Translational Medicine in the Era of Big Data: Hype or Real?

AAHCI MENA Regional Conference
September 27, 2018

AKL FAHED, MD, MPH @aklfahed

MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER
CENTER FOR GENOMIC MEDICINE
HARVARD MEDICAL SCHOOL TEACHING HOSPITAL
BROAD INSTITUTE
Massachusetts Institute of Technology

Disclosures

MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER

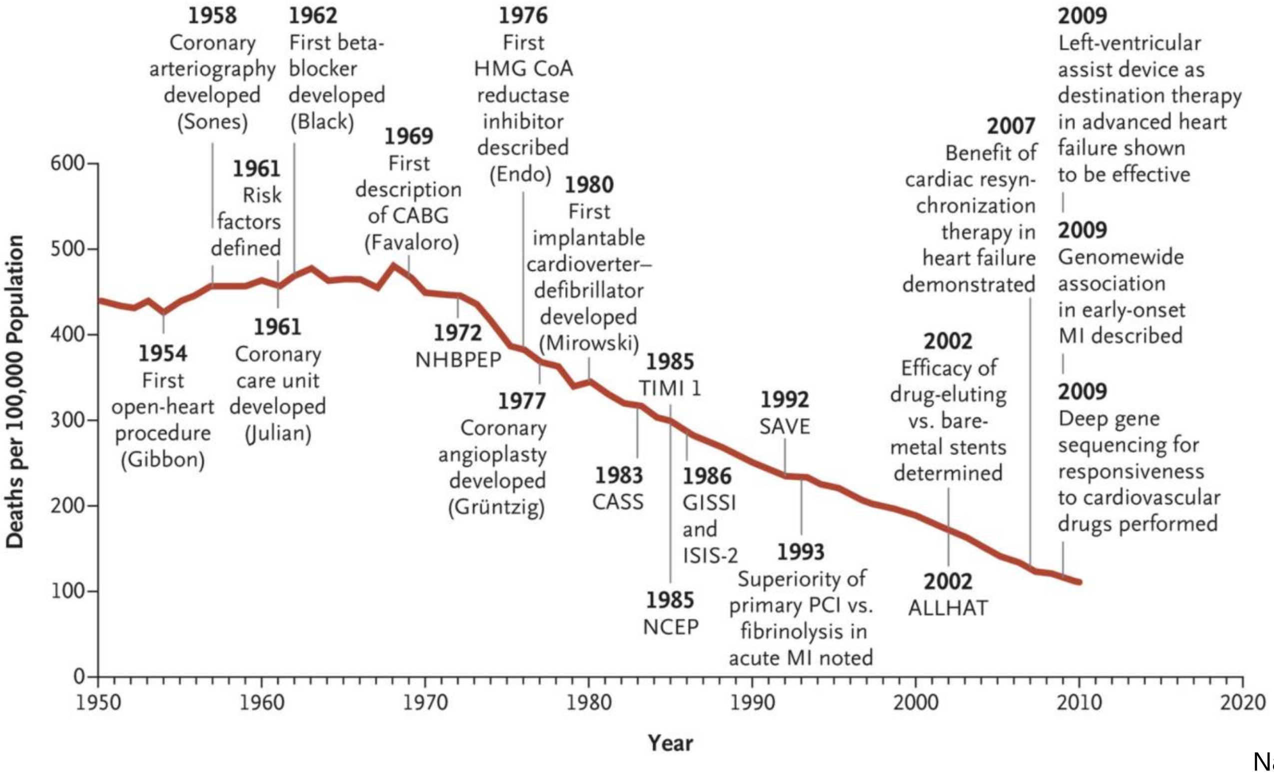
None

Outline

- The Promise of Big Data
- **Genomics**
  - Polygenic Risk Scores
  - Mendelian Randomization
  - Human Knockout Project
  - Phenome-Wide Association Studies
- Challenges and Pitfalls
- Opportunity for Academic Health Centers

Weintraub WS, Fahed AC, Rumsfeld JS. *Circulation Research. In Press*.

