Northwell Health®

Saving hearts using Artificial Intelligence

Khader Shameer, PhD
Program Director, Data Science and Machine Learning
Departments of Medical Informatics and Research Informatics
Northwell Health
E: skhader@northwell.edu | T: @kshameer

Icahn School of Medicine at Mount Sinai

Icahn School of Medicine at Mount Sinai
Icahn Institute and Department of Genetics and Genomic Sciences

DudleyLab
Translational and Biomedical Informatics

Who am I?

CCMB
Centre for Cellular & Molecular Biology
A constituent laboratory of CSIR

INDIAN INSTITUTE OF SCIENCE
Bangalore, India
भारतीय विज्ञान संस्थान
बेंगलूर, भारत

ncbs
national centre for biological sciences

MAYO CLINIC

Mount Sinai

DudleyLab
Translational and Biomedical Informatics

PHILIPS

Northwell Health®

Outline

- Need for AI in healthcare
- Differential diagnoses case-studies
  - Athlete's heart or Hypertrophic cardiomyopathy
  - Diastolic dysfunction
  - Constrictive pericarditis or Restrictive pericarditis
- Heart failure associated readmission prediction
- Predicting disease sequelae
- Discuss implications on cost, outcome and optimization in the setting of learning healthsystem

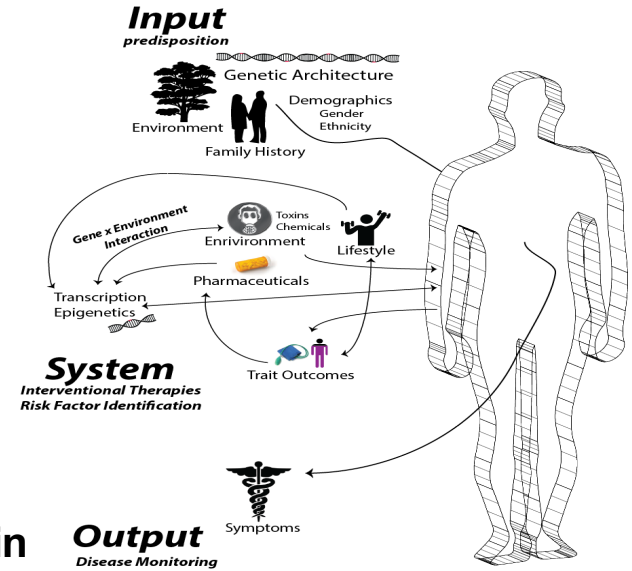
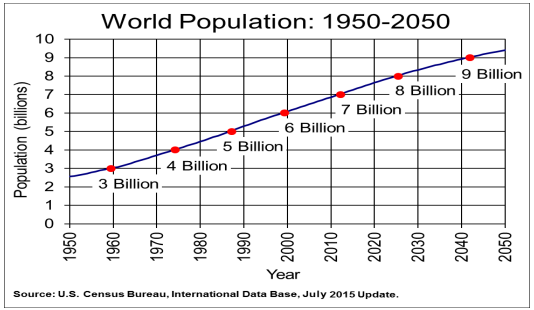


Figure courtesy: B. Glicksberg

We live in a world of growing population, new diseases & resurgence of old diseases

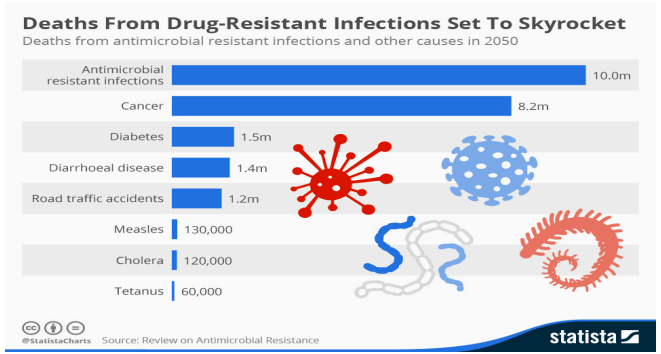


Disease burden varies

Infectious diseases are bigger killers in developing economies, while noncommunicable diseases are more prevalent in advanced economies.
(percent)

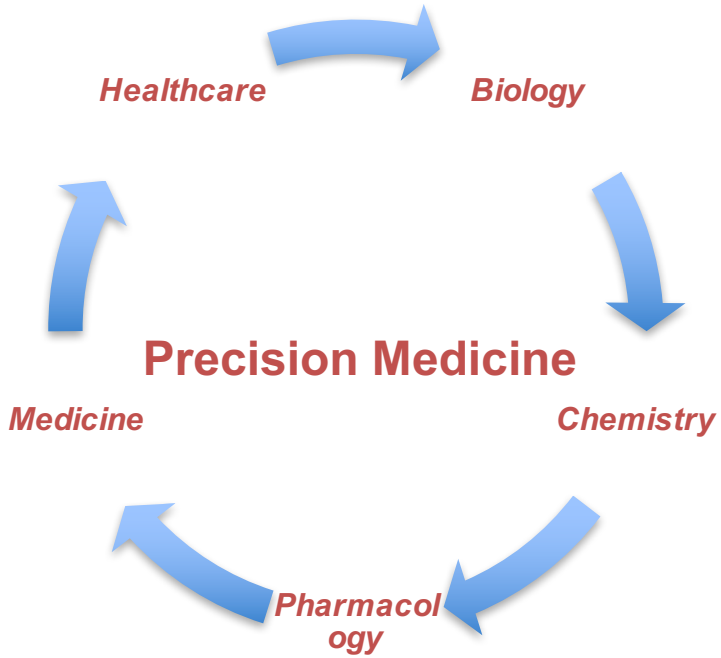
	Share of Disability-Adjusted Life Years			Share of Deaths		
	Global	Developing	Developed	Global	Developing	Developed
Noncommunicable Diseases						
Cardiovascular and circulatory diseases	11.9	10.2	21.3	29.6	25.1	43.4
Neoplasms	7.6	6.2	15.3	15.1	12.3	23.7
Mental and behavioral disorders	7.4	6.7	11.1	0.4	0.3	1
Musculoskeletal disorders	6.7	5.7	12.3	0.3	0.3	0.4
Diabetes, urogenital, blood and endocrine diseases	4.9	4.7	5.8	5.2	5.2	5.1
Chronic respiratory diseases	4.8	4.8	4.5	7.2	7.9	5
Neurological disorders	3	2.7	4.4	2.4	1.9	4.1
Cirrhosis of the liver	1.3	1.2	1.7	1.9	2	2
Digestive diseases	1.3	1.3	1.5	2.1	2.1	2.2
Other noncommunicable diseases	5.1	5.1	5.2	1.2	1.4	0.6
Infectious Diseases						
Diarrhea, lower respiratory infections, and other common infectious diseases	11.4	13	2.5	10	12	4
HIV/AIDS and tuberculosis	5.3	6	1.7	5	6.3	1.1
Neglected tropical diseases and malaria	4.4	5.2	0.1	2.5	3.3	0.03
Other	24.9	27.2	12.6	17.1	19.9	7.37

Source: Institute for Health Metrics and Evaluation, Global Burden of Disease (2010).
Note: Disability-adjusted life years measure the effective life years lost to sickness, disability, or death. The "Other" category includes deaths from such things as injuries, nutritional disorders, and neonatal and birth complications.



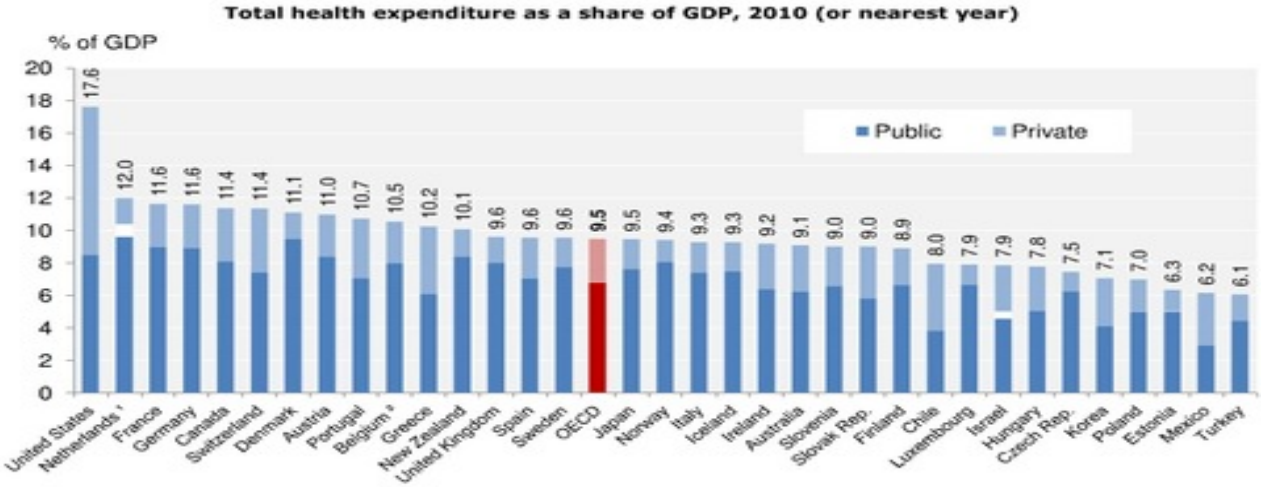
We have automated cars, but what about seamless healthcare and wellness?

From Biology to Therapy to Healthcare



Healthcare is a significant part of our economy

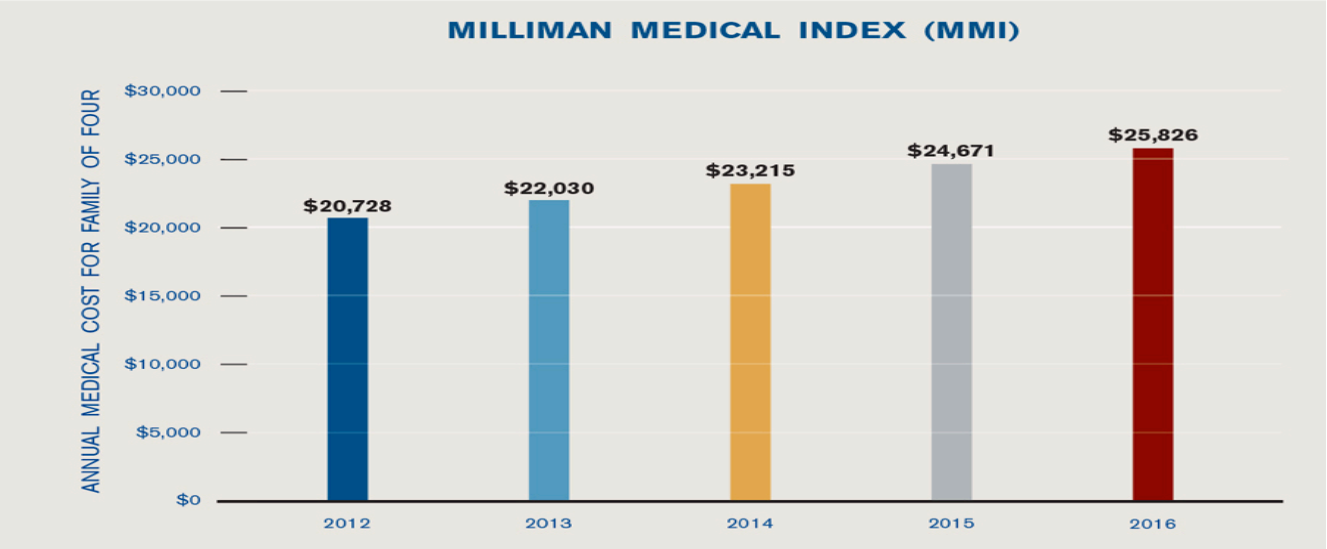
At 17.6% of GDP in 2010, US health spending is one and a half as much as any other country, and nearly twice the OECD average



1. In the Netherlands, it is not possible to clearly distinguish the public and private share related to investments.
2. Total expenditure excluding investments.
Information on data for Israel: http://dx.doi.org/10.1787/888932315602.

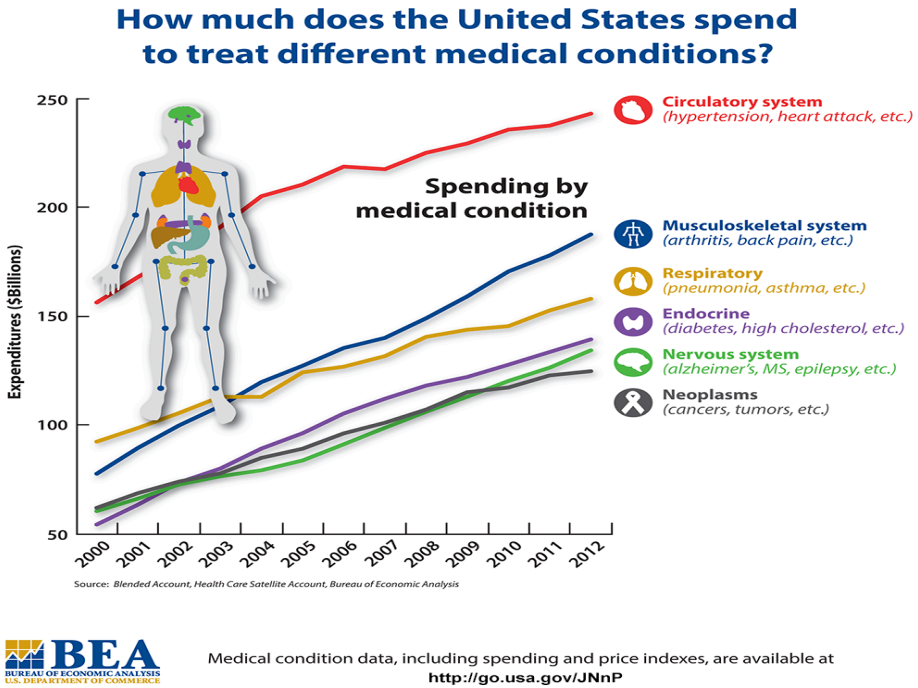
Source: OECD Health Data 2012.

