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

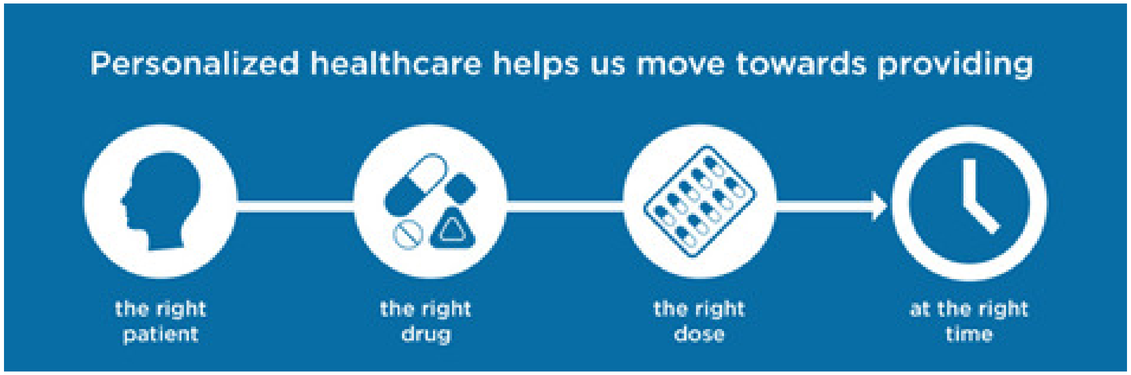
Riccardo Miotto, Ph.D.

New York, NY

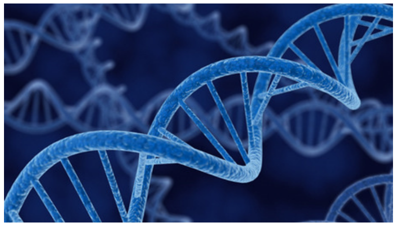
Institute for Next Generation Healthcare
Dept. of Genetics and Genomic Sciences
Icahn School of Medicine

