# Machine Learning for Healthcare

6.7930, HST.956

## Lecture 5:

Feb 23, 2023

Peter Szolovits

(with many slides from David Sontag)

Massachusetts Institute of Technology

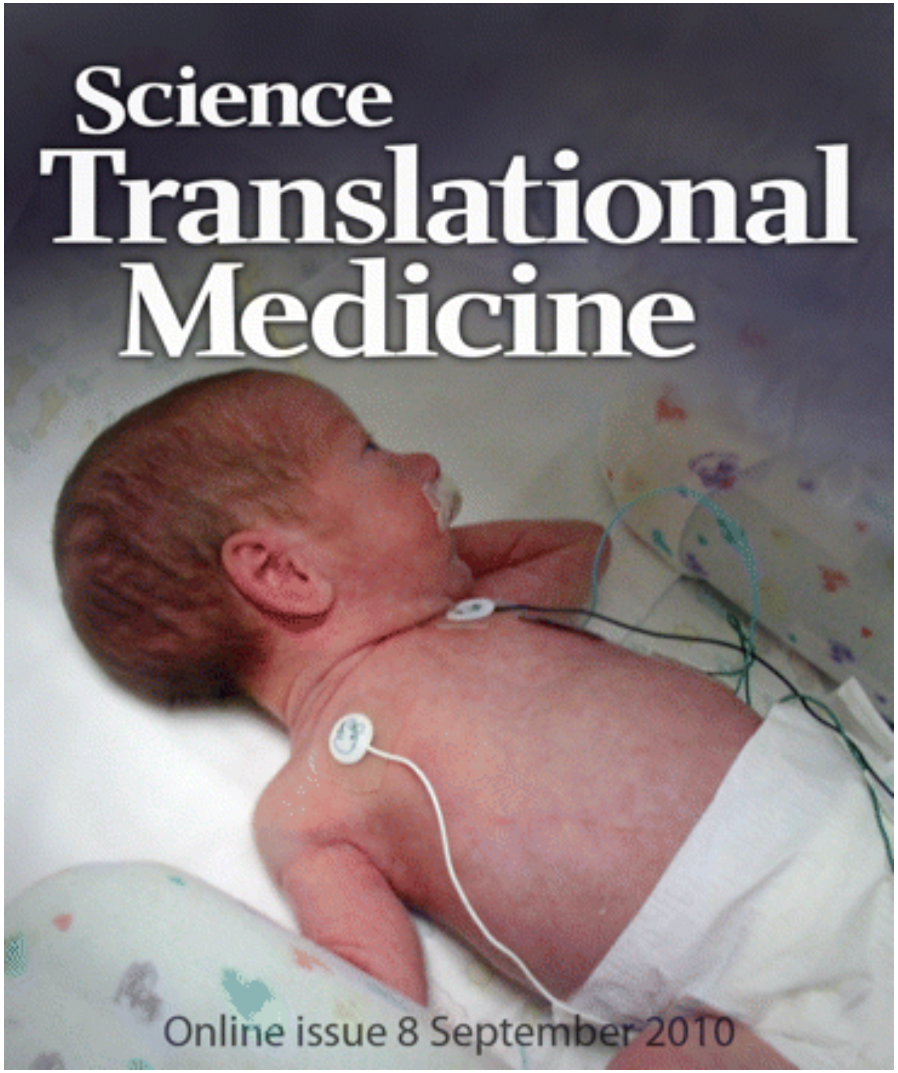
# Outline for today’s class

1. **Introduction to risk stratification**
2. Case study: Early detection of Type 2 diabetes
   - Encoding longitudinal structured health data
3. Framing as supervised learning problem
   - Deriving labels from EHR

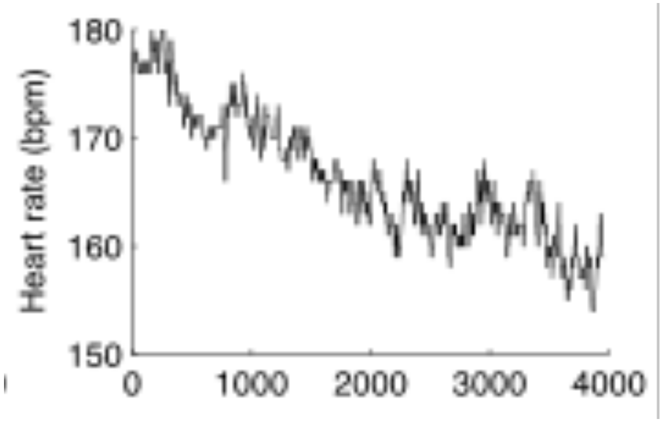
# What *is* risk stratification?

- Separate a patient population into **high-risk** and **low-risk** of having an outcome
  - Predicting something in the future
- Coupled with **interventions** that target high-risk patients
- Goal is typically to reduce cost and improve patient outcomes

# Examples of risk stratification



Preterm infant's risk of severe morbidity?

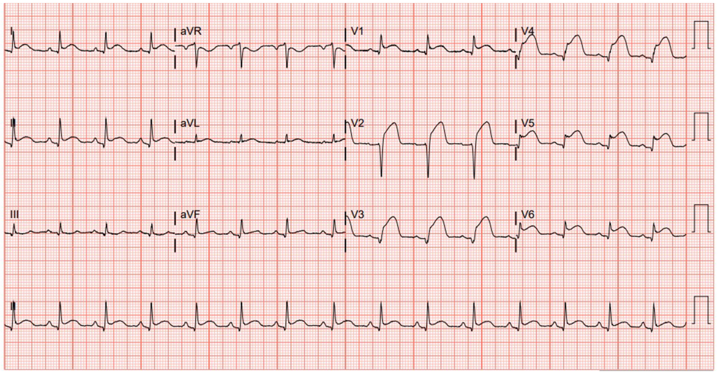


(Saria et al., Science Translational Medicine 2010)

# Examples of risk stratification

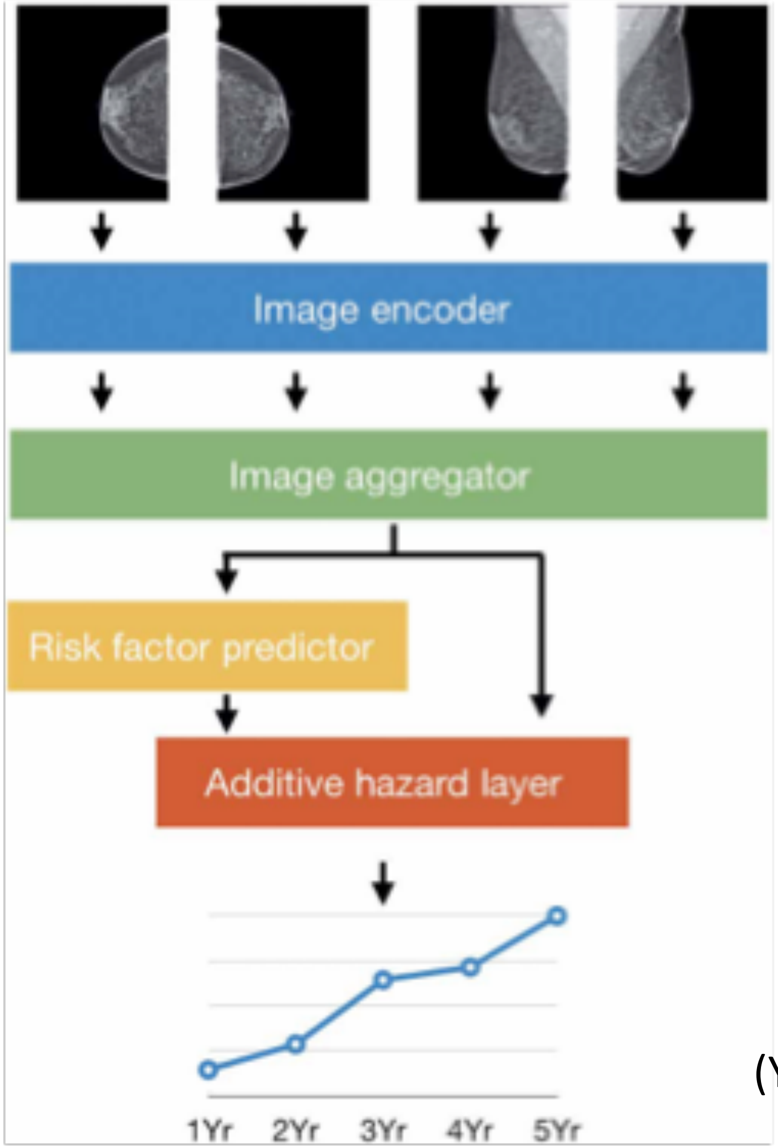
Figure sources:
https://www.drmani.com/heart-attack/ (top)
https://www.emra.org/emresident/article/acute-mi-case-report/ (right)

Does this patient need to be admitted to the coronary-care unit?



(Pozen et al., NEJM 1984)

# Examples of risk stratification

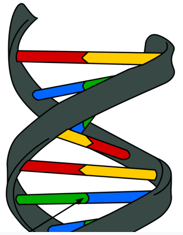


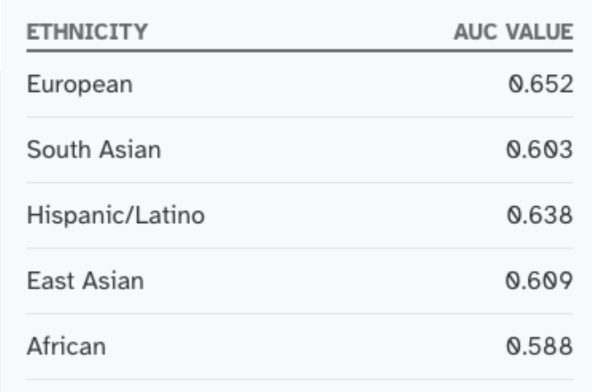
Will this woman develop breast cancer in the next 5 years?