Average cost of care is increasing



Can {Artificial Intelligence, Machine Learning, Cognitive Learning, Deep Learning, Big Data, Predictive Modeling or Data science} help?

Improved technology is not lowering the cost of care



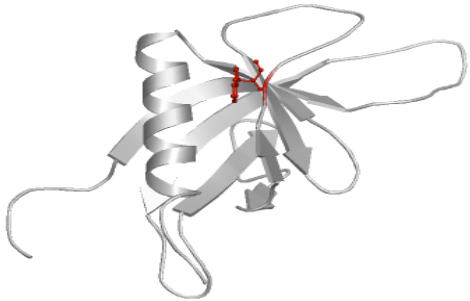
Precision Medicine is *now*.

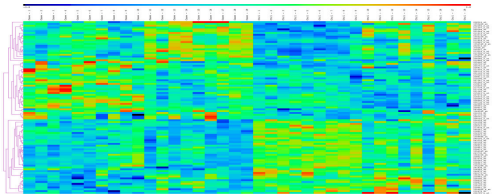
"Tonight, I'm launching a new Precision Medicine Initiative to bring us closer to curing diseases like cancer and diabetes — and to give all of us access to the personalized information we need to keep ourselves and our families healthier."

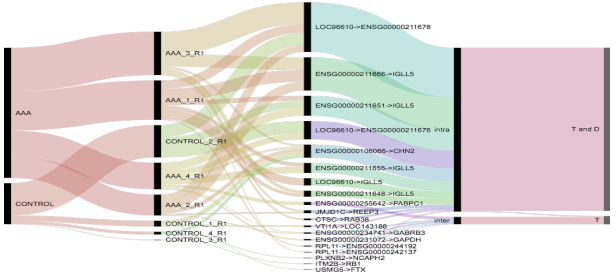
— President Barack Obama,
State of the Union Address, January 20, 2015

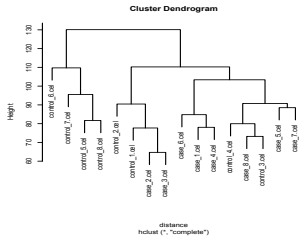
www.nejm.org/doi/full/10.1056/NEJMp1500523

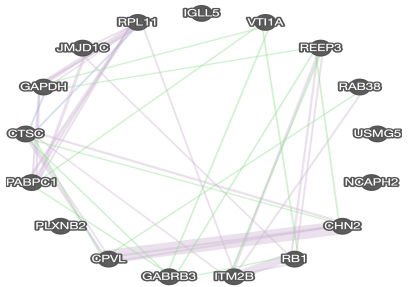
Data rush in biology is NOT new.

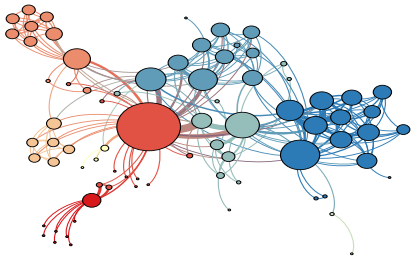


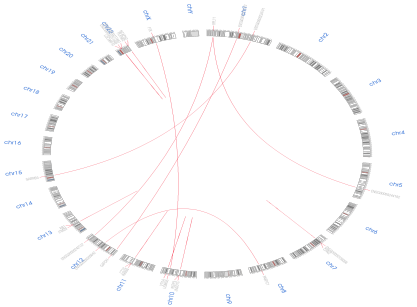












Data rush in medicine is relatively new.

Traditional Medicine

- Traditional data types
- Centralized
- GBs or TBs in size
- Structured
- Stable data model
- Low-dimensional
- Statistical approaches
- Cohort size (~10K)
- Hypothesis-driven

Data-driven Medicine

- Evolving data types
- Decentralized
- Petabytes, exabytes…
- Semi or unstructured
- Evolving, flat data model
- High-dimensional
- Machine learning
- Large cohort size (>10K)
- Data-driven

Figures courtesy: Nature, Science and Popular Science

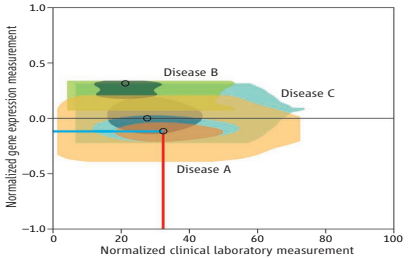
Challenge: Managing the collective data rush in biology and medicine

The Ultimate Model Organism

Atul J. Butte

A deeper understanding of disease requires a database of human traits and disease states that is integrated with molecular information.

Science **320**, 325 (2008);

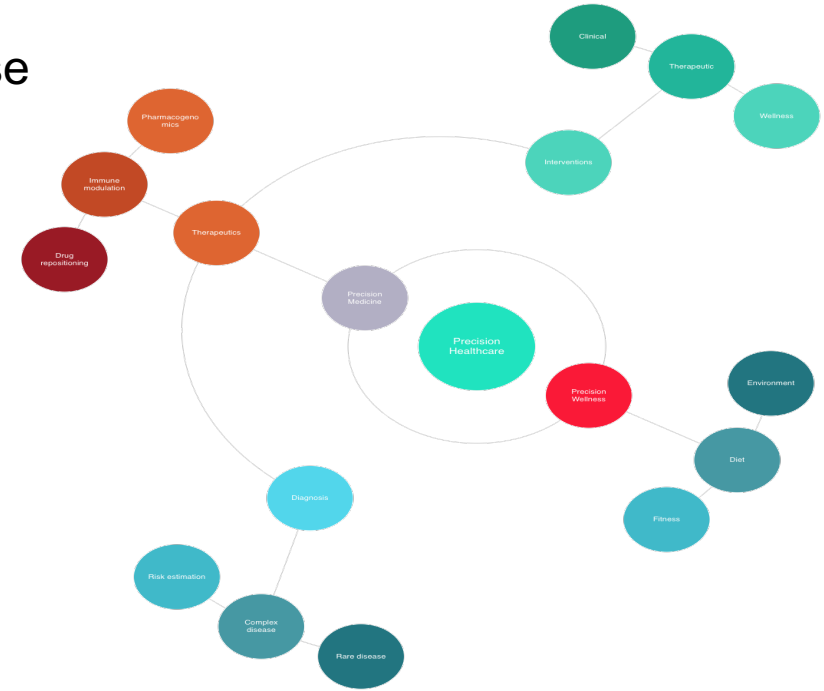


Information intersection. Three diseases may be separately considered by a quantitative clinical laboratory test measurement and a gene expression measurement (from a public repository of gene expression). Associations can be discovered between molecular and clinical measurements, even when these measurements are not made using the same samples or patients. For example, Disease A, when studied across all patients and time points, shows a high average level of a clinical test (red line), and a low level of a gene (blue line). The distribution of gene and clinical measurements are shown by sampling from both independent data sets (colored regions). The trend across the three diseases shown is that as a dis-

Science. 2008 Apr 18;320(5874):325-7. doi: 10.1126/science.1158343.

Application of (small, big, deep, wide and broad) learning in Precision Medicine

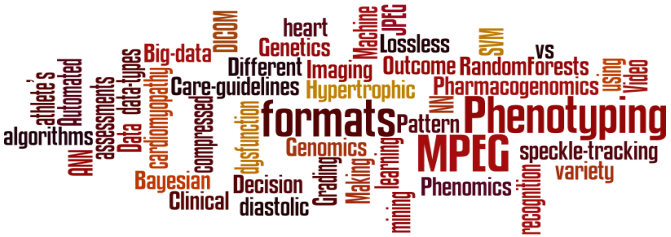
- **Diagnosis**
  - Rare, complex or common disease
  - Disease subtyping
  - Precision phenotyping
- **Interventions**
  - Therapeutic/curative
  - Surgical/Clinical
  - Wellness
- **Outcome assessment**
  - Readmission rate
  - HAI/HAC
  - Mortality rate



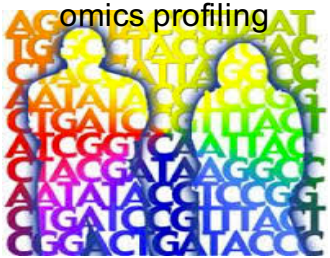
K. Shameer, K Johnson, P. Sengupta and Joel T Dudley
Invited Review, JACC: Cardiovascular Imaging

Converging biomedical & healthcare big data for precision medicine is the need of the hour

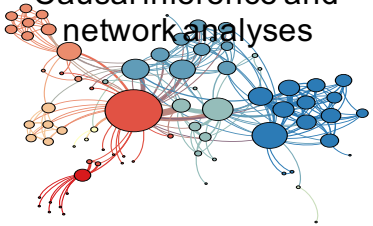
Multi-dimensional clinical data



Genomics & other omics profiling



Data Science, ML, AI, Causal inference and network analyses



Feature modeling, feature recognition, data mining and machine learning

Predictive and prescriptive models

▼

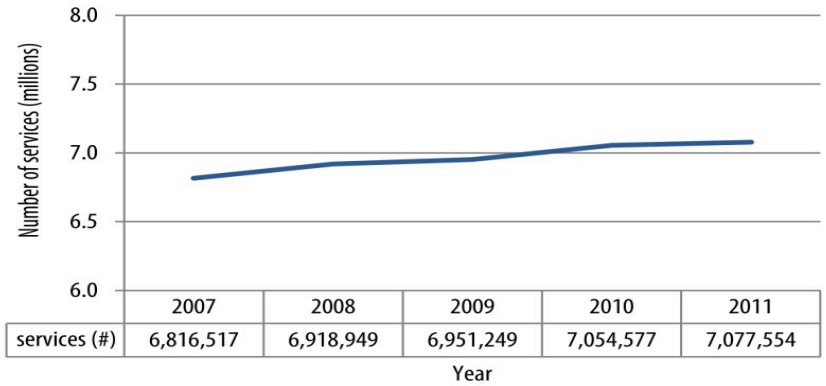
Applications in diagnoses, treatment and rehabilitation:

- Precision phenotyping
- Personalized therapeutic recommendations
- Patient stratification
- Care-pathway modeling

K. Shameer, K Johnson, P. Sengupta and Joel T Dudley
Invited State of The Art Review, JACC: BTS

Big Data in Cardiac Imaging: Echocardiography

(2007-2011 Medicare fee-for-service cohort)



	2007	2008	2009	2010	2011
services (#)	6,816,517	6,918,949	6,951,249	7,054,577	7,077,554

- 7.07 million patients*100MB data/patient = **707 TB data** (compressed)
- 7.07 million patients*2GB data/patient = **14.14 PB** (uncompressed data)
- ASE recommends TEE for general evaluation of cardiac structure and function in the setting of **200 indications ~ 200 algorithms**

Data from AHRQ Report - Data Points Publication Series; http://www.ncbi.nlm.nih.gov/books/NBK63603/ http://www.ncbi.nlm.nih.gov/pubmed/21338862;http://www.ncbi.nlm.nih.gov/pubmed/15746725

Data storage analysis: K. Shameer, P. Sengupta and JT Dudley

Diastolic Dysfunction: challenges

Diastolic heart failure occurs when signs and symptoms of heart failure are present but left ventricular systolic function is preserved (i.e., ejection fraction greater than 45 percent)

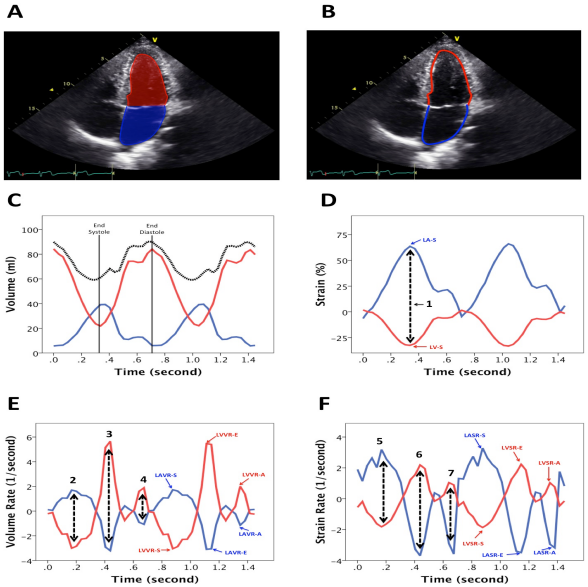
Incidence **increases with age**

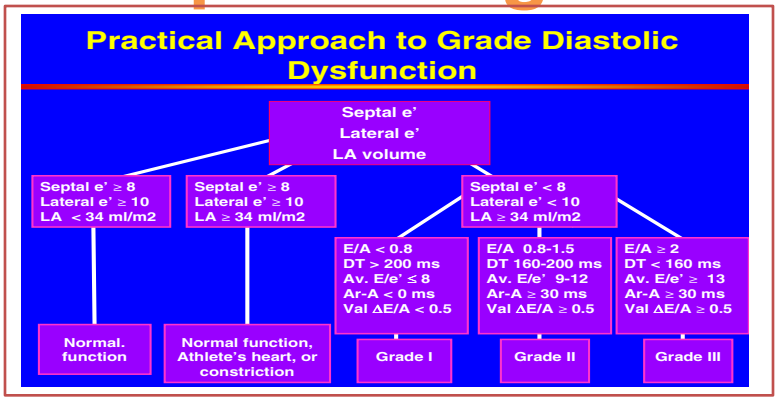
50% of older patients with heart failure may have isolated **diastolic dysfunction**

- Characterized by invasive and noninvasive methods
- Echocardiography is the most practical routine clinical approach
- Comprehensive assessment of diastolic function begins with evaluation of
  - left ventricular (LV) volumes
  - LV ejection fraction (LVEF)
  - LV wall thicknesses
  - Left atrial (LA) volumes
  - Pulmonary artery (PA) pressures
- Reproducible with careful adherence to correct acquisition and analysis techniques

- **Three million Americans** have congestive heart failure (CHF), and **500,000 new cases** are diagnosed each year
- Diastolic heart failure accounts for approximately **40%-60%** of patients with CHF
- The condition is the **most common discharge diagnosis** for patients older than 65 years and is the most expensive disease for Medicare
- With **early diagnosis and proper management** the prognosis of diastolic dysfunction is more favorable than that of systolic dysfunction

Current approaches are guidelines driven, manual and need expert help to read and interpret images.