Decline in Cardiovascular Deaths



Nabel E and Braunwald E. *NEJM* 2012

Evidence-Based Therapies (1980-2000)

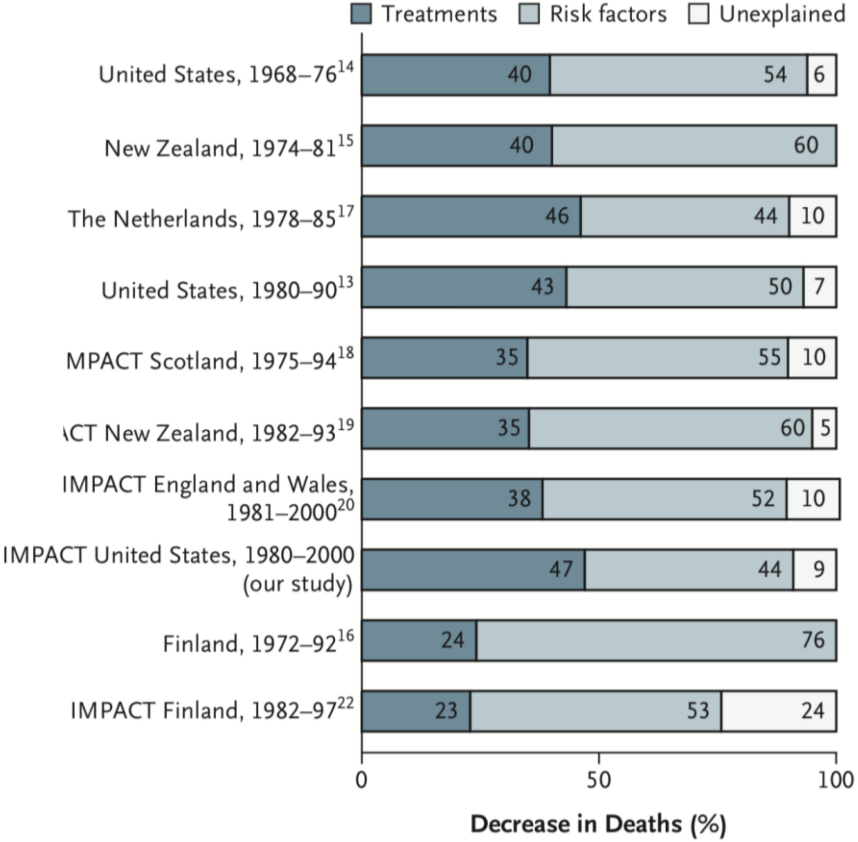
MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER

The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL ARTICLE

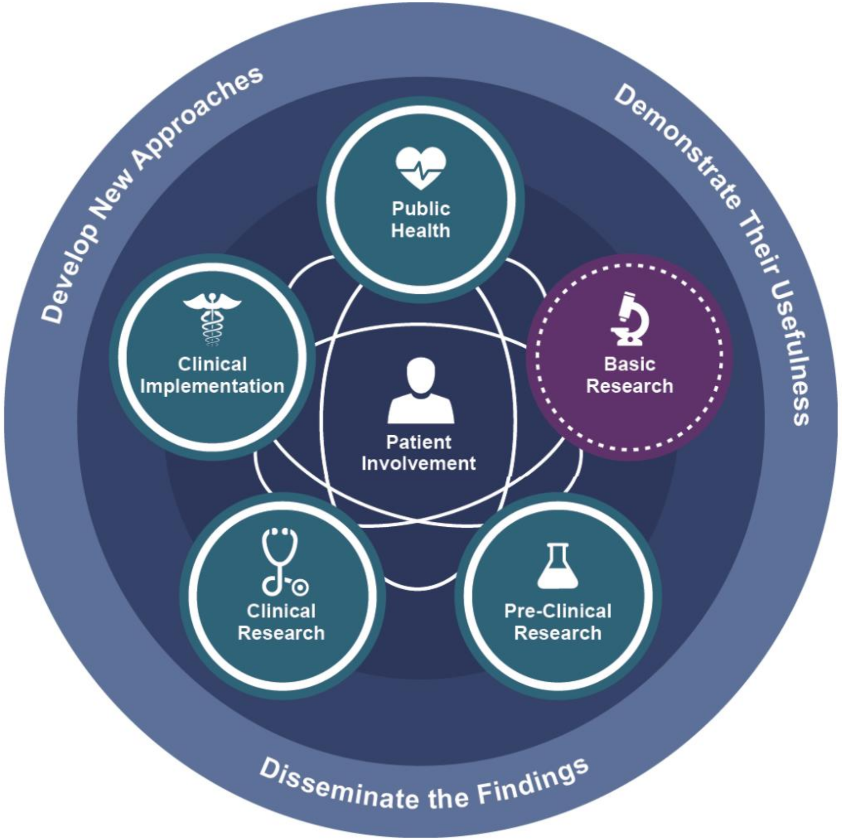
Explaining the Decrease in U.S. Deaths from Coronary Disease, 1980–2000

Earl S. Ford, M.D., M.P.H., Umed A. Ajani, M.B., B.S., M.P.H., Janet B. Croft, Ph.D., Julia A. Critchley, D.Phil., M.Sc., Darwin R. Labarthe, M.D., M.P.H., Ph.D., Thomas E. Kottke, M.D., Wayne H. Giles, M.D., M.S., and Simon Capewell, M.D.