(Yala et al., Science Translational Medicine 2021)

# Examples of risk stratification

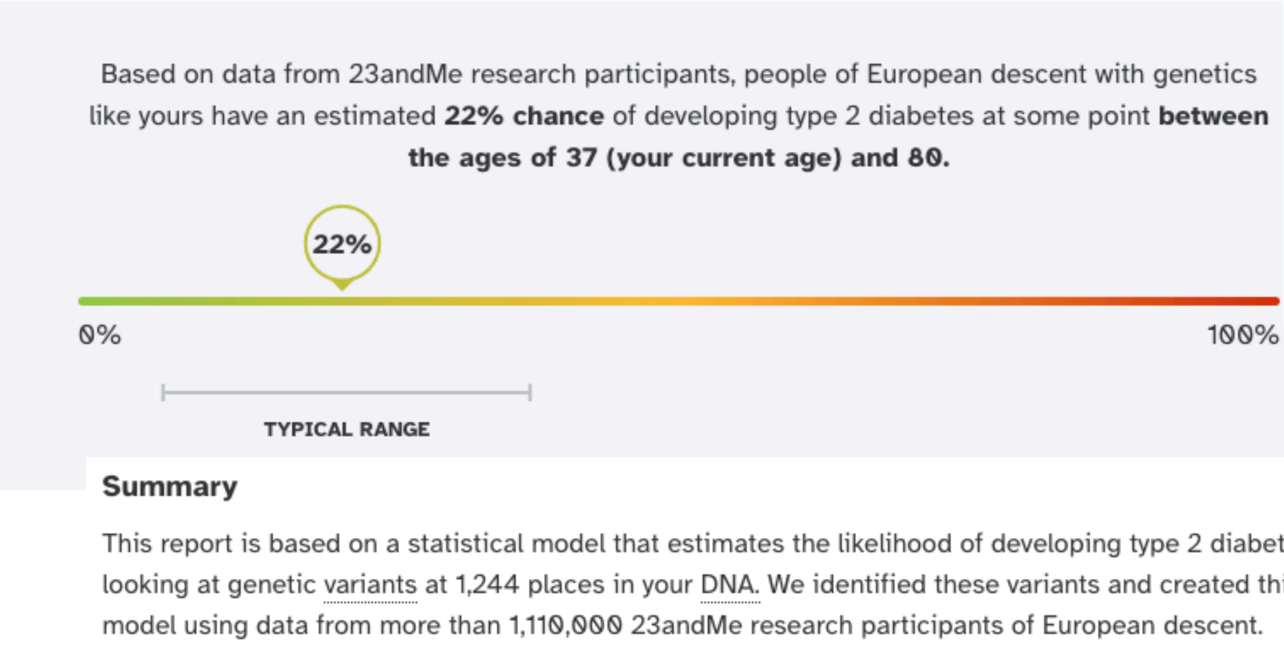
David, your genetics are associated with a **typical likelihood** of developing type 2 diabetes.





| ETHNICITY | AUC VALUE |
|---|---|
| European | 0.652 |
| South Asian | 0.603 |
| Hispanic/Latino | 0.638 |
| East Asian | 0.609 |
| African | 0.588 |

DNA
Deoxyribonucleic acid



Based on data from 23andMe research participants, people of European descent with genetics like yours have an estimated **22% chance** of developing type 2 diabetes at some point **between the ages of 37 (your current age) and 80.**

**Summary**

This report is based on a statistical model that estimates the likelihood of developing type 2 diabetes by looking at genetic variants at 1,244 places in your DNA. We identified these variants and created this model using data from more than 1,110,000 23andMe research participants of European descent.

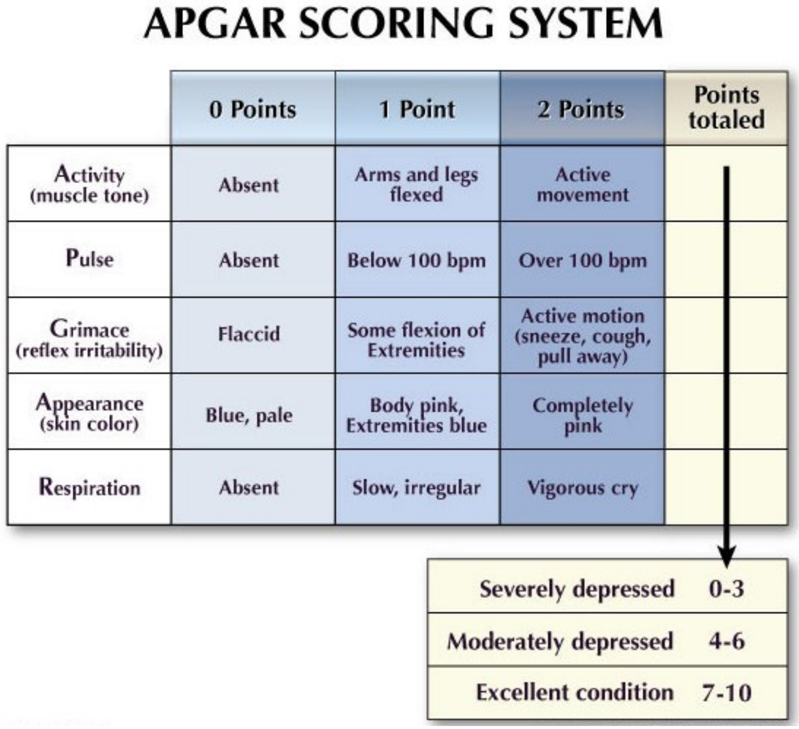
(Source: 23andme. https://permalinks.23andme.com/pdf/23_19-Type2Diabetes_March2019.pdf)

# How does risk stratification differ from differential diagnosis?

| Differential diagnosis | Risk stratification |
|---|---|
| Usually iterative/active | Usually passive |
| Often considers a large set of conditions | Often just one condition |
| Has to consider rare conditions (needs hybrid knowledge/ML approaches) | Often focuses on settings where there is enough training data |

# Old vs. New

- Traditionally, risk stratification was based on simple scores using human-entered data



APGAR SCORING SYSTEM

| | 0 Points | 1 Point | 2 Points | Points totaled |
|---|---|---|---|---|
| Activity (muscle tone) | Absent | Arms and legs flexed | Active movement | |
| Pulse | Absent | Below 100 bpm | Over 100 bpm | |
| Grimace (reflex irritability) | Flaccid | Some flexion of Extremities | Active motion (sneeze, cough, pull away) | |
| Appearance (skin color) | Blue, pale | Body pink, Extremities blue | Completely pink | |
| Respiration | Absent | Slow, irregular | Vigorous cry | |

| | |
|---|---|
| Severely depressed | 0-3 |
| Moderately depressed | 4-6 |
| Excellent condition | 7-10 |

# Old vs. New

- Traditionally, risk stratification was based on simple scores using human-entered data
- Now, based on machine learning on high-dimensional data
  - Fits more easily into workflow
  - Higher accuracy
  - Quicker to derive (can special case)
- **But, ML approach comes with new challenges – to be discussed**

# So, what do we need?

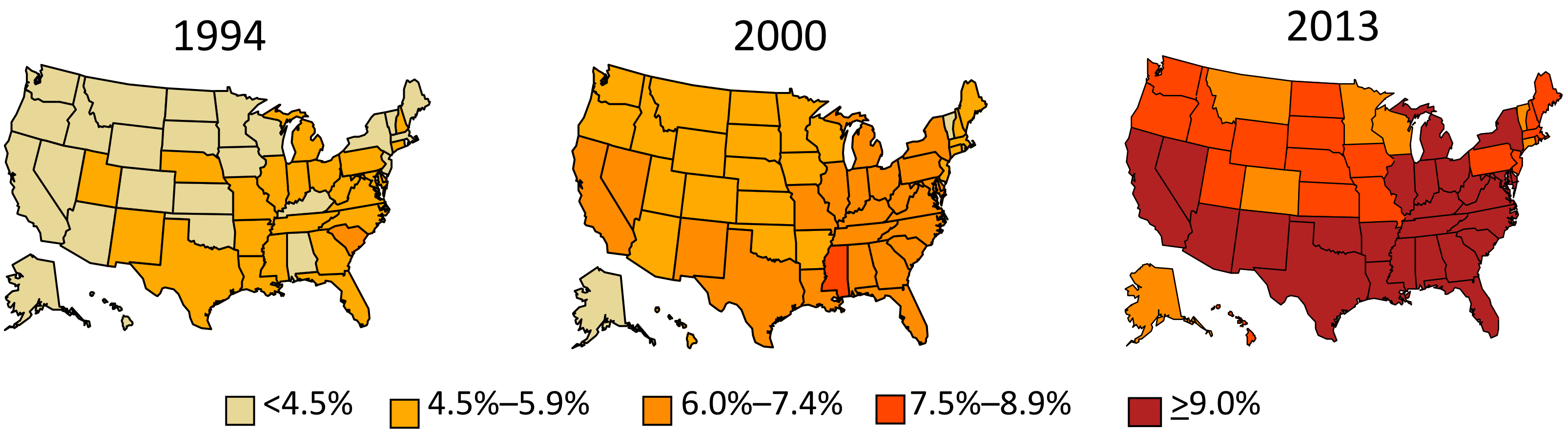
- Specification of prediction time / index date
- A way of encoding the data we have on the patient
  - CNN for images
  - Bag of words for text document
  - Longitudinal structured data …
- A target, typically derived from the EHR
- Choice of appropriate supervised ML algorithm
  - Regression? Classification?

# Outline for today's class

1. Introduction to risk stratification
2. **Case study: Early detection of Type 2 diabetes**
   - Encoding longitudinal structured health data
3. Framing as supervised learning problem
   - Deriving labels from EHR

[Razavian, Blecker, Schmidt, Smith-McLallen, Nigam, Sontag. *Big Data.* '16]

# Type 2 Diabetes: A Major public health challenge



$245 billion: Total costs of diagnosed diabetes in the United States in 2012
$831 billion: Total fiscal year federal budget for healthcare in the United States in 2014

- CDC 2022 estimate:
  - 11.3% of adults: 28.7M diagnosed, 8.5M undiagnosed
- Racial disparities among adults (20+)
  - non-Hisp White: 7.5%
  - non-Hisp Asian: 9.2%
  - non-Hisp Black: 11.7%
  - Hispanic: 12.5%
  - Native American: 14.7%

# Type 2 Diabetes Can Be Prevented *

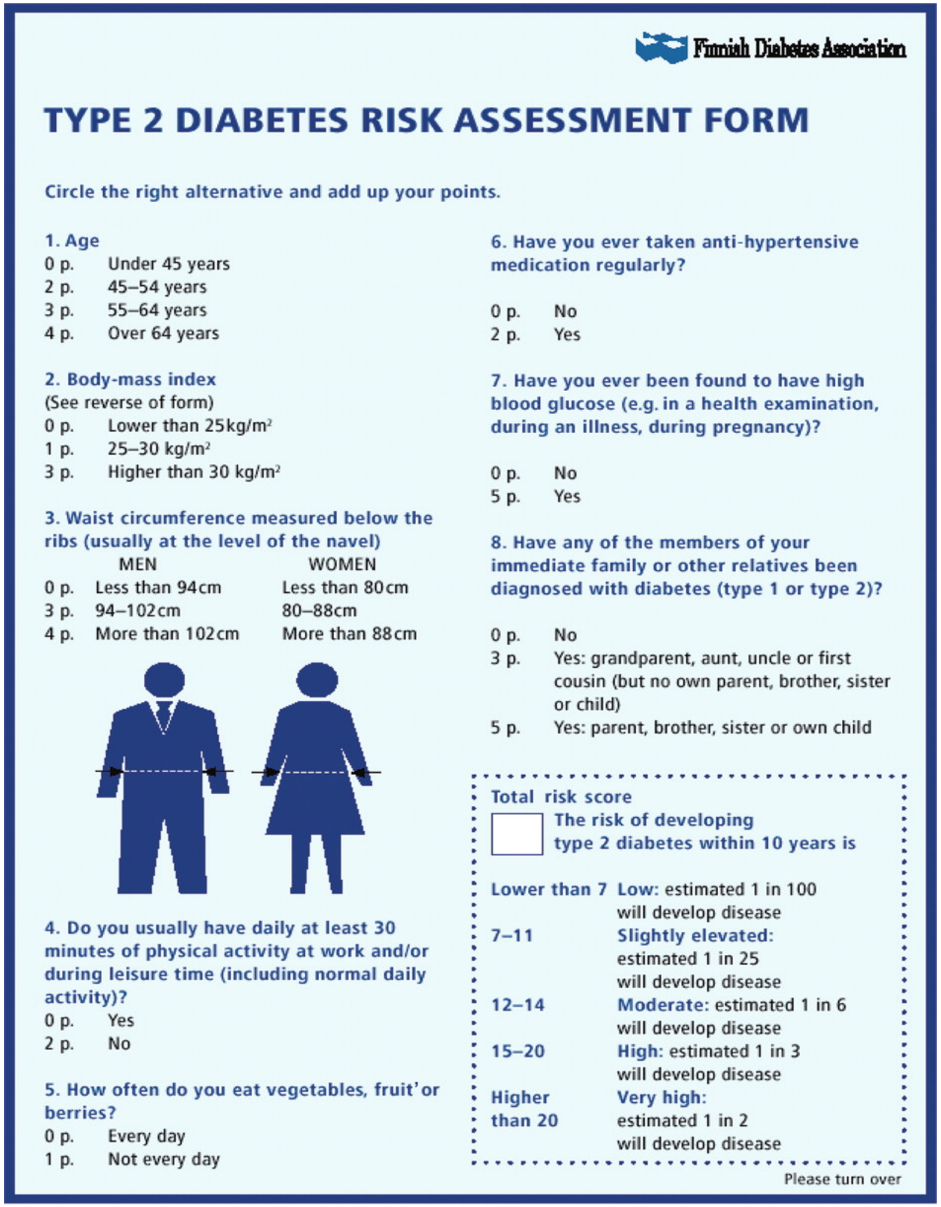
Requirement for successful large scale prevention program:

1. Detect/reach truly at risk population
2. Improve the interventions
3. Lower the cost of intervention

* Diabetes Prevention Program Research Group. "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin." The New England journal of medicine 346.6 (2002): 393.