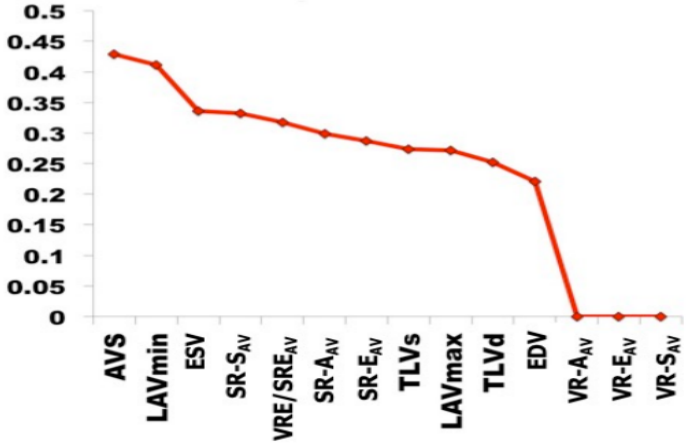


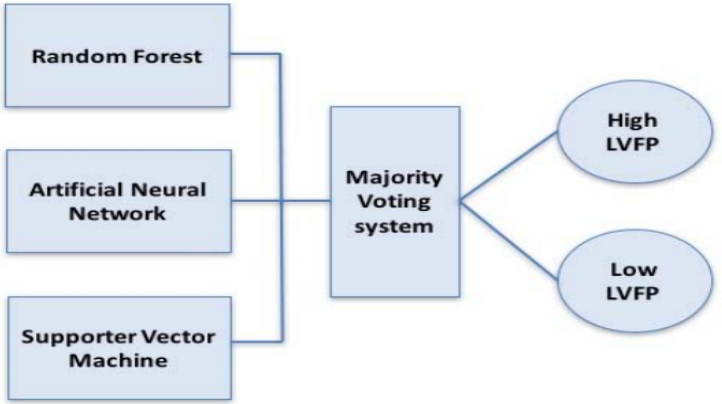


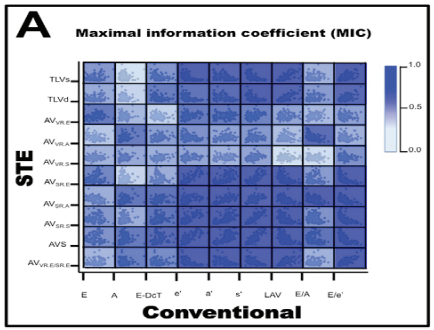

```
## With all variables as predictors
##using all features
rfMod<-randomForest(trainData[,!("DECEASED_INDI
                    trainData$DECEASED_INDICATO
predTest<-predict(rfMod,testData[,!("DECEASED_I
table(predTest,testData$DECEASED_INDICATOR)
impTable<-round(importance(rfMod),2)
pred<-prediction(predTest[,2],testData$DECEASED
perf<-performance(pred,measure="tpr",x.measure=

## using 15 variables as predictors (about 70%)
rfMod<-randomForest(trainData[,names(testData[,
                                with=F],
                    trainData$DECEASED_INDICATO
predTest<-predict(rfMod,testData[,!("DECEASED_I
table(predTest,testData$DECEASED_INDICATOR)
impTable<-round(importance(rfMod),2)
pred<-prediction(predTest[,2],testData$DECEASED
perf1<-performance(pred,measure="tpr",x.measure

## using 10 features only (around 50%)
rfMod<-randomForest(trainData[,names(testData[,
                                with=F],
                    trainData$DECEASED_INDICATO
predTest<-predict(rfMod,testData[,!("DECEASED_I
table(predTest,testData$DECEASED_INDICATOR)
impTable<-round(importance(rfMod),2)
pred<-prediction(predTest[,2],testData$DECEASED
perf2<-performance(pred,measure="tpr",x.measure

par(mfrow=c(1,1))
## plot ROC for three cases
plot(perf, col=rainbow(10))
plot(perf1, col="green",add=T)
plot(perf2, col="blue",add=T)
legend("bottomright",c("All Features","70% Feat
       lwd=c(1.5,1.5,1.5),col=c("red","green","
```

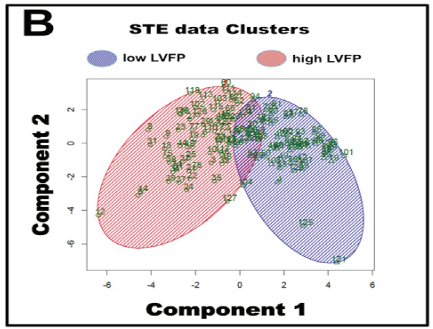
Can an algorithm could learn and perform this classification task?

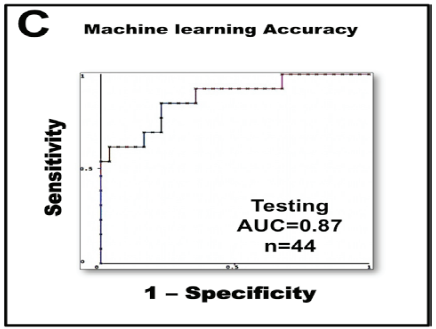
Feature selection and Ensemble machine learning strategy











Cognitive learning approach

Ventricular Structure and Function

Cognitive Machine-Learning Algorithm for Cardiac Imaging

A Pilot Study for Differentiating Constrictive Pericarditis From Restrictive Cardiomyopathy

Partho P. Sengupta, MD*; Yen-Min Huang, PhD*; Manish Bansal, MD; Ali Ashrafi, PhD; Matt Fisher, PhD; Khader Shameer, PhD; Walt Gall, PhD; Joel T. Dudley, PhD

Background—Associating a patient's profile with the memories of prototypical patients built through previous repeat clinical experience is a key process in clinical judgment. We hypothesized that a similar process using a cognitive computing tool would be well suited for learning and recalling multidimensional attributes of speckle tracking echocardiography data sets derived from patients with known constrictive pericarditis and restrictive cardiomyopathy.

Methods and Results—Clinical and echocardiographic data of 50 patients with constrictive pericarditis and 44 with restrictive cardiomyopathy were used for developing an associative memory classifier–based machine-learning algorithm. The speckle tracking echocardiography data were normalized in reference to 47 controls with no structural heart disease, and the diagnostic area under the receiver operating characteristic curve of the associative memory classifier was evaluated for differentiating constrictive pericarditis from restrictive cardiomyopathy. Using only speckle tracking echocardiography variables, associative memory classifier achieved a diagnostic area under the curve of 89.2%, which improved to 96.2% with addition of 4 echocardiographic variables. In comparison, the area under the curve of early diastolic mitral annular velocity and left ventricular longitudinal strain were 82.1% and 63.7%, respectively. Furthermore, the associative memory classifier demonstrated greater accuracy and shorter learning curves than other machine-learning approaches, with accuracy asymptotically approaching 90% after a training fraction of 0.3 and remaining flat at higher training fractions.

Conclusions—This study demonstrates feasibility of a cognitive machine-learning approach for learning and recalling patterns observed during echocardiographic evaluations. Incorporation of machine-learning algorithms in cardiac imaging may aid standardized assessments and support the quality of interpretations, particularly for novice readers with limited experience. (***Circ Cardiovasc Imaging*. 2016;9:e004330. DOI: 10.1161/CIRCIMAGING.115.004330.)**

Key Words: big data ■ cardiovascular imaging ■ cognitive tools ■ machine learning ■ phenomics ■ precision medicine ■ speckle tracking echocardiography

Echocardiography is the most widely used cardiac imaging modality and is indispensable in the management of most patients with a suspected or known cardiac illness. However, echocardiography is highly operator-dependent and requires considerable expertise.[1–4] This is especially relevant in the contemporary clinical environments that demand higher precision in diagnosis while the required high-level diagnostic expertise remains in short supply. Automated techniques using novel machine-learning approaches may potentially help transforming the interpretation process and render clinical imaging much smarter, efficient, and cost effective.

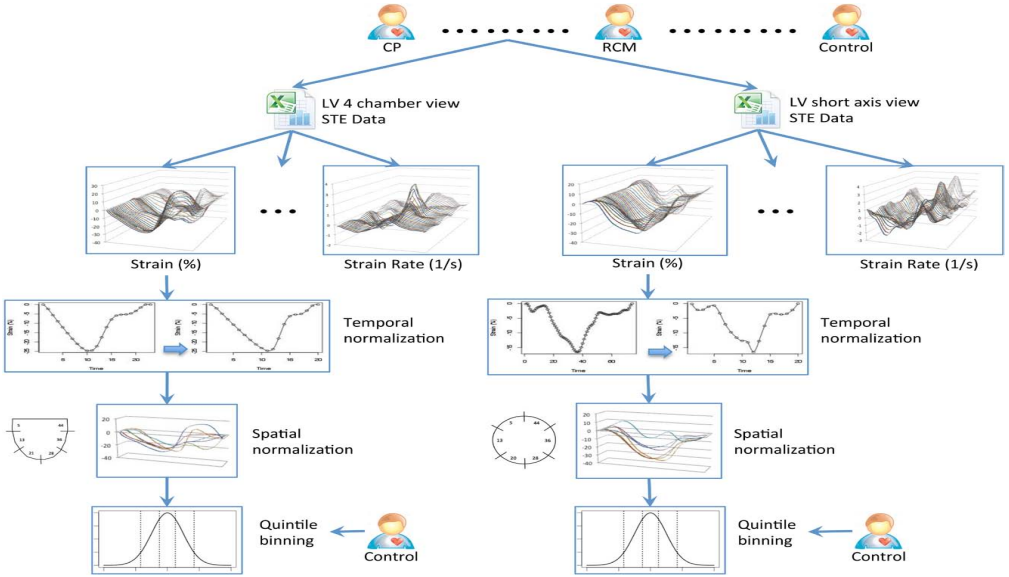
See Editorial by Nagueh
See Clinical Perspective

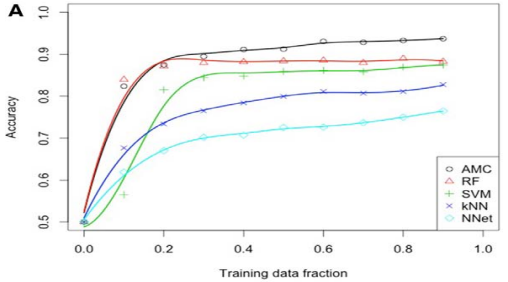
The new generation of so-called Big Data machine-learning techniques has potential applications for nonparametric analysis of cardiac imaging data during routine clinical assessments. However, using traditional statistical model–based or logic/rule-based tools[5–7] would differ from the working of the human brain, which draws relevant inferences by recognizing patterns stored in memories built through previous and repeated experiences in assessing cardiac structure and function.[8] Associative memory can be thought of as conceptually similar to clinical judgment in the medical setting, where the brain of a trained doctor intuitively attempts to connect the dots in search for the best fit or associations for understanding a pattern of medical abnormality. Associative memory-based brain-like machine-learning algorithms have been recently applied successfully in the operational risk intelligence areas of national security and defense; however, their application in clinical medicine or cardiac imaging has not been hitherto reported.[9]

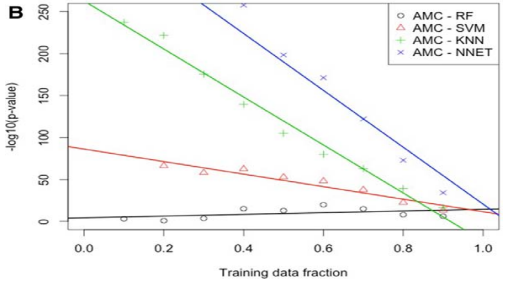
American Heart Association®
Learn and Live

Circulation
Cardiovascular Imaging
JOURNAL OF THE AMERICAN HEART ASSOCIATION

Associative memory classifier: A cognitive learning approach







Differentiating Athlete's heart vs. hypertrophic cardiomyopathy

JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY
© 2016 BY THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION
PUBLISHED BY ELSEVIER

Machine-Learning Algorithms to Automate Morphological and Functional Assessments in 2D Echocardiography

Sukrit Narula, BS,* Khader Shameer, PhD,† Alaa Mabrouk Salem Omar, MD, PhD,*‖
Joel T. Dudley, PhD,† Partho P. Sengupta, MD, DM*

ABSTRACT

BACKGROUND Machine-learning models may aid cardiac phenotypic recognition by using features of cardiac tissue deformation.

