2nd ITU-WHO Workshop on Artificial Intelligence for Health, November 14th 2018

# Disease Prediction Using Machine Learning and Electronic Health Records

Narges Razavian

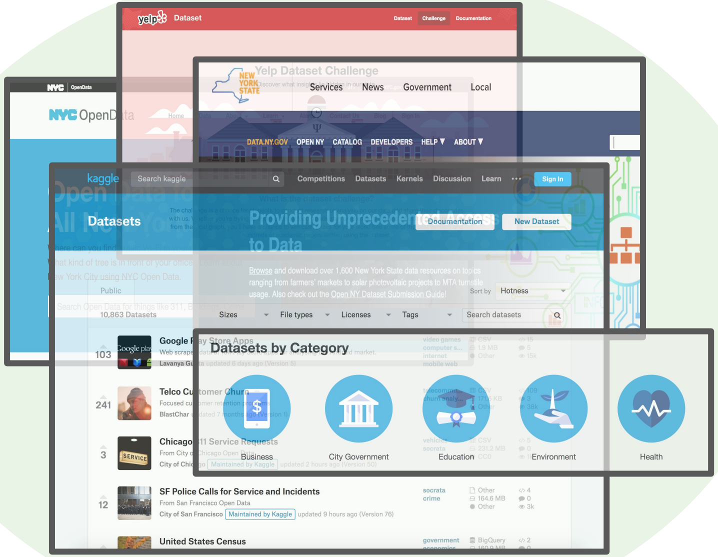
*Assistant Professor*

*NYU Langone Medical Center*

*Departments of Radiology and Population Health*

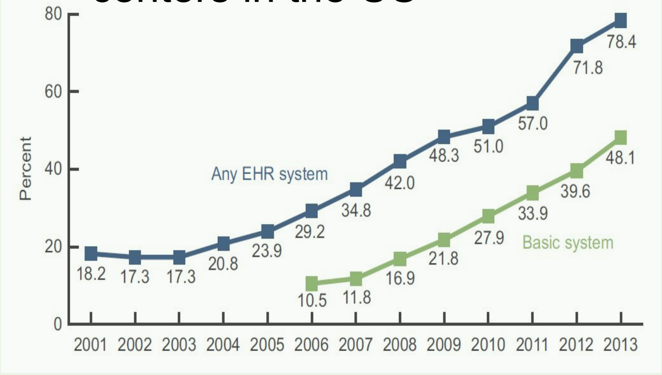
NYU Langone Health

# Changing Landscape of **Digital Data**



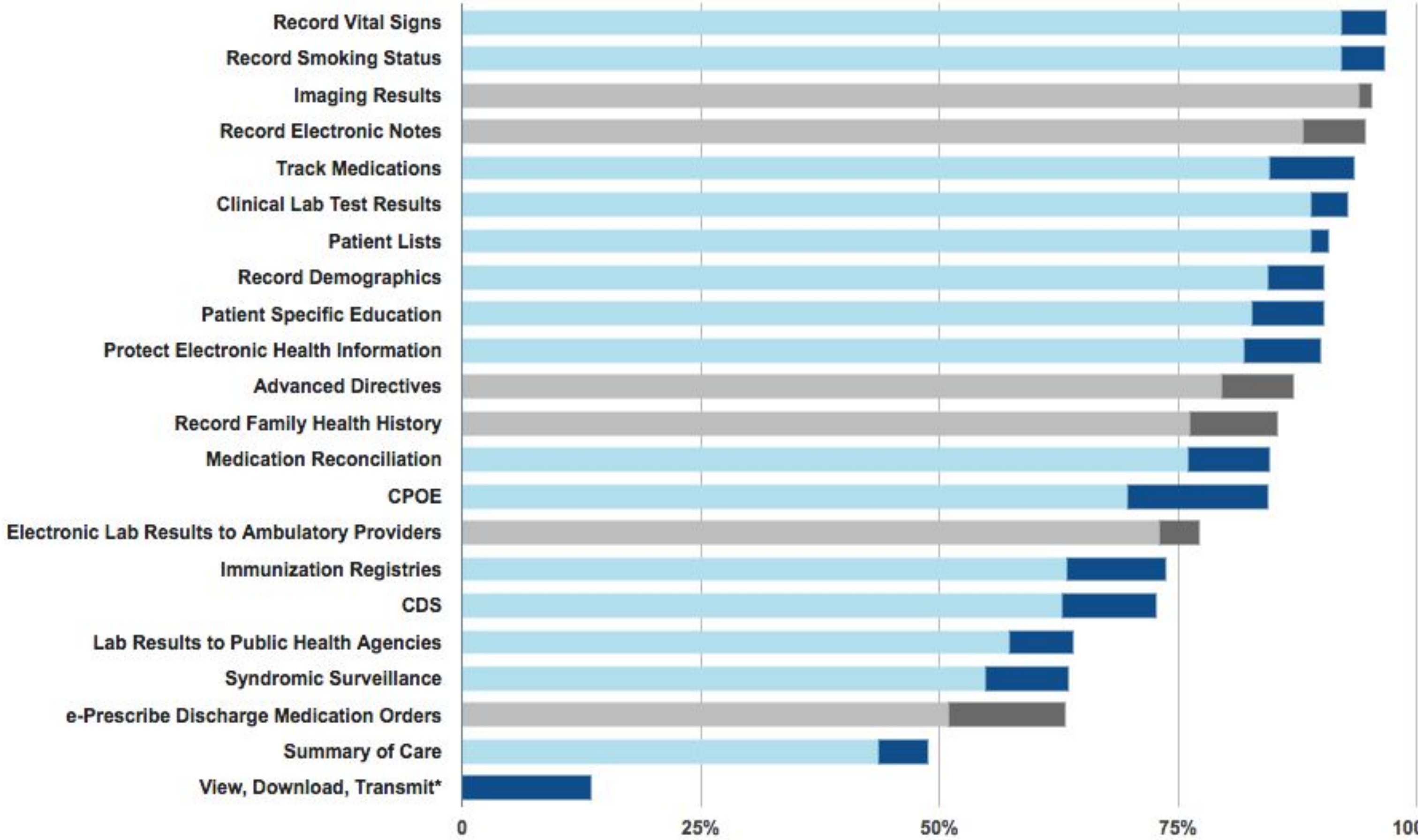

Environment, food, housing, transportation, crime, education, etc. etc.

