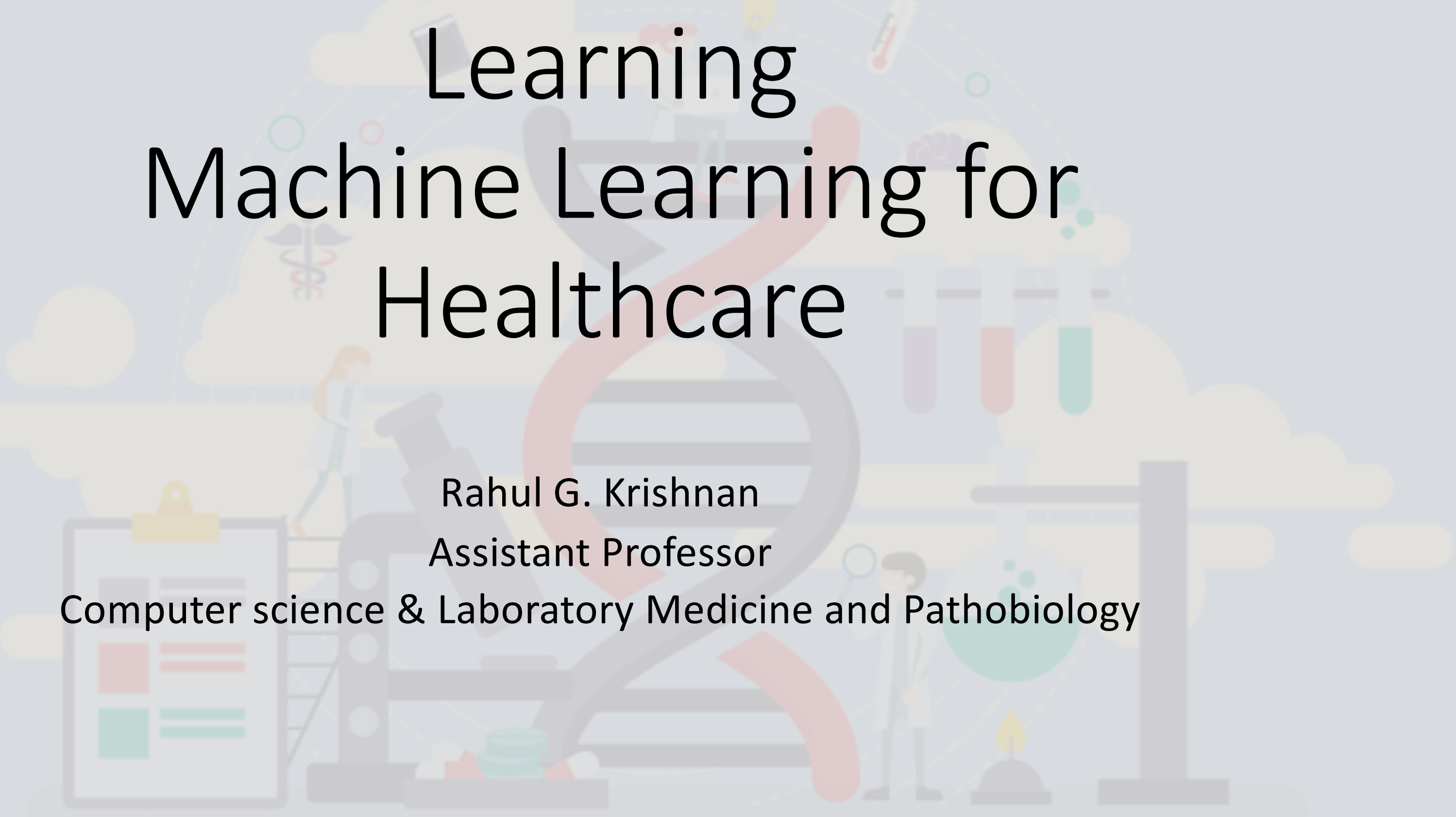
UNIVERSITY OF TORONTO

VECTOR INSTITUTE

Topics in Machine Learning
Machine Learning for Healthcare



Rahul G. Krishnan
Assistant Professor
Computer science & Laboratory Medicine and Pathobiology

Outline

- Recap from last week, supervised machine learning [8 mins]
- Risk stratification: [35 minutes]
  - Stratification as a prediction problem
  - **Case study:** Predicting the onset of diabetes
- Summary and sneak peek of next week [7 mins]

Course grades

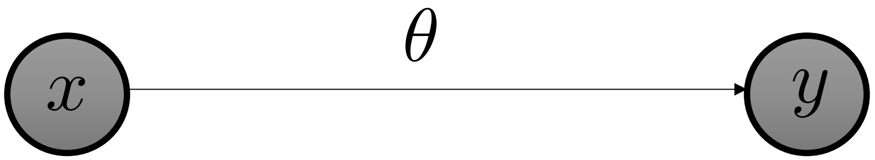
- Individual
  - 5% class participation (attendance and engagement)
  - 15% assignment
    - Paper deconstruction: Summarize four papers of your choosing: highlight the key ideas, what makes them tick, why you think they work and how they could be improved
  - 15% paper presentation:
    - Present (in pairs) one of the papers from the theme of the week,
    - Sign up for papers on a first come first serve basis,
- Groups
  - 10% project proposal
  - 15% project presentation
  - 40% course project report

Questions about course structure or grading
[2 min pause]

Announcements

- Reading list for the first six weeks of class is now available
  - Please do the readings for the class ahead of the week, they will introduce you to research areas and the lectures during the week will provide more context for the theme
- Complete the quiz if you have not done so already

Supervised learning – (1)