OBJECTIVES This study investigated the diagnostic value of a machine-learning framework that incorporates speckle-tracking echocardiographic data for automated discrimination of hypertrophic cardiomyopathy (HCM) from physiological hypertrophy seen in athletes (ATH).

METHODS Expert-annotated speckle-tracking echocardiographic datasets obtained from 77 ATH and 62 HCM patients were used for developing an automated system. An ensemble machine-learning model with 3 different machine-learning algorithms (support vector machines, random forests, and artificial neural networks) was developed and a majority voting method was used for conclusive predictions with further *K*-fold cross-validation.

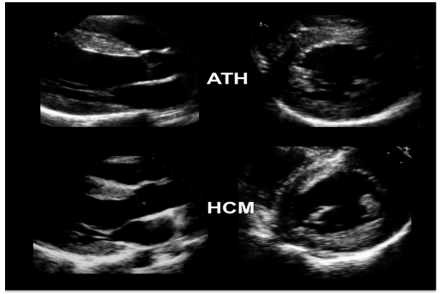
RESULTS Feature selection using an information gain (IG) algorithm revealed that volume was the best predictor for differentiating between HCM ands. ATH (IG = 0.24) followed by mid-left ventricular segmental (IG = 0.134) and average longitudinal strain (IG = 0.131). The ensemble machine-learning model showed increased sensitivity and specificity compared with early-to-late diastolic transmitral velocity ratio ($p < 0.01$), average early diastolic tissue velocity (e′) ($p < 0.01$), and strain ($p = 0.04$). Because ATH were younger, adjusted analysis was undertaken in younger HCM patients and compared with ATH with left ventricular wall thickness >13 mm. In this subgroup analysis, the automated model continued to show equal sensitivity, but increased specificity relative to early-to-late diastolic transmitral velocity ratio, e′, and strain.

CONCLUSIONS Our results suggested that machine-learning algorithms can assist in the discrimination of physiological versus pathological patterns of hypertrophic remodeling. This effort represents a step toward the development of a real-time, machine-learning–based system for automated interpretation of echocardiographic images, which may help novice readers with limited experience. (J Am Coll Cardiol 2016;68:2287–95) © 2016 by the American College of Cardiology Foundation.

From the *Zena and Michael A. Weiner Cardiovascular Institute, Icahn School of Medicine at Mount Sinai, New York, New York; †Institute of Next Generation Healthcare, Department of Genetics and Genomic Sciences, Mount Sinai Health System, New York, New York; and the ‖Department of Internal Medicine, Medical Division, National Research Center, Cairo, Egypt. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript. Drs. Shameer and Dudley have received the grants from National Institutes of Health: National Institute of Diabetes and Digestive and Kidney Diseases (RotDK090042), National Cancer Institute (U54CA189201), Illuminating the Druggable Genome, Knowledge Management Center sponsored by National Institutes of Health Common Fund; National Cancer Institute (U54-CA189201-03); National Center for Advancing Translational Sciences (ULTR000067); and Clinical and Translational Science Award. Dr. Dudley has received consulting fees or honoraria from Janssen Pharmaceuticals, GlaxoSmithKline, AstraZeneca, and Hoffman-La Roche; is a scientific advisor to LAM Therapeutics; and holds equity in NuMedii Inc., Ayasdi Inc., and Ontomics, Inc. Dr. Sengupta is a consultant for TeleHealthRobotics, Heart Test Labs, and Hitachi-Aloka Ltd. All other authors have reported that they have no relationships relevant to the contents of this paper to disclose. Mr. Narula and Dr. Shameer contributed equally to this work. Presented at the American Society of Echocardiography Annual Scientific Sessions 2015 for Arthur Weyman Young Investigator Award. Mr. Narula was declared as the winner of the competition. P.K. Shah, MD, served as Guest Editor-in-Chief for this paper.

Manuscript received May 12, 2016, revised manuscript received July 27, 2016, accepted August 17, 2016.

Listen to this manuscript's audio summary by JACC Editor-in-Chief Dr. Valentin Fuster.

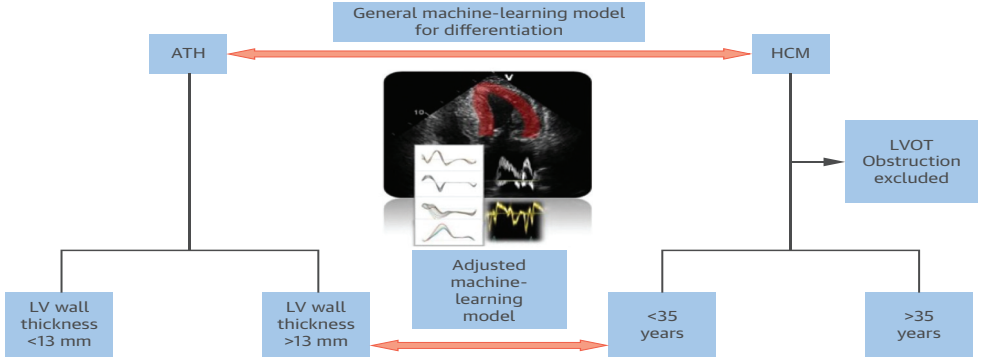


This example shows a morphological resemblance of an athlete's heart with hypertrophic cardiomyopathy (HCM) from the imaging core lab database. ATH = athletes.

	True Positive	True Negative	False Positive	False Negative	Sensitivity	Specificity	p Value*
General model							
ML model	48	37	7	8	87	82	–
E/A	46	30	11	12	80	71	<0.001
e′†	48	32	11	9	84	74	<0.001
LS	38	35	17	10	69	77	0.04
Adjusted model							
ML model	74	17	3	6	96	77	–
E/A	61	17	16	6	79	77	<0.001
e′†	66	18	11	5	86	82	<0.001
LS	53	17	5	25	68	77	0.002

Values are %. *Comparison performed between contingency table of ML model with traditional echocardiographic variables. †e′ represents mean early diastolic tissue velocities measured by speckle-tracking echocardiography.

ML = machine learning; other abbreviations as in Table 1.

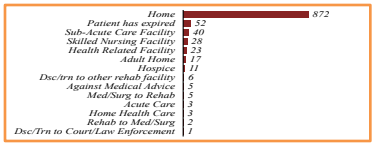


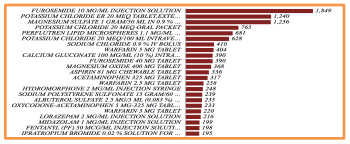
The machine-learning algorithm for differentiating athlete's heart (ATH) from hypertrophic cardiomyopathy (HCM) was primarily developed using 2-dimensional speckle-tracking echocardiography-based parameters. After the initial model was built, a secondary age and phenotype matched analysis was performed.

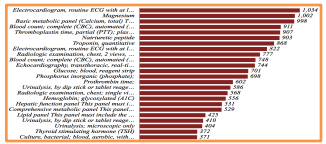
JACC JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY

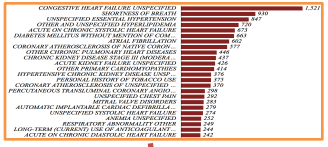
AMERICAN COLLEGE of CARDIOLOGY

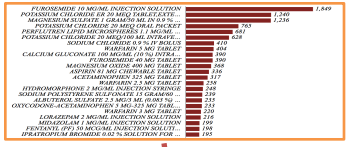
Predicting hospital readmissions











f1(M1) *f2(M2)* *f3(M3)* *f4(M4)* *f5(M5)*

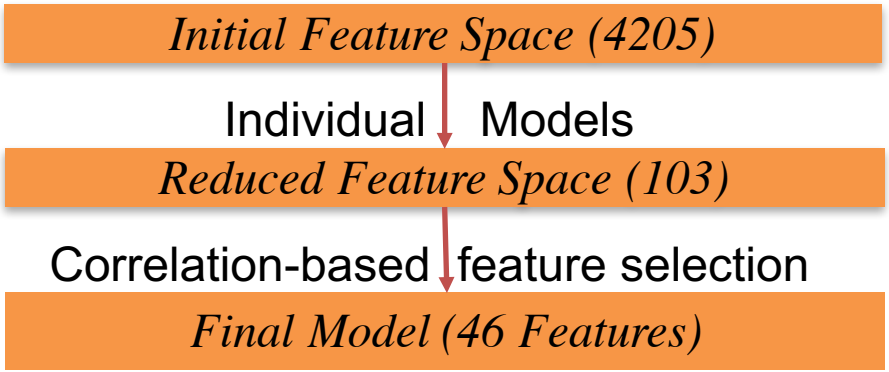
Top features(Mn)

Composite Model

Accuracy and AUC

Clinical validation

Predicting hospital readmissions



- Naïve Bayes Classifier
- Correlation-based feature selection
- 5-fold cross validation
  - Final model
    - Accuracy = 83.19%
    - AUC = 0.781

Data source	Type	Encoding	Accuracy	AUC	Features
Diagnosis	ICD-9 Diagnosis	Binary	70.3297%	0.605	34/1788
Procedures	ICD-9 Procedure Codes	Binary	77.907%	0.505	4/273
Procedure	CPT Procedure Codes	Binary	72.9858%	0.553	8/596
Medications	Medication Name & dosage	Binary	81.9048%	0.615	26/1030
Labs / Vitals	Lab Measurements	Continuous	73.9336%	0.535	29/953

Editorial recommendations from top cardiologists

JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY
© 2016 BY THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION
PUBLISHED BY ELSEVIER

VOL. 68, NO. 21, 2016
ISSN 0735-1097/$36.00
http://dx.doi.org/10.1016/j.jacc.2016

EDITORIAL COMMENT

Machine Learning for Echocardiographic Imaging

Embarking on Another Incredible Journey*

A. Jamil Tajik, MD

Computers are incredibly fast, accurate, and stupid. Human beings are incredibly slow, inaccurate, and brilliant. Together they are powerful beyond imagination.
—Albert Einstein (1)

As I read the exciting work of Narula et al. (2) published in this issue of the *Journal*, I could not help but be transported back to the early days of my career at Mayo Clinic in Rochester, Minnesota. It is always enjoyable to take a journey down memory lane! Ralph E. Smith, MD, was the director of the Mayo electrocardiographic laboratory. He had a passion for the computerization of the electrocar-

of Marquette Electronics (Milwaukee, led to further refinements, miniaturi commercialization—and the rest is histor

Cardiologists of the bygone era alway ipers in their pocket so they could make measurements of P-, Q-, R-, S-, and T-wa and R-R cycle variability. Now these sar ments are accurately and reproducibly p computers. No more calipers! Furthermo gists have witnessed marked improven computer interpretation of abnormal EC testament to machine-learning algori result, cardiologists of today have reduced the time they spend reading EC

Editorial

Unleashing the Potential of Machine-Based Learning for the Diagnosis of Cardiac Diseases

Sherif F. Nagueh, MD

Imaging plays a critical role in the diagnosis and management of patients with cardiovascular disease. Given the need for cost-effective diagnostic and treatment algorithms and the diversity of available imaging modalities, it is important to identify and use the most cost-effective diagnostic strategies. This is particularly true for patients undergoing evaluation for possible pericardial constriction where an equivocal diagnosis after the initial imaging modality, echocardiography, can lead to additional testing, including cardiac catheterization. Thus, the warm welcome for the study by Sengupta et al[1] in this issue of *Circulation: Cardiovascular Imaging*. The authors used clinical and echocardiographic data of 50 patients with constrictive pericarditis and 44 patients with restrictive cardiomyopathy to develop a machine learning–based algorithm: an associative memory classifier (AMC). Findings in 47 controls were also gathered, and then AMC was applied to identify patients with constriction from those with restriction.

See Article by Sengupta et al

Results of the Present Study

Studying the performance of standard criteria was an integral part of the analysis of this study. Notably, standard criteria were applicable to all initial patients. However, 9 patients were excluded because of tracking problems and thus were not included in the analysis using machine learning algorithms. This highlights the lower feasibility of an approach based on strain and strain rate measurements.

The AMC leaning algorithm was based on several variables, including displacement, velocity, strain, and strain rate from the apical 4-chamber view and the parasternal

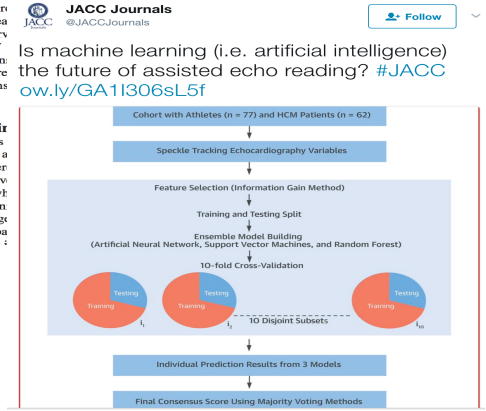
in the training data set. The authors also compared analysis by AMC to other general machine learning algorithms. They observed that AMC performed better than other machine learning approaches and needed fewer training trials.