EHR adoption by healthcare centers in the US



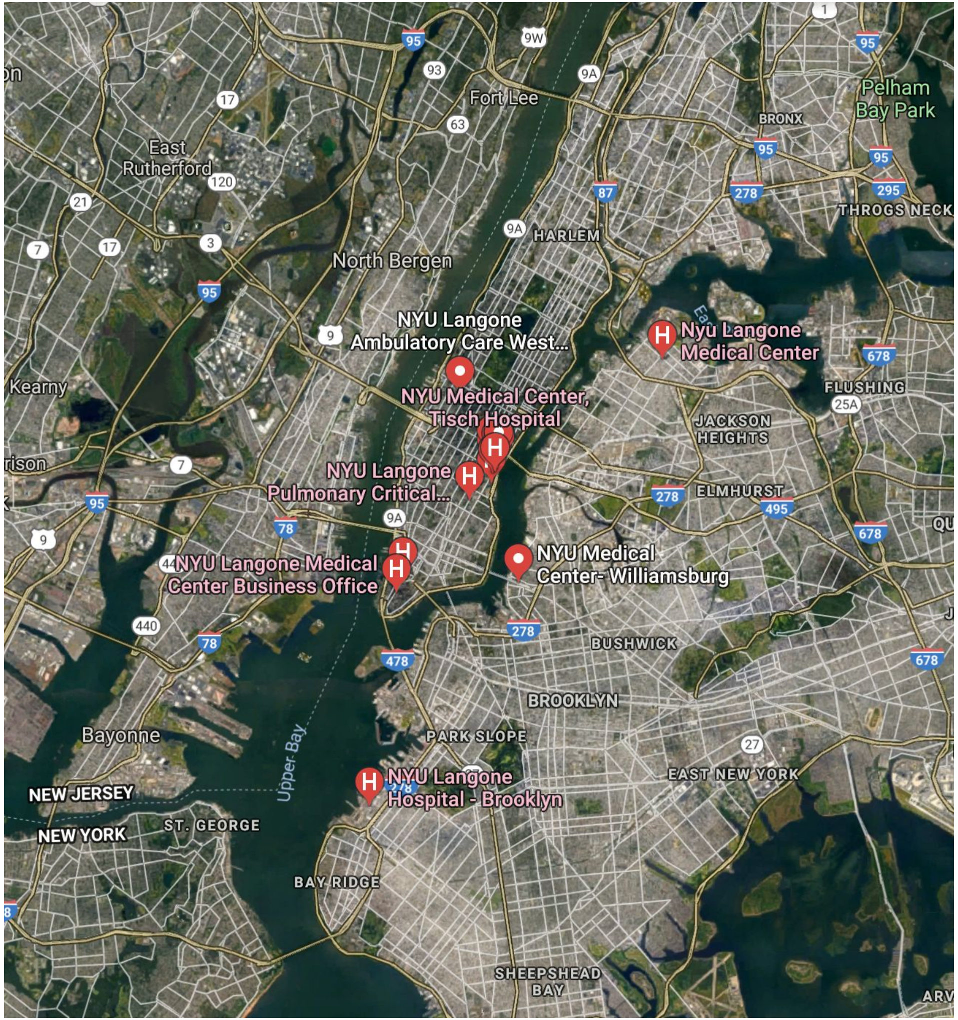

**Electronic Health Records at scale of millions per year**

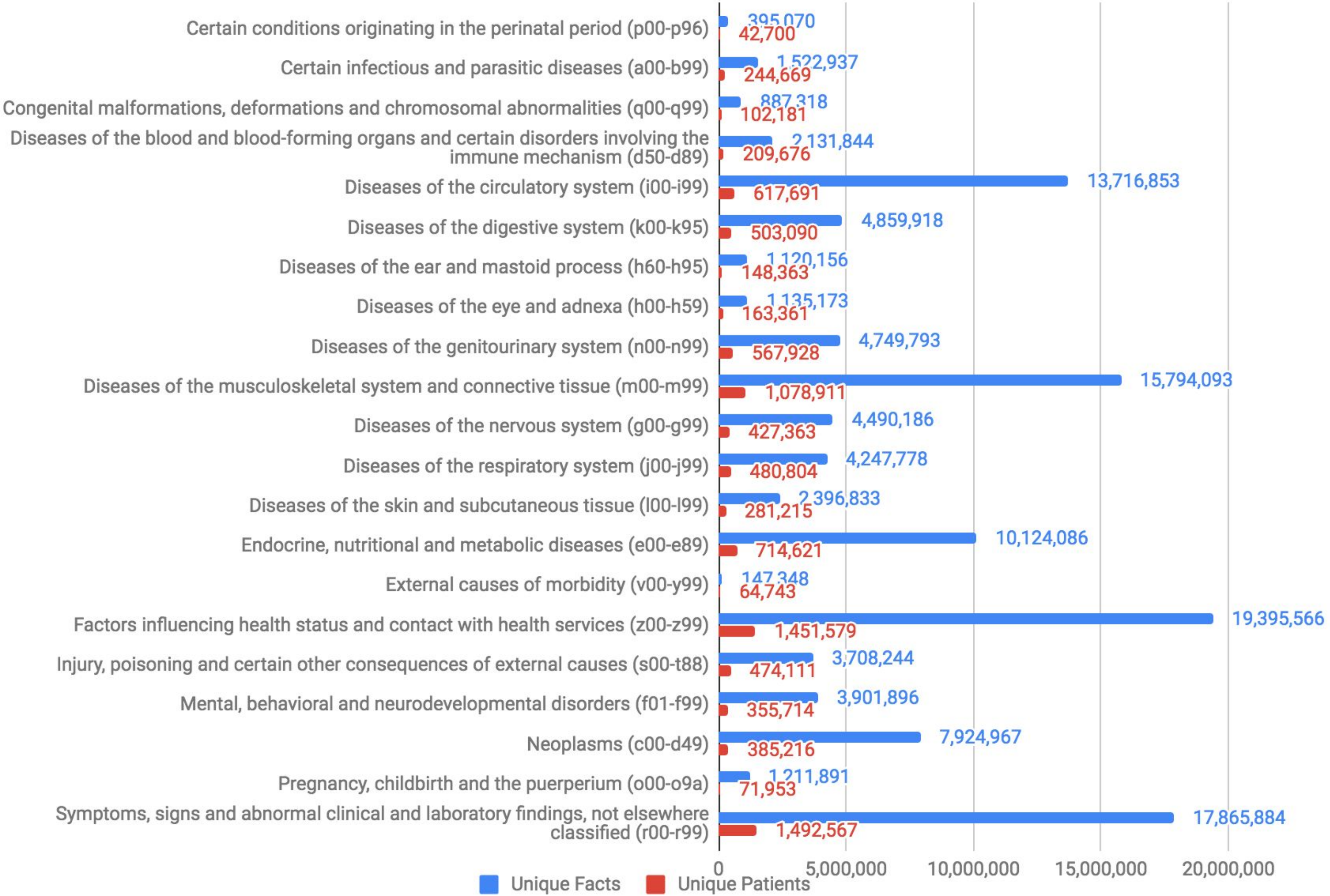
# What is captured in the EHR?




*Source: healthcare.gov*


NYU Langone
MEDICAL CENTER
Main Campus
Main Lobby Entrance
530-550 First Avenue
Tisch Hospital
NYU School of Medicine
Schwartz
Health Care Center
No Smoking Please
Smoke-Free
Environment


Fort Lee
East
Rutherford
North Bergen
HARLEM
BRONX
Pelham
Bay Park
THROGS NECK
NYU Langone
Ambulatory Care West...
Nyu Langone
Medical Center
FLUSHING
NYU Medical Center,
Tisch Hospital
JACKSON
HEIGHTS
Kearny
NYU Langone
Pulmonary Critical...
ELMHURST
NYU Langone Medical
Center Business Office
NYU Medical
Center- Williamsburg
BUSHWICK
BROOKLYN
Bayonne
PARK SLOPE
Upper Bay
NYU Langone
Hospital - Brooklyn
EAST NEW YORK
NEW JERSEY
NEW YORK
ST. GEORGE
BAY RIDGE
SHEEPSHEAD
BAY


Certain conditions originating in the perinatal period (p00-p96)
395,070
42,700
Certain infectious and parasitic diseases (a00-b99)
1,522,937
244,669
Congenital malformations, deformations and chromosomal abnormalities (q00-q99)
887,318
102,181
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism (d50-d89)
2,131,844
209,676
Diseases of the circulatory system (i00-i99)
13,716,853
617,691
Diseases of the digestive system (k00-k95)
4,859,918
503,090
Diseases of the ear and mastoid process (h60-h95)
1,120,156
148,363
Diseases of the eye and adnexa (h00-h59)
1,135,173
163,361
Diseases of the genitourinary system (n00-n99)
4,749,793
567,928
Diseases of the musculoskeletal system and connective tissue (m00-m99)
15,794,093
1,078,911
Diseases of the nervous system (g00-g99)
4,490,186
427,363
Diseases of the respiratory system (j00-j99)
4,247,778
480,804
Diseases of the skin and subcutaneous tissue (l00-l99)
2,396,833
281,215
Endocrine, nutritional and metabolic diseases (e00-e89)
10,124,086
714,621
External causes of morbidity (v00-y99)
147,348
64,743
Factors influencing health status and contact with health services (z00-z99)
19,395,566
1,451,579
Injury, poisoning and certain other consequences of external causes (s00-t88)
3,708,244
474,111
Mental, behavioral and neurodevelopmental disorders (f01-f99)
3,901,896
355,714
Neoplasms (c00-d49)
7,924,967
385,216
Pregnancy, childbirth and the puerperium (o00-o9a)
1,211,891
71,953
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (r00-r99)
17,865,884
1,492,567
0
5,000,000
10,000,000
15,000,000
20,000,000
Unique Facts
Unique Patients

# Electronic Health Records

Time

Demographic and lifestyle

**Medications:**
-NDC code (drug name)
-Quantity
-Date of fill

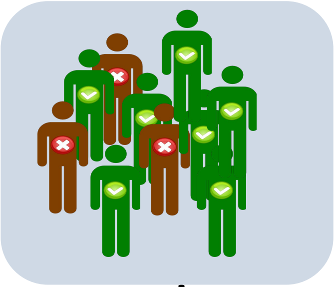
**Encounters**
-Free Text Notes
-Diagnosis code (ICD10s)
-Procedure (CPTs)
-Specialty
-Location of service
-Service Provider ID
-Inpatient/outpatient
-Cost