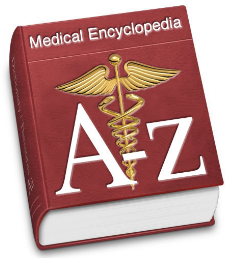
Introduction

- 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



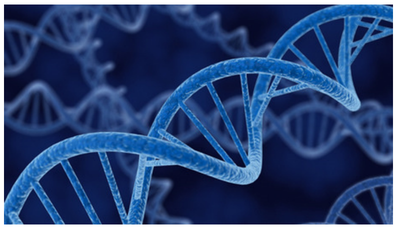
Personalized Medicine Framework

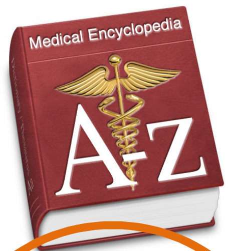




AI

Personalized Medicine Framework





Electronic Health Records

AI

Mount Sinai Medical Center

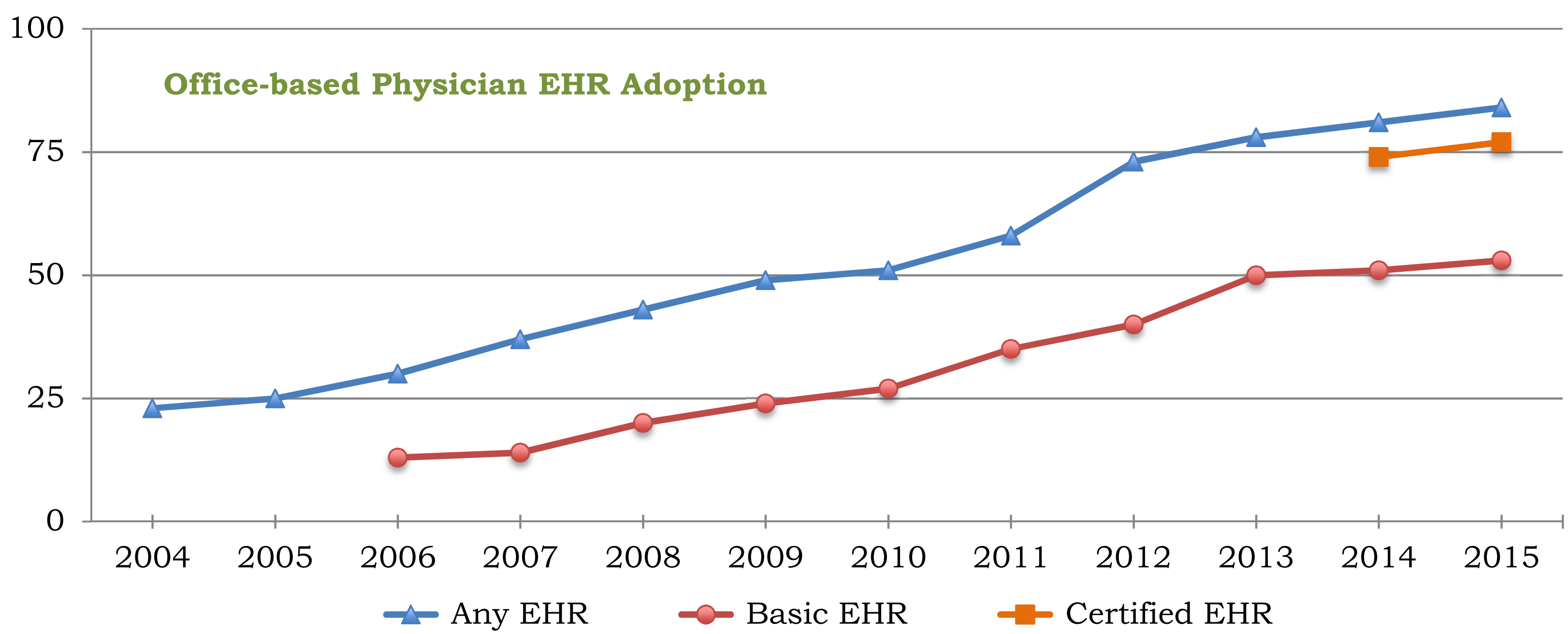
Mount Sinai

Founded in 1852

- 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)

1960	1972	1996	2000s	2003
Initial mention of EHRs to record patient information	Regenstrief Institute develops the first EHR system	Health Insurance Portability and Accountability Act (HIPAA)	Internet revolution and decrease in the cost of information technologies	Mount Sinai starts to implement EHRs



https://dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php

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
  - heterogeneous
  - noisy
  - incomplete
  - structured / unstructured
  - inconsistent
  - redundant
  - subject to random errors
  - subject to systematic errors
  - based on different standards
  - *…and so and so forth*

- The same clinical phenotype can be expressed using different codes and terminologies
  - ✓ patient diagnosed with "**type 2 diabetes mellitus**"
    - laboratory values of hemoglobin A1C greater than 7.0
    - presence of 250.00 ICD-9 code
    - "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

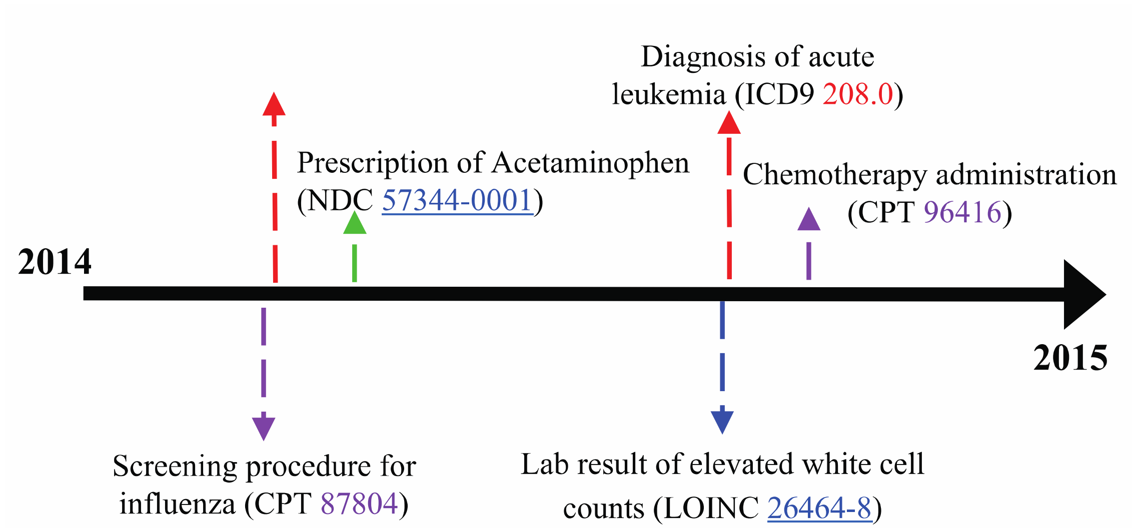
Vector Space Model

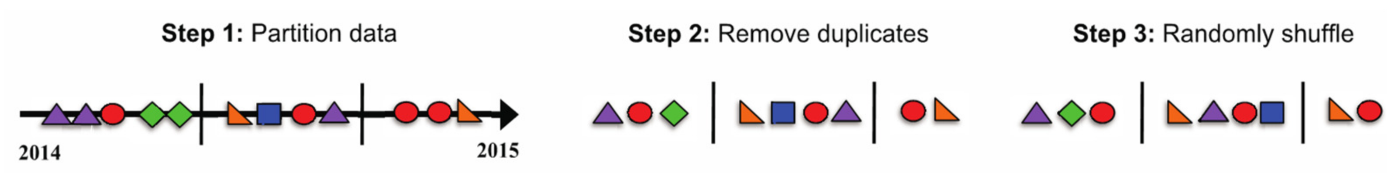
- Learn a dense low-dimensional representation of medical concepts from the EHRs
  - ✓ map different phenotypes in a common metrics space
  - ✓ the closer two concepts are to each other in the embedded space, the more similar their meaning

- 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
Choi Y, Chiu CY, and Sontag D. In the Proceedings of the AMIA Joint Summits Transl Sci Proc, 2016