# Traditional Risk Prediction Models

- Successful Examples
  - ARIC
  - KORA
  - FRAMINGHAM
  - AUSDRISC
  - FINDRISC
  - San Antonio Model
- Easy to ask/measure in the office, or for patients to do online
- Simple model:
  can calculate scores by hand



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/m² |
| 1 p. | 25–30 kg/m² |
| 3 p. | Higher than 30 kg/m² |

**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 |

**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 |

**6. Have you ever taken anti-hypertensive medication regularly?**

|  |  |
|---|---|
| 0 p. | No |
| 2 p. | Yes |

**7. Have you ever been found to have high blood glucose (e.g. 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 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

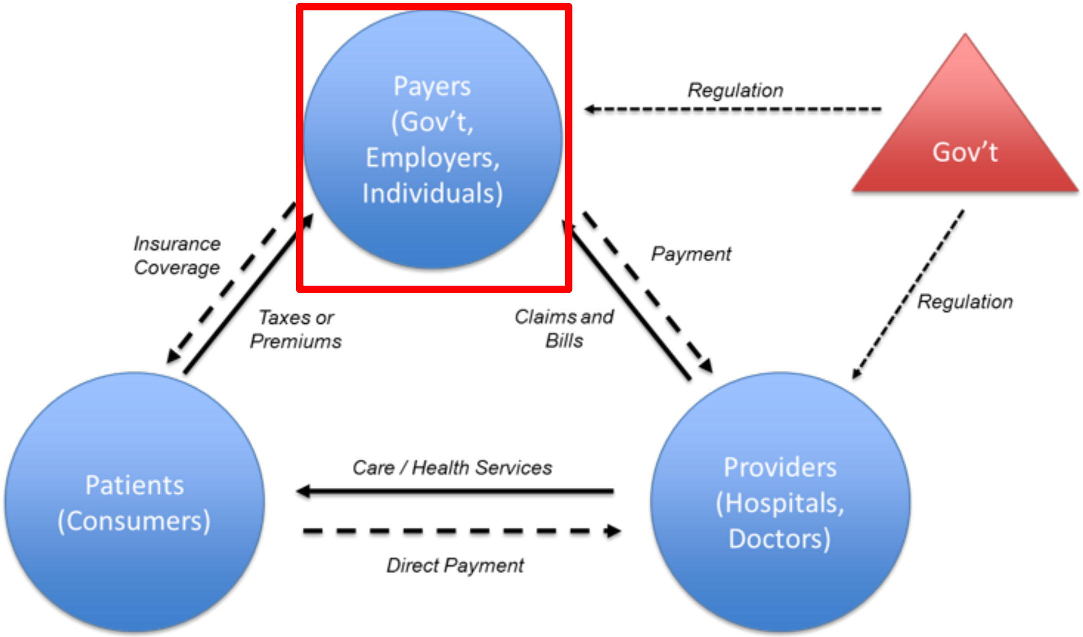
|  |  |
|---|---|
| Lower than 7 | Low: estimated 1 in 100 will develop disease |
| 7–11 | Slightly elevated: estimated 1 in 25 will develop disease |
| 12–14 | Moderate: estimated 1 in 6 will develop disease |
| 15–20 | High: estimated 1 in 3 will develop disease |
| Higher than 20 | Very high: estimated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

# Population-Level Risk Stratification

- Key idea: Use readily available administrative, utilization, and clinical data



Source for figure: http://www.mahesh-vc.com/blog/understanding-whos-paying-for-what-in-the-healthcare-industry

# Population-Level Risk Stratification

- Key idea: Use readily available administrative, utilization, and clinical data
- Machine learning will find surrogates for risk factors that would otherwise be missing
- Perform risk stratification at the population level – millions of patients

# A Data-Driven approach on Longitudinal Data-Based Prediction

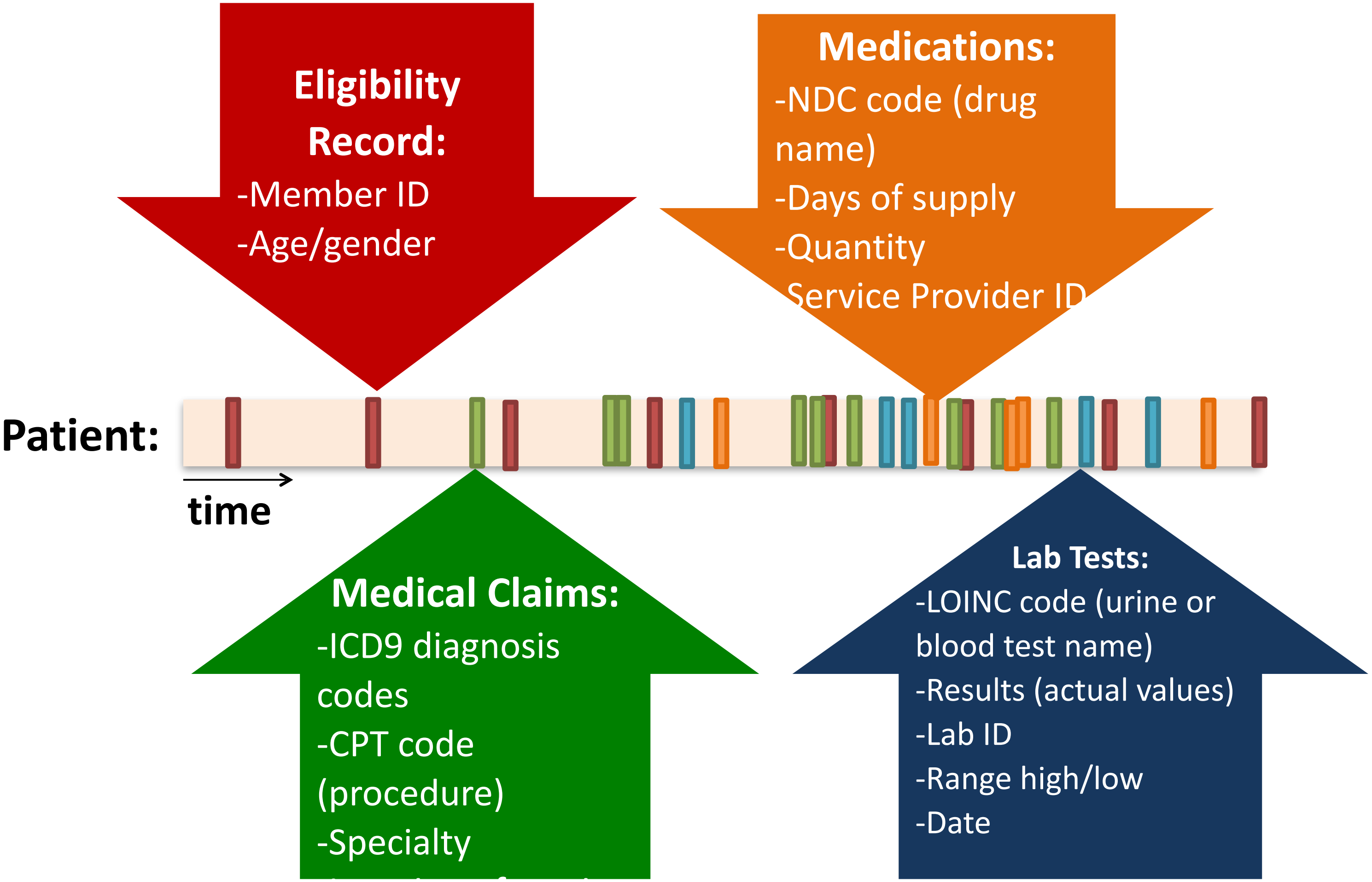
- Looking at individuals who got diabetes *today,* (compared to those who didn't)
  - Can we infer which variables in their record could have predicted their health outcome?

A Few Years Ago

Today

Risk Stratification from Structured Health Data
Reading: Razavian N, Blecker S, Schmidt AM, Smith- McLallen A, Nigam S, Sontag D (2015) Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data 3:4, 277–287, DOI: 10.1089/big.2015.0020.

# Administrative & Clinical Data



**Eligibility Record:**

-Member ID

-Age/gender

**Medications:**

-NDC code (drug name)

-Days of supply

-Quantity

Service Provider ID

**Medical Claims:**

-ICD9 diagnosis codes

-CPT code (procedure)

-Specialty

**Lab Tests:**

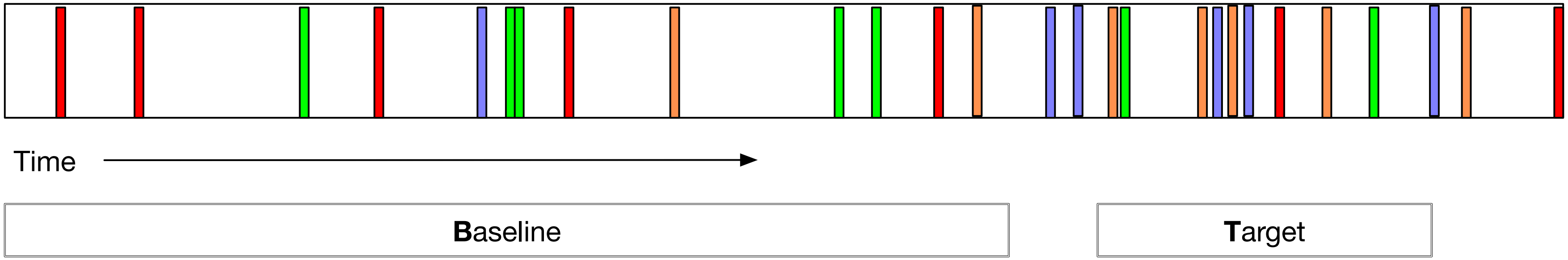
-LOINC code (urine or blood test name)

-Results (actual values)

-Lab ID

-Range high/low

-Date



Target = *f*(Baseline)

- How to represent Baseline and Target?
- What class of models do we consider for *f*?