**Radiology Imaging:**
- MRI, CT, PET, etc.
- Free Text (Radiology notes)
- Assessment codes

**Pathology:**
- Microscopic images (histopathology)
- Genetic test
- Free text assessments

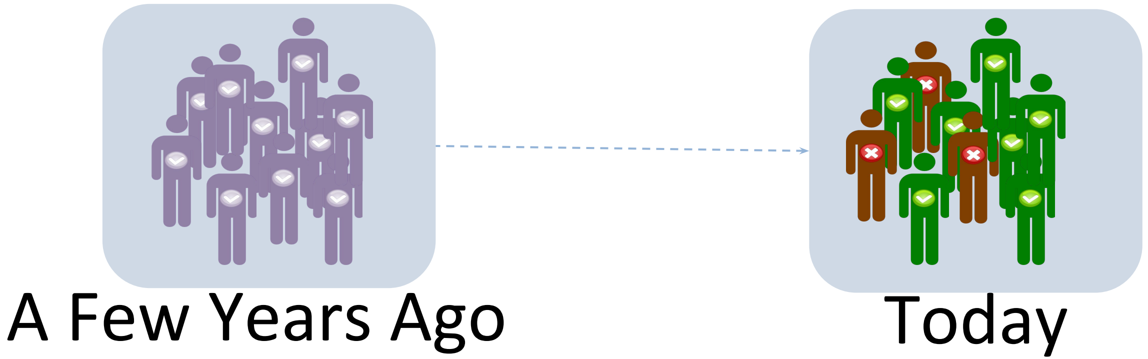
**Lab Tests:**
-LOINC code (urine or blood test name)
-Results (actual values/Flags)
-Date

# Electronic Health Records for future outcome prediction



Today

# Electronic Health Records for future outcome prediction



# Electronic Health Records for future outcome prediction

A Few Years Ago

Today

Demographic and lifestyle

Time

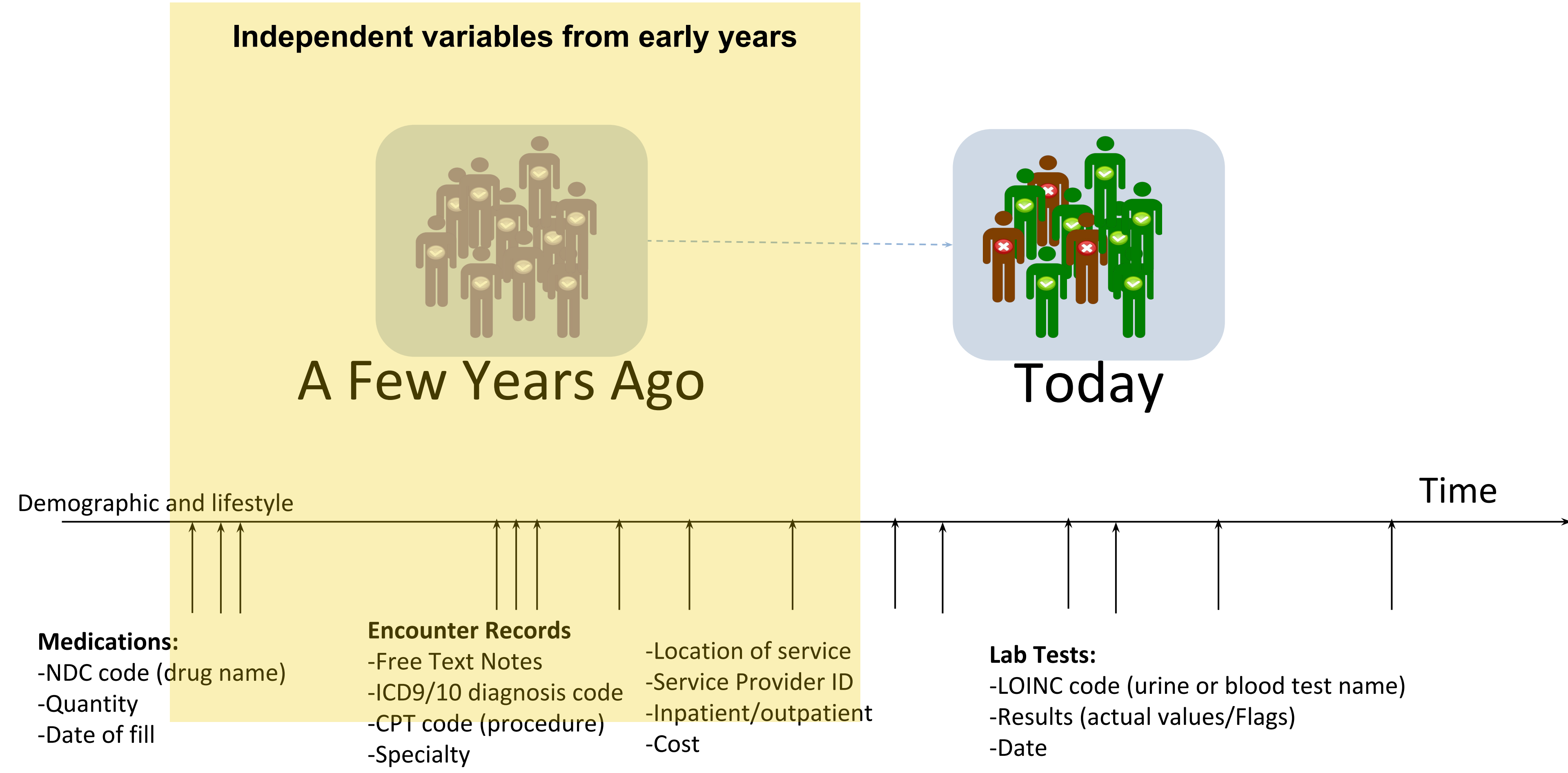
**Medications:**
-NDC code (drug name)
-Quantity
-Date of fill