EHR Phenotype Embedding

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

- EHR Phenotypes
  - ✓ ICD-9s
    - **6,272** codes
    - *normalized to 4 digits*
  - ✓ Medications
    - **4,022** codes
    - *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

Normalized by hand

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

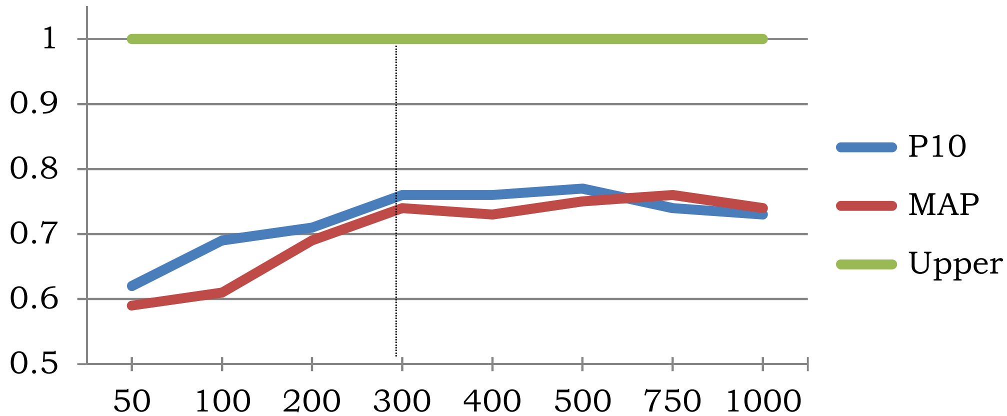
⇒ **14,608** distinct clinical phenotypes

EHR Phenotype Embedding

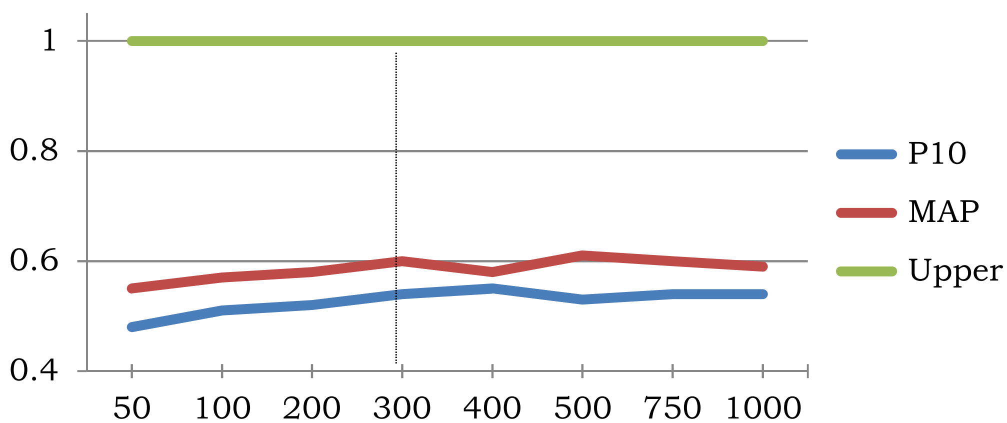
- Every patient was divided in time interval 15 days long
  - ✓ each interval is considered one “sentence” for the embedding algorithm
    - o retained only the sentences with at least 3 phenotypes
- **7,170,200** sentences
  - ✓ used the sentence to train the model
    - o word2vec skip-gram
    - o context window as large as the sentence length