# Claims Data Characteristics

- Sparse data
  - Vast coding spaces for diagnoses, symptoms, procedures, medications, labs
  - Most patients don’t have most of these
- Visit-level temporality
  - Data collected only at interactions with the health care system; highly variable intervals
- Long-term dependencies
  - How to encode these? LSTM, bi-RNN, CNN, attention, …
  - simpler trends

[1] Kodialam RS, Boiarsky R, Sontag D. Deep contextual clinical prediction with reverse distillation. arXiv [Internet]. 2020 Jul; Available from: http://arxiv.org/abs/2007.05611v1

# Top diagnosis codes

| Disease | count |
|---|---|
| **401.1 Benign hypertension** | 447017 |
| 272.4 Hyperlipidemia NEC/NOS | 382030 |
| 401.9 Hypertension NOS | 372477 |
| **250.00 DMII wo cmp nt st uncntr** | 339522 |
| 272.0 Pure hypercholesterolem | 232671 |
| 272.2 Mixed hyperlipidemia | 180015 |
| V72.31 Routine gyn examination | 178709 |
| 244.9 Hypothyroidism NOS | 169829 |
| **780.79 Malaise and fatigue NEC** | 149797 |
| **V04.81 Vaccin for influenza** | 147858 |
| **724.2 Lumbago** | 137345 |
| **V76.12 Screen mammogram NEC** | 129445 |
| **V70.0 Routine medical exam** | 127848 |

| Disease | count |
|---|---|
| **530.81 Esophageal reflux** | 121064 |
| 427.31 Atrial fibrillation | 113798 |
| **729.5 Pain in limb** | 112449 |
| 414.01 Crnry athrscl natve vssl | 104478 |
| 285.9 Anemia NOS | 103351 |
| **786.50 Chest pain NOS** | 91999 |
| **599.0 Urin tract infection NOS** | 87982 |
| V58.69 Long-term use meds NEC | 85544 |
| **496 Chr airway obstruct NEC** | 78585 |
| 477.9 Allergic rhinitis NOS | 77963 |
| 414.00 Cor ath unsp vsl ntv/gft | 75519 |

| Disease | count |
|---|---|
| 719.47 Joint pain-ankle | 28648 |
| 300.4 Dysthymic disorder | 28530 |
| 268.9 Vitamin D deficiency NOS | 28455 |
| V72.81 Preop cardiovsclr exam | 27897 |
| **724.3 Sciatica** | 27604 |
| **787.91 Diarrhea** | 27424 |
| **V2.21 Supervis oth normal preg** | 27320 |
| 365.01 Opn angl brderln lo risk | 26033 |
| 379.21 Vitreous degeneration | 25592 |
| 424.1 Aortic valve disorder | 25425 |
| 616.10 Vaginitis NOS | 24736 |
| 702.19 Other sborheic keratosis | 24453 |
| 380.4 Impacted cerumen | 24046 |

**Out of 135K patients who had laboratory data**

# Top lab test results

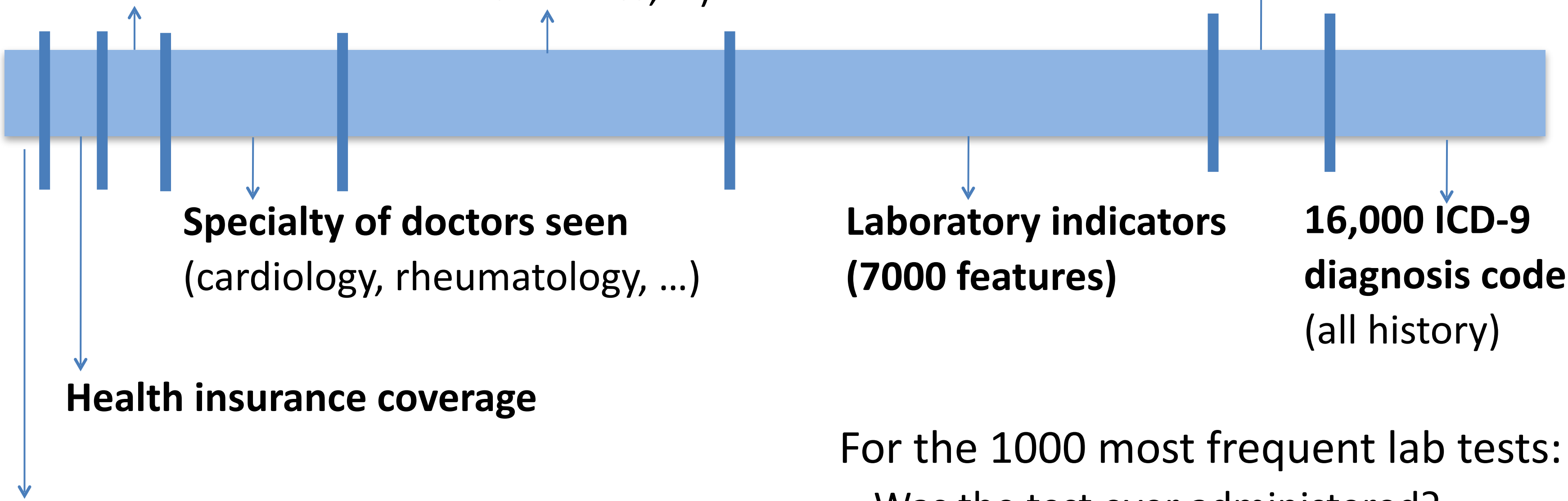
| Lab test | |
|---|---|
| 2160-0 Creatinine | 1284737 |
| 3094-0 Urea nitrogen | 1282344 |
| 2823-3 Potassium | 1280812 |
| 2345-7 Glucose | 1299897 |
| 1742-6 Alanine aminotransferase | 1187809 |
| 1920-8 Aspartate aminotransferase | 1187965 |
| 2885-2 Protein | 1277338 |
| 1751-7 Albumin | 1274166 |
| 2093-3 Cholesterol | 1268269 |
| 2571-8 Triglyceride | 1257751 |
| 13457-7 Cholesterol.in LDL | 1241208 |
| 17861-6 Calcium | 1165370 |
| 2951-2 Sodium | 1167675 |

| Lab test | |
|---|---|
| 2085-9 Cholesterol.in HDL | 1155666 |
| 718-7 Hemoglobin | 1152726 |
| 4544-3 Hematocrit | 1147893 |
| 9830-1 Cholesterol.total/Cholesterol.in HDL | 1037730 |
| 33914-3 Glomerular filtration rate/1.73 sq M.predicted | 561309 |
| 785-6 Erythrocyte mean corpuscular hemoglobin | 1070832 |
| 6690-2 Leukocytes | 1062980 |
| 789-8 Erythrocytes | 1062445 |
| 787-2 Erythrocyte mean corpuscular volume | 1063665 |

| Lab test | |
|---|---|
| 770-8 Neutrophils/100 leukocytes | 952089 |
| 731-0 Lymphocytes | 943918 |
| 704-7 Basophils | 863448 |
| 711-2 Eosinophils | 935710 |
| 5905-5 Monocytes/100 leukocytes | 943764 |
| 706-2 Basophils/100 leukocytes | 863435 |
| 751-8 Neutrophils | 943232 |
| 742-7 Monocytes | 942978 |
| 713-8 Eosinophils/100 leukocytes | 933929 |
| 3016-3 Thyrotropin | 891807 |
| 4548-4 Hemoglobin A1c/Hemoglobin.total | 527062 |

**Count of the test result (ever)**

# Encoding the longitudinal health data



**Service place** (urgent care, inpatient, outpatient, …)

**Medications taken (999 features)** (laxatives, metformin, anti-arthritics, …)

**Procedures performed (457 features)**

**Specialty of doctors seen** (cardiology, rheumatology, …)

**Laboratory indicators (7000 features)**

**16,000 ICD-9 diagnosis codes** (all history)

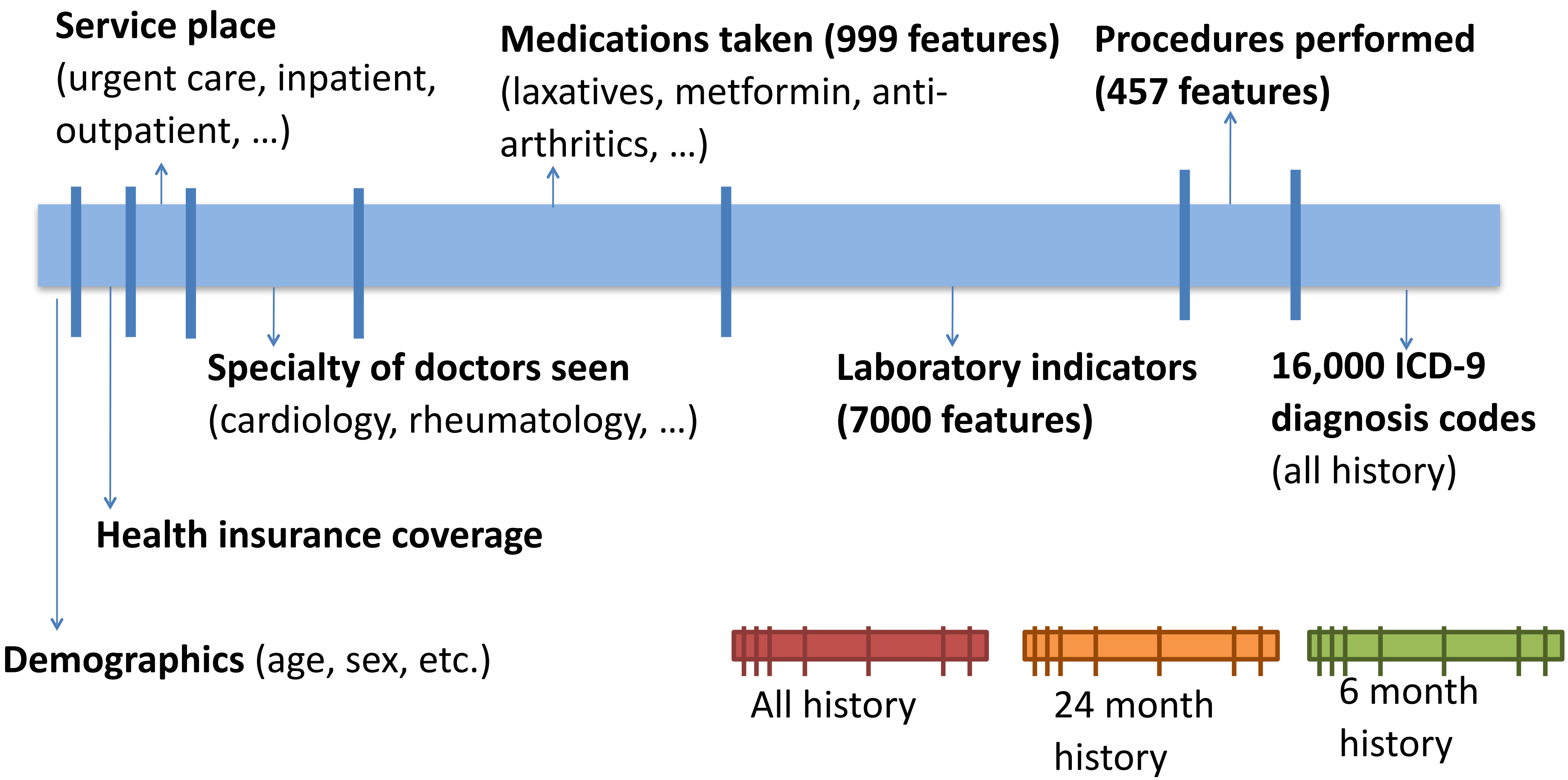
**Health insurance coverage**

**Demographics** (age, sex, etc.)

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?

