Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records

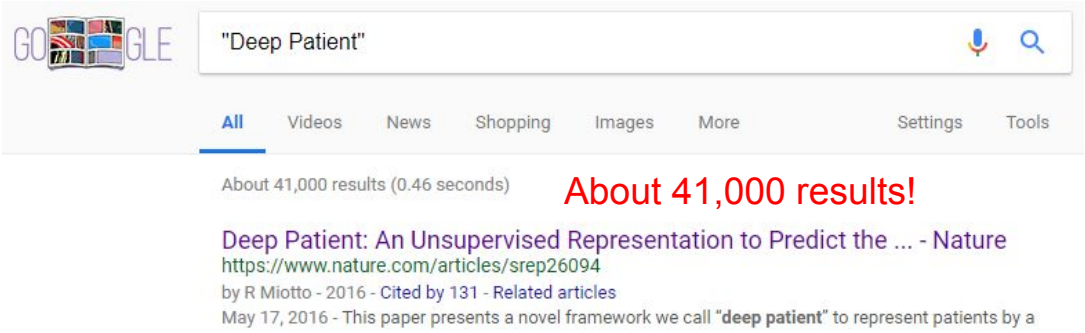
- **Riccardo Miotto et al.**

Presentation By: Jue Wang, Rachneet Kaur

Impact of the paper

- Social Media
- News

Social Media



Online attention

Altmetric score (what's this?)

- Tweeted by **1247**
- Blogged by **2**
- On **10** Facebook pages
- Mentioned in **6** Google+ posts
- Picked up by **14** news outlets
- **7** Reddit
- **1** Video

5 readers on Citeulike

This Altmetric score means that the article is:

- in the 99th percentile (ranked 249th) of the 276,512 tracked articles of a similar age in all journals
- in the 99th percentile (ranked 5th) of the 3,072 tracked articles of a similar age in *Scientific Reports*

Mentions in news, blogs & Google+

News articles (14) | **Scientific blogs (2)** | **Google+ posts (6)**

Can machines really tell us if we're sick?
The Conversation

Can Machines Really Tell Us If We're Sick?
Biotech in Asia

Nvidia says deep learning is about to revolutionize medicine
MIT Technology Review

Nvidia says deep learning is about to revolutionize medicine
MIT Technology Review

From coding to cancer: How AI is changing medicine
CNBC

News report

THE CONVERSATION

Academic rigor, journalistic flair

Search analysis, research, academics...

Arts + Culture | Economy + Business | Education | Environment + Energy | Ethics + Religion | Health + Medicine | Politics + Society | Science + Technology

Can machines really tell us if we're sick?

January 26, 2017 10.47pm EST

Machines don't make the same errors as humans when it comes to decisions based on visual analysis. from www.shutterstock.com

Email | Twitter 36 | Facebook 77 | LinkedIn | Print

This week US scientists announced they have developed an algorithm, or a computerised tool, to identify skin cancers through analysis of photographs.

Rather than relying on human eyes, the new method scans a photo of a patch of skin to look for common and dangerous forms of skin cancer. The authors report their approach performs on par with board-certified dermatologists to distinguish two forms of cancer, keratinocyte carcinoma and malignant melanoma, from benign skin lesions.

The skin cancer diagnostic tool is based on a powerful type of machine learning that extracts information from images. The critical factor in achieving the accuracy and reliability required for a medical diagnostic tool is the large volume of training data the authors have used. This data consists of 129,450 skin images, and a label for each which indicates whether it contains a cancerous region. The machine is trained on this data to make the distinction automatically.

Part of what distinguishes this approach is that it can analyse images taken with a simple hand-held camera, such as the ones on most phones. This means a GP, or even a patient, could take a photo of a patch of skin that presents concerns and receive an indication as to whether it contains a cancerous region.

But translating this research result into a clinical product that can be used for practical diagnosis will require significant further development, documentation, and testing.

Author

Anton van den Hengel
Professor of Computer Science, University of Adelaide

Disclosure statement

Anton van den Hengel receives research funding from the Australian Research Council, LBT Innovations Ltd., Signostics Ltd., and Elex Medical Pty. Ltd.

Partners

THE UNIVERSITY of ADELAIDE

University of Adelaide provides funding as a member of The Conversation AU.

View all partners

Republish this article

NBR PRODUCED BY CNBC

HOME | WATCH NBR | TOPICS | VOICES | TRANSCRIPTS | STATION FINDER | CONTACT US | RETIREMENT

https://www.cnbc.com/video/3000617713

From coding to cancer: How AI is changing medicine

MAY 11, 2017 | *Meg Tirrell, NBR, CNBC.com*

Like 4 | Tweet | G+ | Share 0 | Share 13

Regina Barzilay teaches computers how to learn. A professor at the Massachusetts Institute of Technology, her work focused on natural language processing – training computers to understand human speech – until a breast cancer diagnosis three years ago.