$$p(y|x)$$

Patient features

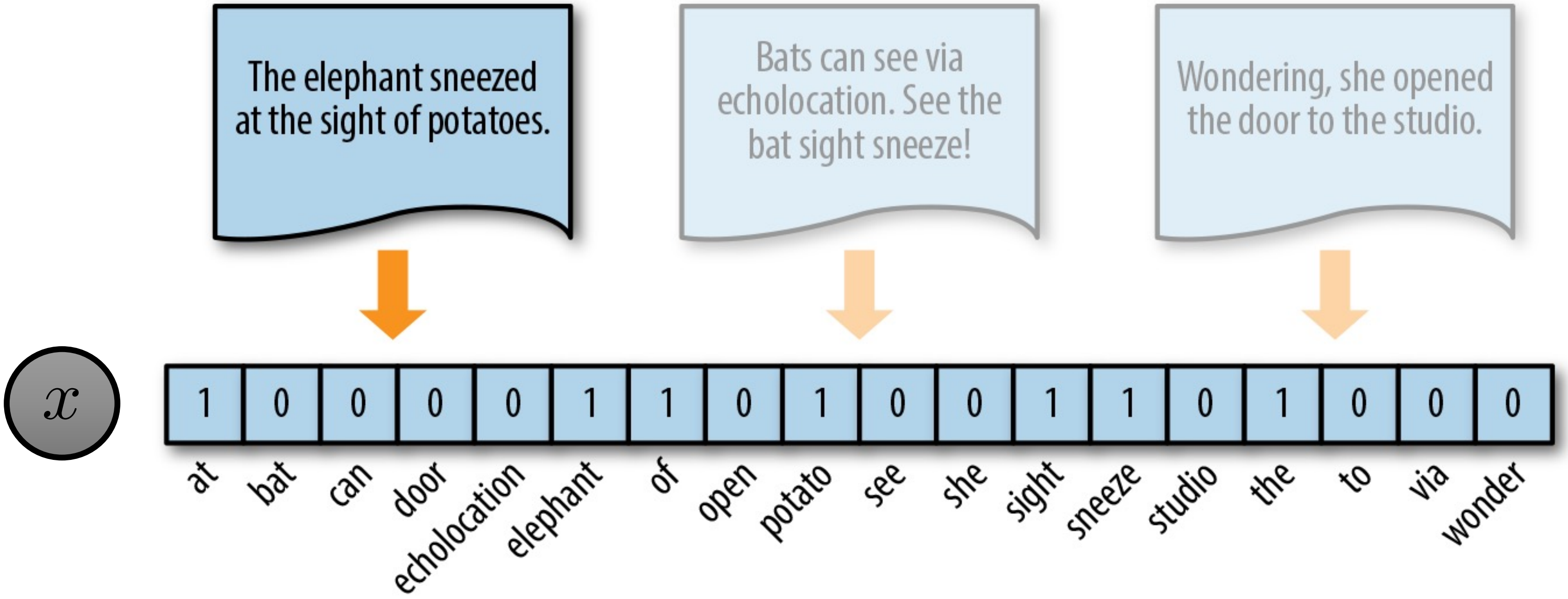
Chest X-ray image

Time to readmission

pneumonia/pneumothorax

- Step 1: Collect a dataset or curate a subset of data with labels from an existing dataset
- Step 2: Learn the model using the dataset
- Step 3: Use the output of the model to build software to help clinicians reach better decisions, faster.
- **Examples**: Logistic regression, random forests, XGBoost, Deep neural networks

Vectorization – Text data



The elephant sneezed at the sight of potatoes.

Bats can see via echolocation. See the bat sight sneeze!

Wondering, she opened the door to the studio.

x

at	bat	can	door	echolocation	elephant	of	open	potato	see	she	sight	sneeze	studio	the	to	via	wonder
1	0	0	0	0	1	1	0	1	0	0	1	1	0	1	0	0	0

Vectorization – Clinical variables in tabular data

x

ICD10 - Diabetes	….	NDC code Metformin	….	CPT – Surgery, Aortic Valve
2	….	4	….	1

Vectorization – Image data

x

Pixel position [0,0]	….	Pixel position [32,15]	….	Pixel position [255,255]
1	….	0	….	1

y Defining labels is challenging

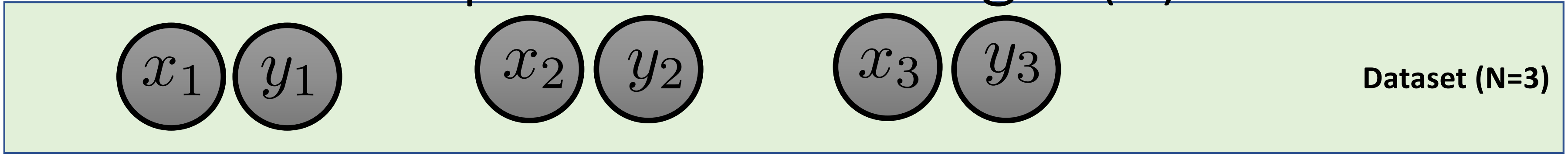
- Examples:
  - Binary: Does the patient die or not
  - Real-valued: When does the patient die
  - Set-valued: The set of complications that a patient has
- Unlike domains such as computer vision, NLP, the true labels in healthcare can be **very** noisy
- We will discuss several kinds of noise in the upcoming lectures that require careful attention to detail

Supervised learning – (2)

- x: random variables
- y: outcome random variable
- θ: model parameters

- x typically high-dimensional [medical images, clinical variables]
- y (typically) low-dimensional [outcomes of interest]
- Model parameters depend on the functional class used:
  - Logistic regression: vector of weights
  - Decision tree: tree where each node is a feature to select and the value to threshold the feature on
  - Random Forest: collection of decision tree parameters

Supervised learning – (2)



x_1 y_1 x_2 y_2 x_3 y_3

Dataset (N=3)

- Given a dataset, the model parameters are learned via **maximum likelihood estimation**

$$\mathcal{L}(y, x) = \log p(y|x;\theta)$$

Score function (high is good, low is bad)

$$\theta = \arg\max_{\theta} \sum_{i=1}^{N} \mathcal{L}(y_i, x_i)$$

Solve this optimization problem to **learn** the model. Often formulated as a minimization of the negative of the log-likelihood function

Coarse-grained control

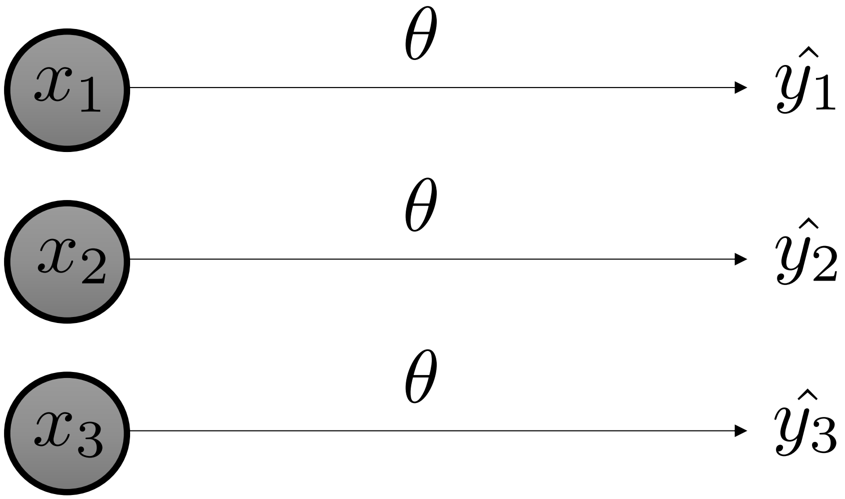
Fine-grained control

scikit learn

PyTorch

Supervised machine learning -- (3)

- The goal of a supervised model is good **generalization**
  - Predict well on data that it has not observed before