**Encounter Records**
-Free Text Notes
-ICD9/10 diagnosis code
-CPT code (procedure)
-Specialty

-Location of service
-Service Provider ID
-Inpatient/outpatient
-Cost

**Lab Tests:**
-LOINC code (urine or blood test name)
-Results (actual values/Flags)
-Date

# Electronic Health Records for future outcome prediction



# Electronic Health Records for future outcome prediction

**Independent variables from early years**

A Few Years Ago

**Target outcome in the future**

Today

Demographic and lifestyle

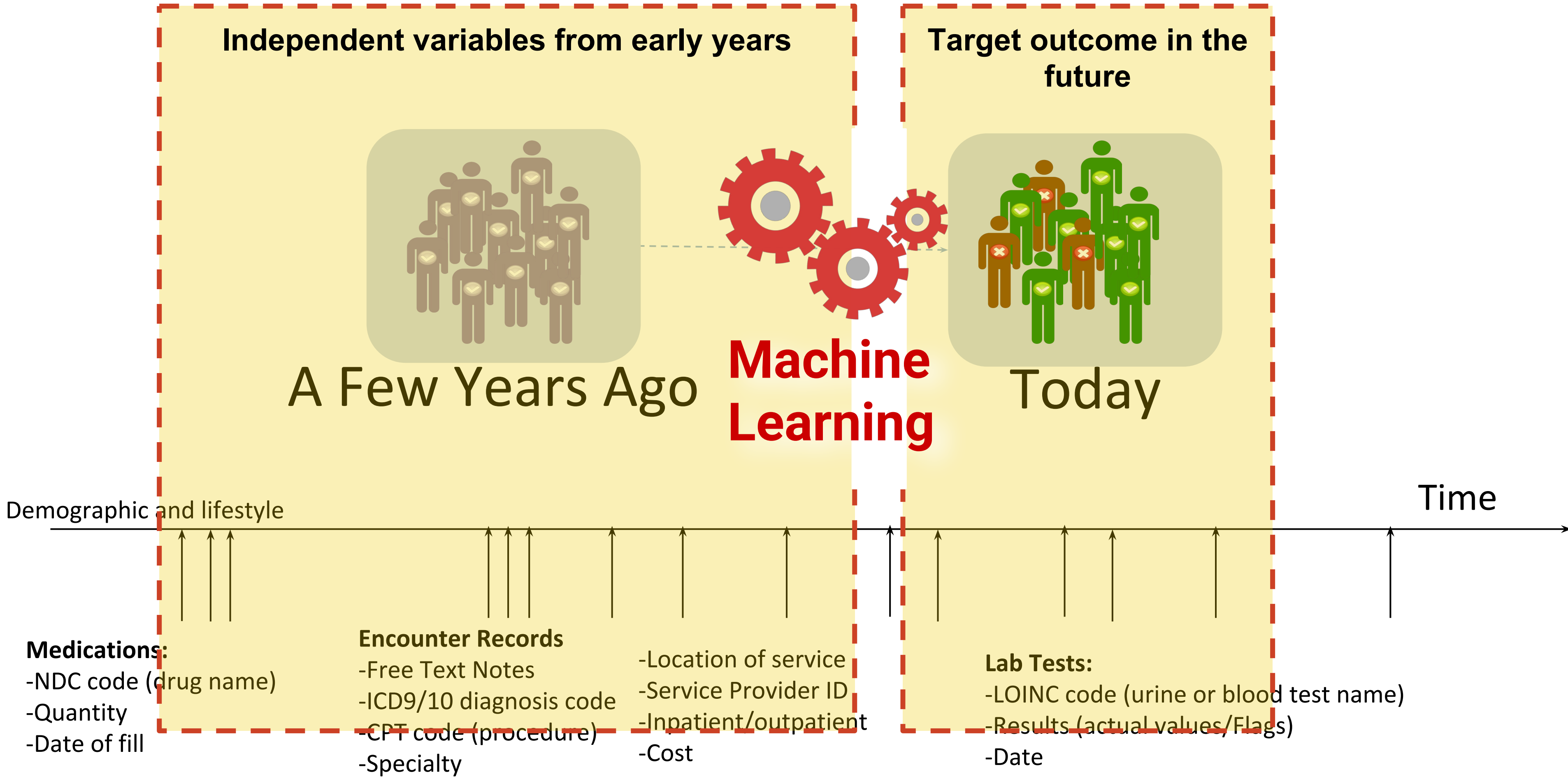
Time

**Medications:**
-NDC code (drug name)
-Quantity
-Date of fill

**Encounter Records**
-Free Text Notes
-ICD9/10 diagnosis code
-CPT code (procedure)
-Specialty
-Location of service
-Service Provider ID
-Inpatient/outpatient
-Cost

**Lab Tests:**
-LOINC code (urine or blood test name)
-Results (actual values/Flags)
-Date

# Electronic Health Records for future outcome prediction



Challenges:

Multi-modal (Time series, Text, Images)
Biased data
Explanations of model decisions

Challenges:

Multi-modal (Time series, Text, Images)

Biased data

Explanations of model decisions

# Using Structured Data for Early Detection

<u>NYU Collaborators</u>
David Sontag
Rahul Krishnan
Uri Shalit
Jake Marcus
YD Choi
Yoni Halpern

<u>NYU Medical School Collaborators</u>
Saul Blecker
Ann Marie Schmidt
Yin Aphinyanaphong
Leora Horwitz

<u>My Students</u>

Anant Gupta

Graduate Student, Courant Institute, Department of Computer Science, NYU

Research: Deep Learning and Machine Learning on Electronic Health Records, Predicting Preventable Diseases

Rob Hammond

Graduate Student, NYU Center for Data Science

Data Scientist, NYU Langone Medical Center

Research: Machine Learning Models for Electronic Health Records, Predicting Childhood Obesity

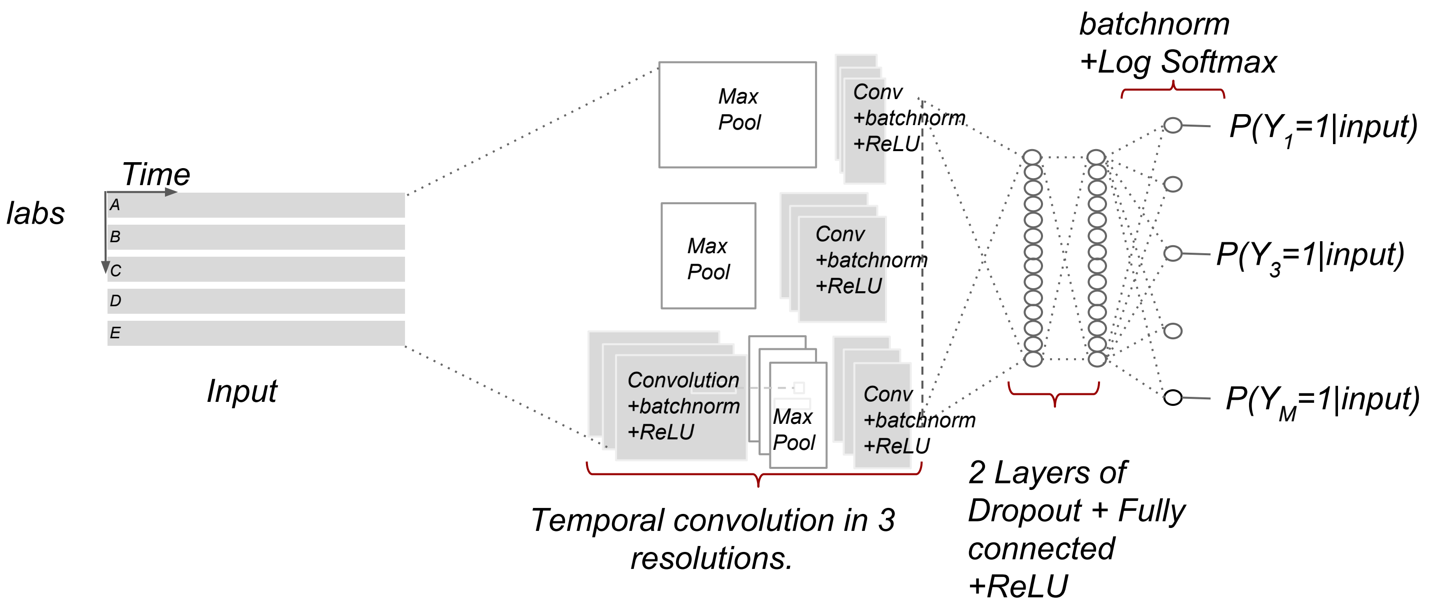
# 42K variables (before) to 280K variables (now), each across time

| 22 | 39 | 990 | 16,632 | 233 | 224 | 7x1000 | 228 | 32 |
|---|---|---|---|---|---|---|---|---|
| Diabetes known risk factors | coverage | indicator for using Medication groups | indicator for each icd9 diagnosis (86K ICD10) | indicator for each ICD-9 procedures group | indicator for each CPT group | Laboratory indicators for:<br>Test request<br>Test value high<br>Test value low<br>Test value normal<br>Test value increasing<br>Test value decreasing<br>Test value fluctuating<br>(80K total Labs) | Indicator for each specialty | Indicator for each service place |

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

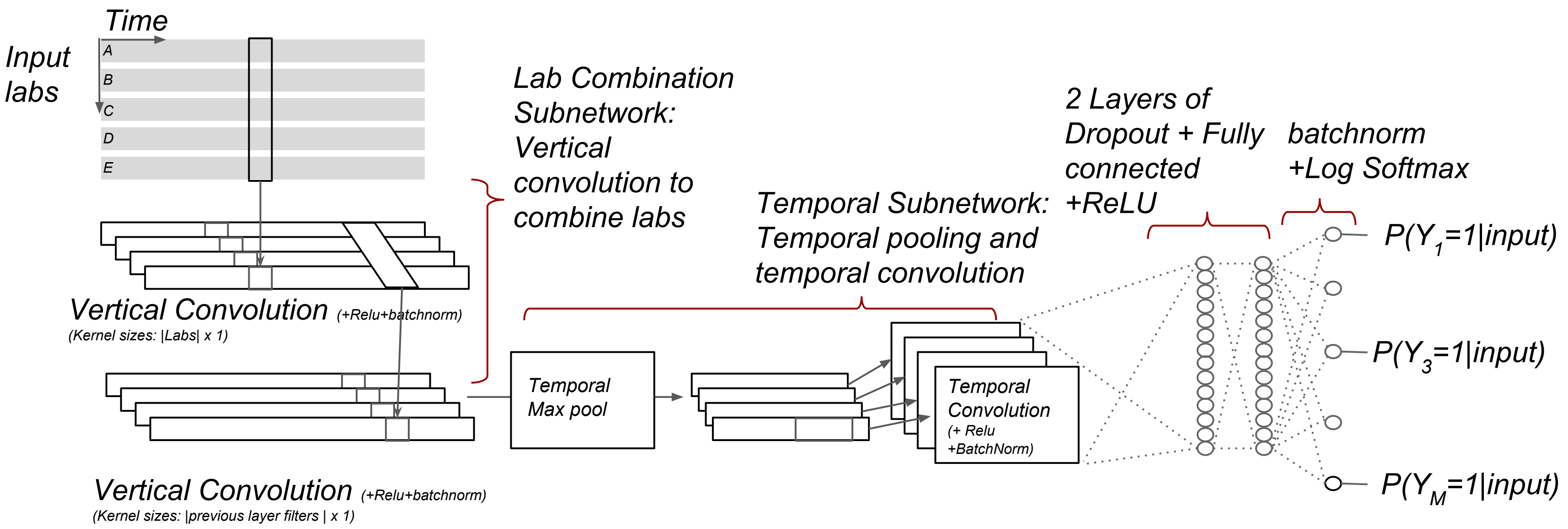
*Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors*
*https://www.liebertpub.com/doi/abs/10.1089/big.2015.0020*

