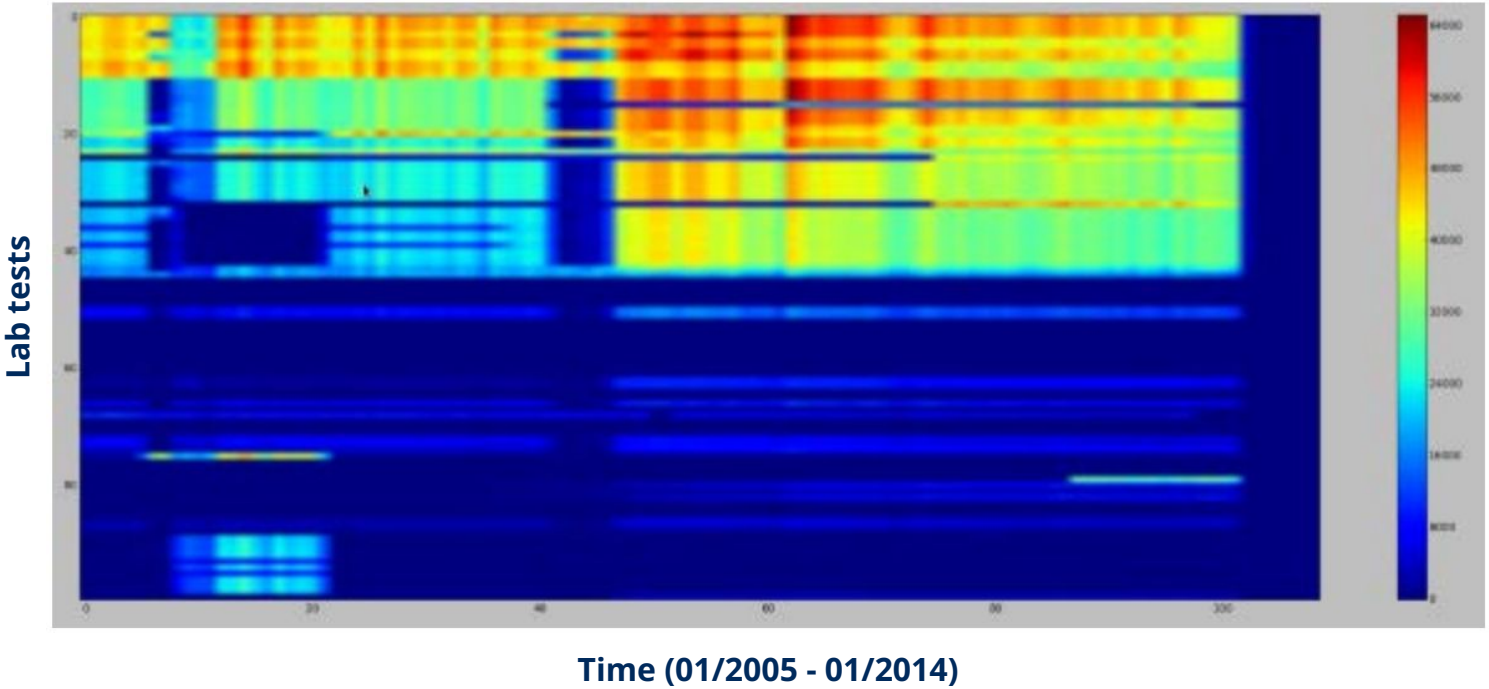
RESULTS: RISK FACTORS AND OR

1-year gap

Variable type	*Variable evaluation period*[a]	*Variable description*	*OR (95% CI)*
NDC medications	Past 2 years	Medication group: antiarthritics	1.43 (1.36 1.50)
	Entire history	Medication group: antiarthritics	1.41 (1.35 1.48)
Healthcare utilization	Entire history	Procedure group: routine chest X-ray	1.96 (1.89 2.03)
	Entire history	Dental coverage = yes	1.47 (1.41 1.53)
	Entire history	Service place: emergency room—hospital	1.32 (1.28 1.37)
	Entire history	Specialty code: independent laboratory	1.18 (1.14 1.22)
	Entire history	Routine medical examination (ICD9 V700)	0.85 (0.82 0.88)
	Entire history	Routine gynecological examination (ICD9 V7231)	0.84 (0.81 0.87)
	Entire history	Routine child health examination (ICD9 V202)	0.10 (0.09 0.12)

Razavian et al., 2015

TOP 100 LAB TESTS OVER TIME (ABS. COUNT)



STRENGTHS AND WEAKNESSES

+ largest Diabetes 2 risk prediction study in terms of # features and cohort size
+ immediate population-level results *and* good performance (as good as before)
+ deployed at Independence Blue Cross
+ allows for prioritization of beneficiaries
- does not work for recently enrolled beneficiaries
- no onset estimation (*when* instead of *if*)
- odds ratio ≠ learned weights (are they identical?)
- performance tested in same time span (susceptible to non-stationarity)
- susceptible to dataset shift also in the future

QUESTIONS