Translational Medicine

MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER



Bench to Bedside to Population

www.ncats.nih.gov
Berwick DW et al. Health Affairs 2008

Yet …

MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER

Even highly efficacious therapies have heterogeneity of effect at the individual level

Significant variation in the use of evidence-based therapies and outcomes in routine clinical practice

Drug development is a very lengthy process

…

…

…

The Promise of Big Data

MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER

Precision Medicine

Artificial Intelligence

Improved Translational Medicine

…

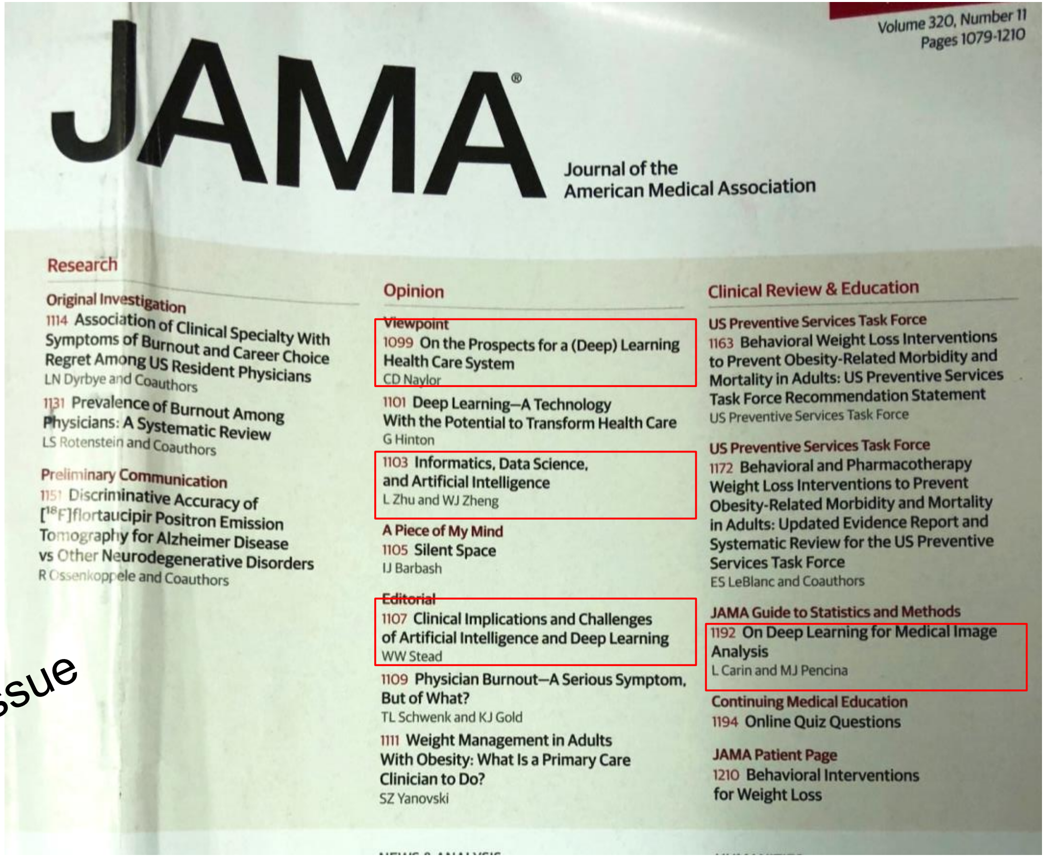
…

…

…

Hype or Real?

Current issue



Volume 320, Number 11
Pages 1079-1210

JAMA®
Journal of the American Medical Association

Research

Original Investigation

1114 Association of Clinical Specialty With Symptoms of Burnout and Career Choice Regret Among US Resident Physicians
LN Dyrbye and Coauthors

1131 Prevalence of Burnout Among Physicians: A Systematic Review
LS Rotenstein and Coauthors

Preliminary Communication

1151 Discriminative Accuracy of [^{18}F]flortaucipir Positron Emission Tomography for Alzheimer Disease vs Other Neurodegenerative Disorders
R Ossenkoppele and Coauthors