# Learning features and Deep Learning/Multitask learning



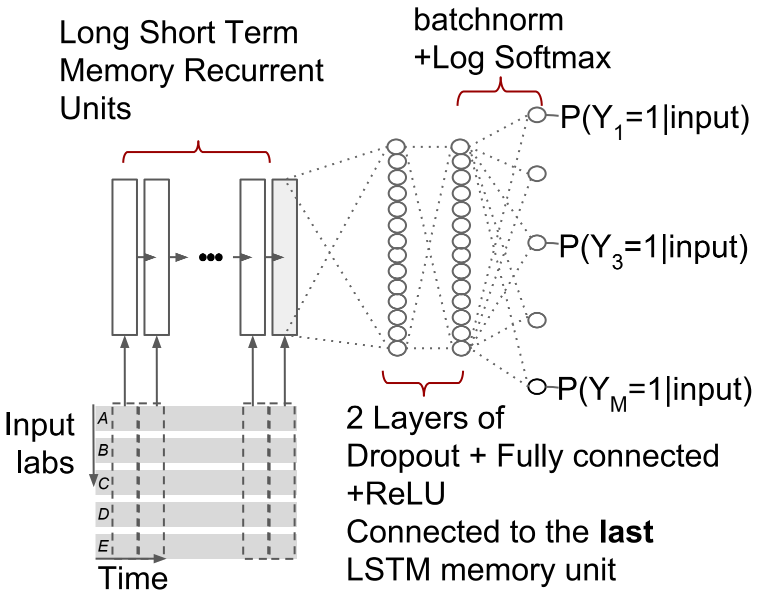

*Temporal convolutional neural networks for diagnosis from lab tests.*
*https://openreview.net/forum?id=ROVmO430RTvnM0J1Ip9z*

# Learning features and Deep Learning/Multitask learning



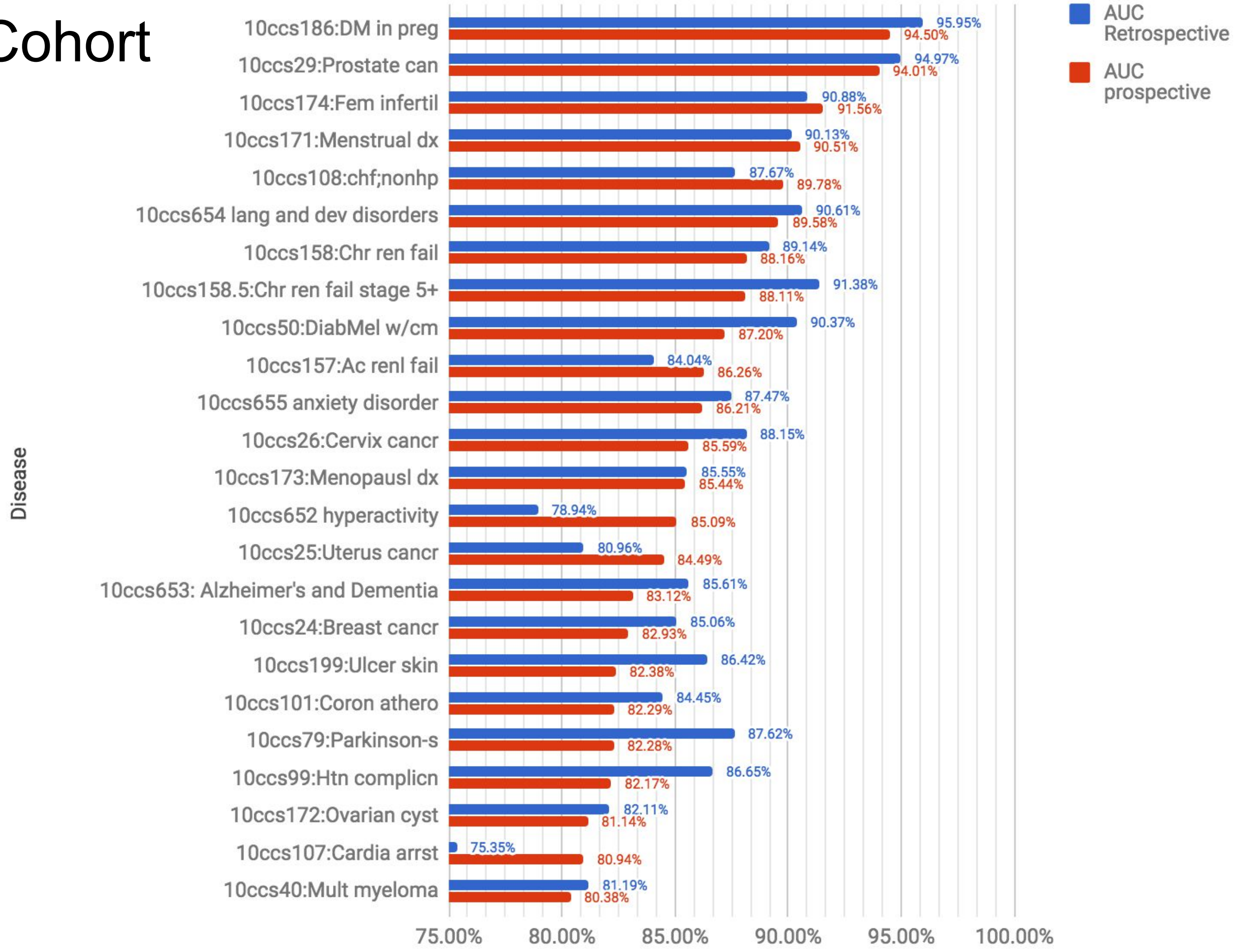

*Multi-task prediction of disease onsets from longitudinal laboratory tests*
*http://www.jmlr.org/proceedings/papers/v56/Razavian16.pdf*

# Learning features and Deep Learning/Multitask learning




*Multi-task prediction of disease onsets from longitudinal laboratory tests*
*http://www.jmlr.org/proceedings/papers/v56/Razavian16.pdf*

# NYUMC Cohort



# Applicable to many more outcomes and tasks

- Early prediction of childhood obesity
- Using environmental factors to predict childhood obesity
- Predicting diabetes and diabetes complications
- Detecting undocumented but existing diseases (all diseases)
- Using lab values only to predict future diseases
- Predicting medication adherence
- Predicting appointment no-shows
- etc. etc. etc.…

# How to handle *Clinical Notes?*

Demographic and lifestyle

Time

**Medications:**
-NDC code (drug name)
-Quantity
-Date of fill

**Encounters**
**-Free Text Notes**
-Diagnosis code (ICD10s)
-Procedure (CPTs)
-Specialty
-Location of service
-Service Provider ID
-Inpatient/outpatient
-Cost

**Radiology Imaging:**
- MRI, CT, PET, etc.
- Free Text (Radiology notes)
- Assessment codes

**Pathology:**
- Microscopic images (histopathology)
- Genetic test
- Free text assessments

**Lab Tests:**
-LOINC code (urine or blood test name)
-Results (actual values/Flags)
-Date

# Using AI + Clinical Notes for Early Detection

Jingshu Liu

Graduate Student, NYU Department of Computer Science

Research: Deep Learning and Natural Language Processing for Clinical Notes

Zachariah Zhang

Graduate Student, NYU Center for Data Science

Research: Deep Learning and Natural Language Processing for Clinical Notes

Sheng Liu

Graduate Student, NYU Center for Data Science

Research: Deep Learning and Natural Language Processing for Clincial Notes on ICU data