# Encoding the longitudinal health data



**Service place**
(urgent care, inpatient, outpatient, …)

**Medications taken (999 features)**
(laxatives, metformin, anti-arthritics, …)

**Procedures performed (457 features)**

**Specialty of doctors seen**
(cardiology, rheumatology, …)

**Laboratory indicators (7000 features)**

**16,000 ICD-9 diagnosis codes**
(all history)

**Health insurance coverage**

**Demographics** (age, sex, etc.)

All history

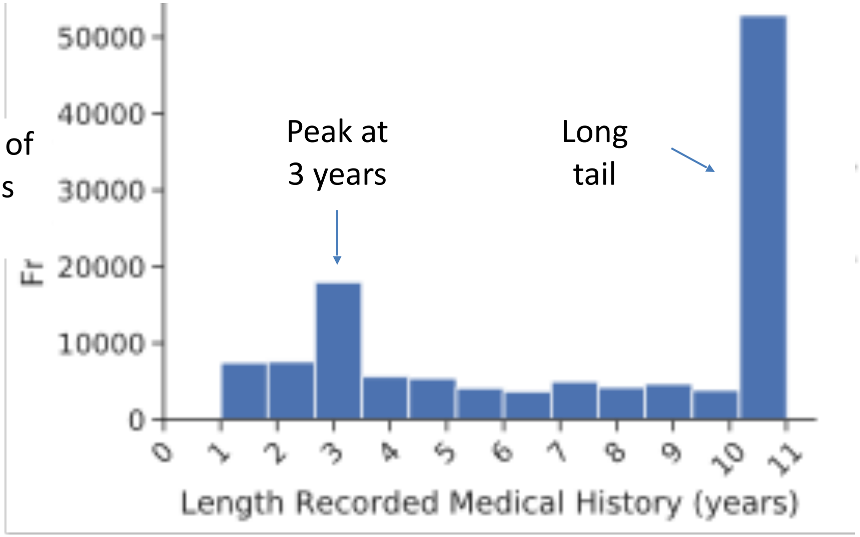
24 month history

6 month history

**10s-100s of thousands of features**

# There may be a varying amount of history per patient

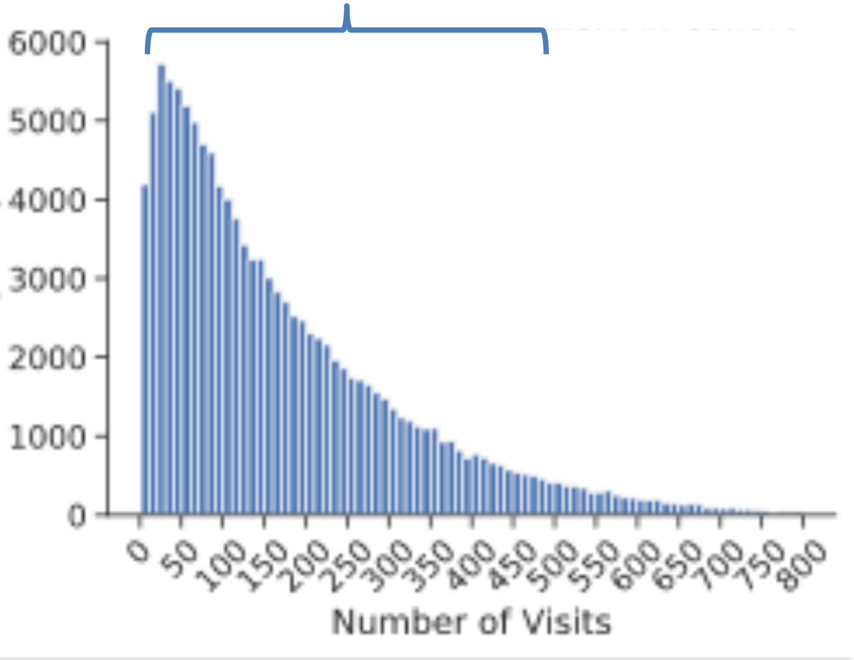
Number of patients



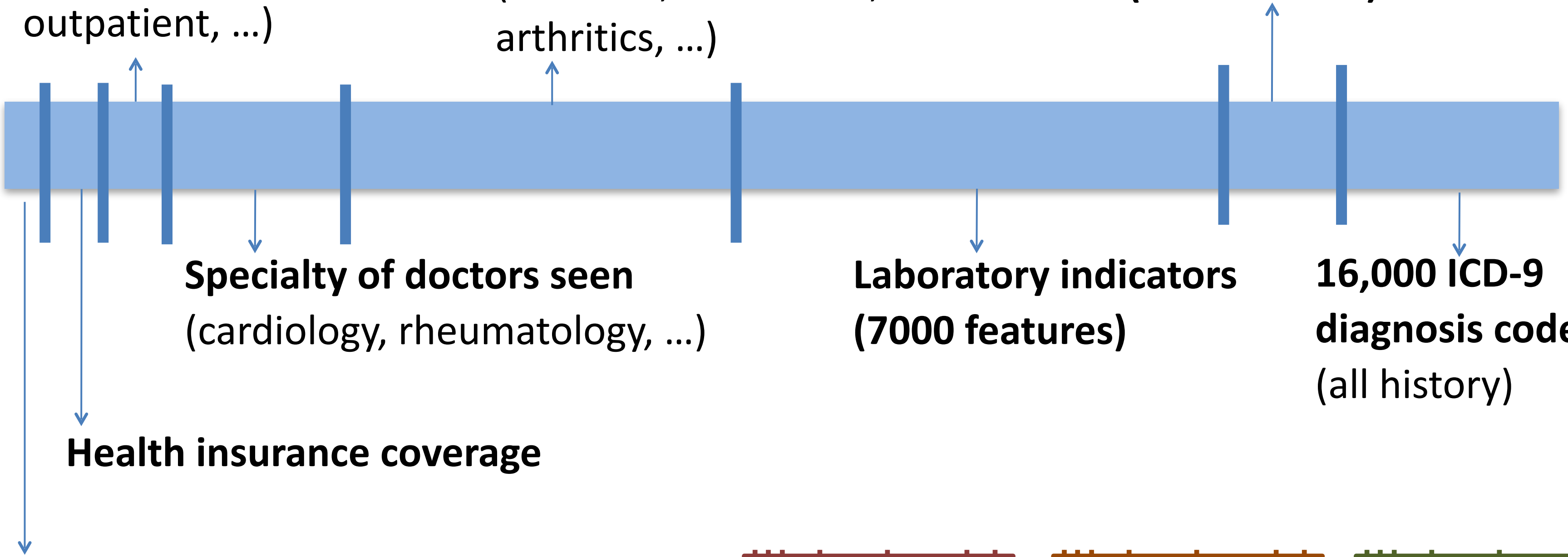
Peak at 3 years

Long tail

Truncate patient's history to include only 512 most recent visits



# Encoding the longitudinal health data



**Service place** (urgent care, inpatient, outpatient, …)

**Medications taken (999 features)** (laxatives, metformin, anti-arthritics, …)

**Procedures performed (457 features)**

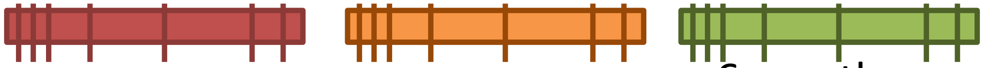
**Specialty of doctors seen** (cardiology, rheumatology, …)

**Laboratory indicators (7000 features)**

**16,000 ICD-9 diagnosis codes** (all history)

**Health insurance coverage**

**Demographics** (age, sex, etc.)



All history

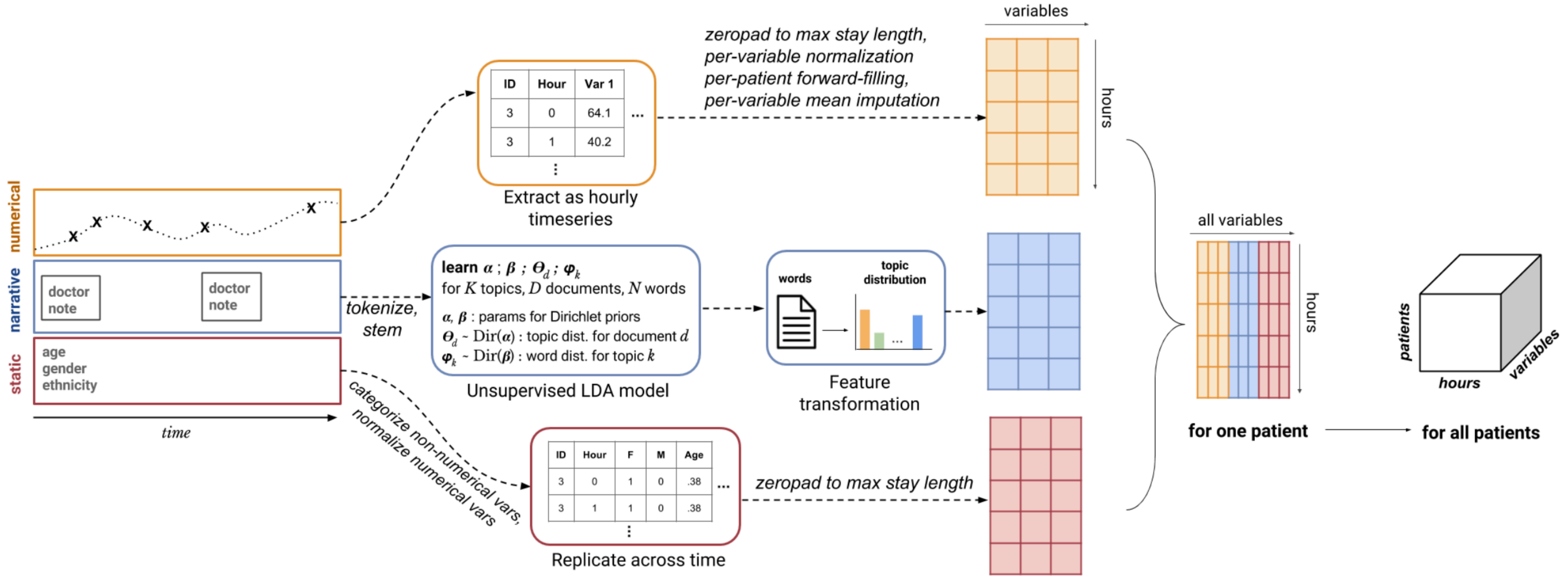
24 month history

6 month history

How does this deal with missing data? What are its limitations?