Embedding Evaluation

Size of the embedded vectors



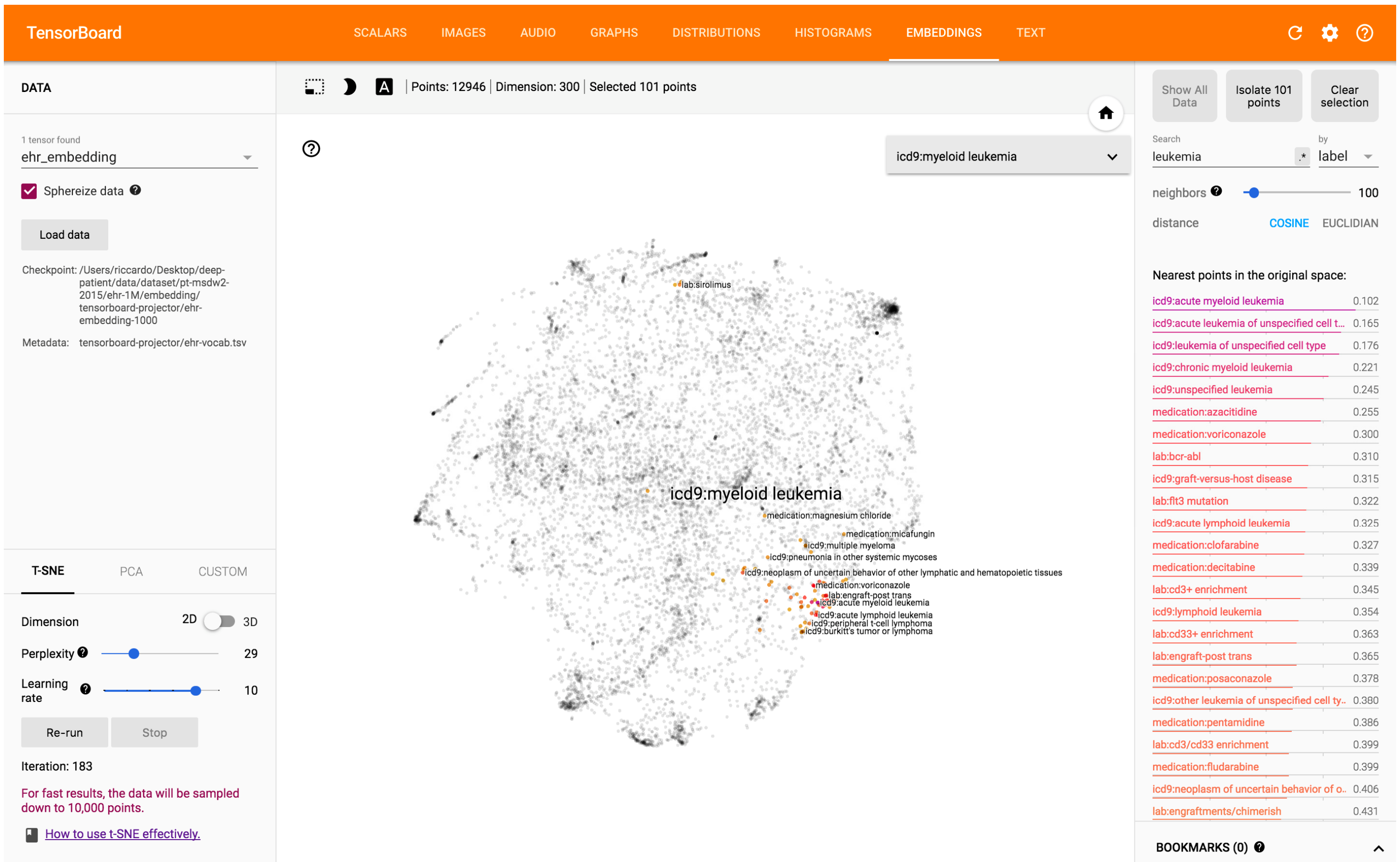
CCS Level 1
17 categories



CCS Single
252 categories

Embedding Evaluation

Myeloid Leukemia



Embedding Evaluation

Myeloid Leukemia

ICD-9s

Acute myeloid leukemia
Acute leukemia of unspecified cell type
Leukemia of unspecified cell type
Chronic myeloid leukemia
Unspecified leukemia

Medications

Azacitidine
Voriconazole
Clofarabine
Decitabine
Posaconazole

Lab Tests

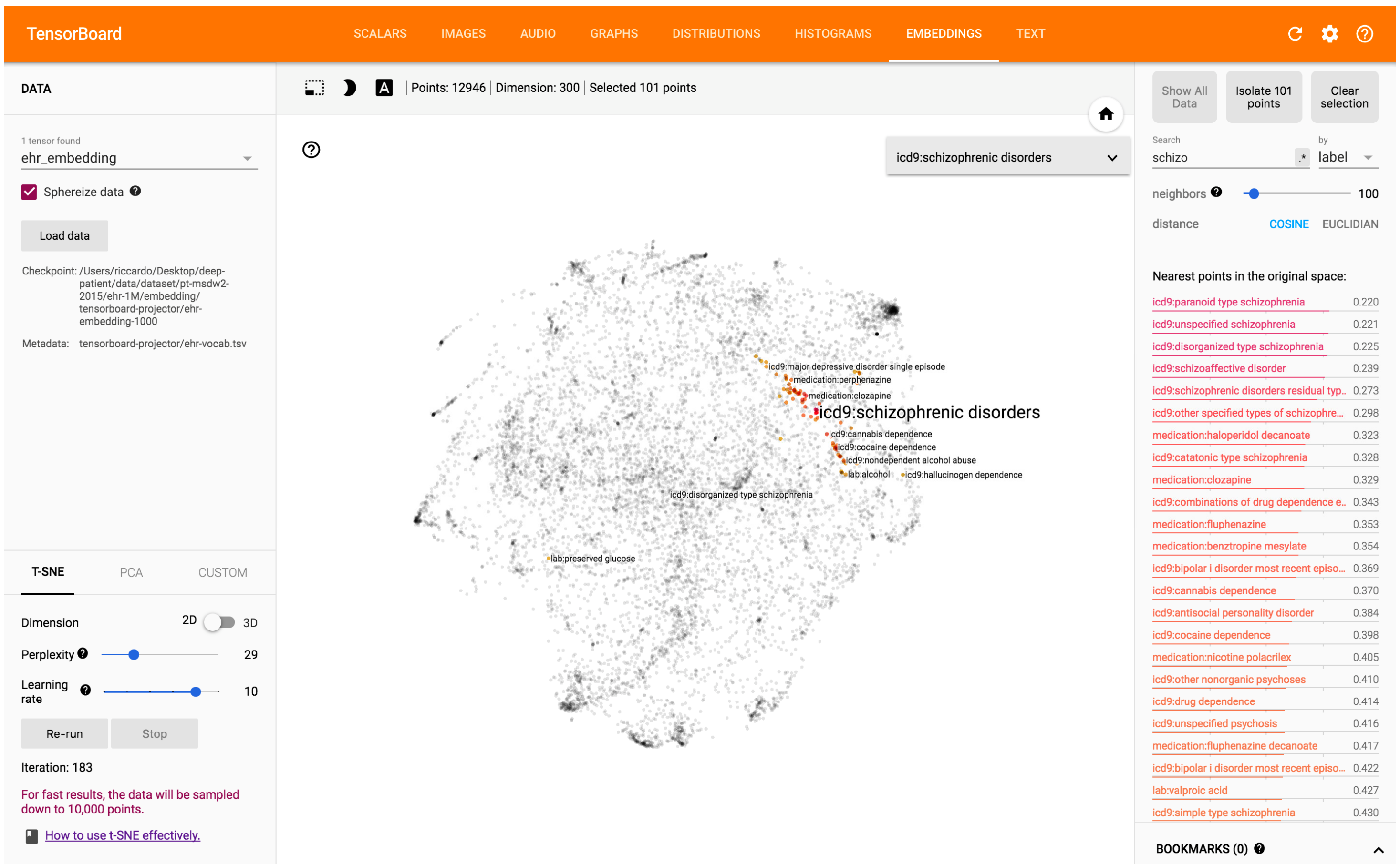
Bcr - Abl
Flt3 mutation
Cd3+ enrichment
Cd33+ enrichment
Engraft-post trans

