Deep Patient: Predict the Medical Future of Patients with Artificial Intelligence and EHRs

Riccardo Miotto, Ph.D.

New York, NY

Institute for Next Generation Healthcare
Dept. of Genetics and Genomic Sciences
Icahn School of Medicine
Introduction

• The increasing cost of healthcare has motivated the drive towards preventive medicine
  ✓ predictive approaches to protect, promote and maintain health and to prevent diseases, disability and death

• Personalized medicine
  ✓ approach for disease treatment and prevention that takes into account all aspects of an individual status
Personalized Medicine Framework
Mount Sinai Medical Center

- 7 Member Hospital Campuses in New York, NY
  - > 3,500 hospital beds
  - > 7,000 physicians

- More than 8 million patients
  - about 7-8 TB of electronic health records
### Electronic Health Records (EHRs)

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960</td>
<td>Initial mention of EHRs to record patient information</td>
</tr>
<tr>
<td>1972</td>
<td>Regenstrief Institute develops the first EHR system</td>
</tr>
<tr>
<td>1996</td>
<td>Health Insurance Portability and Accountability Act (HIPAA)</td>
</tr>
<tr>
<td>2000s</td>
<td>Internet revolution and decrease in the cost of information technologies</td>
</tr>
<tr>
<td>2003</td>
<td>Mount Sinai starts to implement EHRs</td>
</tr>
</tbody>
</table>

#### Office-based Physician EHR Adoption

![Office-based Physician EHR Adoption Graph](https://dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php)

**Legend:**
- Blue triangles: Any EHR
- Red circles: Basic EHR
- Orange squares: Certified EHR

---

Electronic Health Records (EHRs)

Most Common EHR Data Types
- Patient demographics
- Clinical notes
- Vital signs
- Medical histories
- Diagnoses
- Medications
- Clinical images
- Laboratory and test results

EHRs were originally aimed for billing and administrative purposes

Great promise in providing tools to support physicians in their daily activities

Still a promise despite (maybe) 10 – 15 years of research
Electronic Health Records (EHRs)

• EHRs are challenging to represent
  o heterogeneous
  o noisy
  o incomplete
  o structured / unstructured
  o inconsistent
  o redundant
  o subject to random errors
  o subject to systematic errors
  o based on different standards
  o …and so and so forth

• The same clinical phenotype can be expressed using different codes and terminologies
  ✓ patient diagnosed with “type 2 diabetes mellitus”
    o laboratory values of hemoglobin A1C greater than 7.0
    o presence of 250.00 ICD-9 code
    o “type 2 diabetes mellitus” mentioned in the free-text clinical notes, and so on
State of the Art

• Systems focused on one specific disease

• Ad-hoc descriptors manually selected by clinicians
  ✓ not scalable
  ✓ misses the patterns that are not known

• Raw vectors composed of all the clinical descriptors
  ✓ sparse, noisy and repetitive
  ✓ not linearly separable and not robust to distortions
  ✓ linear classifiers split the input space into simple regions

• Simple feature learning algorithms
  ✓ not able to model the hierarchical information in the data
Artificial Intelligence with EHRs

**Feature Engineering**
- Data Normalization
- Phenotype Aggregation
- Clinical Note Understanding

**Clinical Applications**
- Disease Prediction
- Drug Recommendation
- Decision-making Support
- Alert Monitoring

**Clinical Research**
- Patient Stratification
- Dataset Composition
- Clinical Trial Recruitment
Artificial Intelligence with EHRs

• **Feature learning is the key**
  ✓ general-purpose vector-based representations for patients and clinical phenotypes
    o embedding in metric spaces where we can compute relationships between items
  ✓ use this representations for supervised and similarity-based tasks and to
    o build the tools to support clinicians and biomedical researchers

• **Neural networks and deep learning with EHRs**
  ✓ hierarchical deep networks fit the multi-modality of EHRs
  ✓ take inspiration from other domains
  ✓ need a little more data preparation
Artificial Intelligence with EHRs

- Objective
  ✓ general-purpose vector-based representations for patient and clinical phenotypes

- Phenotype embedding

- Deep Patient
  ✓ disease prediction
Phenotype Embedding
Motivations

• Medical concepts are treated as discrete atomic symbols with arbitrary codes
  ✓ no useful information regarding the relationships that may exist between the individual symbols

• Data sparsity
  ✓ high-dimensional one-hot vectors are difficult to process

• Hierarchical ontologies
  ✓ limited by the top-down structure
  ✓ difficult to navigate
Vector Space Model

• Learn a dense low-dimensional representation of medical concepts from the EHRs
  ✓ map different phenotypes in a common metrics space
  ✓ the closer two concepts are to each other in the embedded space, the more similar their meaning