"Going through it, I realized that today we have more sophisticated technology to select your shoes on Amazon than to adjust treatments for cancer patients," Barzilay said in an interview at her office in Cambridge. "I really wanted to make sure that the expertise we have would be used for helping people."

Barzilay's group, in collaboration with Massachusetts General Hospital, is now applying their expertise in artificial intelligence and machine learning to improve cancer diagnosis and treatment. They're asking questions like whether computers can detect signs of breast cancer in mammograms earlier than humans are currently capable of, and whether machine learning can enable doctors to use all the huge quantities of data available on patients to make more personalized treatment decisions.

It's a field some say is on the cusp of changing medicine.

"The potential is perhaps the biggest in any type of technology we've ever had in the field of medicine," said Dr. Eric Topol, director of the Scripps Translational Science Institute. "Computing capability can transcend what a human being could ever do in their lifetime."

Investment is pouring in, from tech giants like IBM's Watson, Alphabet and Philips, to pharmaceutical companies and swiftly proliferating startups. The market for artificial intelligence in health care and the life sciences is projected to grow by 40 percent a year, to $6.6 billion in 2021, according to estimates from Frost & Sullivan.

News report

NVIDIA GPU Technology Conference.

'Deep patient' may point the way to better care

By Eric Barnes, AuntMinnie.com staff writer

May 12, 2017 -- SAN JOSE, CA - With so much research focused on disease diagnosis using artificial intelligence, it's easy to overlook a largely untapped resource that could wield at least as much firepower to reinvent healthcare for the better. That resource is the deep patient, according to a May 11 talk at the NVIDIA GPU Technology Conference.

How do you build the deep patient? In a nutshell, you use deep learning to process patient data and derive representations of the patient that can be used to predict diseases that might develop -- and do it better than current methods, explained Riccardo Miotto, PhD, a data scientist at Icahn School of Medicine at Mount Sinai in New York City.

"The idea ... is basically to use deep learning to process patient data in the electronic health record ... that can be effectively used to predict future patient events and for other unsupervised clinical tasks," Miotto said in his talk.

Predicting outcomes from raw data

The deep-patient framework involves extracting electronic health records (EHRs) from a data warehouse and aggregating them by patient. The warehouse includes different kinds of data, such as structured data in the form of lab tests, medications, and procedures; unstructured data, including clinical notes; and demographic data such as age, gender, and race.

The deep-patient process takes the different data types and normalizes clinically relevant disease phenotypes, grouping similar concepts in the same clinical category to reduce information dispersion, he said. This grouping is known as a "bag of phenotypes." Users can then employ the data to make predictions about patient outcomes, perform drug targeting, predict patient similarities, or do other data modeling tasks.

For the study, Miotto led a team that modeled the data architecture on a pipeline that had worked in the lab. In

Riccardo Miotto, PhD, from Icahn School of Medicine at Mount Sinai.

Background

- Electronic health records
- Journal background
- Previous works

Electronic Health Records

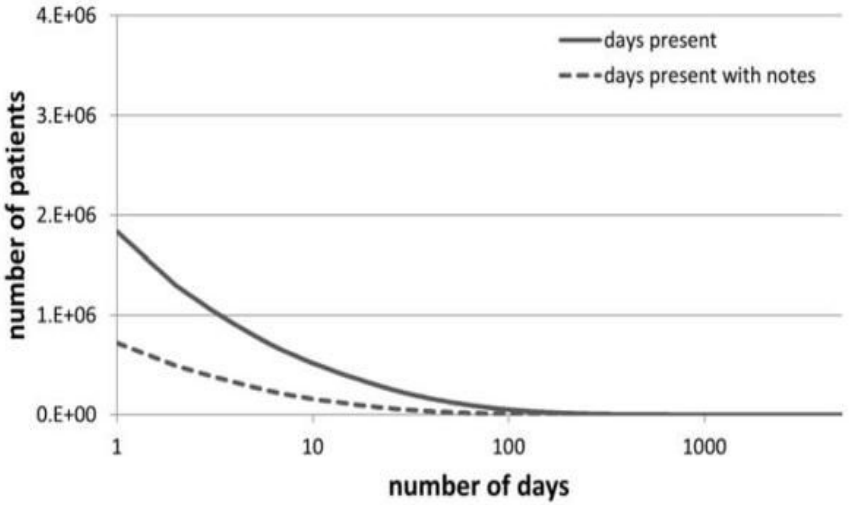
- **Digital Real Time** version of a patient's paper chart
- Information from **all clinicians** involved in a patient's care.
- Eliminates concerns about **illegibility**.
- Offer promise for accelerating Clinical Research and **Predictive Analysis**.
- **Structured** and **Textual** data

Discussion on EHR Data for Machine Learning

- **Empty or Inaccurate data fields**:
  - May reject EHRs for technical malfunctions.
  - Copy and paste for routine or follow-up visits.