Procedures

Nm cardiac blood pool oncology
Ra ir placement of ventral venous catheter
Ra myelogram lumbar puncture
Ra doppler extremity veins complete
Rm ir central access line removal

Embedding Evaluation

Schizophrenic Disorders



Embedding Evaluation

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

Embedding Evaluation

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	**Medications**
Other specified asthma	Proventil
Chronic obstructive asthma	Flovent
Asphyxia and hypoxemia	Proair
Acute bronchiolitis	Atrovent
Asthma unspecified	Singulair

Pregabalin

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

ICD-9s	**Medications**
Neuralgia neuritis and radiculitis unspecified	Gabapentin
Mononeuritis of lower limb	Duloxetine
Mononeuritis of unspecified site	Tramadol
Chronic pain	Tizanidine
Thoracic or lumbosacral neuritis	**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 Patient

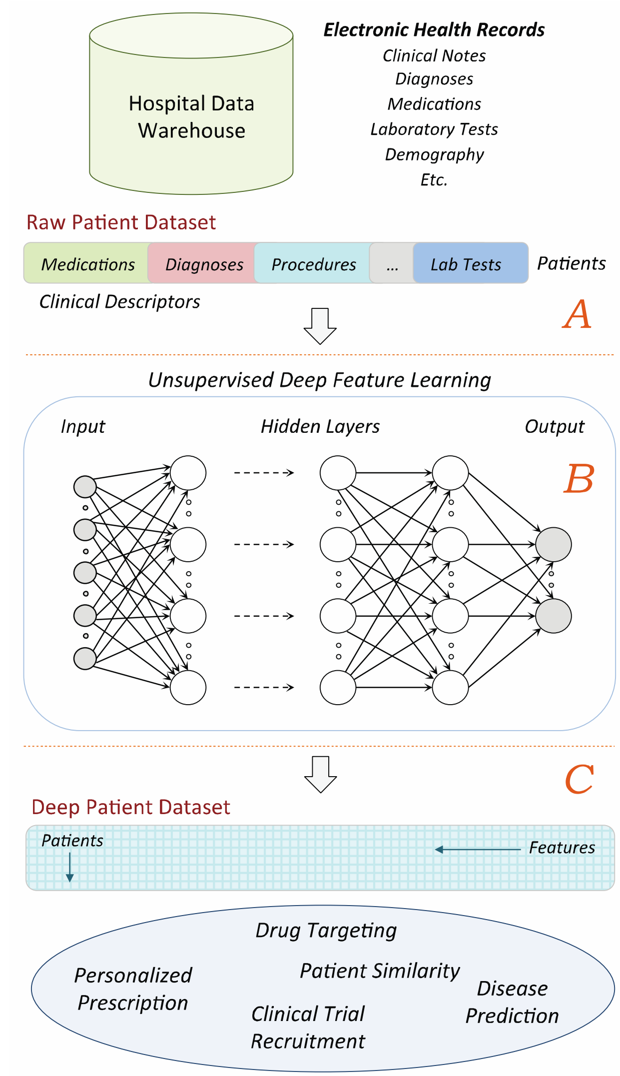
Deep learning to process patient data to derive representations that aim to be domain free, dense, robust, lower-dimensional, and that can be effectively used to predict patient future events