Opinion

Viewpoint

1099 On the Prospects for a (Deep) Learning Health Care System
CD Naylor

1101 Deep Learning—A Technology With the Potential to Transform Health Care
G Hinton

1103 Informatics, Data Science, and Artificial Intelligence
L Zhu and WJ Zheng

A Piece of My Mind

1105 Silent Space
IJ Barbash

Editorial

1107 Clinical Implications and Challenges of Artificial Intelligence and Deep Learning
WW Stead

1109 Physician Burnout—A Serious Symptom, But of What?
TL Schwenk and KJ Gold

1111 Weight Management in Adults With Obesity: What Is a Primary Care Clinician to Do?
SZ Yanovski

Clinical Review & Education

US Preventive Services Task Force

1163 Behavioral Weight Loss Interventions to Prevent Obesity-Related Morbidity and Mortality in Adults: US Preventive Services Task Force Recommendation Statement
US Preventive Services Task Force

US Preventive Services Task Force

1172 Behavioral and Pharmacotherapy Weight Loss Interventions to Prevent Obesity-Related Morbidity and Mortality in Adults: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force
ES LeBlanc and Coauthors

JAMA Guide to Statistics and Methods

1192 On Deep Learning for Medical Image Analysis
L Carin and MJ Pencina

Continuing Medical Education

1194 Online Quiz Questions

JAMA Patient Page

1210 Behavioral Interventions for Weight Loss

Sources of Big Data in Healthcare

Massachusetts General Hospital
Corrigan Minehan Heart Center

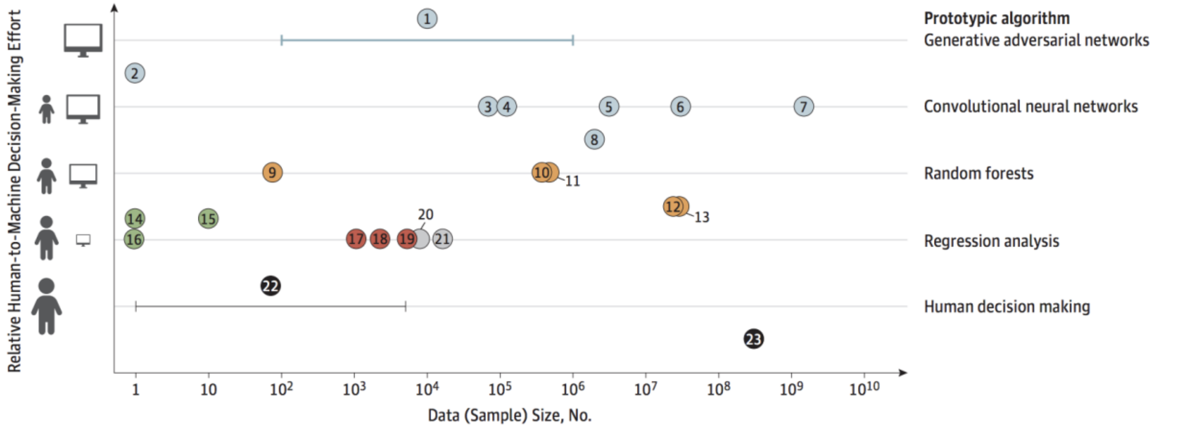
- Electronic Health Records (EHRs)
- Wearables, Apps and Biosensors (IoTs)
- **Genomic data**
- Insurance providers (claims, pharmacies, etc)
- Other clinical data (decision support tools, administrative data, etc)
- Social Media
- Web of knowledge

Zettabyte levels (10^{21})

Lima FV and Fahed AC. Harnessing the Power of Big Data in Healthcare. *Cardiology Magazine* 2018.

Spectrum of Big Data & Machine Learning

Figure. The Axes of Machine Learning and Big Data



Deep learning
1. Generative adversarial networks (2014)
2. Google AlphaGo Zero (2017)
3. ATM check readers (1998)
4. Google diabetic retinopathy (2016)
5. ImageNet computer vision models (2012-2017)
6. Google AlphaGo (2015)
7. Facebook Photo Tagger (2015)
8. Prediction of 1-y all-cause mortality (2017)

Classic machine learning
9. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling (2002)
10. EHR-based CV risk prediction (2017)
11. Netflix Prize winner (2006)
12. Google Search (1998)
13. Amazon product recommendation (2003)

Expert AI systems
14. MYCIN (1975)
15. CASNET (1982)
16. DXplain (1986)

Risk calculators
17. CHA_2DS_2-VASc Score for atrial fibrillation stroke risk (2017)
18. MELD end-stage liver disease risk score (2001)
19. Framingham CV risk score (1998)

Randomized Clinical Trials
20. Celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis (2002)
21. Use of estrogen plus progestin in healthy postmenopausal women (2002)