- **Difficult to analyse:**
  - Record is thousands of pages
  - Textual data is difficult to analyse

- **Some major challenges:**
  - High dimensional data
  - Noisy data
  - Heterogeneity
  - Sparseness
  - Random errors
  - Same representations using different expressions



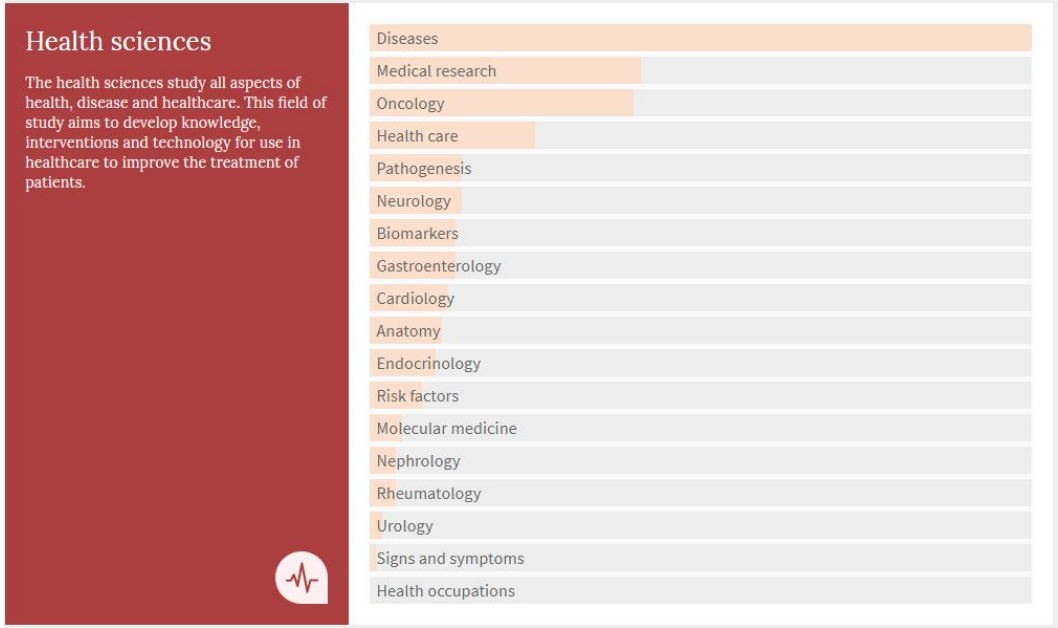
(Abstracted from Weiskopf et al., 2013)

Journal Background

Scientific Reports, a *Nature* journal, publishes in natural and clinical sciences.

Publish scientifically valid primary research from all areas of the natural and clinical sciences

- 5-year impact factor: 4.8
- Article influence score: 1.4
- Considers solely scientific validity



Authors' Background:

- Riccardo Miotto
  - Data Scientist
- Li Li and Brian A. Kidd
  - Genetics and Genomic Sciences
- Joel T. Dudley
  - Genetics and Genomic Sciences
  - Medicine

Previous works and shortcomings

- **Supervised feature selection:**
  - **Experts designate** the patterns and clinical variables
  - Scales poorly, **does not generalize**
  - **Novel patterns** and features not detected
- **Data-driven approaches (RawFeat):**
  - Patient representation:
    - 2D vector with **all available data descriptors**
  - **Sparse, noisy, repetitive** representation
  - Unsuitable for modelling hierarchical or **latent information**

Papers' Approach

- **Unsupervised feature selection:**
  - **Automatically** identifies patterns
  - **General** representation
  - Representation **better understood by machines**
  - **Easier to build classifiers** on
  - **Improves prediction for diverse clinical condition types**
- **Deep Neural Networks:**
  - Captures hierarchical and **latent information** in EHRs.

Previous Approaches:

- Supervised feature selection
- Data Driven Approach

Deep Patient:

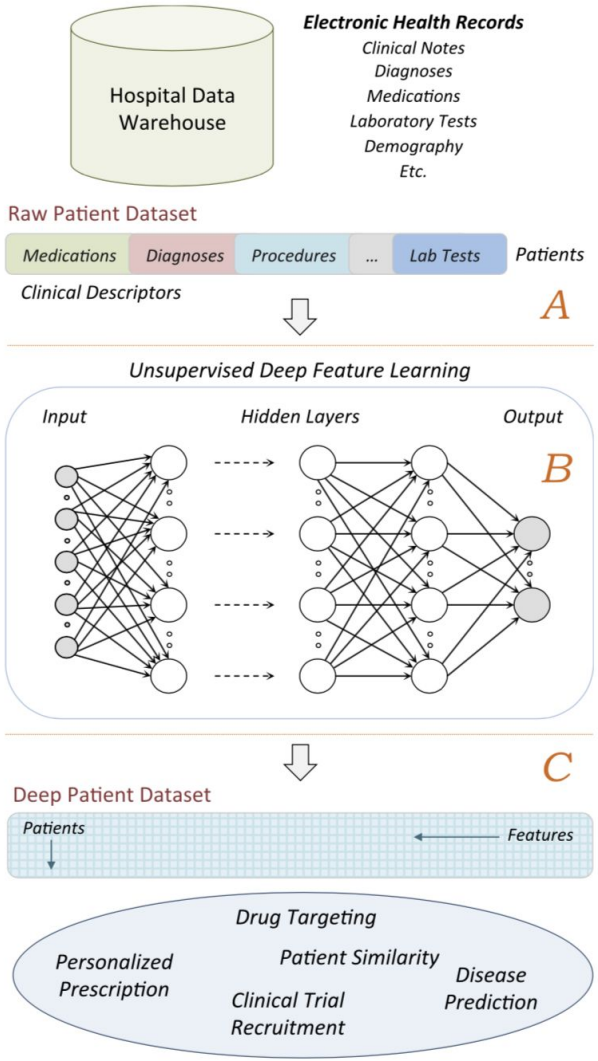
- Unsupervised feature selection
- Deep Learning:
  - Stack of Denoising Autoencoders
- **Domain Free**
- **No human effort** required
- Works for **both supervised and unsupervised** applications

Technical Ideas

Deep Patient Methodology

- Data Preprocessing
- Model:
  - Three layers of Denoising Autoencoder
  - Sigmoid Activation Function
  - 500 hidden units per layer
  - 5% Noise corruption factor
- Loss function:
  - Reconstruction Cross Entropy function
- Optimization:
  - Mini-batch SGD
- Evaluation:
  - Area under ROC curve
  - F-score
  - Accuracy

Framework to derive the Deep Patient Representation



Preprocessing

Raw Patient Representation

Unsupervised Deep Architecture

Robust Features

Applying features to Hospital data

Deep Patient Representation

Data Preprocessing:

- Dataset used:
  - EHR of **Mount Sinai, NY Data Warehouse** with at least 1 record
  - Training Data:
    - 700,000 patients and 78 diseases
    - Records upto Dec 31, 2013
    - **Sampling** 200,000 patients with at least 5 records
  - Test Data:
    - 76, 214 patients and 78 diseases
    - Records in 2014
    - At least 10 records
- Input to the Feature Learning Algorithm:
  - Raw patient vectors using relevant normalized phenotypes

Data Preprocessing:

- EHR Preprocessing:
  - **Open Biomedical Annotator**: Extracts biomedical concepts from text
  - **Negated tag:** Irrelevant and discarded
  - **Family history tag**: Differentiated from the patient related tags
  - **Duplicated information:** Removed
  - Sparseness: Reduced
  - Note summarization:
    - Topic Modelling using **Latent Dirichlet Allocation**
    - 300 topics: Estimation of no. of topics using Perplexity analysis

Semantic abstraction of information

One Topic based representation for all notes of each patient

Data Preprocessing:

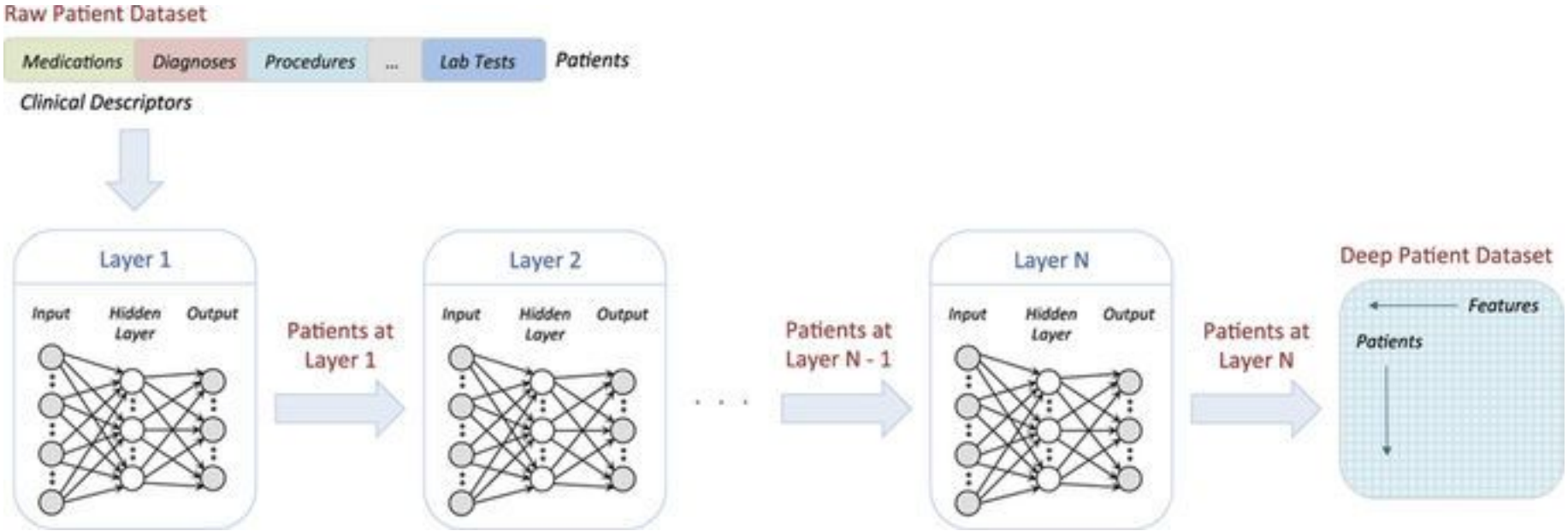
- Disease Selection
  - To state the diagnosis of a disease: ICD 9 codes
    - **Disease labels were used by ICD-9 codes**, which are considered not so accurate.
  - 231 disease definitions: Different codes referring to same disease
  - Further Filtering based on:
    - Diseases with at least 10 training patients
    - Diseases related to:
      - Social behaviour
      - External Life events
      - Too general (Other cancers), were discarded

⇨ 78 Disease Vocabulary

- Descriptor Selection:
  - Less than 80% patients
  - More than 5 patients
  - Raw dataset ~ 1% of entries in Patient Descriptor matrix

⇨ 41072 Descriptors

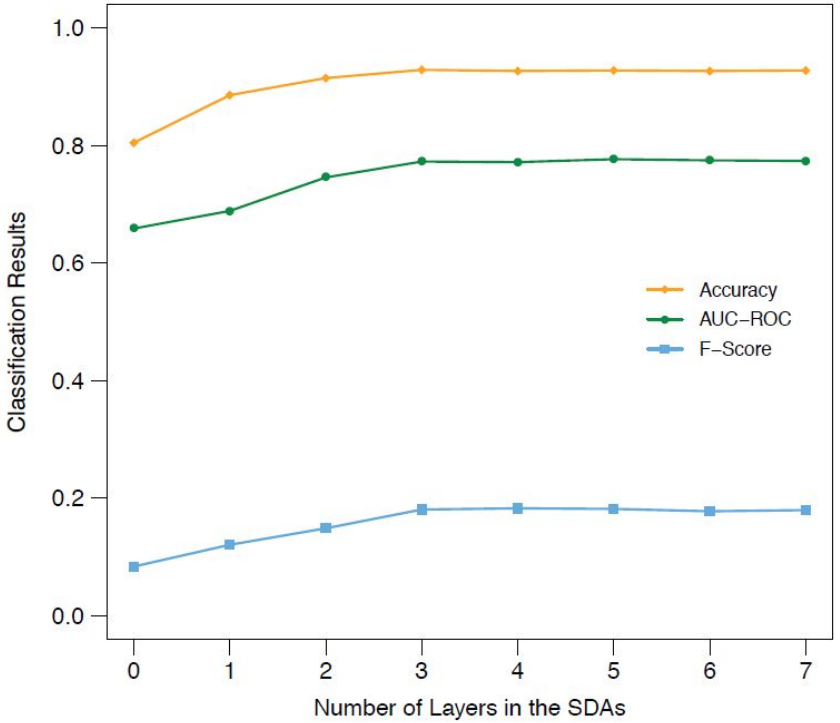
Unsupervised Deep feature learning pipeline:



Exploring the no. of layers in the deep architecture

- Training:
  - 200,000 patients and 78 diseases
- Testing:
  - 76214 patients and 78 diseases
- Evaluation Metrics:
  - AUC-ROC
  - Accuracy
  - F-score
- Classification results stabilize after 3 layers of SDA

⇨ **3 Layers in SDA is optimal**



Model:

- Stack of 3 Denoising Autoencoders
  - Trained independently layer by layer
  - Same structure and functionality for all