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



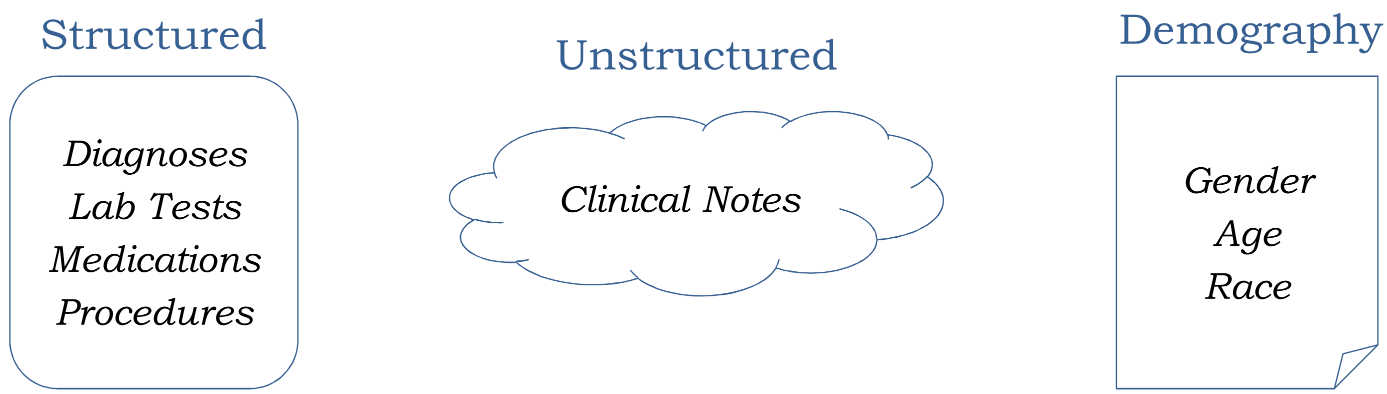
Deep Patient: Learning

- 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

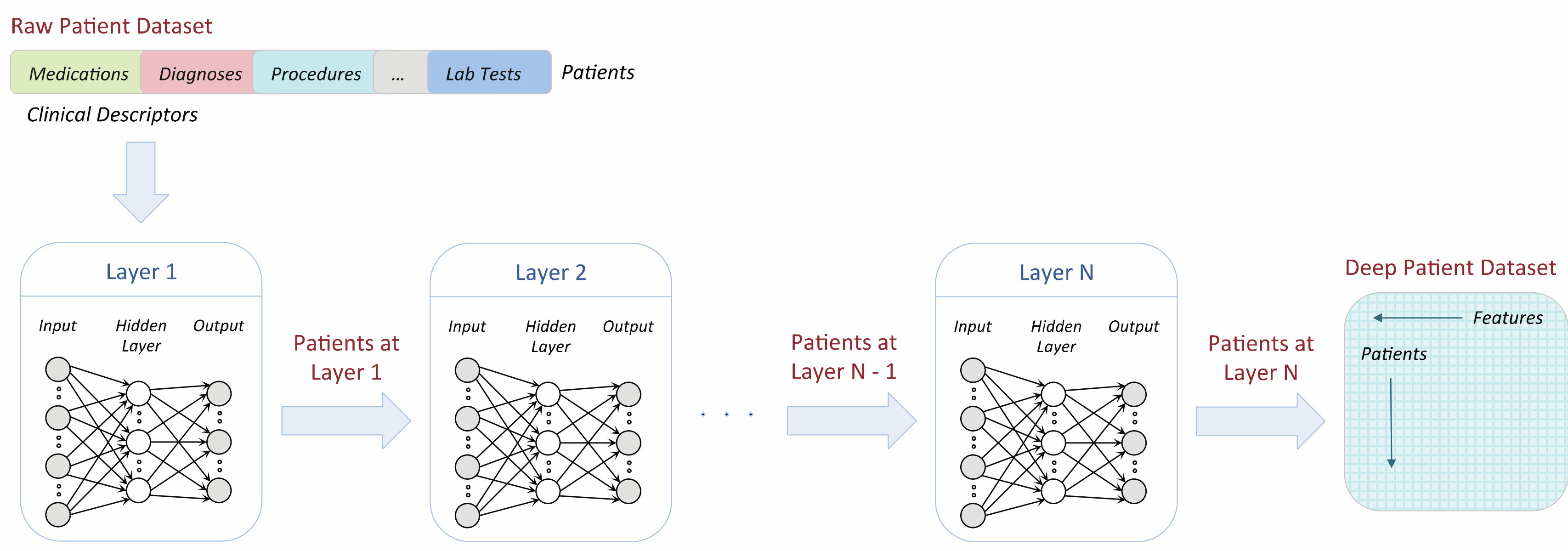
Deep Patient: Data Processing

- 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*

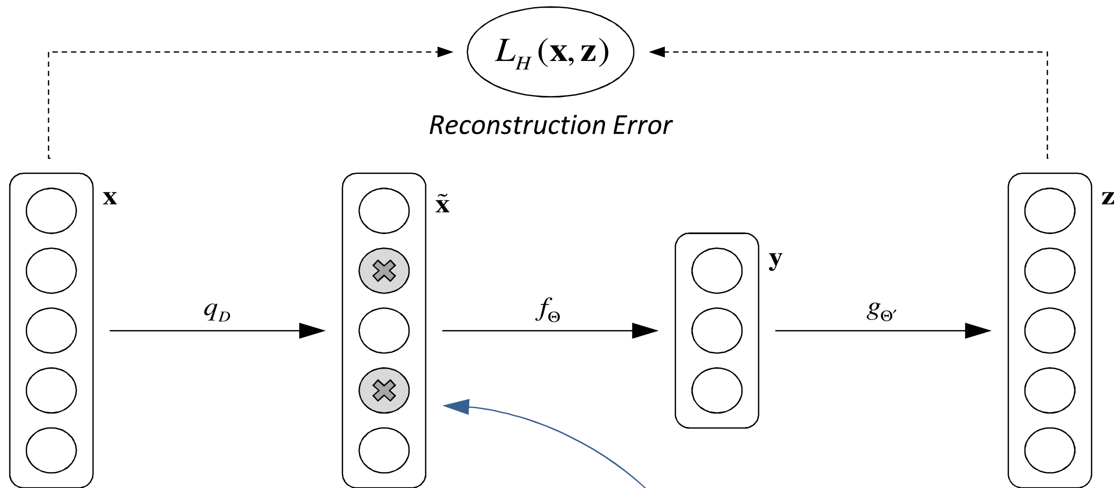
Deep Patient: Architecture