<u>NYU Collaborators</u>
Kyunghyun Cho
Sam Bowman

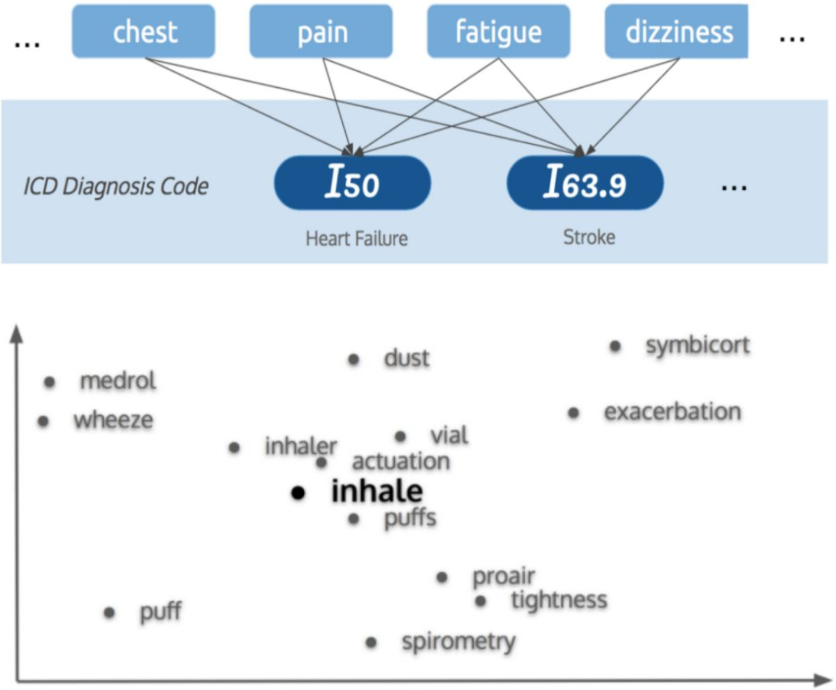
<u>NYU Medical School Collaborators</u>
Yin Aphinyanaphong
Leora Horwitz
Himanshu Grover
Jerko Steiner
Marina Marin

# Clinical Notes: Abbreviated, Messy, Unstructured, *not* English, not grammatical, not intended to be.

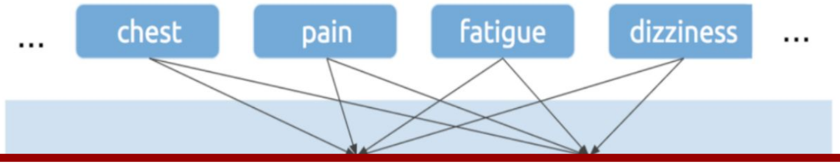
An example of a real note - de-identified:

Patient received via stretcher from ED in NAD. Ambulated without any difficulties. Patient states that he is due for Fentanyl lollipop 600mcg at 1800. PA [**Last Name (un) 1**] made aware of patient's arrival and pain meds. To assess patient at bedside. Handoff given to RN. Patient sleeping comfortably in bed in NAD. Call bell within reach. Safety maintained. Patient off the floor to xray. Patient stable. Handoff report given by [**Name8 (MD) 1**], RN. IV fluids ru nning well. Patient is resting comfortably at this assessment. Call bell within reach. The care of this patient has transferred to PA [**Last Name (un) 1**]. Current disposition: placed in observation. At this time, the care of this patient was transferred to the Emergency Medicine service for ED observation. Reassessment Vital Signs: [**2016-04-29**] 1457 BP: 120/80 Pulse : 75 Temp: 36.2 ?C (97.2 ?F) Resp: 15 SpO2: 100% Temp (24hrs) Max:36.8 ?C (98.3 ?F) Pain Score: 8 - Eight [**First Name8 (NamePattern2) 2**] [**Last Name (NamePattern1) 3**] is a 37 y. o. male placed in observation under the Abdominal Pain Protocol. Pertinent results: Upper GI Series/Abd XR with contrast into the small bowel Please follow-up on: Follow-up abd XR at 8pm to eval f or contrast into the rectum Plan of care in the observation unit: serial abdominal exams, advancement of diet, repeat abdominal XR

# Learning Semantics First - *Learning Embeddings*

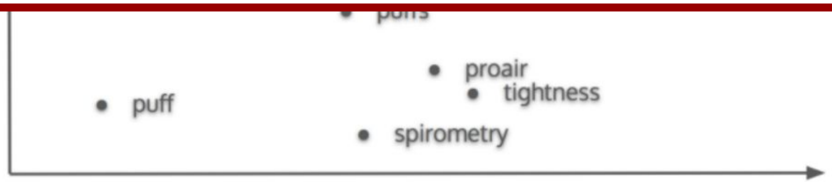


# Learning Semantics First - *Learning Embeddings*

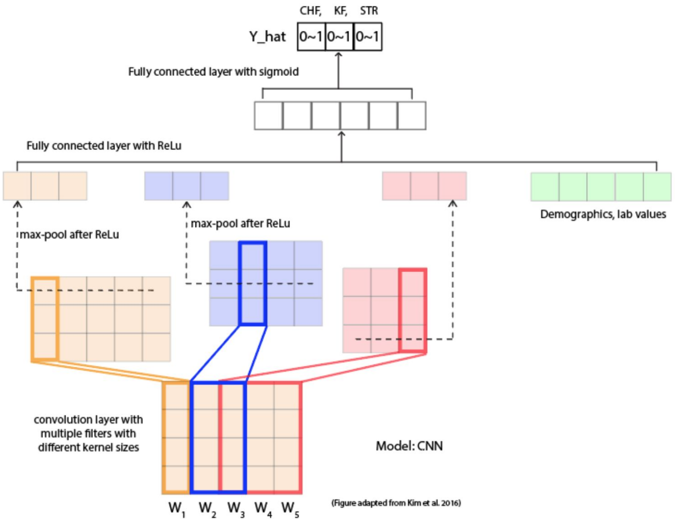


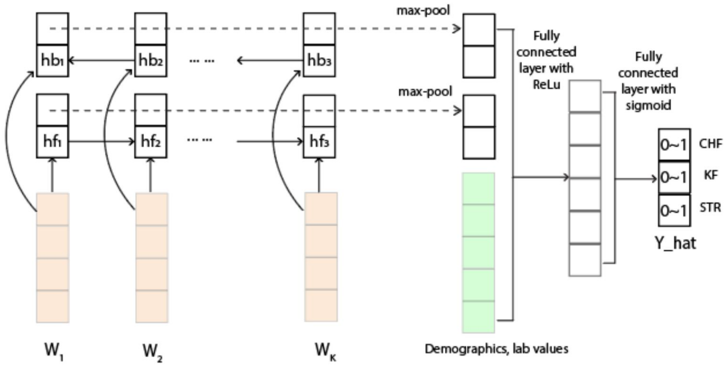

All code & instructions and *trained embeddings for **24,960** clinical terms*:

[https://github.com/**NYUMedML/DeepEHR**](https://github.com/NYUMedML/DeepEHR)
(Fully open-sourced)