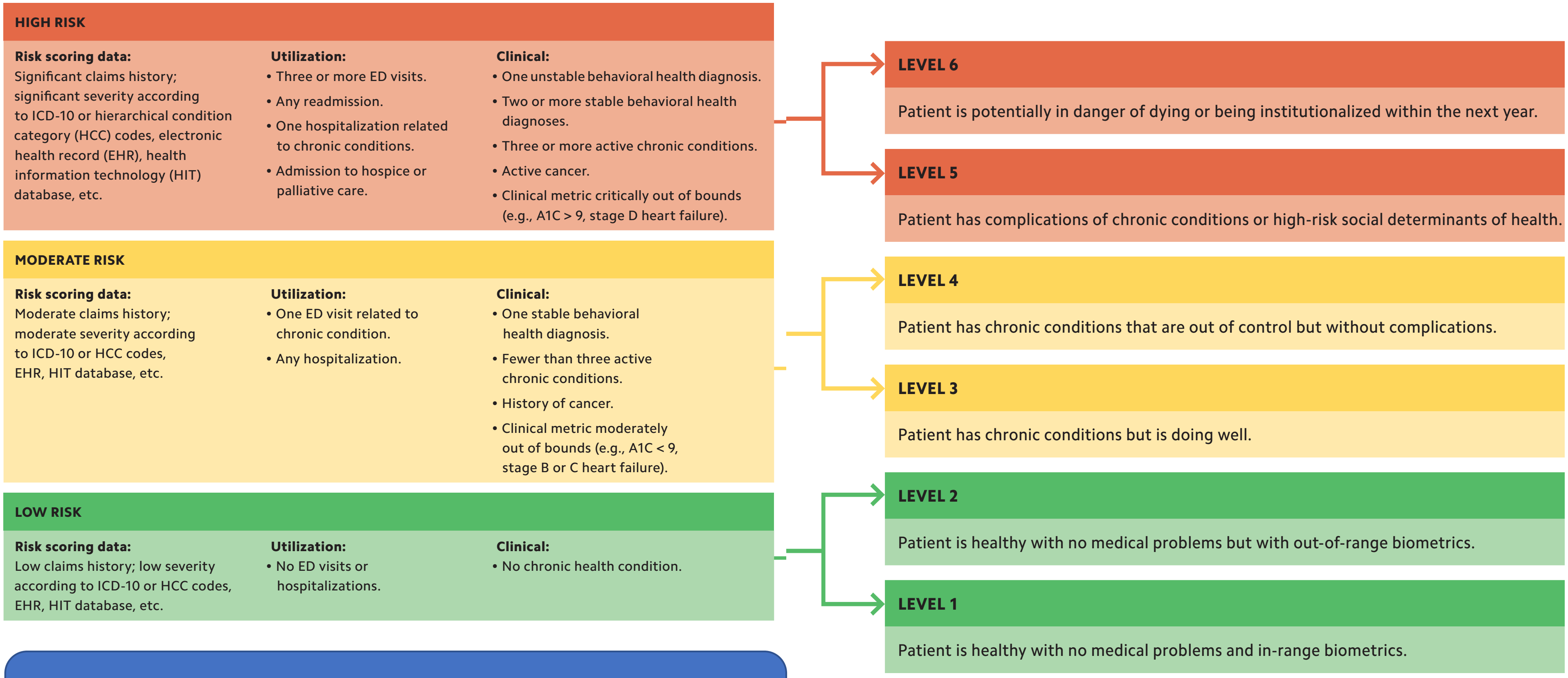
Questions?
[Unmute and ask, ask in chat]

There are no bad questions or questions with obvious answers.

Machine learning for risk stratification

- Clinical task whose goal is to separate patients into high-risk and low-risk of some outcome
- **What do you do with risk:**
  - Choice of interventions prescribed to the patients will vary based on risk
- **Why do this**:
  - Coarse-grained form of personalization
  - Direct clinician attention to patients who need it more

More efficient use of resources



HIGH RISK

Risk scoring data:
Significant claims history; significant severity according to ICD-10 or hierarchical condition category (HCC) codes, electronic health record (EHR), health information technology (HIT) database, etc.

Utilization:
- Three or more ED visits.
- Any readmission.
- One hospitalization related to chronic conditions.
- Admission to hospice or palliative care.

Clinical:
- One unstable behavioral health diagnosis.
- Two or more stable behavioral health diagnoses.
- Three or more active chronic conditions.
- Active cancer.
- Clinical metric critically out of bounds (e.g., A1C > 9, stage D heart failure).

MODERATE RISK

Risk scoring data:
Moderate claims history; moderate severity according to ICD-10 or HCC codes, EHR, HIT database, etc.

Utilization:
- One ED visit related to chronic condition.
- Any hospitalization.

Clinical:
- One stable behavioral health diagnosis.
- Fewer than three active chronic conditions.
- History of cancer.
- Clinical metric moderately out of bounds (e.g., A1C < 9, stage B or C heart failure).

LOW RISK

Risk scoring data:
Low claims history; low severity according to ICD-10 or HCC codes, EHR, HIT database, etc.

Utilization:
- No ED visits or hospitalizations.

Clinical:
- No chronic health condition.

LEVEL 6

Patient is potentially in danger of dying or being institutionalized within the next year.

LEVEL 5

Patient has complications of chronic conditions or high-risk social determinants of health.

LEVEL 4

Patient has chronic conditions that are out of control but without complications.

LEVEL 3

Patient has chronic conditions but is doing well.

LEVEL 2

Patient is healthy with no medical problems but with out-of-range biometrics.

LEVEL 1

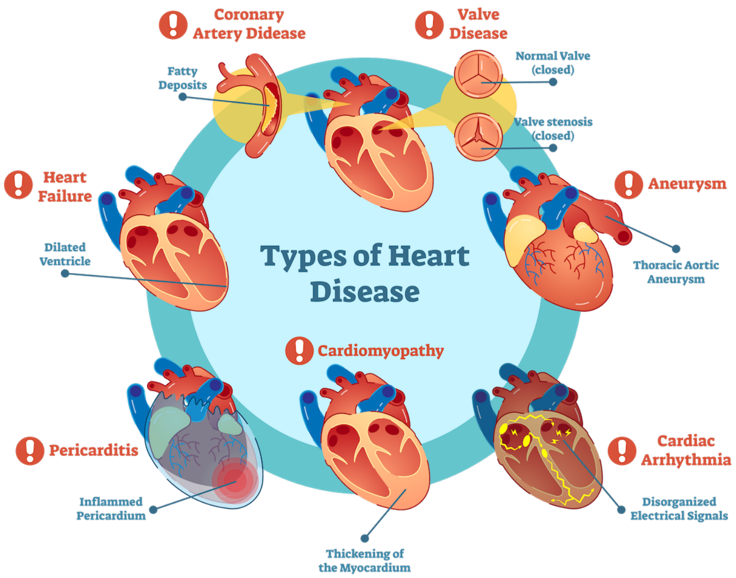
Patient is healthy with no medical problems and in-range biometrics.