• Vector space models
  ✓ long history in natural language processing
    o words that appear in the same context share a semantic mean
    o allows operations on the representations based on similarity measures
EHR Phenotype Embedding

A patient is seen as a sequence of phenotypes

Learning Low-Dimensional Representations of Medical Concepts
EHR Phenotype Embedding

• Patients
  ✓ 1980 – 2015
  ✓ at least one clinical phenotype
  ✓ 1,304,192 unique patients

• EHR Phenotypes
  ✓ ICD-9s
    o 6,272 codes
      o normalized to 4 digits
  ✓ Medications
    o 4,022 codes
      o normalized using the Open Biomedical Annotator
EHR Phenotype Embedding

- Phenotypes
  - Vitals
    - 7 codes
  - Encounter descriptions
    - 10 codes
  - Procedures
    - 2,414 codes
  - Lab tests
    - 1,883 codes

14,608 distinct clinical phenotypes

Normalized by hand

Normalized using an in-house algorithm based on string similarity and sub-string prefixes
EHR Phenotype Embedding

• Every patient was divided in time interval 15 days long
  ✓ each interval is considered one “sentence” for the embedding algorithm
    o retained only the sentences with at least 3 phenotypes

• **7,170,200** sentences
  ✓ used the sentence to train the model
    o word2vec skip-gram
    o context window as large as the sentence length
Embedding Evaluation

Size of the embedded vectors

CCS Level 1
17 categories

CCS Single
252 categories
Embedding Evaluation

Myeloid Leukemia

TensorBoard EMBEDDINGS

DATA

1 tensor found
ehr_embedding

Spheretize data

Load data


Metadata: tensorboard_project/ehr-vocabs.txt

T-SNE

PCA
CUSTOM

Dimension
Perplexity
Learning rate
Re-run
Stop

Iteration 183
For fast results, the data will be sampled down to 10,000 points.

show how to use T-SNE effectively.

Nearest points in the original space:

icd9_myeloid_leukemia 0.102
icd9_acute_myeloid_leukemia 0.165
icd9_myeloid_leukemia_unspecified_cell_type 0.176
icd9_myeloid_leukemia_of_unspecified_cell_type 0.221
icd9_unspecifed_leukemia 0.245
medication_armodilone 0.255
medication_armodilone_sodium 0.300
lactic_acidosis 0.310
lactate_ql 0.315
lactate_graph_vs_host_disease 0.315
lab_rtd_file 0.327
medication_armodilone 0.379
medication_armodilone_sodium 0.379
lactic_acidosis 0.345
medication_armodilone 0.354
lab_rtd_file 0.363
lab_rtd_file 0.365
medication_armodilone_sodium 0.378
medication_armodilone 0.383
lactic_acidosis 0.386
medication_armodilone 0.399
lactic_acidosis 0.409
lab_rtd_file 0.431

BOOKMARKS (0)
# Embedding Evaluation

## Myeloid Leukemia

<table>
<thead>
<tr>
<th><strong>ICD-9s</strong></th>
<th><strong>Medications</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute myeloid leukemia</td>
<td>Azacitidine</td>
</tr>
<tr>
<td>Acute leukemia of unspecified cell type</td>
<td>Voriconazole</td>
</tr>
<tr>
<td>Leukemia of unspecified cell type</td>
<td>Clofarabine</td>
</tr>
<tr>
<td>Chronic myeloid leukemia</td>
<td>Decitabine</td>
</tr>
<tr>
<td>Unspecified leukemia</td>
<td><strong>Posaconazole</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Lab Tests</strong></th>
<th><strong>Procedures</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bcr - Abl</td>
<td>Nm cardiac blood pool oncology</td>
</tr>
<tr>
<td>Flt3 mutation</td>
<td>Ra ir placement of ventral venous catheter</td>
</tr>
<tr>
<td><strong>Cd3+ enrichment</strong></td>
<td>Ra myelogram lumbar puncture</td>
</tr>
<tr>
<td><strong>Cd33+ enrichment</strong></td>
<td>Ra doppler extremity veins complete</td>
</tr>
<tr>
<td>Engraft-post trans</td>
<td>Rm ir central access line removal</td>
</tr>
</tbody>
</table>
Embedding Evaluation

Schizophrenic Disorders
Schizophrenic Disorders

**ICD-9s**
- Paranoid type schizophrenia
- Unspecified schizophrenia
- Disorganized type schizophrenia
- Schizoaffective disorder
- Schizophrenic disorders residual type

