Disease Prediction Using Machine Learning and Electronic Health Records

Narges Razavian
Assistant Professor
NYU Langone Medical Center
Departments of Radiology and Population 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
Certain conditions originating in the perinatal period (p00-p96)
Certain infectious and parasitic diseases (a00-b99)
Congenital malformations, deformations and chromosomal abnormalities (q00-q99)
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism (d50-d89)
Diseases of the circulatory system (i00-i99)
Diseases of the digestive system (k00-k55)
Diseases of the ear and mastoid process (h60-h95)
Diseases of the eye and adnexa (h00-h59)
Diseases of the genitourinary system (n00-n99)
Diseases of the musculoskeletal system and connective tissue (m00-m99)
Diseases of the nervous system (g00-g99)
Diseases of the respiratory system (j00-j99)
Diseases of the skin and subcutaneous tissue (l00-l99)
Endocrine, nutritional and metabolic diseases (e00-e89)
External causes of morbidity (v00-y99)
Factors influencing health status and contact with health services (z00-z99)
Injury, poisoning and certain other consequences of external causes (s00-t88)
Mental, behavioral and neurodevelopmental disorders (f01-f99)
Neoplasms (c00-d49)
Pregnancy, childbirth and the puerperium (o00-o99)
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (r00-r99)
Electronic Health Records

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

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

Pathology:
- Microscopic images (histopathology)
- Genetic test
- Free text assessments
Electronic Health Records for future outcome prediction
Electronic Health Records for future outcome prediction

A Few Years Ago

Today
Electronic Health Records for future outcome prediction

A Few Years Ago

Today

Demographic and lifestyle

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

**Independent variables from early years**

A Few Years Ago

- Demographic and lifestyle
- **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

Today

- **Lab Tests:**
  - LOINC code (urine or blood test name)
  - Results (actual values/Flags)
  - Date
Electronic Health Records for future outcome prediction

A Few Years Ago

Independent variables from early years

- Demographic and lifestyle
- 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

Target outcome in the future

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

Time
Electronic Health Records for future outcome prediction

Independent variables from early years

A Few Years Ago

- Demographic and lifestyle
- Medications: NDC code (drug name), Quantity, Date of fill
- Encounter Records: Free Text Notes, ICD9/10 diagnosis code, CPT code (procedure), Specialty
- Cost

Machine Learning

Target outcome in the future

Today

- Lab Tests: LOINC code (urine or blood test name), Results (actual values/Flags), Date
- Location of service, Service Provider ID, Inpatient/outpatient
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

NYU Collaborators
David Sontag
Rahul Krishnan
Uri Shalit
Jake Marcus
YD Choi
Yoni Halpern

NYU Medical School Collaborators
Saul Blecker
Ann Marie Schmidt
Yin Aphinyanaphong
Leora Horwitz

My Students

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

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

- 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
Temporal convolution in 3 resolutions.

2 Layers of Dropout + Fully connected +ReLU

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
Learning features and Deep Learning/Multitask learning

Multi-task prediction of disease onsets from longitudinal laboratory tests
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**

**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

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

**Pathology:**
- Microscopic images (histopathology)
- Genetic test
- Free text assessments
Using AI + Clinical Notes for Early Detection

NYU Collaborators
Kyunghyun Cho
Sam Bowman

NYU Medical School Collaborators
Yin Aphinyanaphong
Leora Horwitz
Himanshu Grover
Jerko Steiner
Marina Marin
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 running 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
(Fully open-sourced)
Extracting **Structured Data** from the Notes & Combining with Text Data
Dealing with Time

Model: Encounter CNN_LSTM
### Table 1: Number of Records by Target Diseases (Negative Cases : Positive Cases)

<table>
<thead>
<tr>
<th>Target</th>
<th>Training Set</th>
<th>Validation Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Congestive Heart Failure</td>
<td>644K : 4080</td>
<td>93K : 574</td>
<td>184K : 1167</td>
</tr>
<tr>
<td>Kidney Failure</td>
<td>616K : 10051</td>
<td>88K : 1428</td>
<td>176K : 2809</td>
</tr>
<tr>
<td>Stroke</td>
<td>653K : 3195</td>
<td>94K : 406</td>
<td>187K : 916</td>
</tr>
</tbody>
</table>

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

---

Deep EHR: Chronic Disease Prediction Using Medical Notes

How to handle images?