# Combining Multi-Modal Data

## ML approach


numerical
narrative
static
doctor note
doctor note
age
gender
ethnicity
time
Extract as hourly timeseries
zeropad to max stay length,
per-variable normalization
per-patient forward-filling,
per-variable mean imputation
variables
hours
tokenize, stem
learn $\alpha$ ; $\beta$ ; $\Theta_d$ ; $\varphi_k$
for $K$ topics, $D$ documents, $N$ words
$\alpha$, $\beta$ : params for Dirichlet priors
$\Theta_d$ ~ $\mathrm{Dir}(\alpha)$ : topic dist. for document $d$
$\varphi_k$ ~ $\mathrm{Dir}(\beta)$ : word dist. for topic $k$
Unsupervised LDA model
words
topic distribution
...
Feature transformation
categorize non-numerical vars,
normalize numerical vars
zeropad to max stay length
Replicate across time
all variables
hours
for one patient
for all patients
patients
hours
variables

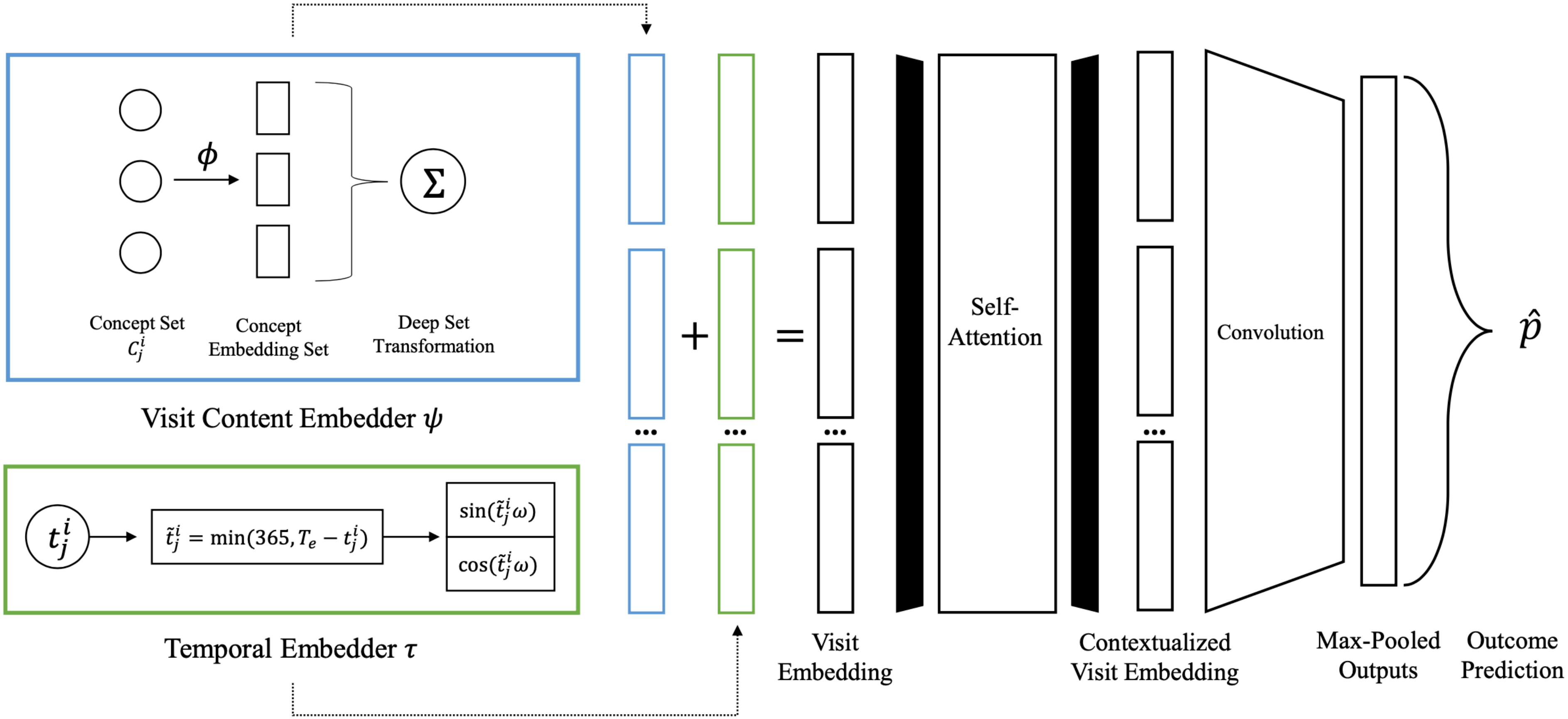
| ID | Hour | Var 1 |
|---|---|---|
| 3 | 0 | 64.1 |
| 3 | 1 | 40.2 |

| ID | Hour | F | M | Age |
|---|---|---|---|---|
| 3 | 0 | 1 | 0 | .38 |
| 3 | 1 | 1 | 0 | .38 |

**Figure 1: A visual representation of the data used. 1) *Numerical data*, including vitals and lab tests. The timestamp for each data point is rounded to the nearest hour, and hours with multiple measurements for a variable are assigned the average of those measurements. Each measurement is normalized according to the min and max for that var and each patient's data are zero-padded to the maximum stay length (240 hours). To fill in missing values, we forward-fill values for each patient, and mean-impute for any remaining missing values. 2) *Narrative data*, which consists of unstructured text notes. After preprocessing, LDA is used to obtain underlying topics and we then represent each note as a distribution over these topics. We forward-fill and aggregate these topic vectors across time, mean-imputing any values that are still missing. 3) *Static Data*, including variables recorded at admission such as sex, age, and ethnicity. Categorical variables such as ethnicity and ICU type are transformed into one-hot vectors containing each possible type. We replicate this data across time so that we are able to feed in this information at every timestep. We normalize numerical values and use forward-filling and imputation as before.**

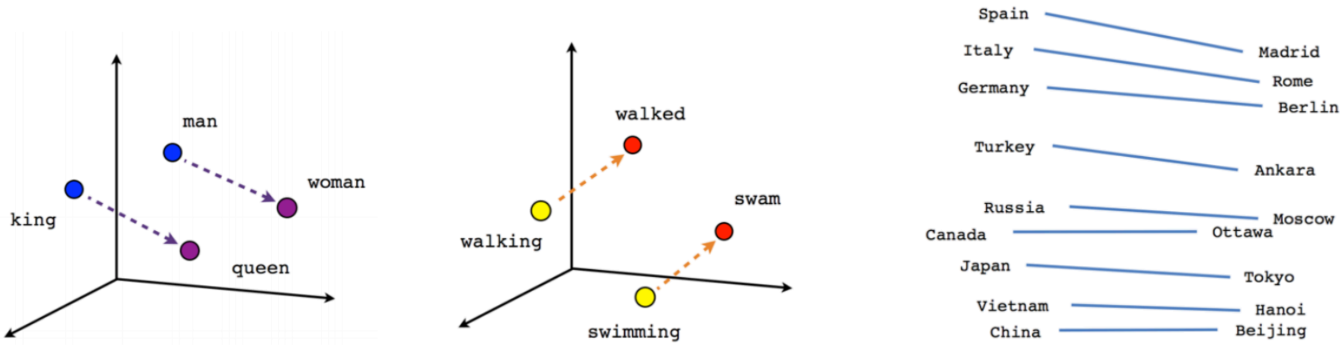
Suresh H, Hunt N, Johnson A, Celi LA, Szolovits P, Ghassemi M. Clinical intervention prediction and understanding with deep neural networks. In: mlhc2017 [Internet]. 2017. p. 1–16. Available from: https://arxiv.org/abs/1705.08498

# Alternative encoding using self-attention / transformers



Li et al., *BEHRT: Transformer for Electronic Health Records*, Scientific Reports ‘20
Kodialam et al., *Deep Contextual Clinical Prediction with Reverse Distillation*, AAAI ‘21

# The latter can make use of unsupervised learning of concept embeddings



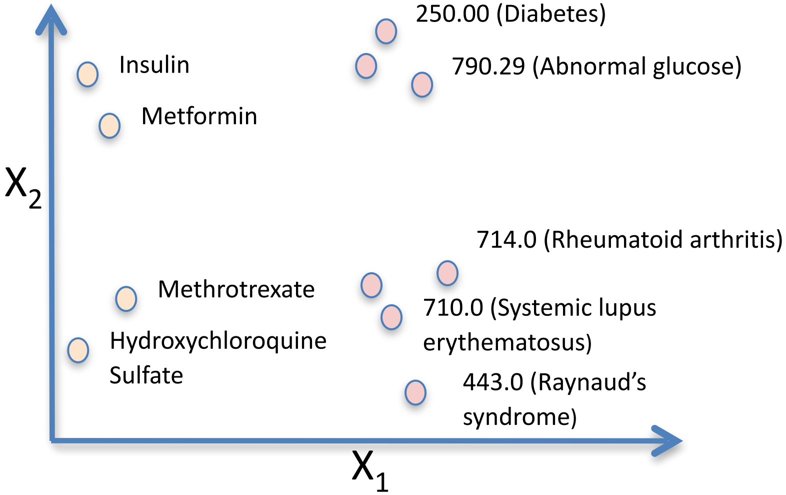
Male-Female

Verb tense

Country-Capital

Mikolov et al., Efficient Estimation of Word Representations in Vector Space, ICLR '13

Figure: https://cbail.github.io/textasdata/word2vec/rmarkdown/word2vec.html



250.00 (Diabetes)
Insulin
790.29 (Abnormal glucose)
Metformin
X2
714.0 (Rheumatoid arthritis)
Methrotrexate
710.0 (Systemic lupus erythematosus)
Hydroxychloroquine Sulfate
443.0 (Raynaud's syndrome)
X1

Choi, Chiu, Sontag. Learning low-dimensional representations of medical concepts. AMIA Summits on Translational Science Proceedings, '16
https://github.com/clinicalml/embeddings

Beam et al., Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. PSB '20

# Outline for today's class

1. Introduction to risk stratification
2. Case study: Early detection of Type 2 diabetes
   - Encoding longitudinal structured health data
3. **Framing as supervised learning problem**
   - Deriving labels from EHR