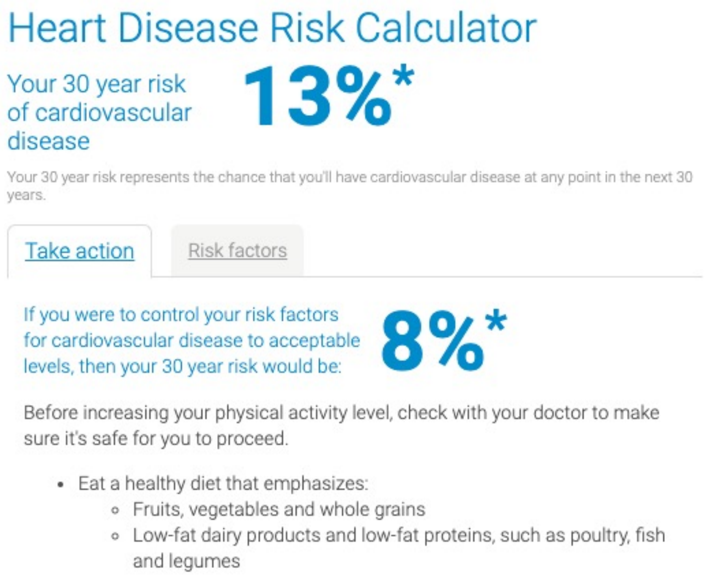
Risk groups via discussion or patients assigned to risk manually (e.g. during rounds)

Source: https://www.aafp.org/fpm/2019/0500/p21.html

Some risk scores are easy to estimate

- Heart disease risk score: https://www.mayoclinichealthsystem.org/locations/cannon-falls/services-and-treatments/cardiology/heart-disease-risk-calculator





Heart Disease Risk Calculator

Your 30 year risk of cardiovascular disease **13%***

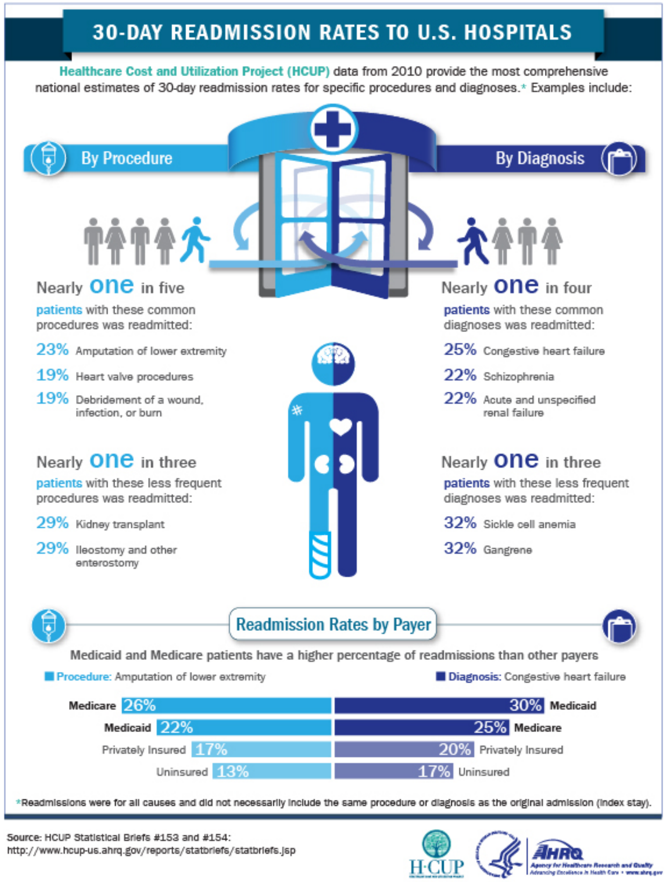
Your 30 year risk represents the chance that you'll have cardiovascular disease at any point in the next 30 years.

If you were to control your risk factors for cardiovascular disease to acceptable levels, then your 30 year risk would be: **8%***

Before increasing your physical activity level, check with your doctor to make sure it's safe for you to proceed.

- Eat a healthy diet that emphasizes:
  - Fruits, vegetables and whole grains
  - Low-fat dairy products and low-fat proteins, such as poultry, fish and legumes

Other risk scores are much harder



30-DAY READMISSION RATES TO U.S. HOSPITALS

Healthcare Cost and Utilization Project (HCUP) data from 2010 provide the most comprehensive national estimates of 30-day readmission rates for specific procedures and diagnoses.* Examples include:

By Procedure

Nearly one in five patients with these common procedures was readmitted:

- 23% Amputation of lower extremity
- 19% Heart valve procedures
- 19% Debridement of a wound, infection, or burn

Nearly one in three patients with these less frequent procedures was readmitted:

- 29% Kidney transplant
- 29% Ileostomy and other enterostomy

By Diagnosis

Nearly one in four patients with these common diagnoses was readmitted:

- 25% Congestive heart failure
- 22% Schizophrenia
- 22% Acute and unspecified renal failure

Nearly one in three patients with these less frequent diagnoses was readmitted:

- 32% Sickle cell anemia
- 32% Gangrene

Readmission Rates by Payer

Medicaid and Medicare patients have a higher percentage of readmissions than other payers

■ Procedure: Amputation of lower extremity		■ Diagnosis: Congestive heart failure	
Medicare	26%	30%	Medicaid
Medicaid	22%	25%	Medicare
Privately Insured	17%	20%	Privately Insured
Uninsured	13%	17%	Uninsured

*Readmissions were for all causes and did not necessarily include the same procedure or diagnosis as the original admission (index stay).

Source: HCUP Statistical Briefs #153 and #154:
http://www.hcup-us.ahrq.gov/reports/statbriefs/statbriefs.jsp

H·CUP
AHRQ
Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

Re-admissions are costly to the hospital and to the patient but difficult to detect.

How do I know what is particularly relevant for an increased risk of readmission?

Figure source: https://www.air.org/project/revolving-door-u-s-hospital-readmissions-diagnosis-and-procedure

Risk stratification as supervised learning

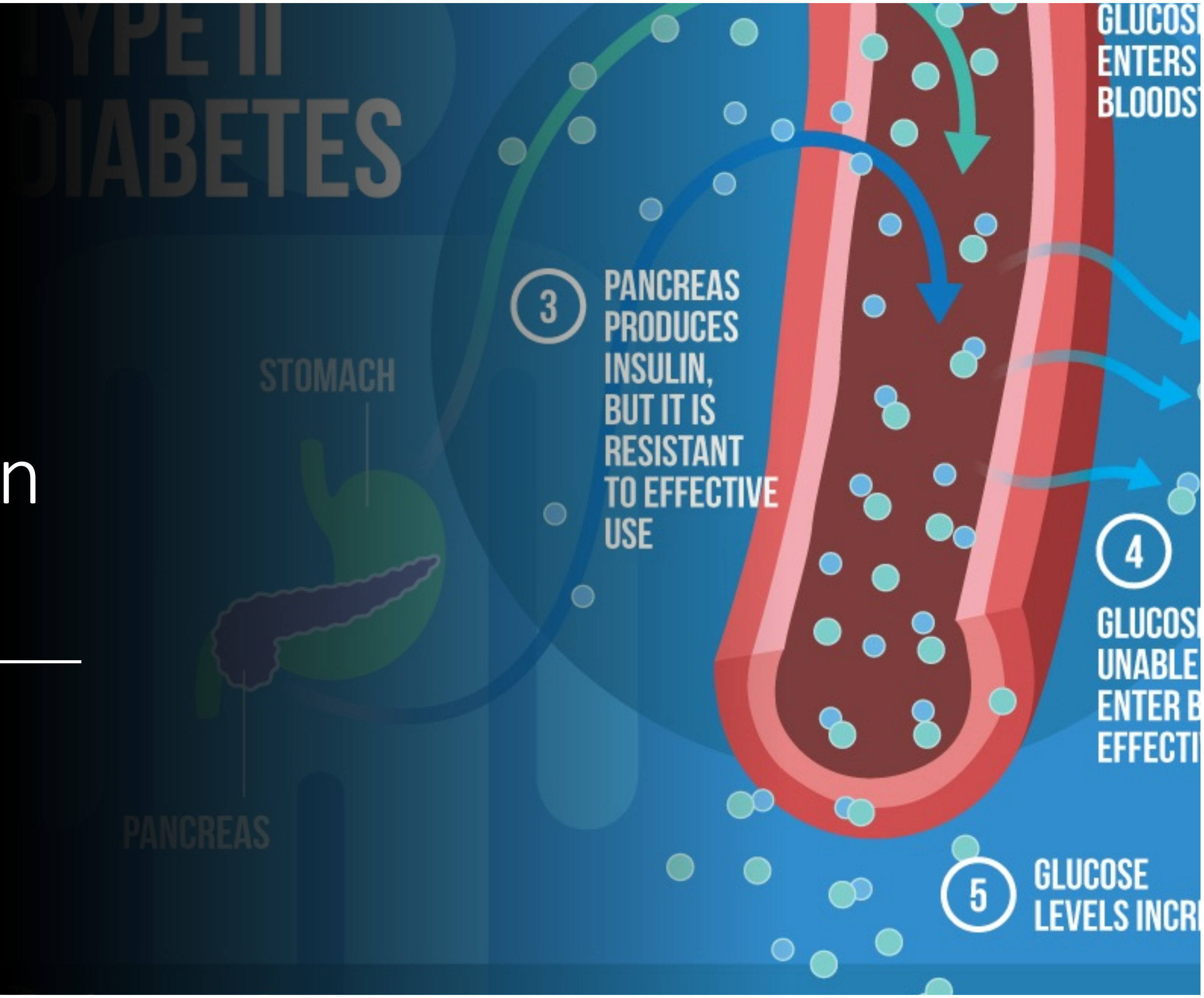
- **Key idea:**
  - Outcome [y] needs to be an adverse outcome (or strongly correlated with it)
  - Train a predictive model to predict **y** from clinical variable **x**
  - Threshold/bin the predicted outcome of the model to assess risk