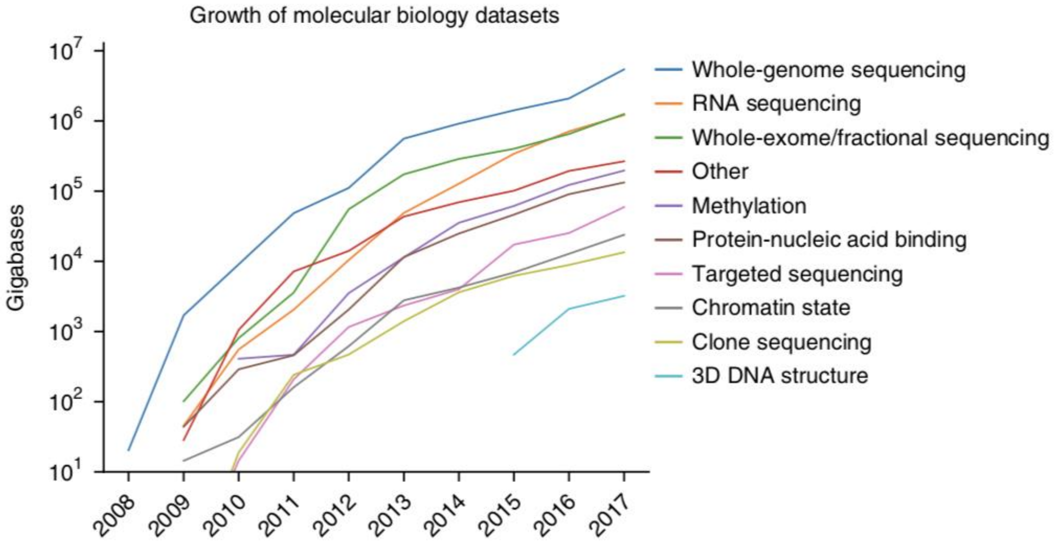
Other
22. Clinical wisdom
23. Mortality rate estimates from US Census (2010)

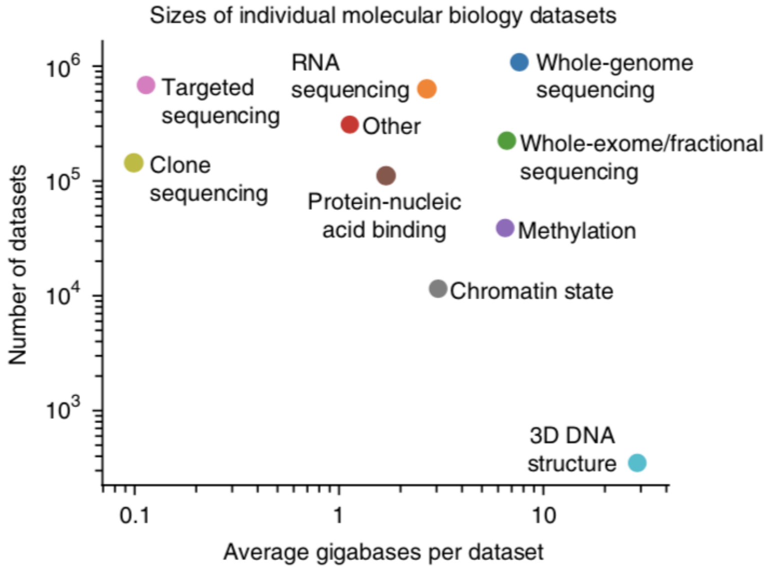
Beam A and Kohane I. JAMA 2018

Types of Genomic Data

Whole-Genome Genotyping	*Whole-Exome Sequencing*	*Whole-Genome Sequencing*
Array with 100,000- 1 millions SNPs Imputation: >90 million SNPs	Coding part (1%) of the genome All exons of all genes	Entirety of the genome
Common variation (allele frequency >1%)	Rare coding variation	Rare and Common disease Noncoding rare variation
GWAS, Mendelian Randomization, Polygenic Risk Scores	Rare disease diagnosis, discovery of novel rare loss of function	Role of noncoding DNA
~ 50 USD	~ 400 USD	~ 1500 USD
Public data +++++++	Public data ++++++++	Public data emerging