- 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

masking noise corruption strategy

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
    - 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

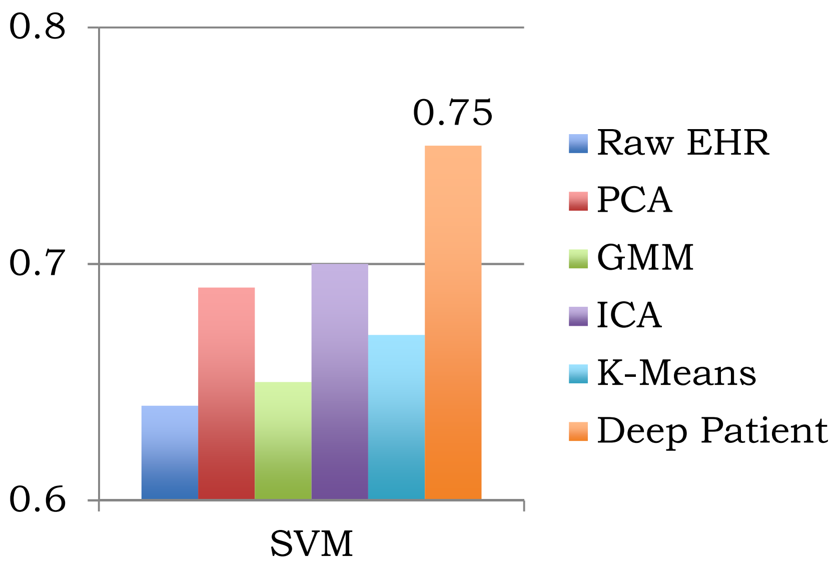
Support Vector Machine (SVM)

Deep Logistic Regression Network

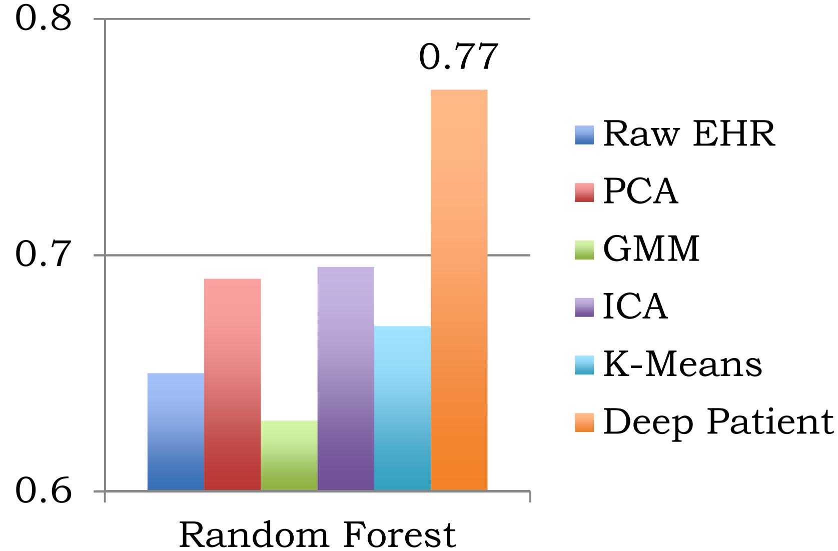
Disease Prediction: Evaluation by Disease