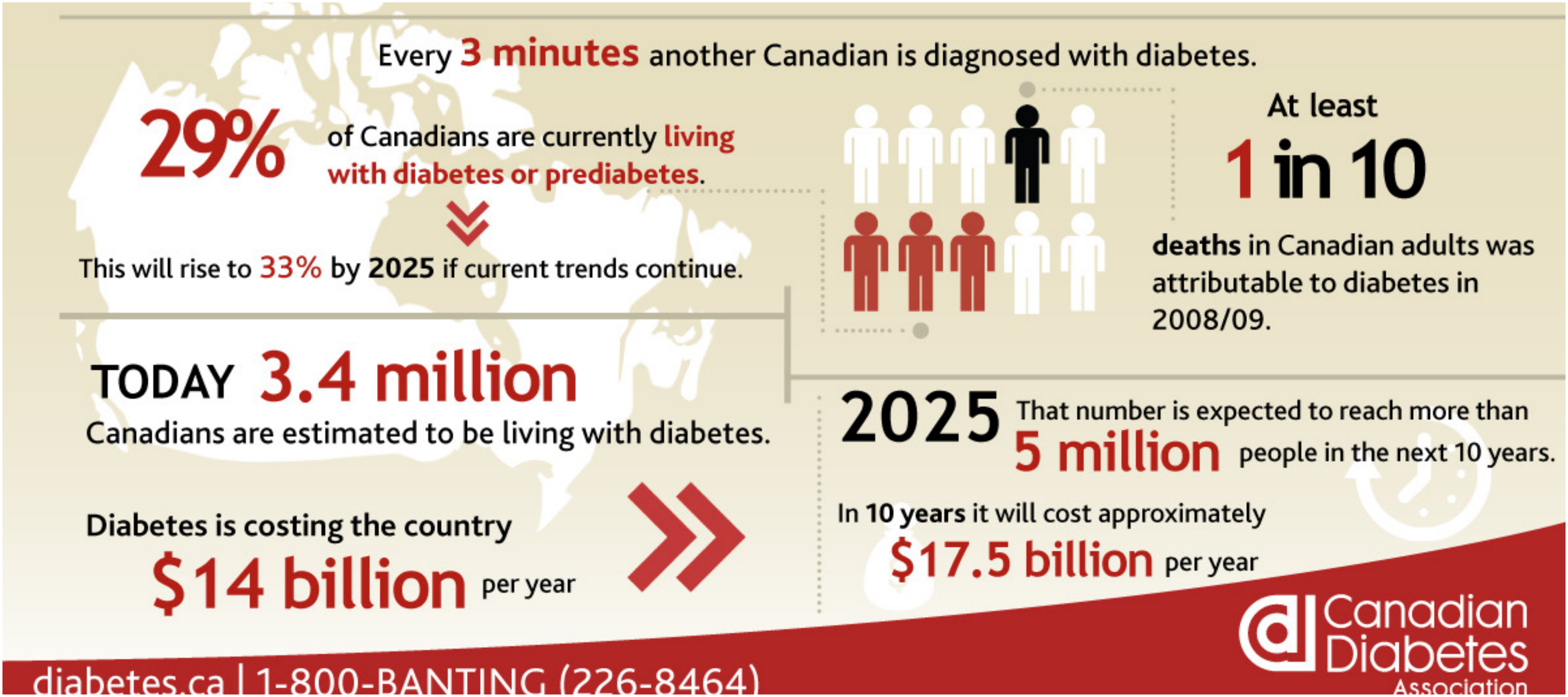
$0 \leq p(y|x;\theta) \leq 0.3$ $0.7 \leq p(y|x;\theta) \leq 1$

Questions?
[Unmute and ask, ask in chat]

There are no bad questions or questions with obvious answers.

A case study in diabetes





Every **3 minutes** another Canadian is diagnosed with diabetes.

29% of Canadians are currently **living with diabetes or prediabetes.**

This will rise to 33% by **2025** if current trends continue.

At least

1 in 10

deaths in Canadian adults was attributable to diabetes in 2008/09.

TODAY 3.4 million

Canadians are estimated to be living with diabetes.

2025 That number is expected to reach more than **5 million** people in the next 10 years.

Diabetes is costing the country

$14 billion per year

In **10 years** it will cost approximately

$17.5 billion per year

diabetes.ca | 1-800-BANTING (226-8464)

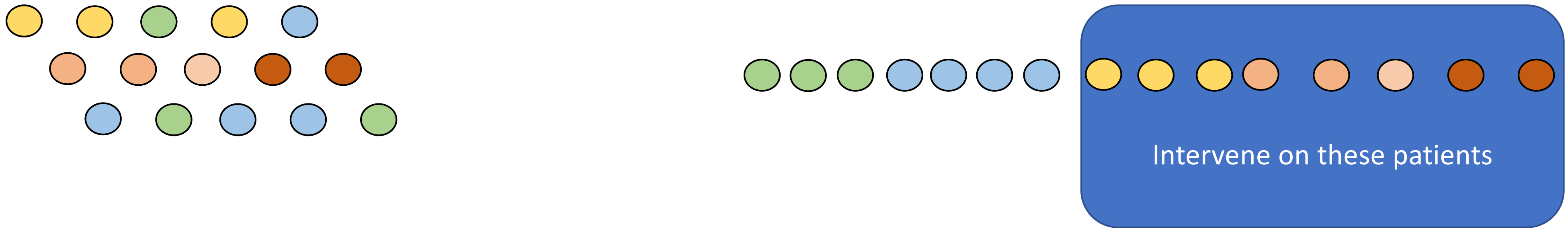
Canadian Diabetes Association

The costs of diabetes

Early detection of T2diabetes

- Early detection of undiagnosed diabetes mellitus: a US perspective, Harris et. al, 2000
- “There is a latent phase before diagnosis of Type 2 diabetes….. risk factors for diabetic micro- and macrovascular complications are markedly elevated and diabetic complications are developing.”
- “define a group of individuals with significant hyperglycemia who also have a high frequency of risk factors for micro- and macrovascular disease.”
- “treating hyperglycemia to prevent complications is more effective than treating these complications after they have developed”

What is that paper saying?