Growth and Size of Molecular Data

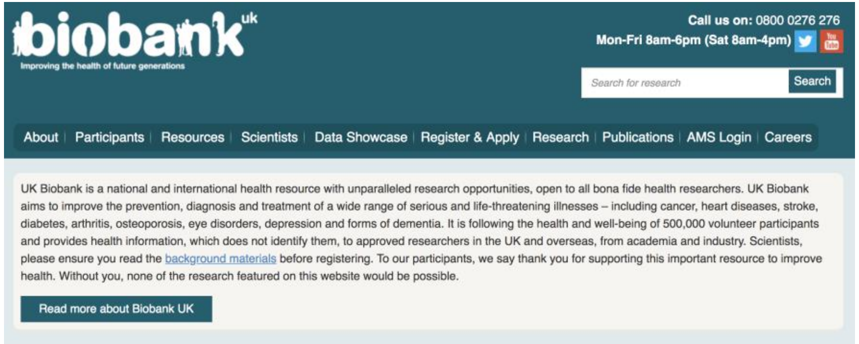




Wainberg et al. *Nature Biotechnology*. 2018

The Rise of the Biobanks

UK Biobank 500,0000



USA 1,000,000

USA 1,000,000

Biobank	Enrollment locations	Initial enrollment	Enrollment to date	Target enrollment
Commercial funding				
deCODE Genetics (Amgen) (http://www.decode.com/)	Iceland	1996	>200,000	Unknown
Geisinger MyCode® Community Health (Regeneron Pharmaceuticals and Others)	Geisinger Health System (Danville, PA)	2007	>50,000	Unknown
Government funding				
China Kadoorie Biobank (http://www.ckbiobank.org/site/)	China	2004	>500,000	Enrollment Completed
UK Biobank (https://www.ukbiobank.ac.uk/)	United Kingdom	2006	>500,000	Enrollment Completed
Electronic Medical Records and Genomics (eMERGE) Network (https://emerge.mc.vanderbilt.edu/about-emerge/)	United States Hospital Sites	2007	>50,000	Unknown
Million Veterans Program (http://www.research.va.gov/mvp/)	Veterans Affairs Hospital	2011	>500,000	~1,000,000
Precision Medicine Initiative (https://www.nih.gov/precision-medicine-initiative-cohort-program)	United States	Early 2017	--	~1,000,000
Institutional funding				
BioVu Biorepository (https://victr.vanderbilt.edu/pub/biovu/)	Vanderbilt University Medical Center (Nashville, TN)	2007	>215,000	Unknown
Kaiser Permanente Research Bank (http://researchbank.kaiserpermanente.org/)	United States	2016	>250,000	~500,000
Partners Healthcare Biobank (https://biobank.partners.org/)	Partners Health Care (Boston, MA)	2010	>50,000	~100,000