As expected, there were significant differences in routinely analyzed imaging parameters between patients with pericardial constriction and those with restrictive cardiomyopathy (the majority of which had cardiac amyloidosis). Thus, most patients with pericardial constriction had normal left ventricular wall thickness. In comparison, patients with restriction had higher *E/e′* ratio, lower *e′* velocity, and more impaired left ventricular longitudinal strain

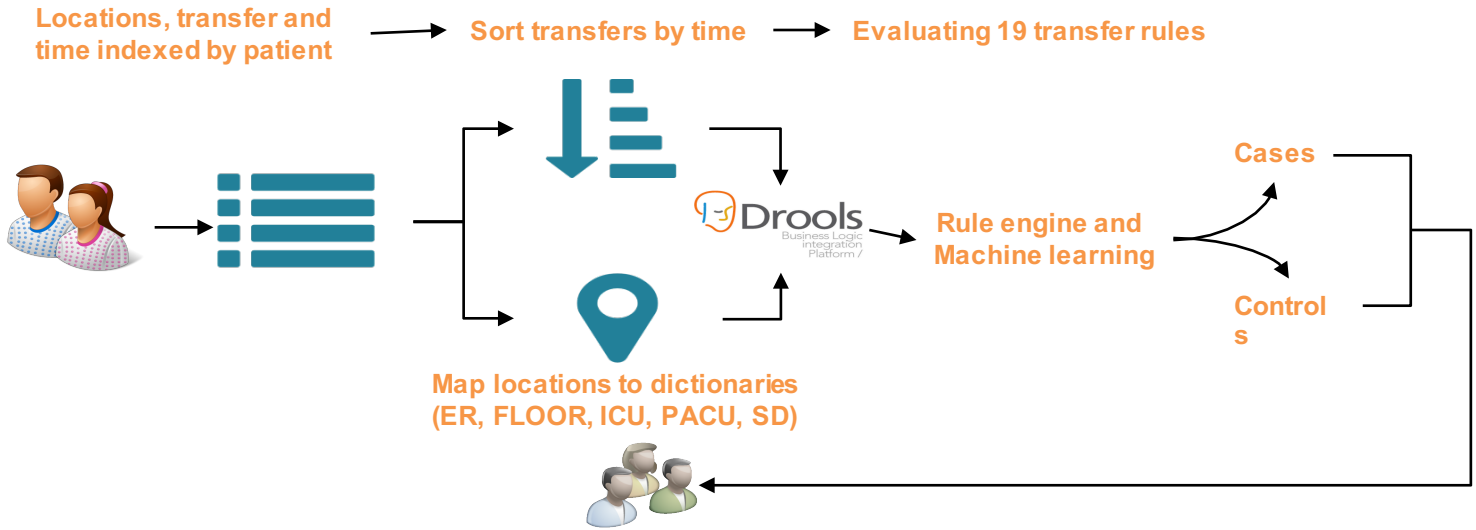
wide overlap was present between the 2 gr
tricular longitudinal strain based on the mea
in Table 2, and thus, the area under the curv
for the latter variable. The frequency of
motion (septal bounce) in patients with con
presented. Likewise, the performance of a re
simple algorithm for the diagnosis of cons
included in the analysis.[2]

Machine-Based Learnir

The concept of machine-based learning has
for years, and the authors have successfully a
nose 2 challenging diseases. Although inter
serves as an initial step toward a well-dev
because it has several limitations, some of wh
acknowledged by the authors. First, the train
obtained from a select group of patients image
The accuracy of the algorithm depends in pa

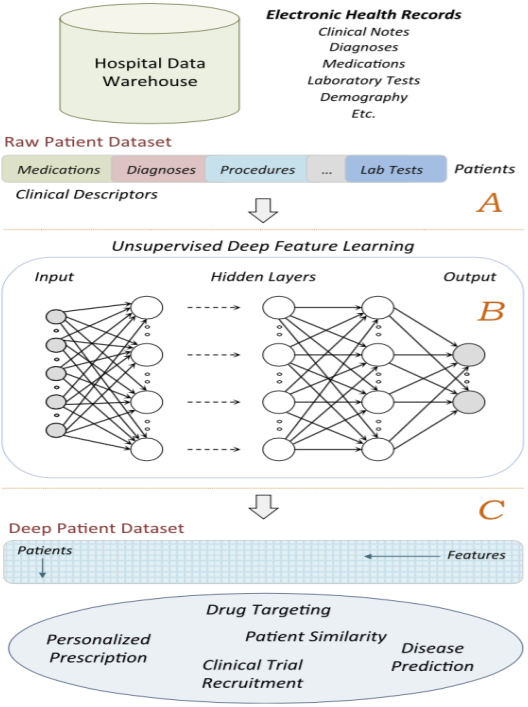


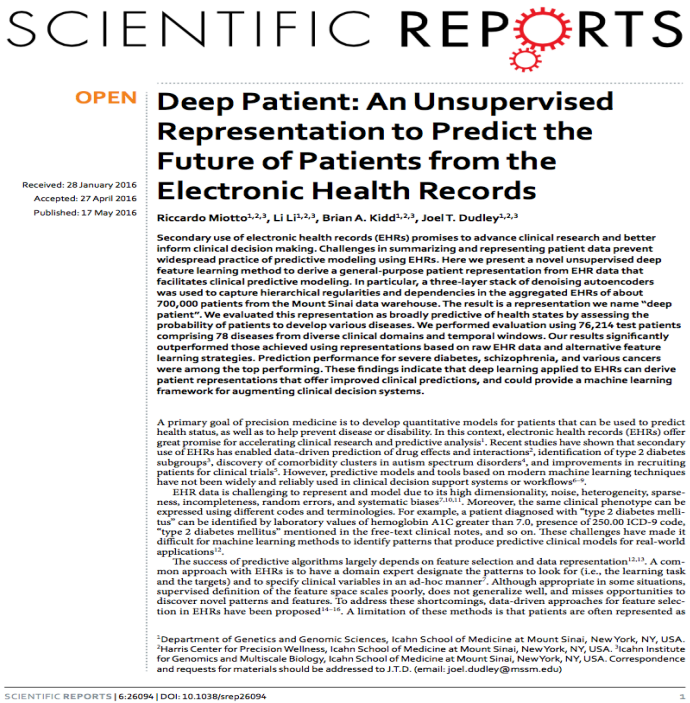
Unified acuity score: Predicting factors and events of unexpected escalation of care



Unexpected ICU transfer phenotyping algorithm using Rule Engines and Machine Learning ; Unpublished data

Going deep: Deep learning for automated patient abstraction ("Deep Patient")





SCIENTIFIC REPORTS

OPEN

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

Received: 28 January 2016
Accepted: 27 April 2016
Published: 17 May 2016

Riccardo Miotto[1,2,3], Li Li[2,2,3], Brian A. Kidd[2,2,3], Joel T. Dudley[1,2,3]

Secondary use of electronic health records (EHRs) promises to advance clinical research and better inform clinical decision making. Challenges in summarizing and representing patient data prevent widespread practice of predictive modeling using EHRs. Here we present a novel unsupervised deep feature learning method to derive a general-purpose patient representation from EHR data that facilitates clinical predictive modeling. In particular, a three-layer stack of denoising autoencoders was used to capture hierarchical regularities and dependencies in the aggregated EHRs of about 700,000 patients from the Mount Sinai data warehouse. The result is a representation we name "deep patient". We evaluated this representation as broadly predictive of health states by assessing the probability of patients to develop various diseases. We performed evaluation using 76,214 test patients comprising 78 diseases from diverse clinical domains and temporal windows. Our results significantly outperformed those achieved using representations based on raw EHR data and alternative feature learning strategies. Prediction performance for severe diabetes, schizophrenia, and various cancers were among the top performing. These findings indicate that deep learning applied to EHRs can derive patient representations that offer improved clinical predictions, and could provide a machine learning framework for augmenting clinical decision systems.

A primary goal of precision medicine is to develop quantitative models for patients that can be used to predict health status, as well as to help prevent disease or disability. In this context, electronic health records (EHRs) offer great promise for accelerating clinical research and predictive analysis[1]. Recent studies have shown that secondary use of EHRs has enabled data-driven prediction of drug effects and interactions[2], identification of type 2 diabetes subgroups[3], discovery of comorbidity clusters in autism spectrum disorders[4], and improvements in recruiting patients for clinical trials[5]. However, predictive models and tools based on modern machine learning techniques have not been widely and reliably used in clinical decision support systems or workflows[6–9].

EHR data is challenging to represent and model due to its high dimensionality, noise, heterogeneity, sparseness, incompleteness, random errors, and systematic biases[7,10,11]. Moreover, the same clinical phenotype can be expressed using different codes and terminologies. For example, a patient diagnosed with "type 2 diabetes mellitus" can be identified by 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. These challenges have made it difficult for machine learning methods to identify patterns that produce predictive clinical models for real-world applications[12].

The success of predictive algorithms largely depends on feature selection and data representation[12,13]. A common approach with EHRs is to have a domain expert designate the patterns to look for (i.e., the learning task and the targets) and to specify clinical variables in an ad-hoc manner[7]. Although appropriate in some situations, supervised definition of the feature space scales poorly, does not generalize well, and misses opportunities to discover novel patterns and features. To address these shortcomings, data-driven approaches for feature selection in EHRs have been proposed[14–16]. A limitation of these methods is that patients are often represented as

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2]Harris Center for Precision Wellness, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. Correspondence and requests for materials should be addressed to J.T.D. (email: joel.dudley@mssm.edu)

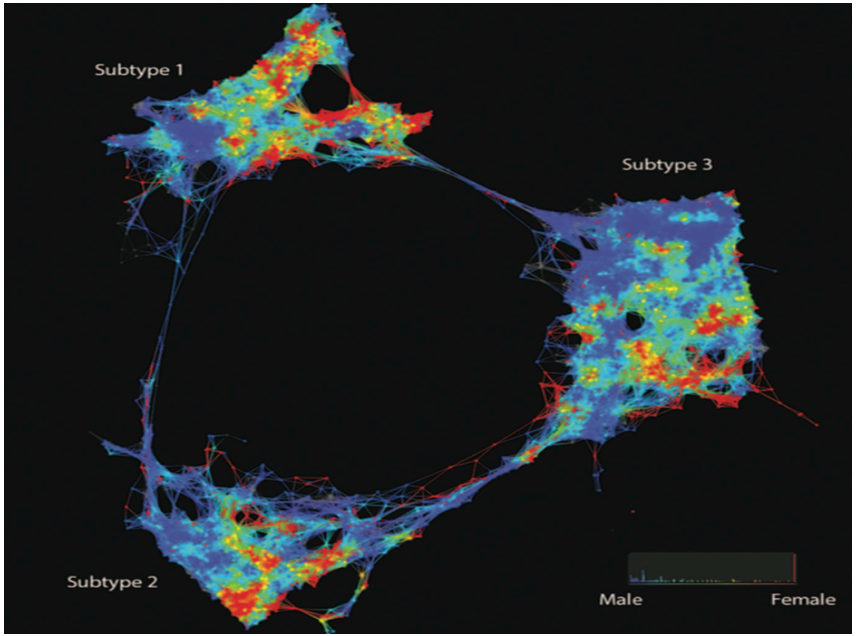
Miotto R, Li L, Kidd BA, Dudley JT., Sci Rep. 2016 May 17;6:26094. doi: 10.1038/srep26094.

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

Deep Learning to Predict Patient Future Diseases from the Electronic Health Records

Riccardo Miotto, Li Li, Joel T Dudley, European Conference on Information Retrieval, 2016

Subtyping Diabetes Cohort: Application of TDA for Population Health Data



RESEARCH ARTICLE

PRECISION MEDICINE

Identification of type 2 diabetes subgroups through topological analysis of patient similarity

Li Li,[1] Wei-Yi Cheng,[1] Benjamin S. Glicksberg,[1] Omri Gottesman,[2] Ronald Tamler,[3] Rong Chen,[1] Erwin P. Bottinger,[2] Joel T. Dudley[1,4]*

Type 2 diabetes (T2D) is a heterogeneous complex disease affecting more than 29 million Americans alone with a rising prevalence trending toward steady increases in the coming decades. Thus, there is a pressing clinical need to improve early prevention and clinical management of T2D and its complications. Clinicians have understood that patients who carry the T2D diagnosis have a variety of phenotypes and susceptibilities to diabetes-related complications. We used a precision medicine approach to characterize the complexity of T2D patient populations based on high-dimensional electronic medical records (EMRs) and genotype data from 11,210 individuals. We successfully identified three distinct subgroups of T2D from topology-based patient-patient networks. Subtype 1 was characterized by T2D complications diabetic nephropathy and diabetic retinopathy; subtype 2 was enriched for cancer malignancy and cardiovascular diseases; and subtype 3 was associated most strongly with cardiovascular diseases, neurological diseases, allergies, and HIV infections. We performed a genetic association analysis of the emergent T2D subtypes to identify subtype-specific genetic markers and identified 1279, 1227, and 1338 single-nucleotide polymorphisms (SNPs) that mapped to 425, 322, and 437 unique genes specific to subtypes 1, 2, and 3, respectively. By assessing the human disease–SNP association for each subtype, the enriched phenotypes and biological functions at the gene level for each subtype matched with the disease comorbidities and clinical differences that we identified through EMRs. Our approach demonstrates the utility of applying the precision medicine paradigm in T2D and the promise of extending the approach to the study of other complex, multifactorial diseases.

INTRODUCTION

Type 2 diabetes (T2D) is a complex, multifactorial disease that has emerged as an increasing prevalent worldwide health concern associated with high economic and physiological burdens. An estimated 29.1 million Americans (9.3% of the population) were estimated to have some form of diabetes in 2012—up 13% from 2010—with T2D representing up to 95% of all diagnosed cases (*1*, *2*). Risk factors for T2D include obesity, family history of diabetes, physical inactivity, ethnicity, and advanced age (*1*, *2*). Diabetes and its complications now rank among the leading causes of death in the United States (*2*). In fact, diabetes is the leading cause of nontraumatic foot amputation, adult blindness, and need for kidney dialysis, and multiplies risk for myocardial infarction, peripheral artery disease, and cerebrovascular disease (*3*–*6*). The total estimated direct medical cost attributable to diabetes in the United States in 2012 was $176 billion, with an estimated $76 billion attributable to hospital inpatient care alone. There is a great need to improve understanding of T2D and its complex factors to facilitate prevention, early detection, and improvements in clinical management.