**Medications**
- Haloperidol decanoate
- Clozapine
- Fluphenazine
- Benztropine mesylate
- **Nicotine polacrilex**

**Lab Tests**
- Valproic acide
- Drug abuse
- Lithium
- Clozapine
- Norclozapine

**Procedures**
- Screening mammography
# Diabetes Mellitus

## ICD-9s
- Unspecified essential hypertension
- Essential hypertension
- Diabetes with unspecified complications
- Other and unspecified hyperlipidemia
- Diabetes with renal manifestations

## Medications
- Sitagliptin phosphate
- Alcohol
- Glipizide
- Insulin Glargine
- Glimepiride

## Lab Tests
- Microalbumin / Creatine
- Microalbumin
- Fructosamine
- Cholesterol ratio
- Triglycerides

## Procedures
- Electrocardiogram tracing
- Electrocardiogram complete
Embedding Evaluation

Ventolin
It can treat or prevent bronchospasm

**ICD-9s**
- Other specified asthma
- Chronic obstructive asthma
- Asphyxia and hypoxemia
- Acute bronchiolitis
- Asthma unspecified

**Medications**
- Proventil
- Flovent
- Proair
- Atrovent
- Singulair

Pregabalin
It can treat nerve and muscle pain, including fibromyalgia. It can also treat seizures

**ICD-9s**
- Neuralgia neuritis and radiculitis unspecified
- Mononeuritis of lower limb
- Mononeuritis of unspecified site
- Chronic pain
- Thoracic or lumbosacral neuritis

**Medications**
- Gabapentin
- Duloxetine
- Tramadol
- Tizanidine
- **Nortriptyline**
Embedding Evaluation

Combined query

Cocaine dependence (ICD-9) + Drug dependence (ICD-9)
|↓|
|Cannabis dependence (ICD-9)

Cannabis dependence (ICD-9) + Cocaine dependence (ICD-9)
|↓|
|Antisocial personality disorder (ICD-9)

Cannabis dependence (ICD-9) - Cocaine dependence (ICD-9)
|↓|
|Disturbance of emotions specific to childhood and adolescence (ICD-9)

Sciatica (ICD-9) + Chronic pain (ICD-9)
|↓|
|Cyclobenzaprine (Medication)
Potential Use Cases

- Expand a query by medical concept to include nearby concepts
  - ✓ search for patients eligible for clinical trials

- Speed up inclusion criteria
  - ✓ facilitate the compositions of specific patient datasets

- Low-dimensional and dense representations to use in machine learning applications

- Knowledge resource
  - ✓ computer scientist and engineers approaching the medical domain without prior knowledge
Deep Patient
Deep learning to process patient data to derive representations that aim to be domain free, dense, robust, lower-dimensional, and that can be effectively used to predict patient future events
Deep Patient: Overall Framework

EHRs are extracted from the clinical data warehouse and are aggregated by patient.

Unsupervised deep feature learning to derive the patient representations.

Predict patient future events from the deep representations.
Deep Patient: Learning

• Multi-layer neural network
  ✓ each layer of the network produces a higher-level representation of the observed patterns, based on the data it receives as input from the layer below, by optimizing a local unsupervised criterion

• Hierarchically combine the clinical descriptors into a more compact, non-redundant and unified representation through a sequence of non-linear transformations
Deep Patient: Data Processing

- Patients data available in the data warehouse

- Normalize the clinically relevant phenotypes
  ✓ group together the similar concepts in the same clinical category to reduce information dispersion

- Aggregate data by patients in a vector form
  ✓ *bag of phenotypes*
Deep Patient: Architecture

- The first layer receives as input the EHR bag of phenotypes
- Every intermediate level is fed with the output of the previous layer
- The last layer outputs the Deep Patient representations
Deep Patient: Implementation

Denoising Autoencoder

1. Corruption strategy
2. Encoder
3. Decoder
4. Minimize the difference between the original input and the reconstruction
Deep Patient: Application

• Apply the deep system to the patient EHRs
  ✓ deep patient data warehouse

• Analytics
  ✓ clustering
  ✓ similarity
  ✓ topology analysis

• Predict future events
  ✓ standalone classifier
  ✓ fine-tuned neural network
    o e.g., logistic regression layer
Disease Prediction
Disease Prediction: Experiment

• Disease Prediction
  ✓ predict the probability that patients might develop a certain new disease within a certain amount of time given their current clinical status

• Training Set
  ✓ patient data between 1980 – 2013, inclusive
    o about 1.6 millions patients

• Test Set
  ✓ 100k different patients
    o evaluation on the new diagnoses of 2014
  ✓ 79 diseases
    o oncology, endocrinology, cardiology, etc.
Disease Prediction: Evaluation

- Pipeline
  - train the feature learning models
  - train one-vs.-all classifier per each disease
  - apply the models to each patient in the test dataset and predict their probability to develop every disease in the vocabulary

- Each patient is represented by a vector of disease risk probabilities

- Evaluate the quality of the predictions over different temporal windows
Disease Prediction: Models