Input: $x \in [0,1]^d$

Output: Hidden representation

$$y = s(Wx + b) \quad y \in [0,1]^{d'}$$

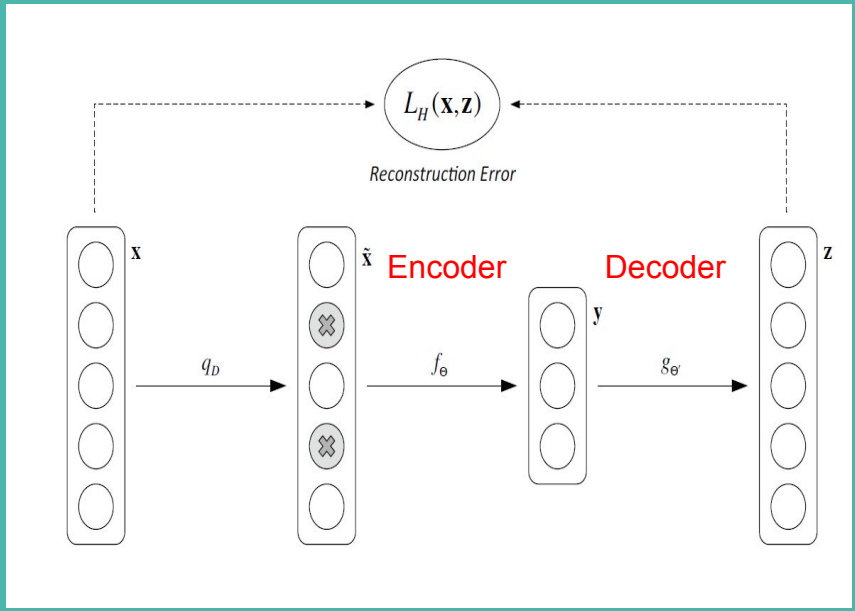
Reconstructed Vector:

$$z = s(W'x + b')$$

- Minimize Average Reconstruction Error

$$L_H(x, z)$$

Denoising Autoencoder Architecture



- 3 layers of Denoising Autoencoders
- Sigmoid Activation Function
- 500 hidden units per layer
- 5% Noise corruption factor

Denoising Autoencoders

- **Denoising:**
  - Noisy version of Initial data used
  - Prevents Overfitting
  - **Fills artificially introduced blanks**
- Procedure to Denoise:
  - Corrupt the input through stochastic mapping $x^* \sim q_D(x^* | x)$
    - **Fraction of randomly selected elements are zeroed - Reflects Missing EHR Data**
  - Map the corrupted input to hidden code $y* = s(Wx^* + b)$
  - Map back to the reconstructed vector $z^* = s(W'y^* + b')$
- **Each patient is a dense vector of 500 features**

Future Disease Prediction:

- Random Forest Classifier (one vs all learning)
  - Assumed that:
    - Better performing
    - Easy to tune
    - Robust to overfitting
  - 100 Trees

No Verification Provided

- **Each Patient is represented as a vector of disease probabilities**

GIT for an implementation of “Deep Patient”: https://github.com/natoromano/deep-patient

Preprocessing and Model Discussion Points:

- Records only from one region (Mount Sinai, NY) were collected both in training and test datasets
- Random sampling used for the training set
- Cross validation (how many folds)
- At least 10 records for each test case and 5 records for each training case
- Same structure and functionality for all autoencoders in the SDA model
- Disease Selection Procedure
- Descriptor Selection Procedure
- Missing Data filled using Denoising of Autoencoders
- Use of the reconstruction cross entropy loss function (justify)
- Use of the Mini Batch Stochastic Gradient Descent Optimization (justify)
- A negated tag was considered non relevant and discarded
- Use of Family history tag in the analysis
- Use of Random Forest Classifier

Results

- Evaluation by disease
  - To measure **prediction of patient developing new diseases within one year**
- Evaluation by Patient
  - Disease predictions with score > 0.6
  - Temporal windows (30, 60, 90 days)
- Model comparison:
  - PCA with 100 PCs
  - K - means with 500 clusters
  - GMM with 200 mixtures
  - ICA with 100 PCs
  - Deep Patient
- Score Comparison Metrics:
  - AUC-ROC
  - Accuracy
  - F-score

Evaluation by disease

- Prediction accuracy of each **Feature Extraction strategy**:
  - AUC-ROC
  - Accuracy
  - F-score
    - Threshold: 0.6
  - t -test for Statistical Significance

Time Interval = 1 year (76,214 patients)			
		Classification Threshold = 0.6	
Patient Representation	**AUC-ROC**	**Accuracy**	**F-Score**
RawFeat	0.659	0.805	0.084
PCA	0.696	0.879	0.104
GMM	0.632	0.891	0.072
K-Means	0.672	0.887	0.093
ICA	0.695	0.882	0.101
DeepPatient	**0.773***	**0.929***	**0.181***

Evaluation by disease

- **Disease Wise accuracy:**
  - AUC-ROC
    - Raw Feat
    - PCA
    - Deep Patient