- If you have a patient population and can predict those at high-risk
- You can intervene on those high-risk patients **early**
- Effects of early intervention
  - “preventive interventions should start as early as possible in order to allow a wide variety of relatively low- and moderate-intensity programs”
  - Source: https://care.diabetesjournals.org/content/39/Supplement_2/S115.full-text.pdf

Traditional risk prediction models

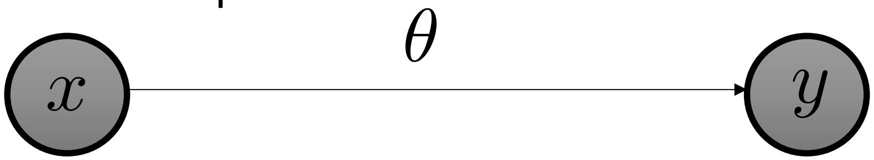
- Risk prediction models:
  - ARIC [Atherosclerosis **Risk** in Communities]
  - FRAMINGHAM [Coronary heart disease]
- Easy to ask questions, or measure in clinics when patients come in
- Simple model (typically rule based list or equations)

Challenges of traditional risk models

- Difficult to scale
  - Requires patient to come in to know they are high-risk
  - Can be time-consuming
- Existing risk scores do not work well if all values in risk calculator are not observed

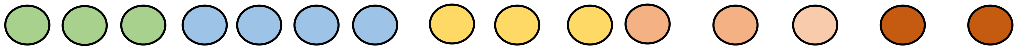
Automated population-level risk scoring

- **Key idea:**
  - The early detection of progression onto diabetes gives clinicians opportunities for early intervention.
  - Develop a predictive model from population level data
  - Use the predictive model to estimate risk
  - Scale up to millions of patients



If y>threshold:
The patient and the clinician have a conversation about how to reduce downstream complications

y: probability of contracting diabetes in the future



Who cares about this?

- Patients (us!) & disease registries:
  - Better outcomes for patients
- Provincial governments that pay for clinician time
  - Frees up clinician time
- Taxpayers (us!)
  - Lower costs for healthcare

Predicting diabetic onset from claims data

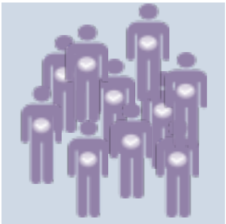
Using administrative claims data from the United States to predict diabetic onset

[Population level prediction of T2 diabetes from health claims and analysis of risk factors, Razavian, Blecker, Schmidt, Smith-McLallen, Nigam, Sontag. Big Data. ‘16]

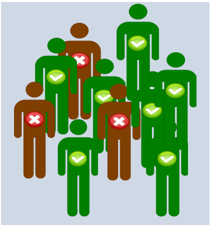
[Early detection of diabetes from health claims, Krishnan, N Razavian, Y Choi, S Nigam, S Blecker, A Schmidt, D. Sontag, Neurips Workshop 2013]

Learning from retrospective data

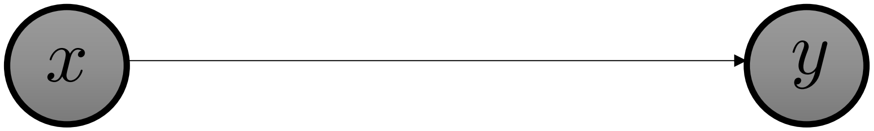
5 years ago



Now



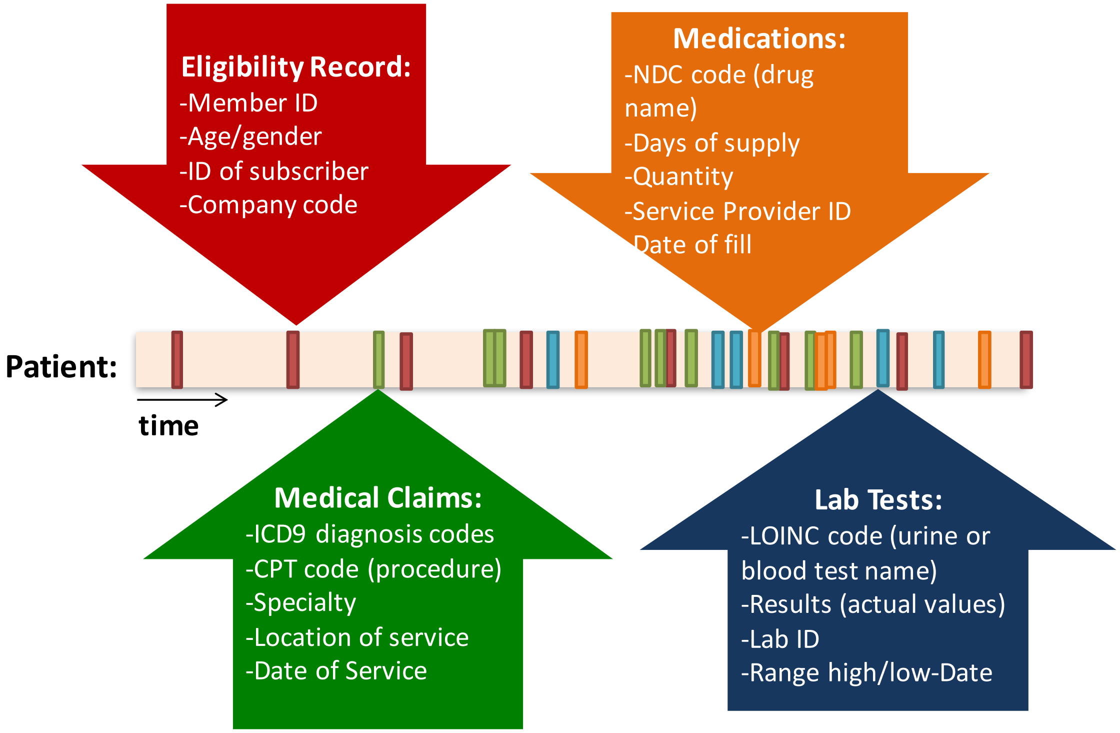
- **Idea:** retrospective data to build predictive models that we can use right now?