Khera AV and Kathiresan S. *Nature Reviews Genetics*. 2017

Democratization of Genomic Data

MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER

Ben Neale

HOME RESEARCH PEOPLE MEDIA BLOG UK BIOBANK JOBS CONTACT

RAPID GWAS OF THOUSANDS OF PHENOTYPES FOR 337,000 SAMPLES IN THE UK BIOBANK

September 20, 2017

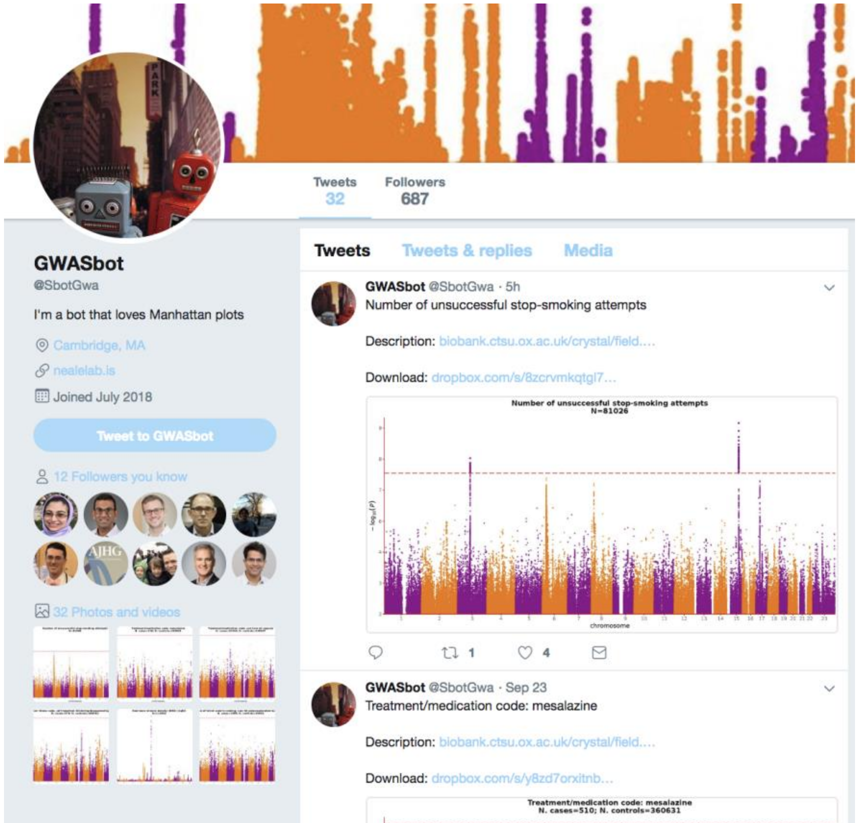
biobank uk

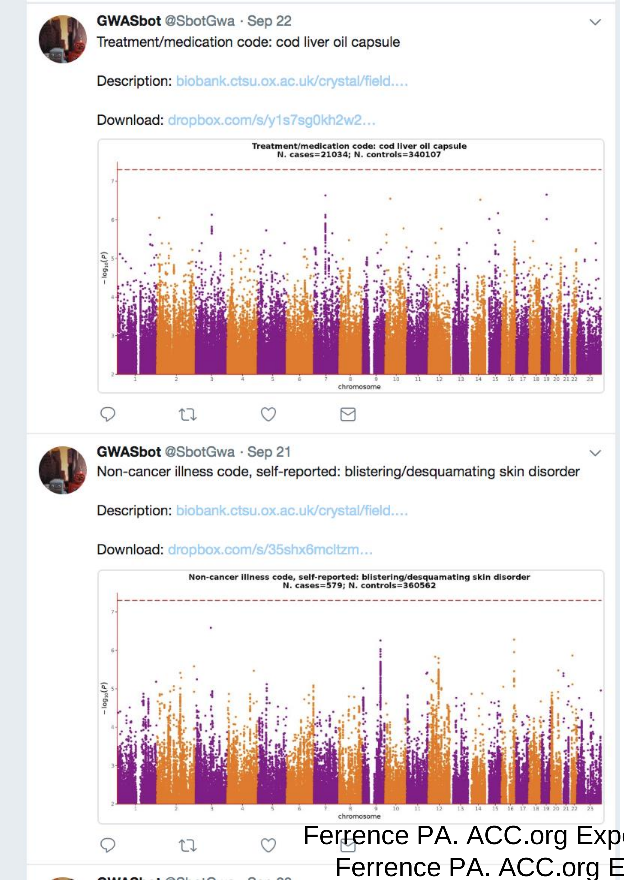
The UK Biobank recently released genome-wide association data on ~500,000 individuals. The genotype data for these samples have been cleaned, imputed and released to the scientific community. This public release of data represents an extraordinary advance for genetics, pushing the envelope for data sharing and rapid uptake by the research community. These data will be used for novel discovery of disease-associated genes, in the development of new methods, and to serve as an example for how future efforts in genetics and biology ought to proceed.

To further enhance the value of this resource, we have performed a basic association test on ~337,000 unrelated individuals of British ancestry for over 2,000 of the available phenotypes. We're making these results available for browsing through several portals, including the Global Biobank Engine where they will appear soon. They are also available for download here.

UKBB GWAS bot

MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER





Ferrence PA. ACC.org Expert Analysis

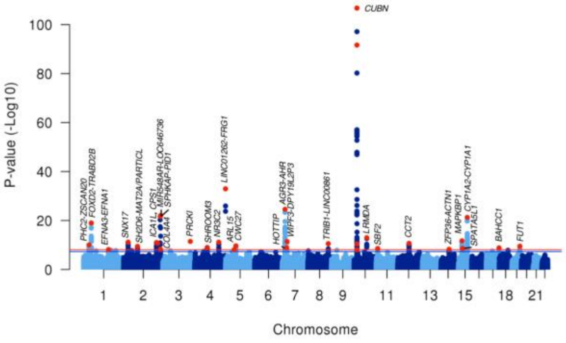
Ferrence PA. ACC.org Expert Analysis

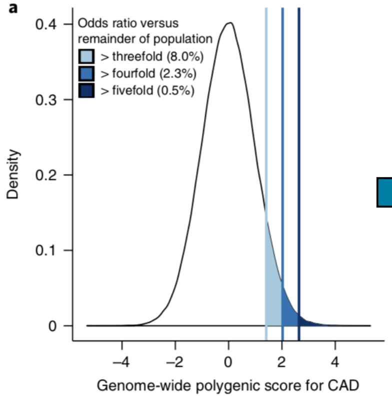
Polygenic Risk Scores (PRS)