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

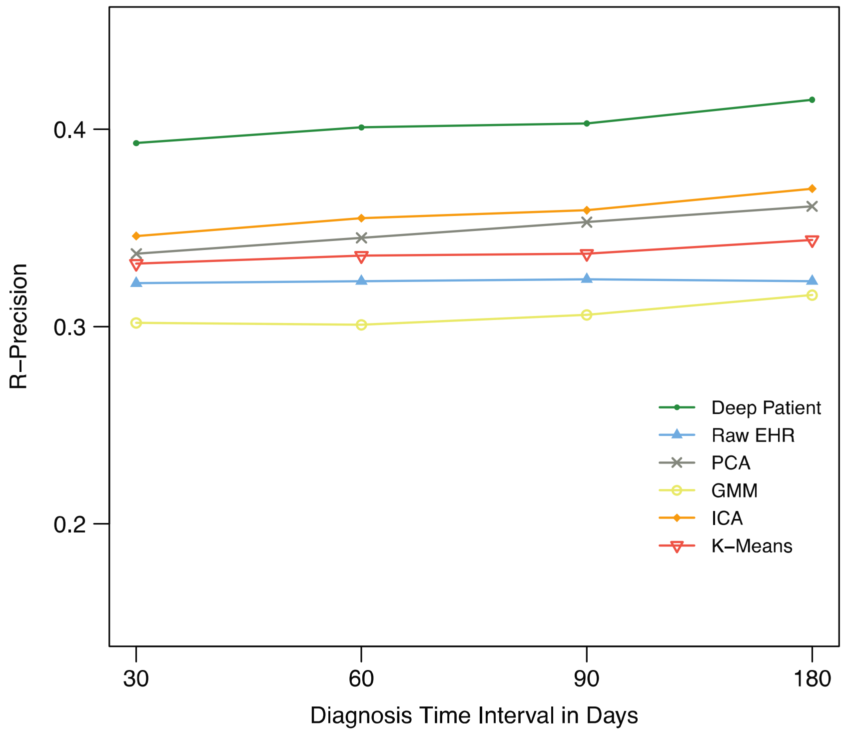
Deep Logistic Regression Network

Disease	AUC-ROC
Cancer of Liver	0.93
Regional Enteritis and Ulcerative Colitis	0.91
Type 2 Diabetes Mellitus	0.91
Congestive Heart Failure	0.90
Chronic Kidney Disease	0.89
Personality Disorders	0.89
Schizophrenia	0.88
Multiple Myeloma	0.87
Delirium and Dementia	0.85
Coronary Atherosclerosis	0.84

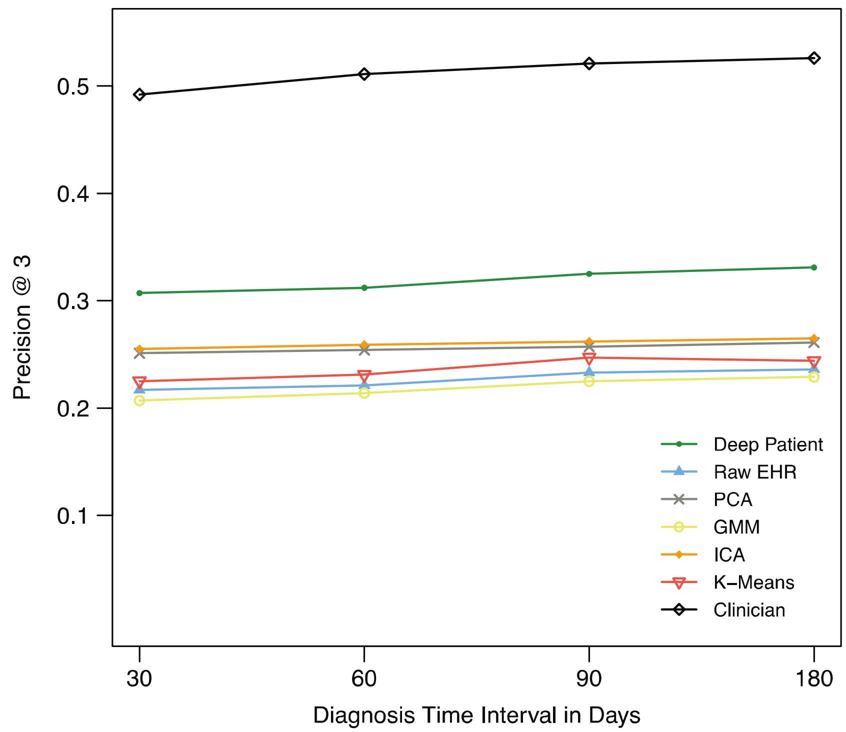
Disease Prediction: Evaluation by Patient

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, {schueta1,stewarwf}@sutterhealth.org, jsun@cc.gatech.edu

DeepCare: A Deep Dynamic Memory Model for Predictive Medicine

Trang Pham, Truyen Tran, Dinh Phung and Svetha Venkatesh

Disease progression modeling, future risk prediction, intervention recommendation

Deepr: A Convolutional Net for Medical Records

Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, Svetha 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

Miotto, R., Wang F., Wang, S., Jiang, X., and Dudley J.T. **Deep Learning for Healthcare: Review, Opportunities and Challenges.** Briefings in Bioinformatics, 2017 (to appear).

Miotto, R., Li, L., Kidd, B.A., and Dudley, J.T. **Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records**. *Nature Scientific Reports*, 6: 26094, 2016.

Miotto, R., Li, L. and Dudley, J.T. **Deep Learning to Predict Patient Future Diseases from the Electronic Health Records**. *In the Proceedings of the European Conference on Information Retrieval (ECIR).* Springer International Publishing, pp. 768 – 774, 2016.

Contacts: riccardo.miotto@mssm.edu