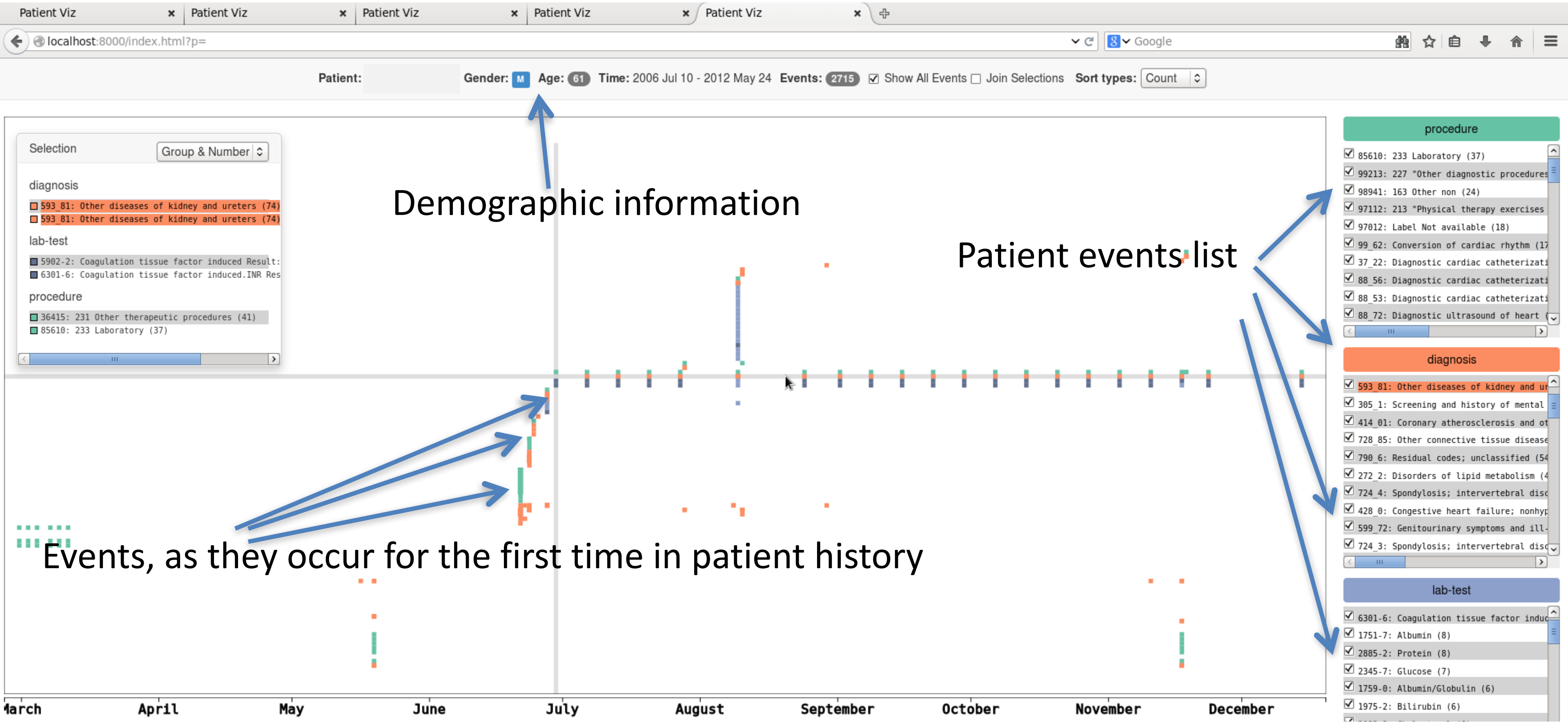
# Where do the labels come from?

Typical pipeline:

1. Manually label several patients’ data by “chart review”
2. A) Come up with a simple rule to automatically derive label for all patients, **or**

   B) Use machine learning to get the labels themselves

# Step 1:
# Visualization of individual patient data is an important part of chart review



**Patient:** **Gender:** M **Age:** 61 **Time:** 2006 Jul 10 - 2012 May 24 **Events:** 2715 ☑ Show All Events ☐ Join Selections **Sort types:** Count

Selection Group & Number

diagnosis

🟧 593_81: Other diseases of kidney and ureters (74)
🟧 593_81: Other diseases of kidney and ureters (74)

lab-test

■ 5902-2: Coagulation tissue factor induced Result:
■ 6301-6: Coagulation tissue factor induced.INR Res

procedure

🟩 36415: 231 Other therapeutic procedures (41)
🟩 85610: 233 Laboratory (37)

procedure

☑ 85610: 233 Laboratory (37)
☑ 99213: 227 "Other diagnostic procedures
☑ 98941: 163 Other non (24)
☑ 97112: 213 "Physical therapy exercises
☑ 97012: Label Not available (18)
☑ 99_62: Conversion of cardiac rhythm (17
☑ 37_22: Diagnostic cardiac catheterizati
☑ 88_56: Diagnostic cardiac catheterizati
☑ 88_53: Diagnostic cardiac catheterizati
☑ 88_72: Diagnostic ultrasound of heart (

diagnosis

☑ 593_81: Other diseases of kidney and ur
☑ 305_1: Screening and history of mental
☑ 414_01: Coronary atherosclerosis and ot
☑ 728_85: Other connective tissue disease
☑ 790_6: Residual codes; unclassified (54
☑ 272_2: Disorders of lipid metabolism (4
☑ 724_4: Spondylosis; intervertebral disc
☑ 428_0: Congestive heart failure; nonhyp
☑ 599_72: Genitourinary symptoms and ill-
☑ 724_3: Spondylosis; intervertebral disc

lab-test

☑ 6301-6: Coagulation tissue factor induc
☑ 1751-7: Albumin (8)
☑ 2885-2: Protein (8)
☑ 2345-7: Glucose (7)
☑ 1759-0: Albumin/Globulin (6)
☑ 1975-2: Bilirubin (6)

Demographic information

Patient events list

Events, as they occur for the first time in patient history

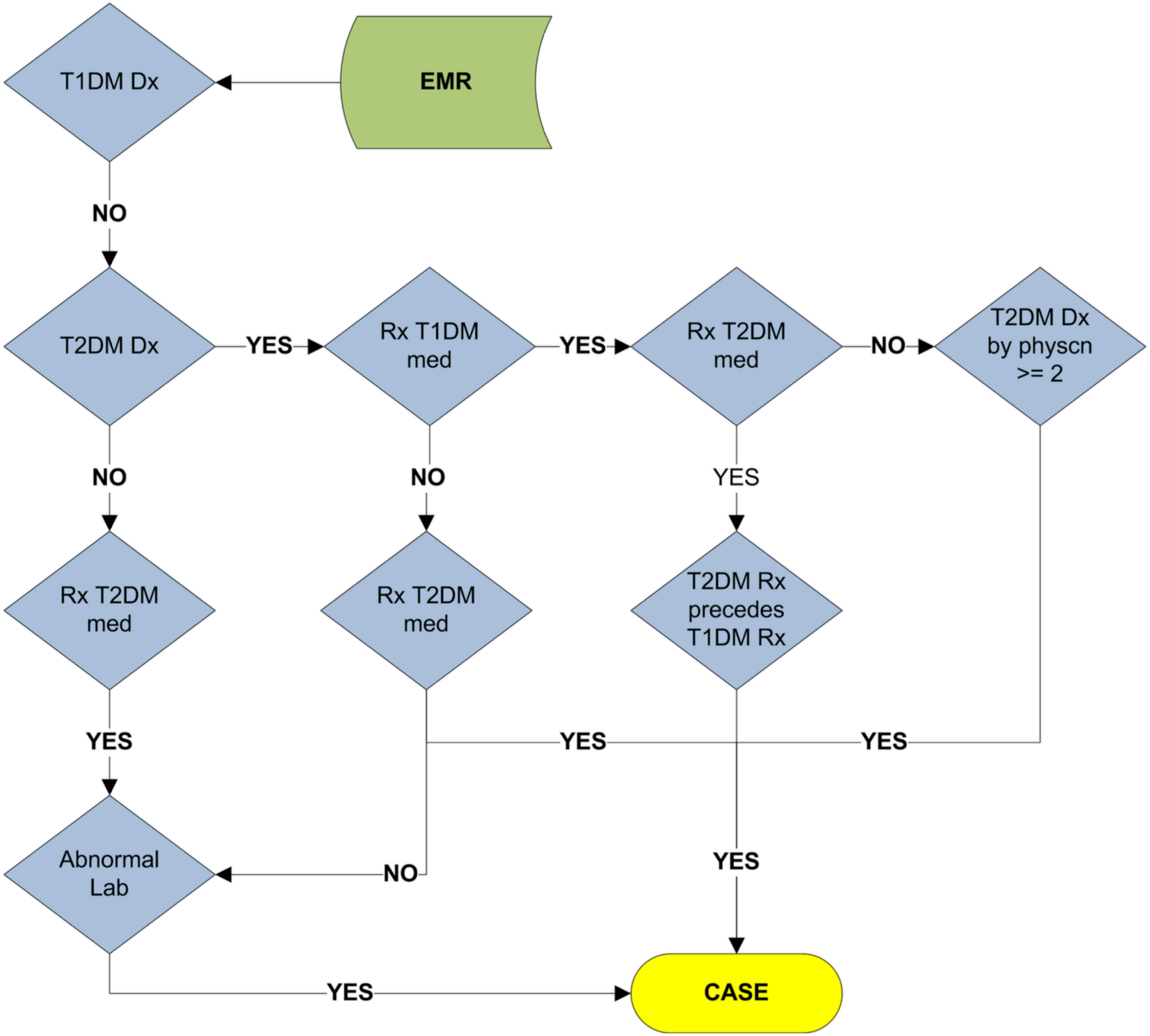
https://github.com/nyuvis/patient-viz

https://github.com/BenGlicksberg/PatientExploreR

# Step 2: Example of a rule-based phenotype

(Northwestern U.)

Figure 1: Algorithm for identifying T2DM cases in the EMR.



Source: https://phekb.org/sites/phenotype/files/T2DM-algorithm.pdf

# Step 2: Example of a rule-based phenotype

Coverage of Different Diabetes Outcome Definitions on Claims Data

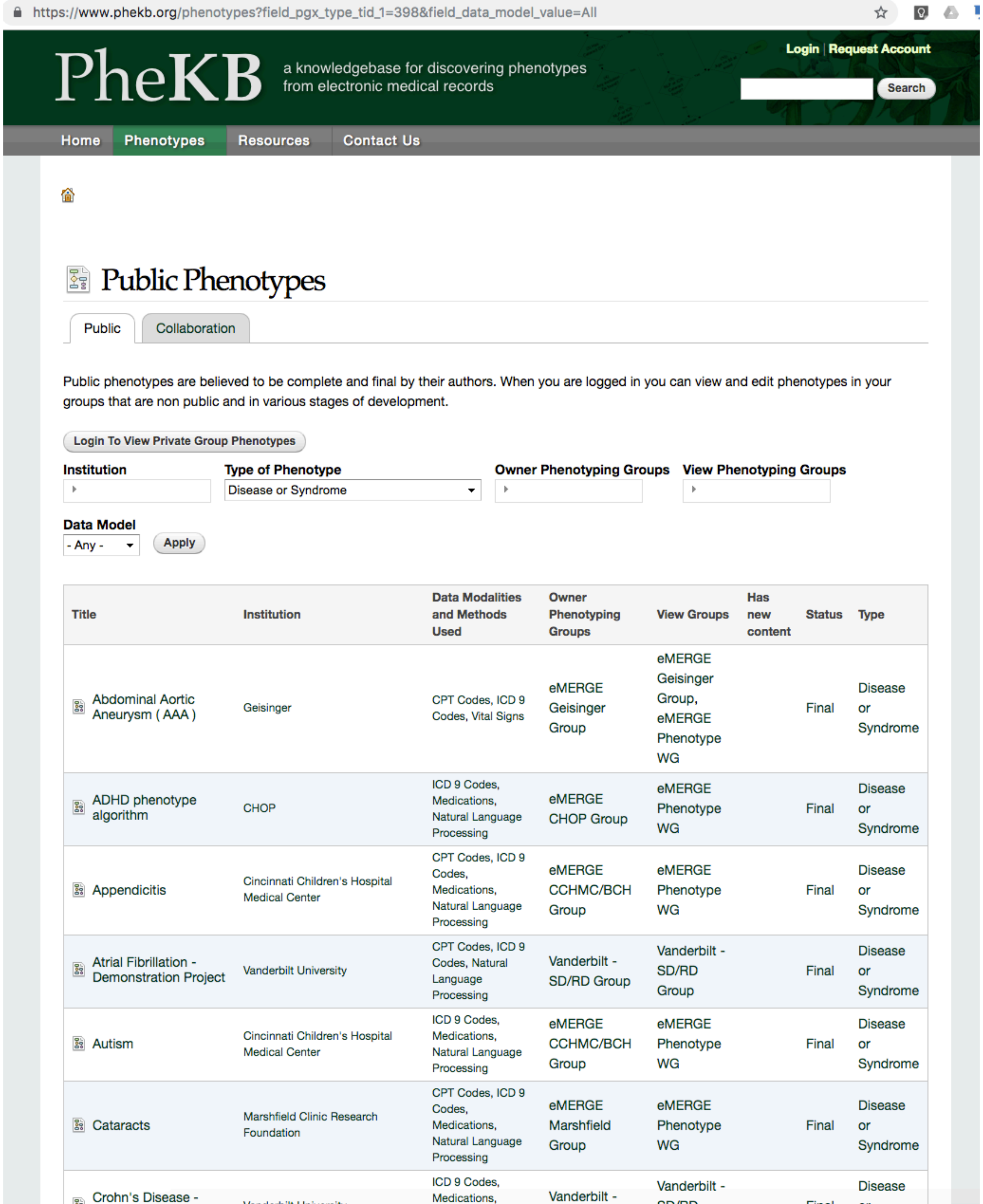
| Condition | Percentage |
|---|---|
| Have 250.x diagnosis, or have been on diabetic medication, or have any HbA1c ≥ 6.5 | **100 %** |
| Have been diagnosed 250.xx | **89.9 %** |
| Have been on diabetic medications | **15.0 %** |
| Have HbA1c values ≥ 6.5 | **20.9 %** |
| Have 250.xx diagnosis on more than one distinct date | **40.0 %** |
| (Have 250.xx diagnosis, or have been on diabetic medication, or have any HbA1c ≥ 6.5) on more than one distinct date | **44.0 %** |
| (Have 250.xx diagnosis, or have been on diabetic medication, or have any HbA1c ≥ 6.5) on two dates separated by at least a week | **41.1 %** |