Clinical features

Diabetic status in the future

Administrative claims data



Eligibility Record:
-Member ID
-Age/gender
-ID of subscriber
-Company code

Medications:
-NDC code (drug name)
-Days of supply
-Quantity
-Service Provider ID
-Date of fill

Medical Claims:
-ICD9 diagnosis codes
-CPT code (procedure)
-Specialty
-Location of service
-Date of Service

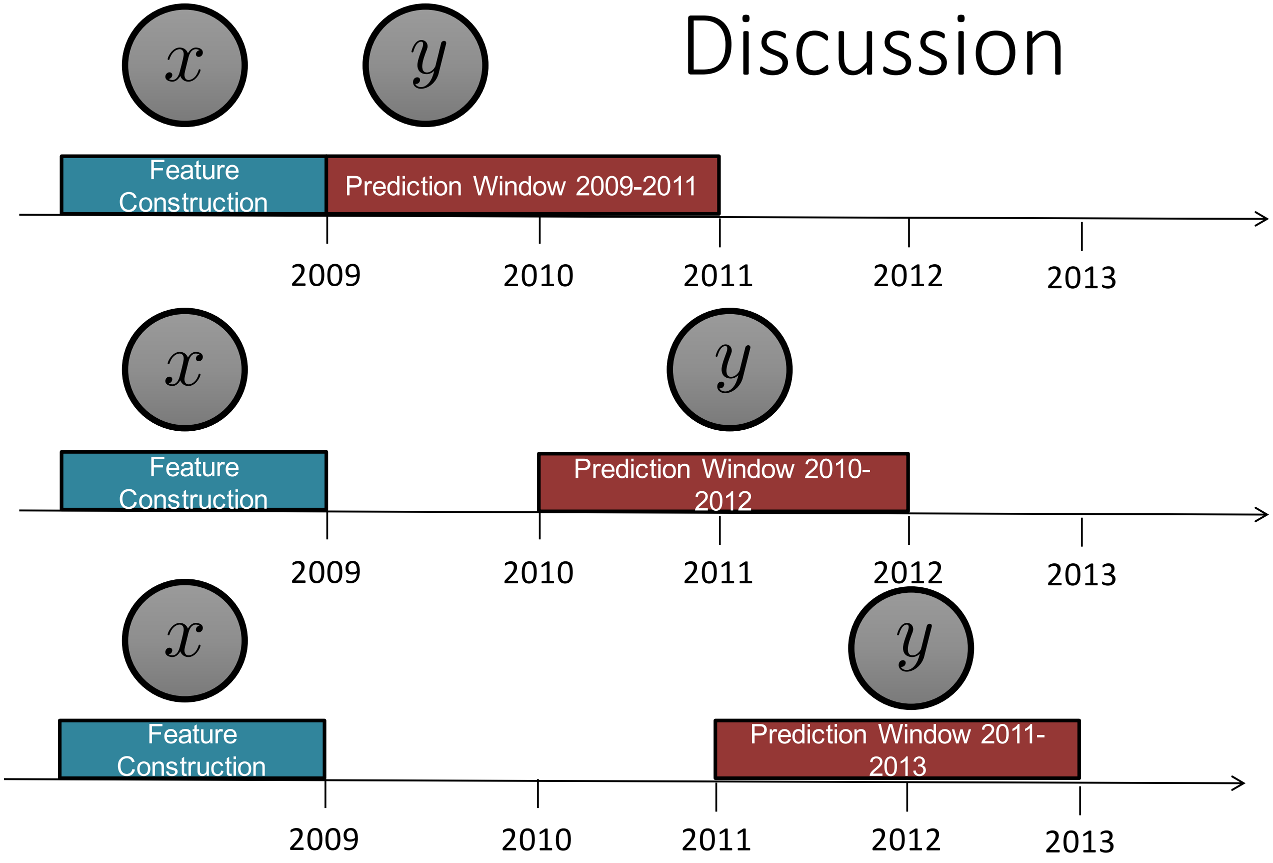
Lab Tests:
-LOINC code (urine or blood test name)
-Results (actual values)
-Lab ID
-Range high/low-Date

x

y

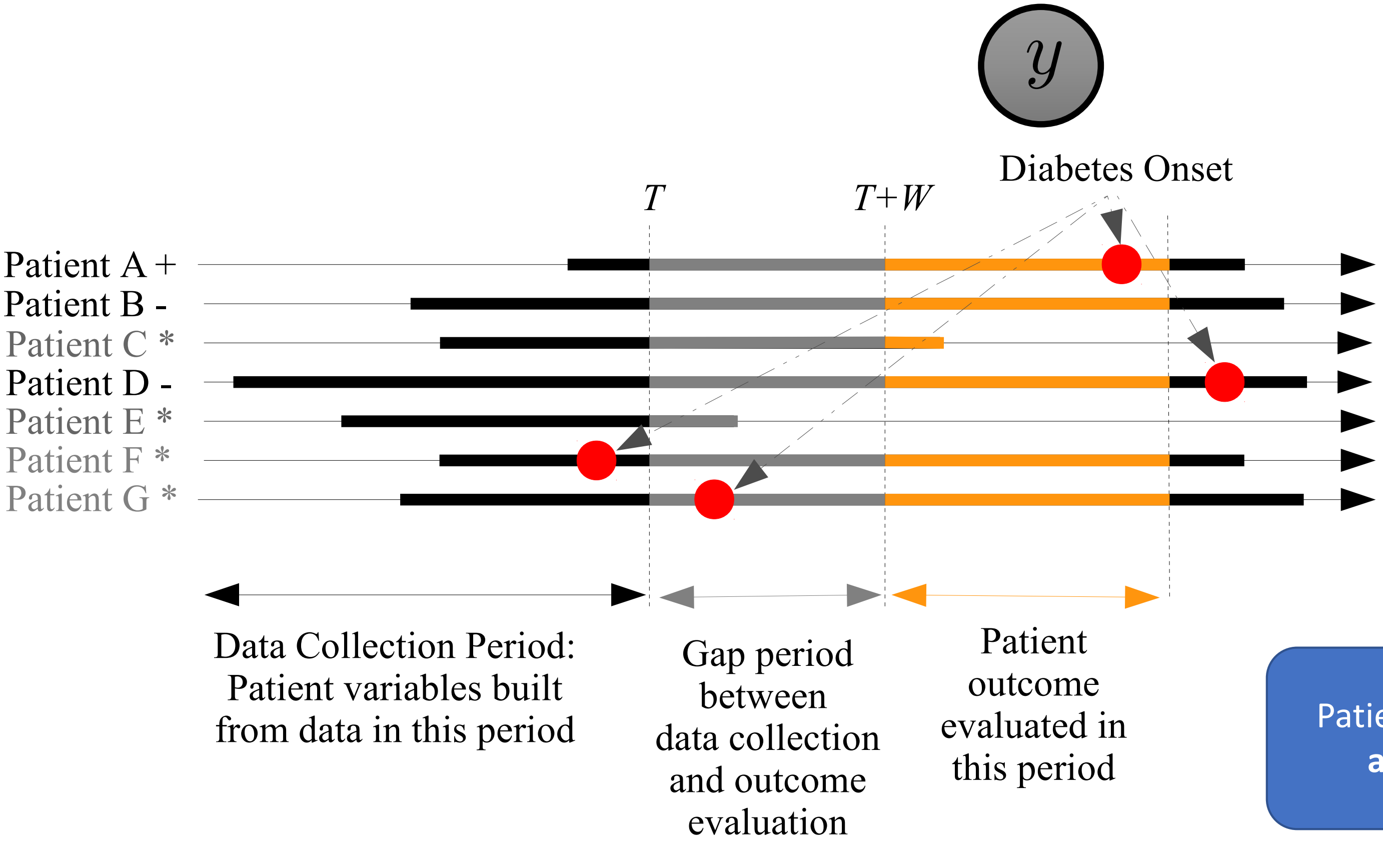
This data is used to define both x and y!