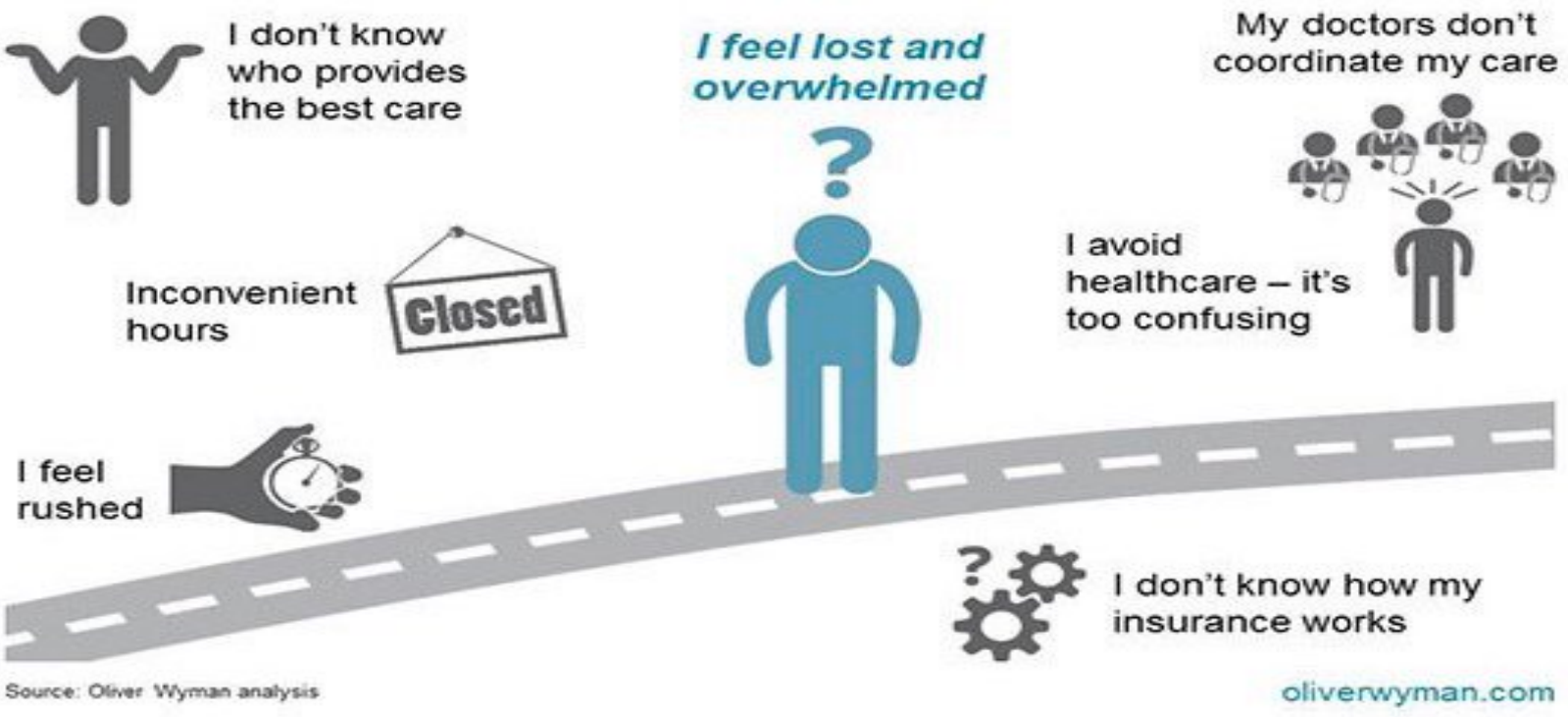
A more precise characterization of T2D patient populations can enhance our understanding of T2D pathophysiology (*7*, *8*). Current clinical definitions classify diabetes into three major subtypes: type 1 diabetes (T1D), T2D, and maturity-onset diabetes of the young. Other subtypes based on phenotype bridge the gap between T1D and T2D, for example, latent autoimmune diabetes in adults (LADA) (*7*) and ketosis-prone T2D. The current categories indicate that the traditional definition of diabetes, especially T2D, might comprise additional subtypes with distinct clinical characteristics. A recent analysis of the longitudinal Whitehall II cohort study demonstrated improved assessment of cardiovascular risks when subgrouping T2D patients according to glucose concentration criteria (*9*). Genetic association studies reveal that the genetic architecture of T2D is profoundly complex (*10*–*12*). Identified T2D-associated risk variants exhibit allelic heterogeneity and directional differentiation among populations (*13*, *14*). The apparent clinical and genetic complexity and heterogeneity of T2D patient populations suggest that there are opportunities to refine the current, predominantly symptom-based, definition of T2D into additional subtypes (*7*).

Because etiological and pathophysiological differences exist among T2D patients, we hypothesize that a data-driven analysis of a clinical population could identify new T2D subtypes and factors. Here, we develop a data-driven, topology-based approach to (i) map the complexity of patient populations using clinical data from electronic medical records (EMRs) and (ii) identify new, emergent T2D patient subgroups with subtype-specific clinical and genetic characteristics. We apply this approach to a data set comprising matched EMRs and genotype data from more than 11,000 individuals. Topological analysis of these data revealed three distinct T2D subtypes that exhibited distinct patterns of clinical characteristics and disease comorbidities. Further, we identified genetic markers associated with each T2D subtype and performed gene- and pathway-level analysis of subtype genetic associations. Biological and phenotypic features enriched in the genetic analysis corroborated clinical disparities observed among subgroups. Our findings suggest that data-driven, topological analysis of patient cohorts has utility in precision medicine efforts to refine our understanding of T2D toward improving patient care.

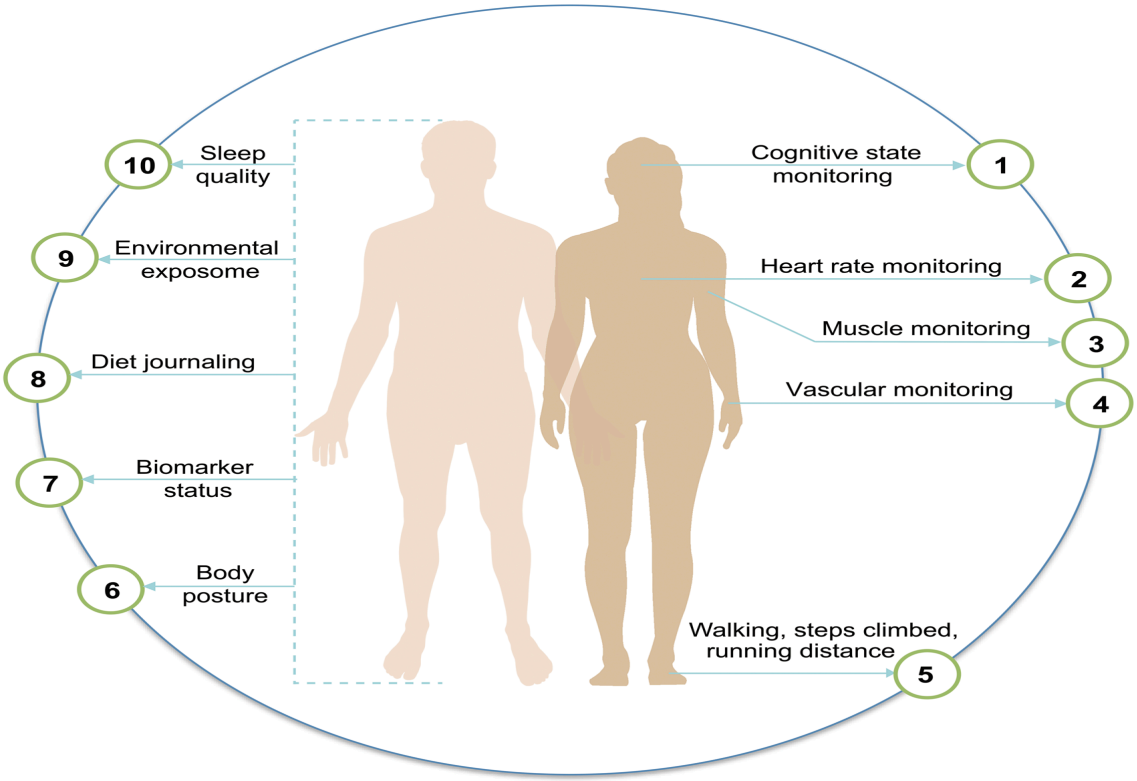
[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, 700 Lexington Ave., New York, NY 10065, USA. [2]Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, NY 10029, USA. [3]Division of Endocrinology, Diabetes, and Bone Diseases, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [4]Department of Health Policy and Research, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.
*Corresponding author. E-mail: joel.dudley@mssm.edu

Sci Transl Med. 2015 Oct 28;7(311):311ra174. doi: 10.1126/scitranslmed.aaa9364.
Identification of type 2 diabetes subgroups through topological analysis of patient similarity.
Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, Bottinger EP, Dudley JT.

Healthcare is about All of US.



Limited data captured in healthcare setting



1: psychiatric and neurological disease, cerebrovascular disease, stress responses/autonomic reactivity, chronic pain;

2: cardiac arrest, myocardial infarction, coronary heart disease, anxiety, aerobic fitness levels;

3: chronic back pain, movement disorders (Parkinsonism), tremors, rehabilitation recovery, agility testing, dystonia, myalgia, chronic fatigue syndrome;

4: hypertension, orthostatic hypotension, chronic kidney disease, peripheral arterial disease, vasculitis (e.g. Lupus, Raynaud's disease);

5: movement disorders, rehabilitation, epilepsy, myalgia;

6: chronic and acute lung diseases, obstructive sleep apnea, sleep disorders, narcolepsy, synucleopathies;

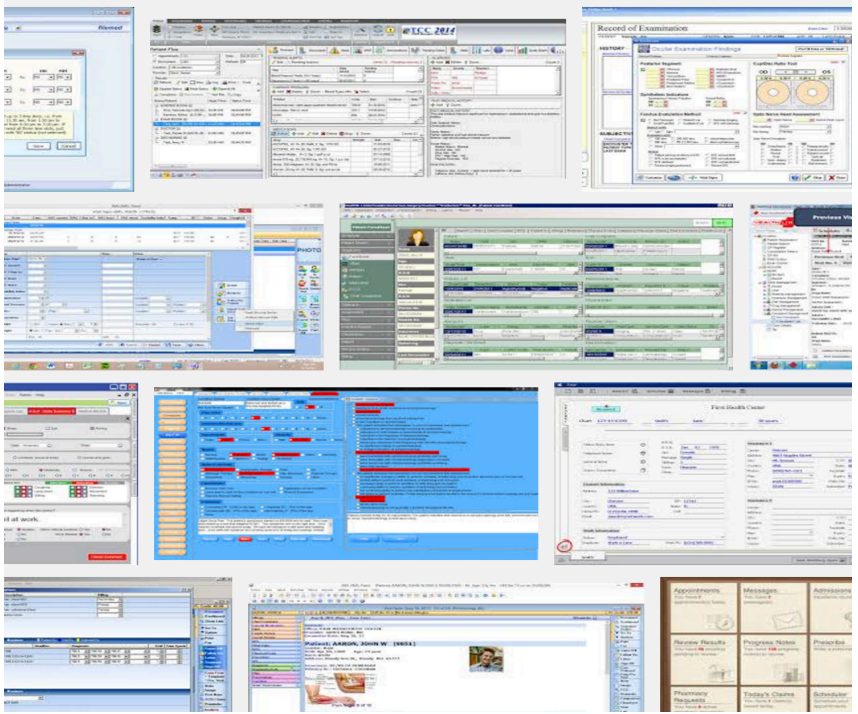
7: insulin level (Type 1 or Type 2 diabetes);

8: diabetes, cardiovascular disease, inflammatory bowel disease, irritable bowel syndrome, gluten sensitivity, eating disorders;

Sensors are cheaper, real-time and provides far more data than a typical hospital visit associated with a clinical trial

EHR: A rate-limiting factor for machine learning in healthcare?

- EMR is all about human “biology.”
- Point and click interface
- Customizations are costly
- Summarizing multiple patients for reporting is an “optional” feature
- Integrations are mostly BPAs, forms, and more forms



Figures courtesy: https://www.google.com/search?q=emr+screenshots

The need for a learning health system: an integration challenge

Harlan Krumholz
@hmkyale

On rounds last week almost every patient had vital information somewhere out of our team's view. Typical. Critical.

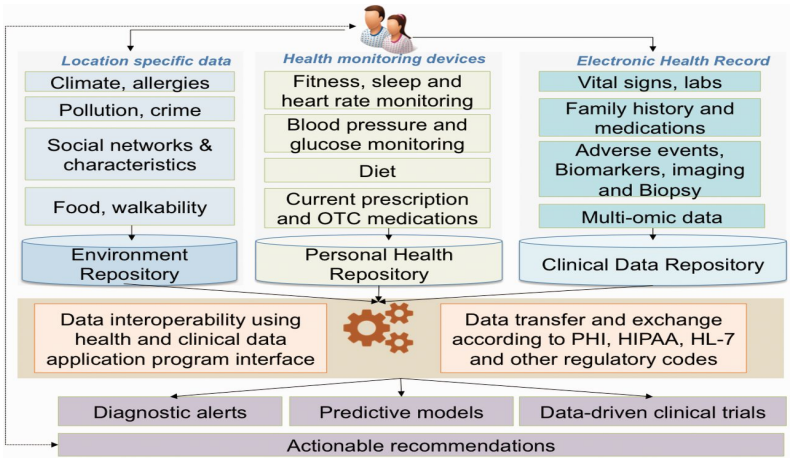
> **Qualcomm Life** @QualcommLife
> 65% of doctors say they don't have all the health care info needed for patients. How can we streamline #healthdata? ln.is/modernmedicine...