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

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

Pathology:
- Microscopic images (histopathology)
- Genetic test
- Free text assessments
Using Histopathology Images for Lung Cancer subtype and mutation detection

NYU Medical School Collaborators

Aristotelis Tsirigos
David Fenyo
Paulo Ocampo
Matija Snuderl
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 Fenyő, 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

1. Download from GDC database
2. Separate in 3 datasets
3. Tile and filter out background tiles
4. Per-tile training
5. Testing and per-slide tile aggregation
## Results

<table>
<thead>
<tr>
<th>Classification</th>
<th>Information</th>
<th>AUC after aggregation by...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>... average predicted probability</td>
</tr>
<tr>
<td>Normal vs Tumor (20x tiles)</td>
<td>a) Inception v3, fully-trained</td>
<td>0.993 [0.974-1.000]</td>
</tr>
<tr>
<td></td>
<td>b) Inception v3, transfer learning</td>
<td>0.847 [0.782-0.906]</td>
</tr>
<tr>
<td></td>
<td>c) Inception v3, fully-trained</td>
<td>0.950 [0.913-0.980]</td>
</tr>
<tr>
<td></td>
<td>d) Same as (c) but aggregation done solely on tiles classified as &quot;tumor&quot; by A</td>
<td>0.952 [0.915-0.981]</td>
</tr>
<tr>
<td>LUAD vs LUSC (20x tiles)</td>
<td>Inception v3, fully-trained</td>
<td>0.942 [0.907-0.971]</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>0.964 [0.947-1.000]</td>
</tr>
<tr>
<td></td>
<td>LUAD</td>
<td>0.969 [0.933-0.994]</td>
</tr>
<tr>
<td></td>
<td>LUSC</td>
<td>0.966 [0.935-0.990]</td>
</tr>
<tr>
<td></td>
<td>Micro-average</td>
<td>0.970 [0.950-0.986]</td>
</tr>
<tr>
<td></td>
<td>Macro-average</td>
<td>0.976 [0.949-0.993]</td>
</tr>
<tr>
<td>3 classes. Normal vs LUAD vs LUSC at 20x</td>
<td>Normal</td>
<td>0.997 [0.993-0.998]</td>
</tr>
<tr>
<td></td>
<td>LUAD</td>
<td>0.965 [0.942-0.983]</td>
</tr>
<tr>
<td></td>
<td>LUSC</td>
<td>0.977 [0.960-0.991]</td>
</tr>
<tr>
<td></td>
<td>Micro-average</td>
<td>0.980 [0.972-0.987]</td>
</tr>
<tr>
<td></td>
<td>Macro-average</td>
<td>0.981 [0.968-0.991]</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>Mutations</th>
<th>Per-tile AUC</th>
<th>Per-slide AUC after aggregation by...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>... average predicted probability ... percentage of positively classified tiles</td>
</tr>
<tr>
<td>STK11</td>
<td>0.845 (0.838-0.852)</td>
<td>0.856 (0.709-0.964)</td>
</tr>
<tr>
<td>EGFR</td>
<td>0.754 (0.746-0.761)</td>
<td>0.826 (0.626-0.979)</td>
</tr>
<tr>
<td>SETBP1</td>
<td>0.785 (0.776-0.794)</td>
<td>0.775 (0.595-0.931)</td>
</tr>
<tr>
<td>TP53</td>
<td>0.674 (0.666-0.681)</td>
<td>0.760 (0.626-0.872)</td>
</tr>
<tr>
<td>FAT1</td>
<td>0.739 (0.732-0.746)</td>
<td>0.750 (0.512-0.940)</td>
</tr>
<tr>
<td>KRAS</td>
<td>0.814 (0.807-0.829)</td>
<td>0.733 (0.580-0.857)</td>
</tr>
<tr>
<td>KEAP1</td>
<td>0.684 (0.670-0.694)</td>
<td>0.675 (0.466-0.865)</td>
</tr>
<tr>
<td>LRP1B</td>
<td>0.640 (0.633-0.647)</td>
<td>0.656 (0.513-0.797)</td>
</tr>
<tr>
<td>FAT4</td>
<td>0.768 (0.760-0.775)</td>
<td>0.642 (0.470-0.799)</td>
</tr>
<tr>
<td>NFI</td>
<td>0.714 (0.704-0.723)</td>
<td>0.640 (0.419-0.845)</td>
</tr>
</tbody>
</table>

n = 62 slides from 199 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)
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