Time Interval = 1 year (76,214 patients)			
	Area under the ROC curve		
Disease	RawFeat	PCA	DeepPatient
Diabetes mellitus with complications	0.794	0.861	**0.907**
Cancer of rectum and anus	0.863	0.821	**0.887**
Cancer of liver and intrahepatic bile duct	0.830	0.867	**0.886**
Regional enteritis and ulcerative colitis	0.814	0.843	**0.870**
Congestive heart failure (non-hypertensive)	0.808	0.808	**0.865**
Attention-deficit and disruptive behavior disorders	0.730	0.797	**0.863**
Cancer of prostate	0.692	0.820	**0.859**
Schizophrenia	0.791	0.788	**0.853**
Multiple myeloma	0.783	0.739	**0.849**
Acute myocardial infarction	0.771	0.775	**0.847**

Discussion on Disease Classification Results:

Best Predicted Diseases:

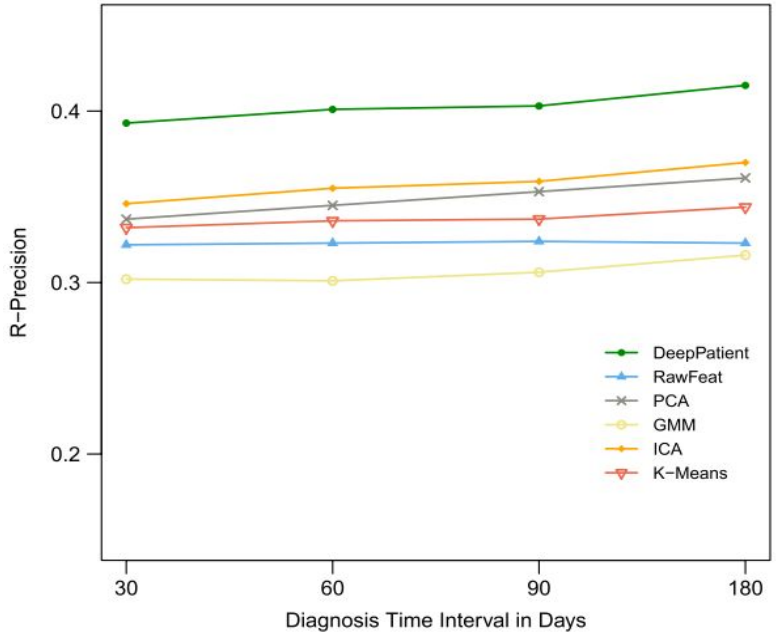
Deep Patient attains highest accuracy on every disease but "*Cancer of brain and nervous system*" - **Why?**

Worst Predicted Diseases:

	Area under the ROC curve		
Disease	**RawFeat**	**PCA**	**DeepPatient**
Diabetes mellitus with complications	0.794	0.861	**0.907**
Cancer of rectum and anus	0.863	0.821	**0.887**
Cancer of liver and intrahepatic bile duct	0.830	0.867	**0.886**
Regional enteritis and ulcerative colitis	0.814	0.843	**0.870**
Congestive heart failure (non-hypertensive)	0.808	0.808	**0.865**
Anxiety disorders	0.572	0.564	**0.605**
Deficiency and other anemia	0.567	0.576	**0.603**
Diabetes mellitus without complications	0.564	0.552	**0.586**
Hypertension	0.536	0.528	**0.574**
Disorders of lipid metabolism	0.549	0.527	**0.561**

Analysis of what leads to the distinction!

Evaluation by patient



Time Interval	Metrics	UppBnd	Patient Representation			
			RawFeat	PCA	ICA	DeepPatient
30 days *(16,374 patients)*	Prec@1	1.000	0.319	0.343	0.345	**0.392***
	Prec@3	0.492	0.217	0.251	0.255	**0.277***
	Prec@5	0.319	0.191	0.214	0.215	**0.226***
60 days *(21,924 patients)*	Prec@1	1.000	0.329	0.349	0.353	**0.402***
	Prec@3	0.511	0.221	0.254	0.259	**0.282***
	Prec@5	0.335	0.199	0.216	0.219	**0.230***
90 days *(25,220 patients)*	Prec@1	1.000	0.332	0.353	0.360	**0.404***
	Prec@3	0.521	0.243	0.257	0.262	**0.285***
	Prec@5	0.345	0.201	0.219	0.220	**0.232***
180 days *(33,607 patients)*	Prec@1	1.000	0.331	0.361	0.363	**0.418***
	Prec@3	0.549	0.246	0.261	0.265	**0.290***
	Prec@5	0.370	0.207	0.221	0.224	**0.236***

- Evaluation of Precision for different time intervals
- UppBnd - Best results achievable

Discussion

- Significance/ Advantages
  - Not Optimized for any specific task
- Applications
  - Patient Clustering and similarity
  - Treatment recommendations etc.
- Limitations/Future Work
  - Including lab test results etc.
- Piazza Points
  - Time series data etc.