Discussion



Q: Which of these is preferrable and why?

Reduction to binary classification


y
Diabetes Onset
T
$T+W$
Patient A +
Patient B -
Patient C *
Patient D -
Patient E *
Patient F *
Patient G *
Data Collection Period: Patient variables built from data in this period
Gap period between data collection and outcome evaluation
Patient outcome evaluated in this period

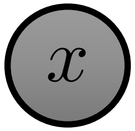
Patient alignment by **absolute time**

Patient alignment

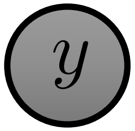
- Absolute time: Collect features based on their status as of 2008
- Relative time:
  - Collect features based on their first visit to their family doctor
  - Collect features by aligning based on each patient's first major comorbidity
- The choice of alignment will depend on how you want to use your model.

Best practices for creating clinical cohorts

- **Cohort design:** For patients that have more than one datapoint, make sure they appear either in the train, the validate or test set
- **Label leakage:**
  - Work with clinicians to understand their practice,
  - There are often subtle, easy to miss signs of disease indication
    - Errors in coding
    - Prescription of a drug
- **Selection bias:**
  - Ensure that the cohort is representative of the population you want to test it on.



Methods



- X: patient features, Y: did the patient have diabetes in the window of time
- L1 regularized logistic regression:
  - Optimize for predictive performance while
  - Doing aggressive feature selection
- Penalize the L1 norm of the weight vector.

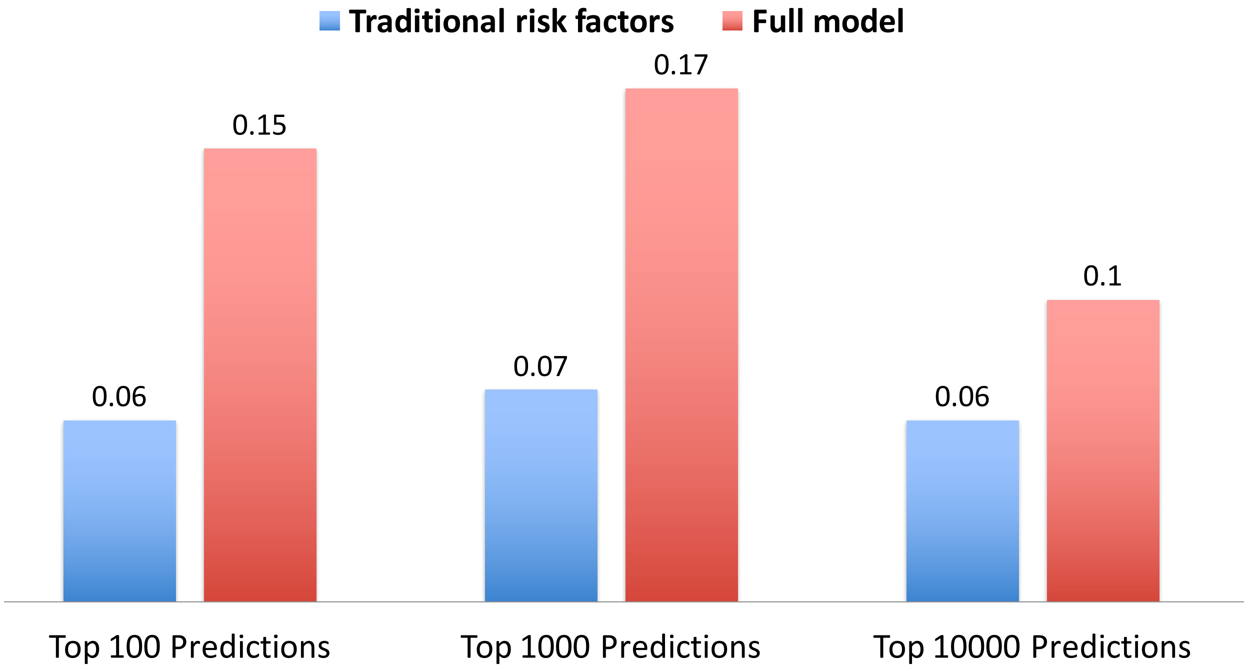
L1 regularized regression

$$\min_{\theta} \sum_{i=1}^{N} -\mathcal{L}(y_i, x_i) + \lambda ||\theta||_1 \quad ||\theta||_1 = \sum_{d} |w_d|$$

- X: patient features, Y: did the patient have diabetes in the window of time
- L1 regularized logistic regression:
  - Optimize for predictive performance while
  - Doing aggressive feature selection
- Penalize the L1 norm of the weight vector.
  - d: dimension of feature vector

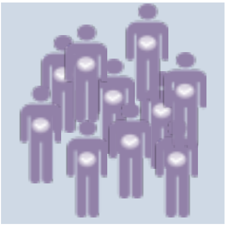
Highlights of results

- Total number of features in model: 42000

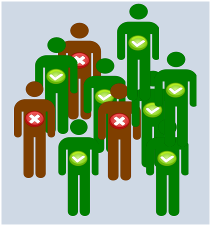


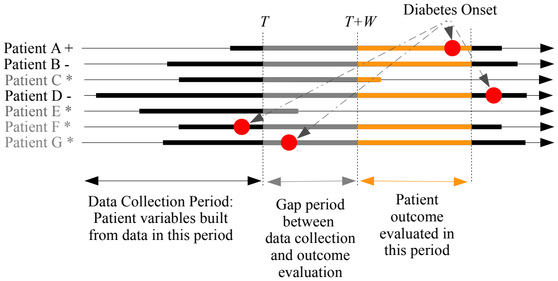
Recap and summary

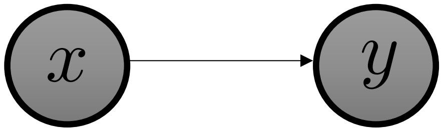
5 years ago



Now




Diabetes Onset
T
$T+W$
Patient A +
Patient B -
Patient C *
Patient D -
Patient E *
Patient F *
Patient G *
Data Collection Period: Patient variables built from data in this period
Gap period between data collection and outcome evaluation
Patient outcome evaluated in this period



Clinical features

Diabetic status
in the future

Discussion : What are the limitations of this approach?

Class poll

Feedback welcome!

- Key advantage of this style of class – feedback!
- Are there interesting topics you'd like to learn more about?
- Send me & course staff an email.

TODOs

- Finish the online questionnaire if you have not already
- Friday will be a project planning session. It will involve a discussion with Alistair Johnson (lead creators of the MIMIC dataset)
  - Please watch the following posted videos **before** the class on Friday
- Introducion to MIMIC
- MIMIC analysis tutorial
- MIMIC data tutorial