The earliest date the rule triggers is defined as the date of diabetes diagnosis

Definition selected

[Razavian, Blecker, Schmidt, Smith-McLallen, Nigam, Sontag. *Big Data.* ‘16]

# Step 2: Example of a rule-based phenotype



https://www.phekb.org/phenotypes?field_pgx_type_tid_1=398&field_data_model_value=All

PheKB a knowledgebase for discovering phenotypes from electronic medical records

Login | Request Account

Search

Home Phenotypes Resources Contact Us

## Public Phenotypes

Public Collaboration

Public phenotypes are believed to be complete and final by their authors. When you are logged in you can view and edit phenotypes in your groups that are non public and in various stages of development.

Login To View Private Group Phenotypes

Institution:

Type of Phenotype: Disease or Syndrome

Owner Phenotyping Groups:

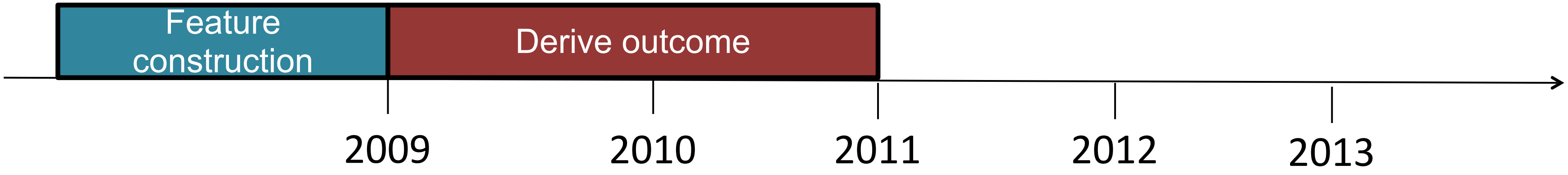
View Phenotyping Groups:

Data Model: - Any -

Apply

| Title | Institution | Data Modalities and Methods Used | Owner Phenotyping Groups | View Groups | Has new content | Status | Type |
|---|---|---|---|---|---|---|---|
| Abdominal Aortic Aneurysm ( AAA ) | Geisinger | CPT Codes, ICD 9 Codes, Vital Signs | eMERGE Geisinger Group | eMERGE Geisinger Group, eMERGE Phenotype WG | | Final | Disease or Syndrome |
| ADHD phenotype algorithm | CHOP | ICD 9 Codes, Medications, Natural Language Processing | eMERGE CHOP Group | eMERGE Phenotype WG | | Final | Disease or Syndrome |
| Appendicitis | Cincinnati Children's Hospital Medical Center | CPT Codes, ICD 9 Codes, Medications, Natural Language Processing | eMERGE CCHMC/BCH Group | eMERGE Phenotype WG | | Final | Disease or Syndrome |
| Atrial Fibrillation - Demonstration Project | Vanderbilt University | CPT Codes, ICD 9 Codes, Natural Language Processing | Vanderbilt - SD/RD Group | Vanderbilt - SD/RD Group | | Final | Disease or Syndrome |
| Autism | Cincinnati Children's Hospital Medical Center | ICD 9 Codes, Medications, Natural Language Processing | eMERGE CCHMC/BCH Group | eMERGE Phenotype WG | | Final | Disease or Syndrome |
| Cataracts | Marshfield Clinic Research Foundation | CPT Codes, ICD 9 Codes, Medications, Natural Language Processing | eMERGE Marshfield Group | eMERGE Phenotype WG | | Final | Disease or Syndrome |
| Crohn's Disease - | Vanderbilt University | ICD 9 Codes, Medications, | Vanderbilt - | Vanderbilt - SD/RD | | Final | Disease or |

# Framing for supervised machine learning



Exclusion criteria:

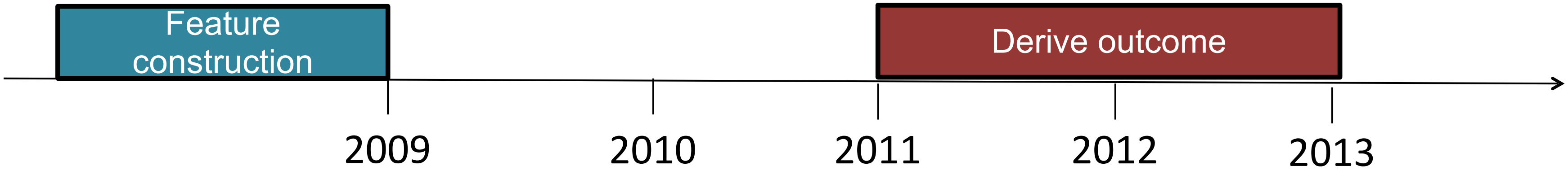
- Diabetes diagnosis (according to our rule) observed prior to January 1, 2009
- Less than 6 months of enrollment in feature construction window
- Member left health insurance prior to Jan. 1, 2011

What if someone is diagnosed with diabetes in 2012?

Why not model as “patient develops diabetes anytime after 2009”?

[Razavian, Blecker, Schmidt, Smith-McLallen, Nigam, Sontag. *Big Data.* ‘16]

# Framing for supervised machine learning

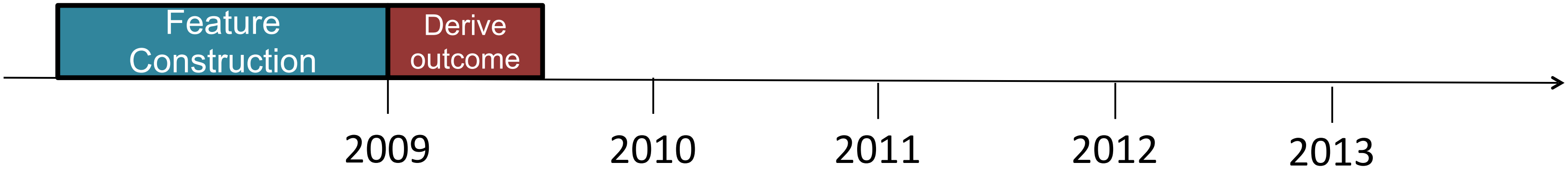


Exclusion criteria:

- Diabetes diagnosis (according to our rule) observed prior to January 1, ~~2009~~ 2011
- Less than 6 months of enrollment in feature construction window
- Member left health insurance prior to Jan. 1, ~~2011~~ 2013

[Razavian, Blecker, Schmidt, Smith-McLallen, Nigam, Sontag. *Big Data.* ‘16]

# Framing for supervised machine learning

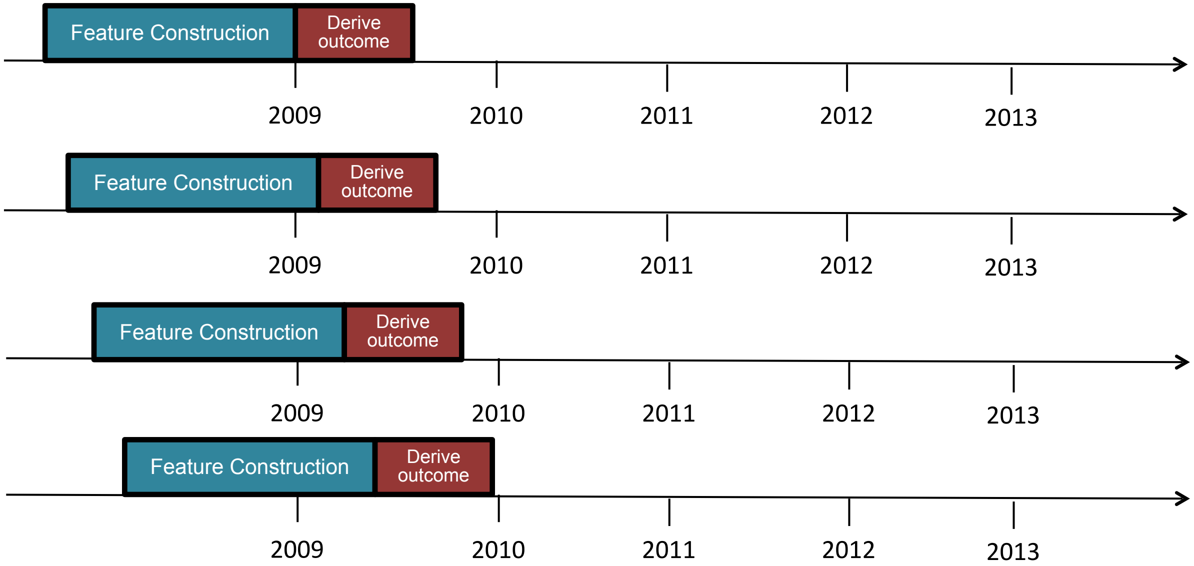


- Suppose we want to run the above model in August 2009. It may not have good performance due to *non-stationarity* in the data

- We now have data through 2021. Using a fixed prediction time / index date of Jan. 1, 2009 is ignoring most of the diabetes onsets!

# Framing for supervised machine learning

- We can instead create ***many*** data points from each patient, using e.g. every month as an index date:



- **Important:** If multiple data points per patient, make sure each patient's data is in *only* train, validate, or test