Feature Learning

Deep Patient (3 layers)
Raw EHRs
Principal Component Analysis (PCA)
K-Means
Gaussian Mixture Model (GMM)
Independent Component Analysis (ICA)

Classification

Random Forest
Support Vector Machine (SVM)
Deep Logistic Regression Network
Disease Prediction: Evaluation by Disease

Determine if a disease is likely to be diagnosed to patients within one-year interval
Disease Prediction: Evaluation by Disease

Results in terms of the Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

Deep Logistic Regression Network

**AUC-ROC = 0.79**
### Disease Prediction: Evaluation by Disease

#### Deep Logistic Regression Network

<table>
<thead>
<tr>
<th>Disease</th>
<th>AUC-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer of Liver</td>
<td>0.93</td>
</tr>
<tr>
<td>Regional Enteritis and Ulcerative Colitis</td>
<td>0.91</td>
</tr>
<tr>
<td>Type 2 Diabetes Mellitus</td>
<td>0.91</td>
</tr>
<tr>
<td>Congestive Heart Failure</td>
<td>0.90</td>
</tr>
<tr>
<td>Chronic Kidney Disease</td>
<td>0.89</td>
</tr>
<tr>
<td>Personality Disorders</td>
<td>0.89</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>0.88</td>
</tr>
<tr>
<td>Multiple Myeloma</td>
<td>0.87</td>
</tr>
<tr>
<td>Delirium and Dementia</td>
<td>0.85</td>
</tr>
<tr>
<td>Coronary Atherosclerosis</td>
<td>0.84</td>
</tr>
</tbody>
</table>
Disease Prediction: Evaluation by Patient

Evaluate the risk to develop diseases for each patient over different temporal windows
Disease Prediction: Evaluation by Patient

Classification based on Random Forest models

Deep Patient significantly outperforms the other models
Conclusions
Deep Patient: summary

• Pros
  ✓ Deep Patient enables to leverage EHRs towards improved patient representations
  ✓ The model requires the same input format as simpler feature learning models
    o Deep Patient can help to improve previous medical studies based on EHRs

• Cons
  ✓ representations are not interpretable
    o interpretability is a key only on predictive tasks
  ✓ time is not modeled
Deep Patient: vs. the Others

Predict diagnoses and medications for the subsequent visit

Doctor AI: Predicting Clinical Events via Recurrent Neural Networks
Edward Choi, Mohammad Taha Bahadori
College of Computing
Georgia Institute of Technology
Atlanta, GA, USA
Andy Schuetz, Walter F. Stewart
Research Development & Dissemination
Sutter Health
Walnut Creek, CA, USA
Jimeng Sun
College of Computing
Georgia Institute of Technology
Atlanta, GA, USA

Heart failure prediction

RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism

Edward Choi*, Mohammad Taha Bahadori*, Joshua A. Kulas*,
Andy Schuetz†, Walter F. Stewart†, Jimeng Sun*
* Georgia Institute of Technology † Sutter Health
{mp2893, bahadori, jkulas3}@gatech.edu,
{schuetzal, stewartf}@sutterhealth.org, jsun@cc.gatech.edu

DeepCare: A Deep Dynamic Memory Model for Predictive Medicine
Trang Pham, Truyen Tran, Dinh Phung and Svetla Venkatesh

Disease progression modeling, future risk prediction, intervention recommendation

Deepr: A Convolutional Net for Medical Records
Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, Svetla Venkatesh

Predict unplanned readmission after discharge
Artificial Intelligence with EHRs

• Feature enrichment
  ✓ use as many descriptors as possible from the EHRs

• Temporal modeling
  ✓ timing is important for a better understanding of the patient condition and for providing timely clinical decision support

• Interpretable predictions
  ✓ the clinician needs to trust the machine predictions

• Federated inference
  ✓ building a deep learning model by leveraging the patients from different sites without leaking their sensitive information
Artificial Intelligence with EHRs

• Patient representations are the key towards better AI models for EHRs
  ✓ from the representations, you can build the tools to support clinician activities

• Deep learning can be used to leverage the information in the EHRs

• Exciting opportunities for AI with EHRs
  ✓ generative models
    o GANs for missing values
  ✓ lab results processing
  ✓ genetics and EHRs
  ✓ wearable data and EHRs
Acknowledgements

• Joel T. Dudley
• Brian A. Kidd
• Li Li

This work is supported by funding from the NIH National Center for Advancing Translational Sciences (NCATS) Clinical and Translational Science Awards (UL1TR001433), National Cancer Institute (NCI) (U54CA189201), and National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (R01DK098242) to J.T.D.

References


Contacts: riccardo.miotto@mssm.edu