# Extracting Structured Data from the Notes & Combining with Text Data



# Dealing with Time

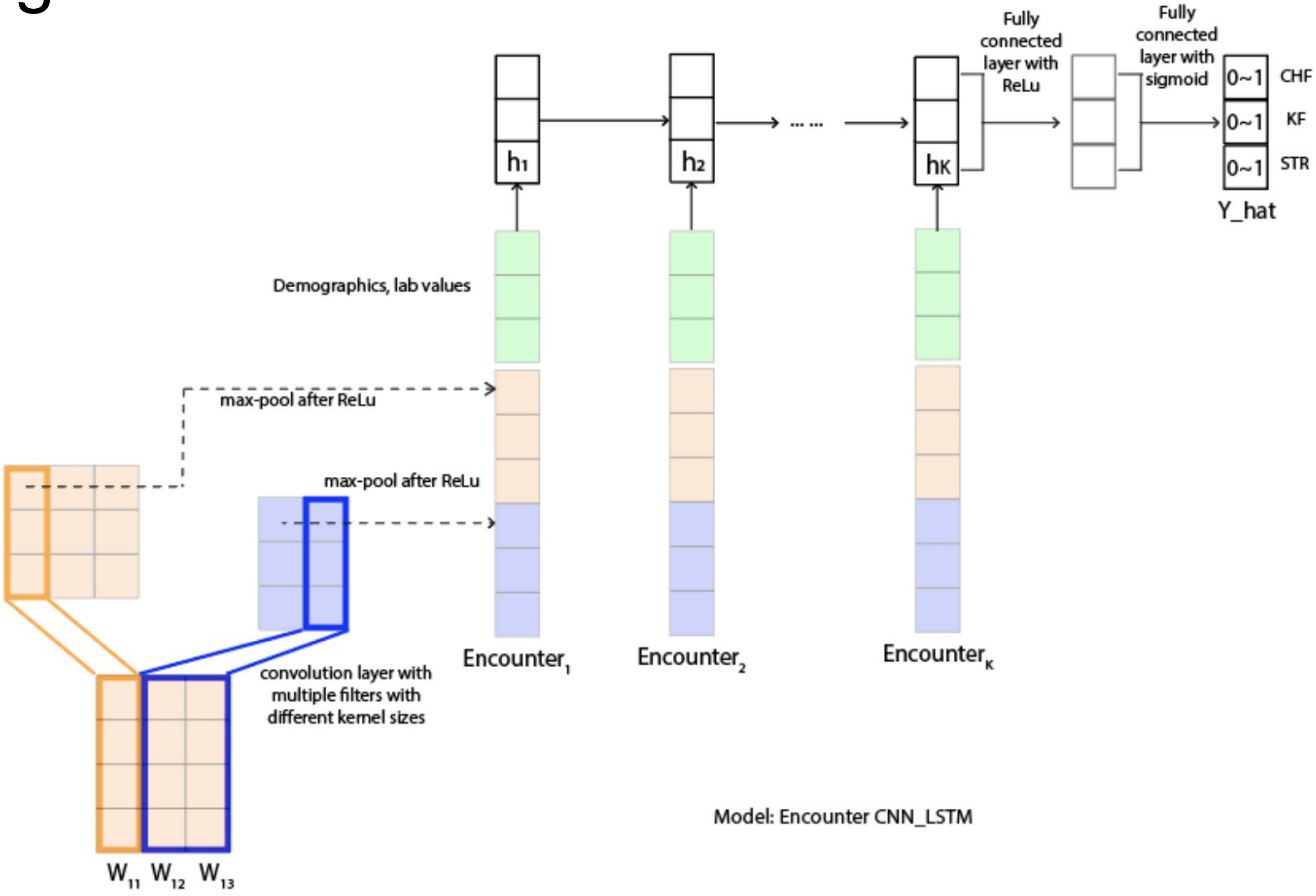


Table 1: Number of Records by Target Diseases (Negative Cases : Positive Cases)

| Target | Training Set | Validation Set | Test Set |
|---|---|---|---|
| Congestive Heart Failure | 644K : 4080 | 93K : 574 | 184K : 1167 |
| Kidney Failure | 616K : 10051 | 88K : 1428 | 176K : 2809 |
| Stroke | 653K : 3195 | 94K : 406 | 187K : 916 |

| | **Heart Failure** | **Kidney Failure** | **Stroke** |
|---|---|---|---|
| Logistic Reg Lab/Demo | 0.781 | 0.724 | 0.70 |
| LSTM Lab/Demo | 0.813 | 0.743 | 0.699 |
| Logistic Reg Notes | 0.810 | 0.752 | 0.708 |
| CNN PubMed Embeddings | 0.844 | 0.799 | 0.711 |
| CNN Single Task | 0.847 | 0.796 | 0.706 |
| CNN | 0.854 | 0.802 | 0.714 |
| CNN + Neg Tag | 0.867 | 0.811 | 0.727 |
| CNN + Neg Tag + Dense | 0.880 | 0.812 | 0.733 |
| CNN + Neg Tag + Dense + Lab/Demo | 0.893 | 0.822 | 0.749 |
| BiLSTM | 0.869 | 0.807 | 0.738 |
| BiLSTM + Neg Tag | 0.875 | 0.811 | 0.745 |
| BiLSTM + Neg Tag + Dense | 0.892 | 0.823 | 0.739 |
| BiLSTM + Neg Tag + Dense + Lab/Demo | **0.900** | **0.833** | **0.753** |
| Enc CNN-LSTM | 0.859 | 0.797 | 0.727 |
| Enc CNN-LSTM + Lab/Demo | 0.885 | 0.812 | 0.740 |

Deep EHR: Chronic Disease Prediction Using Medical Notes
https://arxiv.org/abs/1808.04928

# How to handle images?

Demographic and lifestyle

Time

**Medications:**
-NDC code (drug name)
-Quantity
-Date of fill

**Encounters**
-Free Text Notes
-Diagnosis code (ICD10s)
-Procedure (CPTs)
-Specialty
-Location of service
-Service Provider ID
-Inpatient/outpatient
-Cost

**Radiology Imaging:**
**- MRI, CT, PET, etc.**
- Free Text (Radiology notes)
- Assessment codes

**Pathology:**
**- Microscopic images (histopathology)**
- Genetic test
- Free text assessments

**Lab Tests:**
-LOINC code (urine or blood test name)
-Results (actual values/Flags)
-Date

# Using Histopathology Images for Lung Cancer subtype and mutation detection

Nicolas Coudray

Image Analysis Specialist, Applied Bioinformatics Laboratories, NYU Lagnone Medical Center

Research: Deep Learning for Histopathology and Medical Imaging

Shaivi Kochar

Graduate Student, NYU Tandon School of Engineering

Research: Visualization of Deep Learning Models and Generative Adversarial Networks for Histopathology Models

Xianzhi (Viola) Cao

Graduate Student, NYU Center for Data Science

Research: Deep Learning and Natural Language Processing for Clinical Notes

<u>NYU Medical School Collaborators</u>

**Aristotelis Tsirigos**

David Fenyo

Paulo Ocampo

Matija Snuderl

Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning https://www.nature.com/articles/s41591-018-0177-5

MENU

nature medicine

Article | Published: 17 September 2018

# Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning

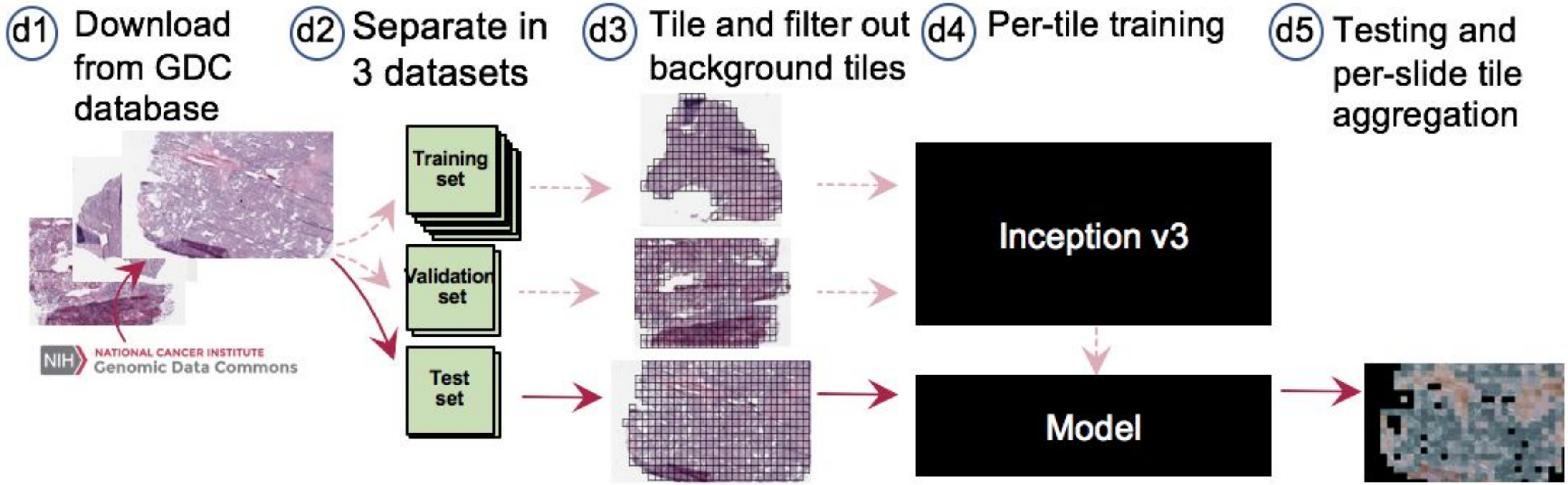
Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian & Aristotelis Tsirigos

*Nature Medicine* **24**, 1559–1567 (2018) | Download Citation

## Abstract

Visual inspection of histopathology slides is one of the main methods used by pathologists to assess the stage, type and subtype of lung