The Good

- **Large Dataset (**34 years and 1.2 million patients)
- Captures hierarchical regularities and **patterns**
- Achieves **low dimensional** representation of EHR data
- Outperforms **original EHR** representation
- Helps **scaling Hospital** data warehouse
- **Domain Free** (Not optimized for any specific task)
    - Can fit different clinical applications
- **No human effort** (Unsupervised)

- Applicable to **supervised and unsupervised**
- Outperforms **other feature extraction schemes:**
    - PCA, ICA, K means, GMM.
- Representation **better understood by machines**
    - Easier to build classifiers on
- **Evaluation using 3 different metrics:**
    - AUC-ROC, F score, Accuracy

The Good

- Comparison with **different number of layers** for deep architecture selection
- **Topic based representation** of a patient averaged over all the notes
- **EHR Pre processing**: Topic modelling, Open Biomedical Annotator
- SDAs and feature learning do not focus on a particular clinical descriptor
    - Learns **Non Domain specific descriptors**

Applications

- General Feature learning
  - Personalized Prescriptions
  - Treatment Recommendations
- Unsupervised vector oriented representation
  - Patient Clustering and Similarity
- Scales for billion records
- Single Representative distribution
- Updating model corresponding to change in patient population
- Safe exchange of Data between hospitals
  - Joint Feature learning between hospitals
- Early prediction may help alert care providers

Limitations

- Classification results **only for Random Forest**
- Dataset used only from a **single region**
- No comparison **with other optimization schemes or activation functions**
- No analysis on **factors affecting the performance** of model on various diseases
- Not all diseases had high predictive power
- **Didn't compare with T. Lasko, PLoS ONE, 2013 [7]** with similar methodology and dataset.
- **Disease labels were used by ICD-9 codes**, which are considered not so accurate.

Future Improvements

- Inclusion of **lab test values** may improve the situation for the diseases with low prediction power
- **Temporal sequence** of vectors representation instead of one vector is expected to improve results
- **Additional features** to the EHR dataset
  - Insurance, Family history and behavioral details
- **Preprocessing using PCA and then Deep** Modelling
- Application to a **specific clinical domain** to qualitatively evaluate outcomes
- Inclusion of data from more institutions in the model

Some Points to discuss

- Use of **EHR dataset**
- **78 diseases** out of 231 general diseases
- **Distribution of the diseases** in the dataset
- Dataset used from **only one region** for both training and testing
- **Negated tags** were considered non relevant
- **Family tags** were separated
- Choice of at **least 5 records per patient for training**
- Choice of at **least 10 records per patient for test**
- **Descriptor selection** Procedure
- Additional features to the EHR dataset
    - **Insurance, Family history** and behavioral details

Some other interesting Piazza Ideas

- Use of **Time Series** data
- Utilize only the **Random Forests** for the disease classification task.
- Their train/test split being before **2013/after 2014** (try to predict the potential diseases in the futures for any patient given their health history)
- Comparison with a **classifier using Feature Selection** or **network embedding** based approaches
- **Values of lab tests** instead of frequency
- **LSTM instead of LDA**
- **Kernel PCA, Spectral Clustering** comparison
- **Cross-entropy function as the loss function**, is it the best choice?
- Diseases that could not be predicted from the EHR labels **were filtered**
- Diseases may have an **effect on one another**
- **Autoencoders?**
- **Optimization Scheme?**

References:

1. Miotto, R. et al. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci. Rep. 6, 26094; doi: 10.1038/srep26094 (2016)

2. Deep Learning A-Z: Hands-On Artificial Neural Networks

3. https://www.auntminnie.com/index.aspx?sec=ser&sub=def&pag=dis&ItemID=117351

4. CNBC: https://www.cnbc.com/2017/05/11/from-coding-to-cancer-how-ai-is-changing-medicine.html

5. The Conversation https://theconversation.com/can-machines-really-tell-us-if-were-sick-71798

6. Weiskopf, Nicole G., George Hripcsak, Sushmita Swaminathan, and Chunhua Weng. "Defining and measuring completeness of electronic health records for secondary use." Journal of biomedical informatics 46, no. 5 (2013): 830-836.

7. Lasko, Thomas A., Joshua C. Denny, and Mia A. Levy. "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data." PloS one 8, no. 6 (2013): e66341.

Thank you

Questions?

Faraz Notes

1. Classifier with Feature Selection comparison: Did not perform the comparison well
2. Is the data open source, can we replicate the paper
3. Get insights from the Twitter, News post and discuss
4. Why encoder, decoder
5. Some issues with collection of data, noisy
6. Put Limitations and future work within the slides

Limitations

- **Didn’t compare with T. Lasko, PLoS ONE, 2013 [7]** with similar methodology and dataset.

Future Improvements

- **Preprocessing using PCA and then Deep** Modelling
- I