10:47 AM · 10 Jul 16

3 RETWEETS **3** LIKES

Individualome—a biomedical, healthcare data model for the precision medicine era

Briefings in Bioinformatics



Briefings in Bioinformatics Advance Access published February 14, 2016

OXFORD

Briefings in Bioinformatics, 2016, 1–20

doi: 10.1093/bib/bbv118
Paper

Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams

Khader Shameer*, Marcus A. Badgeley*, Riccardo Miotto, Benjamin S. Glicksberg, Joseph W. Morgan and Joel T. Dudley

Corresponding author. Joel T. Dudley, Department of Genetics and Genomics, Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, Mount Sinai Health System, 770 Lexington Avenue, 15th Floor, New York, NY 10065. Tel.: 212-731-7073; Fax: 212-731-7099. E-mail: joel.dudley@msmn.edu
*These authors contributed equally to this work.

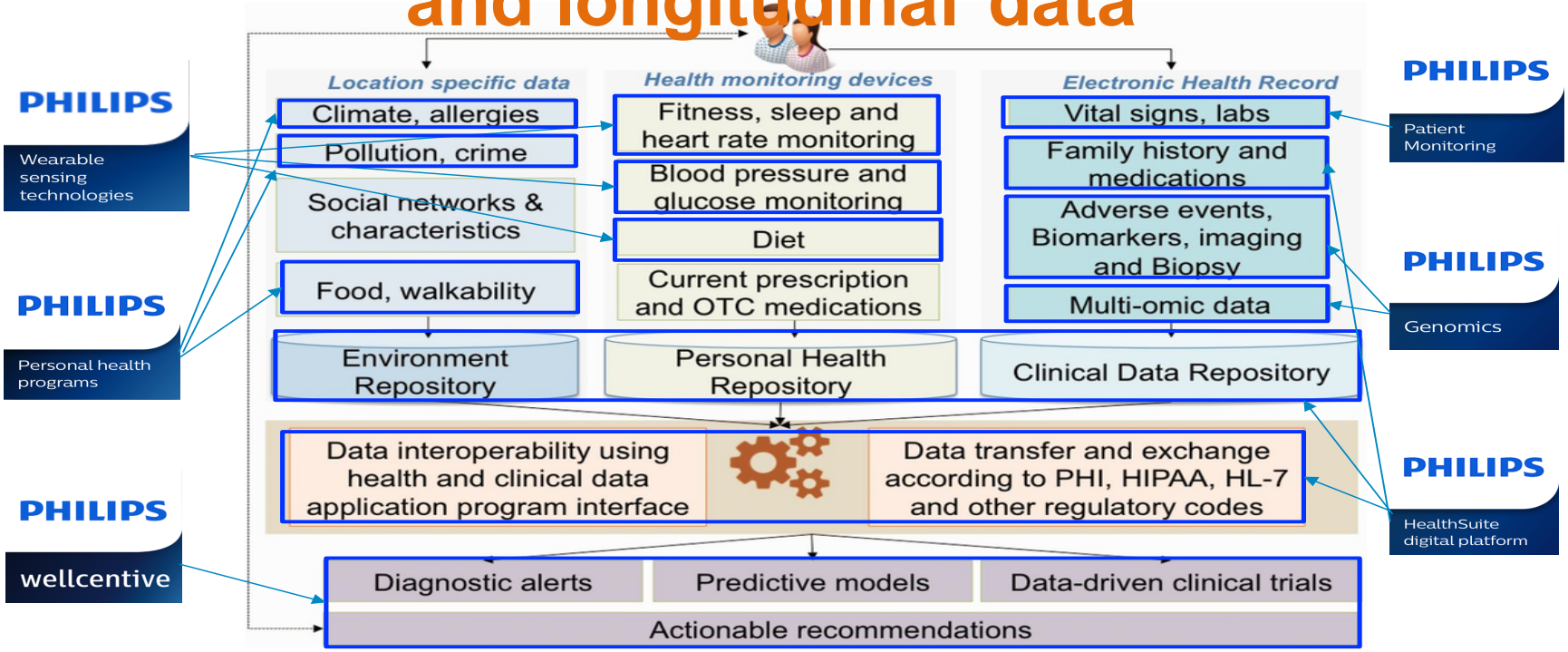
Abstract

Monitoring and modeling biomedical, health care and wellness data from individuals and converging data on a population scale have tremendous potential to improve understanding of the transition to the healthy state of human physiology to disease setting. Wellness monitoring devices and companion software applications capable of generating alerts and sharing data with health care providers or social networks are now available. The accessibility and clinical utility of such

Khader Shameer is a senior biomedical and health care informatics scientist in Dr Joel Dudley's laboratory and the Harris Center for Precision Wellness of the Mount Sinai Health System. Dr Khader is working at the interface of biomedical informatics, drug repositioning and genomic medicine and pioneering efforts to translate advances in translational bioinformatics to population health management and precision medicine. He obtained his MS in Bioinformatics from Mahatma Gandhi University, Kerala and PhD in Computational Biology from National Centre for Biological Sciences, Bangalore. He completed his postdoctoral training at the Mayo Clinic in the Departments of Biomedical Informatics and Statistics, Health Science Research and Cardiovascular Diseases and worked on developing applications for integrating genomic information into the electronic medical records, development of genome analyses pipelines and phenomics.
Marcus A. Badgeley is an MD-PhD student at Icahn School of Medicine at Mount Sinai, NYC. Badgeley is interested in high-dimensional biomedical data analytics and pursuing his PhD at the interface of biosensors, wellness science and real-time data analytics in Dr Dudley laboratory and Harris Center for Precision Wellness.
Riccardo Miotto is a senior data scientist at Dr. Dudley laboratory and Harris Center for Precision Wellness at Icahn School of Medicine at Mount Sinai, NYC. Dr Miotto's research interests encompass the design of algorithms for information retrieval, machine learning and data mining applied to clinical data for personalized medicine. Previously, Dr Miotto worked on clinical trial search engines through free-text eligibility criteria processing, and music information retrieval, in the particular semantic discovery and recommendation, automatic tagging and cover identification. Dr Miotto obtained his PhD in Information Engineering from the University of Padova, Italy.
Benjamin S. Glicksberg is a PhD student at Icahn School of Medicine at Mount Sinai, NYC. His research interests are in connecting genetic, epidemiological, environmental and clinical data into network analytics. He uses these to discover novel associations between diseases in the context of personalized health predictions. His work also focuses on uncovering the effects of genetic variation in clinical traits and disease risk. He works in the laboratories of Dr Joel Dudley and Dr Rong Chen as well as at the Harris Center for Precision Wellness.
Joseph W. Morgan practices Anesthesiology for Watauga Anesthesia Associates, within the Appalachian Regional Healthcare System located in Boone, NC and he is also a senior consultant for the Harris Center for Precision Wellness of the Mount Sinai Health System. Dr Morgan focuses on the use of wearable health technology, the physiological interpretation of wearable data, and is pioneering efforts to leverage these technologies to inform better clinical decision-making and patient-centered care. He obtained his MD from Upstate Medical University, Syracuse, NY and completed his residency in Anesthesiology at the Cleveland Clinic, Cleveland, OH.
Joel T. Dudley is an assistant professor of the department of genetics and genomics at Icahn Institute of Multiscale Biology, health policy at Icahn School of Medicine at Mount Sinai, NYC. He currently directs the biomedical informatics programs of Department of Genetics and Genomics, The Icahn Institute for Genomics and Multiscale Biology at Mount Sinai, Clinical and Translational Science Activities, Mount Sinai Health System and Harris Center for Precision Wellness. Dr Dudley obtained his PhD in Biomedical Informatics from Stanford University, CA. Dr Dudley's research focuses on the development and use of translational informatics approaches to address critical challenges in systems medicine. Dudley laboratory is involved in various projects for identification and development of novel therapeutic and diagnostic approaches for human disease through integration and analysis of health care, biomedical and wellness data, and also perform research to develop and evaluate methods to incorporate multiscale data including genomic sequencing data and wellness science into clinical practice.
Submitted: 14 September 2015; **Received (in revised form):** 27 November 2015

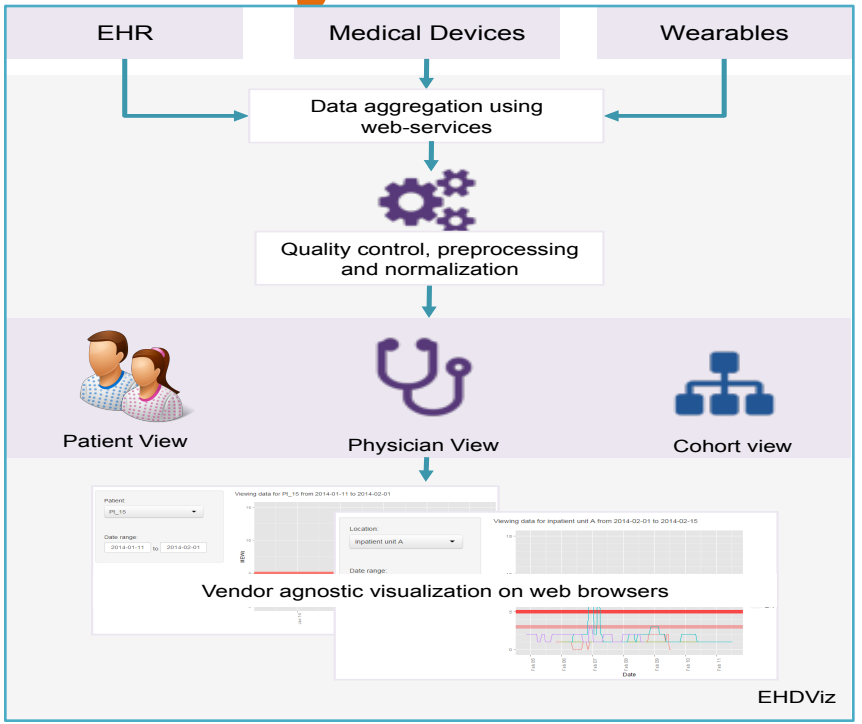
Shameer K, Badgeley MA, Miotto R, Glicksberg BS, Morgan JW, Dudley JT. Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. Brief Bioinform. 2016 Feb 14. pii: bbv118. PubMed PMID: 26876889.

Case Study: Predictors and patterns of T1D and comorbidities from sensors, wearables and longitudinal data



Design principles for a new healthcare visual analytics tool.

- Healthcare first
- Modern UI
- Responsive UI
- Real-time visualization
- Interactive UI
- Extensible
- Cross-platform rendering
- Enables visual analytics
- Scriptable
- …



EHDViz: Extensible toolkit for visual analytics of electronic healthcare data

Open Access Research

BMJ Open

EHDViz: clinical dashboard development using open-source technologies

Marcus A Badgeley,[1,2] Khader Shameer,[1,2] Benjamin S Glicksberg,[1,2] Max S Tomlinson,[1,2] Matthew A Levin,[2,3] Patrick J McCormick,[3] Andrew Kasarskis,[2] David L Reich,[3] Joel T Dudley[1,2,4]

To cite: Badgeley MA, Shameer K, Glicksberg BS, *et al*. EHDViz: clinical dashboard development using open-source technologies. *BMJ Open* 2016;**6**:e010579. doi:10.1136/bmjopen-2015-010579

► Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2015-010579).

MAB and KS equally contributed.

Received 25 November 2015
Revised 29 February 2016
Accepted 3 March 2016

For numbered affiliations see end of article.

Correspondence to
Dr Joel T Dudley;
joel.dudley@mssm.edu

ABSTRACT

Objective: To design, develop and prototype clinical dashboards to integrate high-frequency health and wellness data streams using interactive and real-time data visualisation and analytics modalities.

Materials and methods: We developed a clinical dashboard development framework called electronic healthcare data visualization (EHDViz) toolkit for generating web-based, real-time clinical dashboards for visualising heterogeneous biomedical, healthcare and wellness data. The EHDViz is an extensible toolkit that uses *R* packages for data management, normalisation and producing high-quality visualisations over the web using *R/Shiny* web server architecture. We have developed use cases to illustrate utility of EHDViz in different scenarios of clinical and wellness setting as a visualisation aid for improving healthcare delivery.

Results: Using EHDViz, we prototyped clinical dashboards to demonstrate the contextual versatility of EHDViz toolkit. An outpatient cohort was used to visualise population health management tasks (n=14 221), and an inpatient cohort was used to visualise real-time acuity risk in a clinical unit (n=445), and a quantified-self example using wellness data from a fitness activity monitor worn by a single individual was also discussed (n-of-1). The back-end system retrieves relevant data from data source, populates the main panel of the application and integrates user-defined data features in real-time and renders output using modern web browsers. The visualisation elements can be customised using health features, disease names, procedure names or medical codes to populate the visualisations. The source code of EHDViz and various prototypes developed using EHDViz are available in the public domain at http://ehdviz.dudleylab.org.

Conclusions: Collaborative data visualisations, wellness trend predictions, risk estimation, proactive acuity status monitoring and knowledge of complex disease indicators are essential components of implementing data-driven precision medicine. As an open-source visualisation framework capable of integrating health assessment, EHDViz aims to be a valuable toolkit for rapid design, development and implementation of scalable clinical data visualisation dashboards.