# Our Approach



# Results

| Classification | Information | AUC after aggregation by… ... average predicted probability | AUC after aggregation by… ... percentage of positively classified tiles |
|---|---|---|---|
| **Normal vs Tumor (20x tiles)** | a) Inception v3, fully-trained | 0.993 [0.974-1.000] | 0.990 [0.969-1.000] |
| **LUAD vs LUSC (20x tiles)** | b) Inception v3, transfer learning | 0.847 [0.782-0.906] | 0.844 [0.777-0.904] |
| | c) Inception v3, fully-trained | 0.950 [0.913-0.980] | 0.947 [0.911-0.978] |
| | d) Same as (c) but aggregation done solely on tiles classified as "tumor" by A | 0.952 [0.915-0.981] | 0.949 [0.912-0.980] |
| **LUAD vs LUSC (5x tiles)** | Inception v3, fully-trained | **0.942 [0.907-0.971]** | 0.906 [0.851-0.951] |
| **3 classes. Normal vs LUAD vs LUSC at 20x** | Normal | 0.984 [0.947-1.000] | 0.985 [0.953-1.000] |
| | LUAD | **0.969 [0.933-0.994]** | 0.970 [0.937-0.993] |
| | LUSC | 0.966 [0.935-0.990] | 0.964 [0.932-0.989] |
| | Micro-average | 0.970 [0.950-0.986] | 0.969 [0.949-0.985] |
| | Macro-average | 0.976 [0.949-0.993] | 0.976 [0.950-0.993] |
| **3 classes. Normal vs LUAD vs LUSC at 5x** | Normal | **0.997 [0.993-0.998]** | 0.988 [0.962-1.000] |
| | LUAD | 0.965 [0.942-0.983] | 0.938 [0.896-0.971] |
| | LUSC | **0.977 [0.960-0.991]** | 0.964 [0.937-0.986] |
| | Micro-average | **0.980 [0.972-0.987]** | 0.966 [0.948-0.980] |
| | Macro-average | **0.981 [0.968-0.991]** | 0.964 [0.939-0.980] |

*n=244 slides for LUAD vs LUSC classifiers and n=170 slides for the others, all from 137 patients.*

# Predicting gene mutational status from whole-slide images

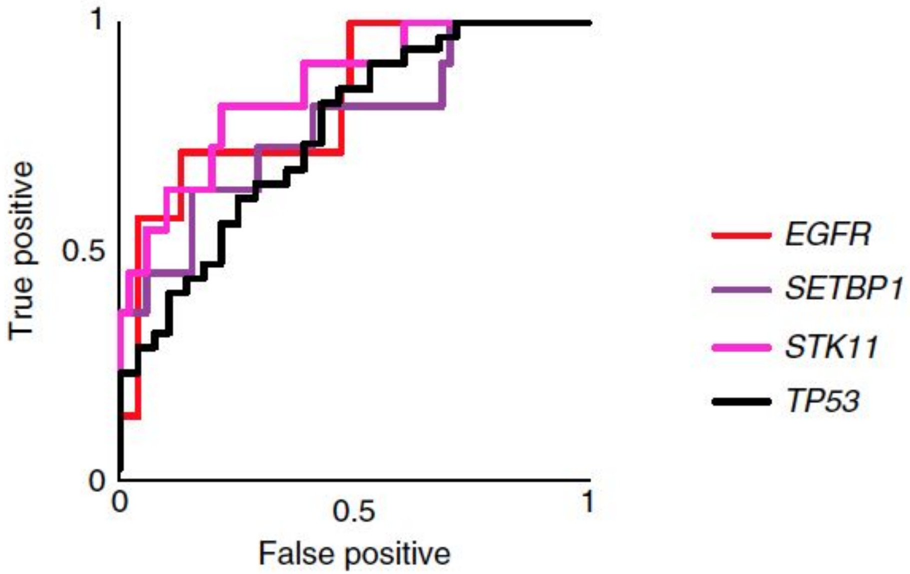
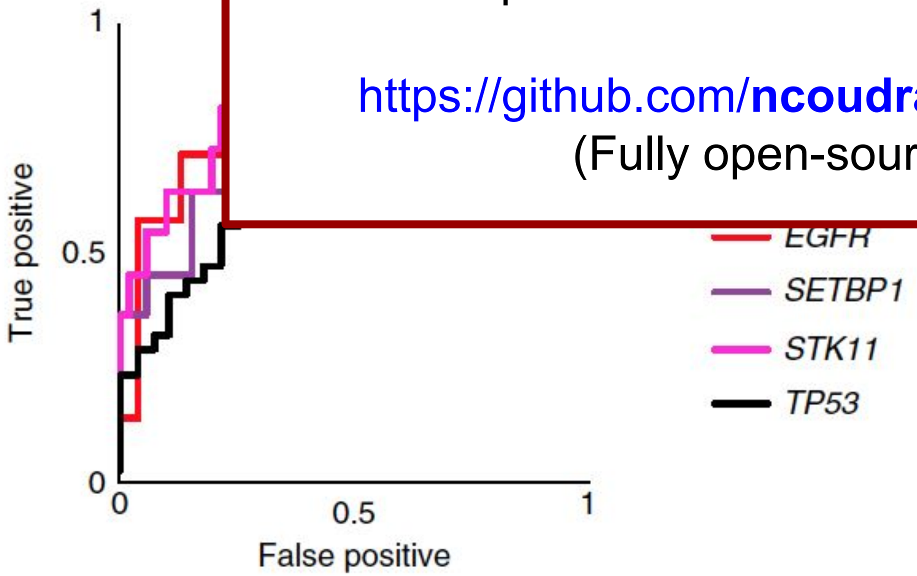



**Table 1 | AUC achieved by the network trained on mutations (with 95% CIs)**

| Mutations | Per-tile AUC | Per-slide AUC after aggregation by… | |
|---|---|---|---|
| | | … average predicted probability | … percentage of positively classified tiles |
| *STK11* | 0.845 (0.838–0.852) | 0.856 (0.709–0.964) | 0.842 (0.683–0.967) |
| *EGFR* | 0.754 (0.746–0.761) | 0.826 (0.628–0.979) | 0.782 (0.516–0.979) |
| *SETBP1* | 0.785 (0.776–0.794) | 0.775 (0.595–0.931) | 0.752 (0.550–0.927) |
| *TP53* | 0.674 (0.666–0.681) | 0.760 (0.626–0.872) | 0.754 (0.627–0.870) |
| *FAT1* | 0.739 (0.732–0.746) | 0.750 (0.512–0.940) | 0.750 (0.491–0.946) |
| *KRAS* | 0.814 (0.807–0.829) | 0.733 (0.580–0.857) | 0.716 (0.552–0.854) |
| *KEAP1* | 0.684 (0.670–0.694) | 0.675 (0.466–0.865) | 0.659 (0.440–0.856) |
| *LRP1B* | 0.640 (0.633–0.647) | 0.656 (0.513–0.797) | 0.657 (0.512–0.799) |
| *FAT4* | 0.768 (0.760–0.775) | 0.642 (0.470–0.799) | 0.640 (0.440–0.856) |
| *NF1* | 0.714 (0.704–0.723) | 0.640 (0.419–0.845) | 0.632 (0.405–0.845) |

$n$ = 62 slides from 59 patients.

# Predicting gene mutational status from whole-slide images




All code and instructions to get the data and reproduce results are available:

https://github.com/**ncoudray/DeepPATH**
(Fully open-sourced)

**Table 1 | AUC achieved by the network trained on mutations (with 95% CIs)**

| Mutations | Per-tile AUC | Per-slide AUC after aggregation by… | |
|---|---|---|---|
| | | … average predicted probability | … percentage of positively classified tiles |
| *STK11* | 0.845 (0.838– | 0.856 (0.709–0.964) | 0.842 (0.683-0.967) |
| | | 8– | 0.782 (0.516-0.979) |
| | | 5– | 0.752 (0.550-0.927) |
| | | 6– | 0.754 (0.627-0.870) |
| | | 2– | 0.750 (0.491-0.946) |
| | (0.807–0.829) | 0– 0.857) | 0.716 (0.552-0.854) |
| *KEAP1* | 0.684 (0.670–0.694) | 0.675 (0.466–0.865) | 0.659 (0.440-0.856) |
| *LRP1B* | 0.640 (0.633–0.647) | 0.656 (0.513–0.797) | 0.657 (0.512-0.799) |
| *FAT4* | 0.768 (0.760–0.775) | 0.642 (0.470–0.799) | 0.640 (0.440-0.856) |
| *NF1* | 0.714 (0.704–0.723) | 0.640 (0.419–0.845) | 0.632 (0.405-0.845) |

*n* = 62 slides from 59 patients.

# Implications and Summary

- AI can fundamentally change how we
  - Screen for Conditions
  - Generate hypotheses
  - Recruit for clinical trials
  - Develop treatments

- Many many supervised learning tasks for next few years
  - Predicting *current* and *future* diseases
  - Predicting from Time series, Text and Images and Between them to save time/costs

- *Deployment* and *workflow changes* remain challenging

# Thank you

Questions and Comments:

narges.razavian@nyumc.org

https://github.com/**ncoudray/DeepPATH**

https://github.com/**NYUMedML/DeepEHR**