Strengths and limitations of this study

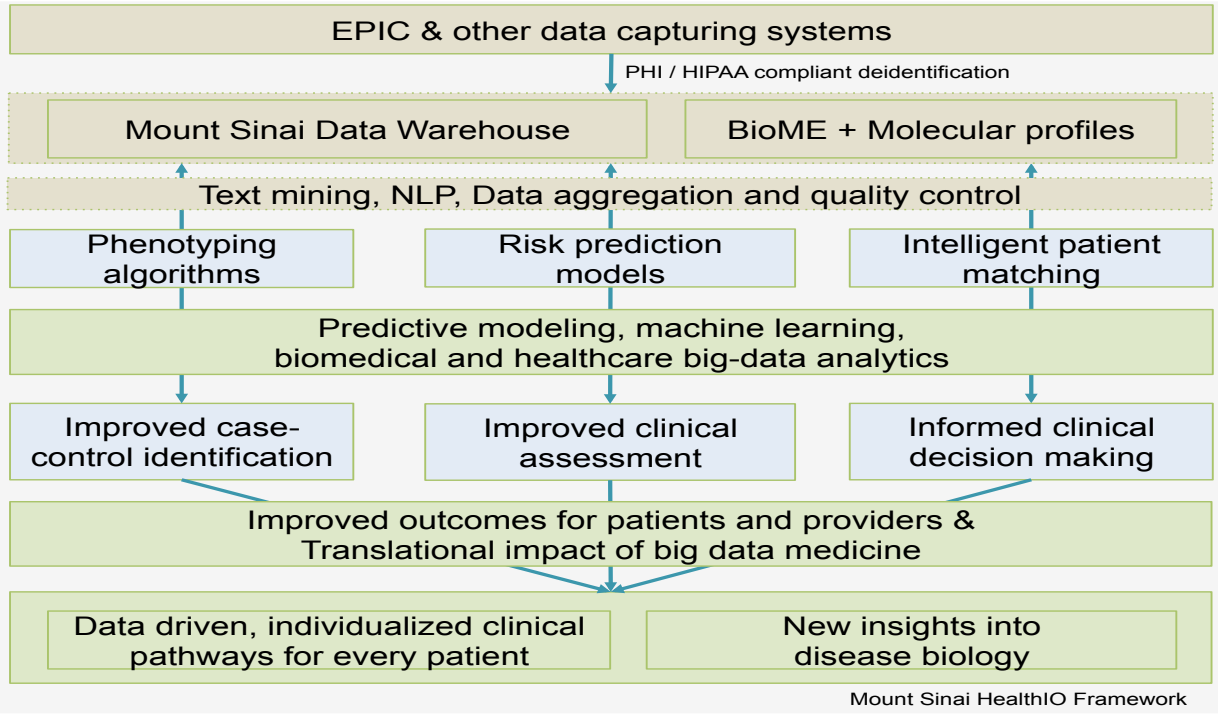
- Developing scalable and sustainable healthcare information technology (healthIT) solutions for data management, visual analytics and predictive modeling are critical to improving the quality of affordable healthcare delivery.
- We developed electronic healthcare data visualization (EHDViz) as a cost-effective, open-source, extensible toolkit for rapid design, development and implementation of clinical dashboards to address the need to improve data visualisation in different aspects of healthcare delivery, including population health management, patient engagement and simulation-based learning.
- A limitation of current version of EHDViz is that developers need skills in *R* and web development; also, extensive data cleaning and quality control steps are a priority before importing large quantity data for visualisation.
- While EHDViz is designed as vendor agnostic framework, importing data from external systems such as sensor devices and fitness monitors needs authorisations for access to the data and technical support from the device manufacturers or data integration services.
- To understand the benefits and limitations of the user experience, EHDViz and other open-source or commercial solutions with similar capabilities must be compared.

INTRODUCTION

The subclinical features and symptoms vary for individual patients as diseases progress and can be affected by lifestyle and medical interventions. These variations deviate greatly even among people with similar demographics, clinical profiles, family history and disease burdens. The patient-specific intervention a physician incorporates into a treatment plan relies heavily on the course of the illness. Electronic health record (EHR) software is widely used to capture longitudinal data and record vital signs,

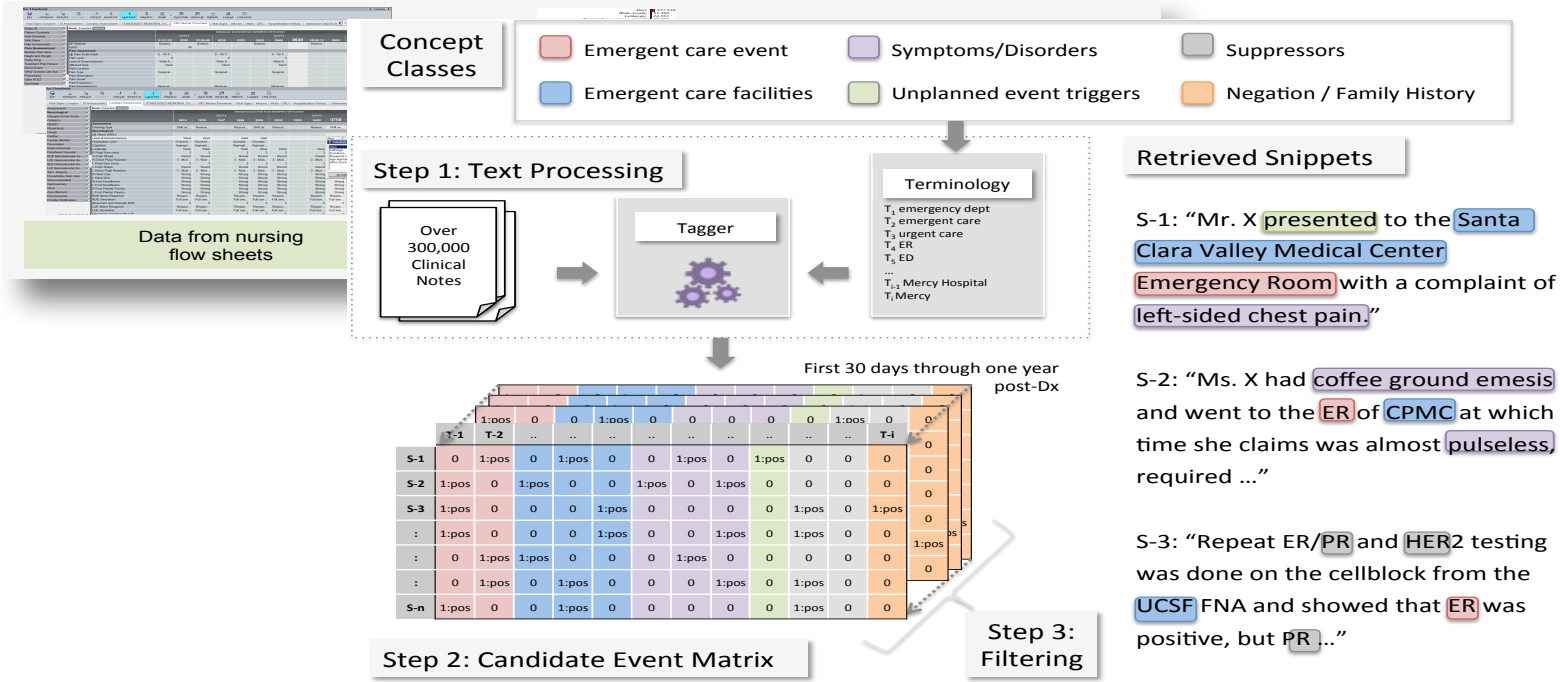
Badgeley MA, Shameer K, Glicksberg BS, Tomlinson MS, Levin MA, McCormick PJ, Kasarskis A, Reich DL, Dudley JT. EHDViz: clinical dashboard development using open-source technologies. BMJ Open. 2016 Mar 24;6(3):e010579. doi: 10.1136/bmjopen-2015-010579. PubMed PMID: 27013597; PubMed Central PMCID:PMC4809078.

HealthIO: Architecting a real-time modelling and predictive analytics data lake



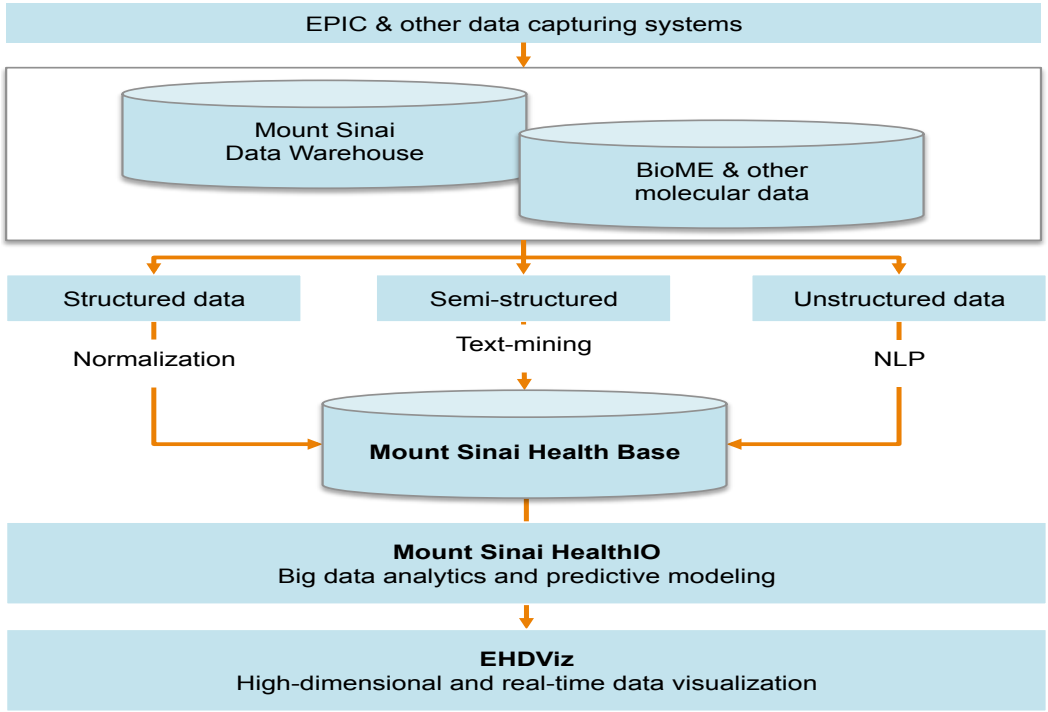
Mount Sinai Health Base/Mount Sinai HealthIO: Manuscript in preparation

Text mining and Medical NLP



Mount Sinai Health Base/Mount Sinai HealthIO: Manuscript in preparation
NLP project in collaboration with Stanford University, http://clever.stanford.edu

Integrating analytics and machine learning with healthcare delivery



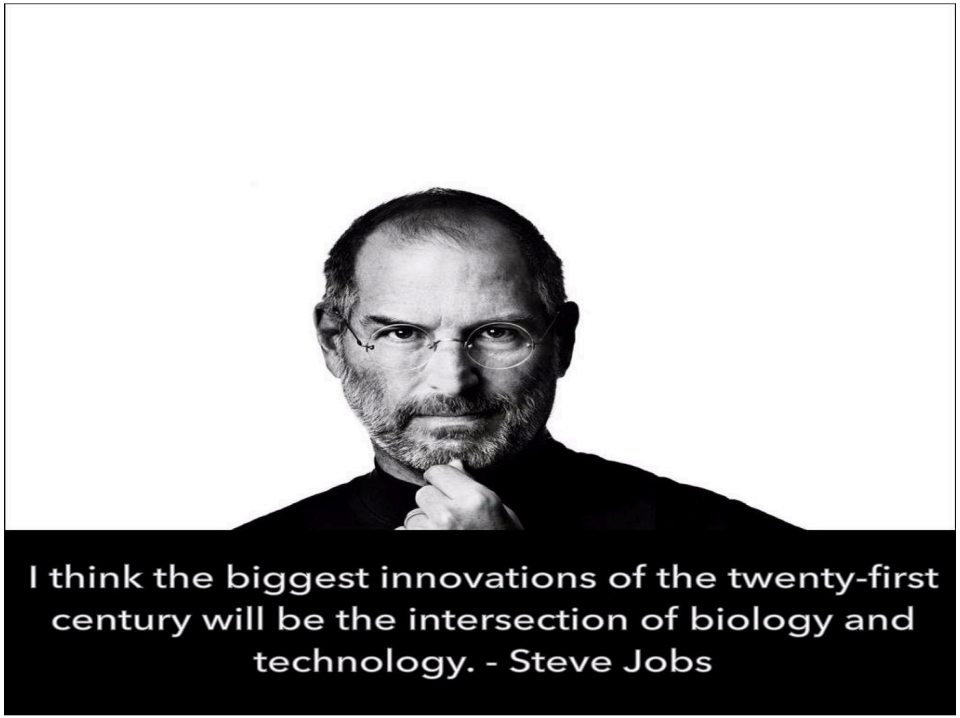
Mount Sinai Health Base/Mount Sinai HealthIO: Manuscript in preparation

Conclusions

- Implementing **data-driven methods that use real-time clinical variables** in a hypothesis-free approach could help us to find new features
- Designing predictive and prescriptive models would help to **accelerate stratification of patients at risk for improved care**
- Major implications in **quality of healthcare delivery and impact on patient outcomes**
- **Precision medicine is poised to optimize care and reduce cost**
- AI in healthcare is a **team sport** - Need team work (clinical research, experimental biology, bioinformatics, data science)
- **Everything is connected** — hence important to ask "bigger" and "bolder" questions using all available data

http://www.nejm.org/doi/pdf/10.1056/NEJMp1401111
http://jama.jamanetwork.com/article.aspx?articleid=1674245
http://www.thelancet.com/journals/langlo/article/PIIS2214-109X%2814%2970356-0/abstract

One more thing…



https://twitter.com/kshameer/status/575170163072499712