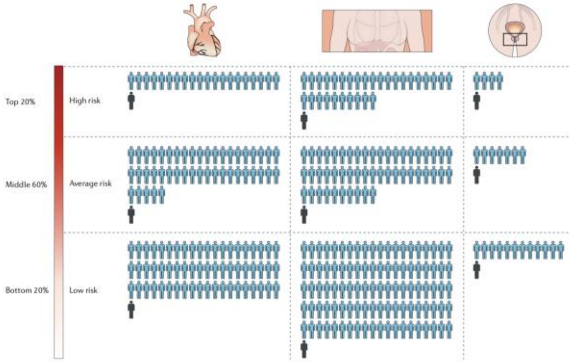
Weighted sum of number of risk alleles carried by an individual

- Sum of the risk alleles (X)
- Measured effects as detected by GWAS (β)

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots.$$

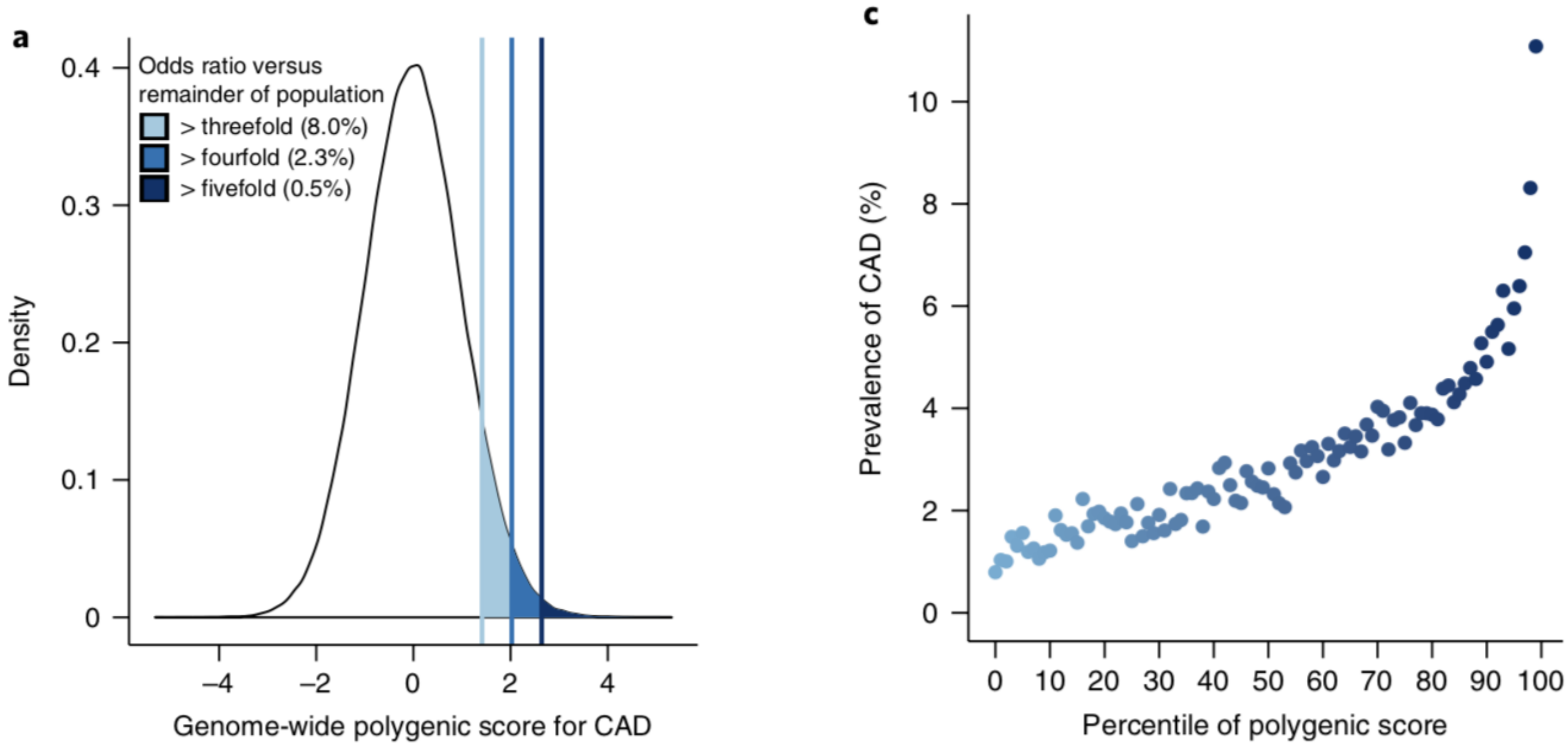






CAD Polygenic Risk Score
LDpred method(>6 million alleles)

MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER



Khera AV et al. *Nature Genetics* 2018

Atrial Fibrillation, Type 2 Diabetes, Inflammatory Bowel Disease, Breast Cancer

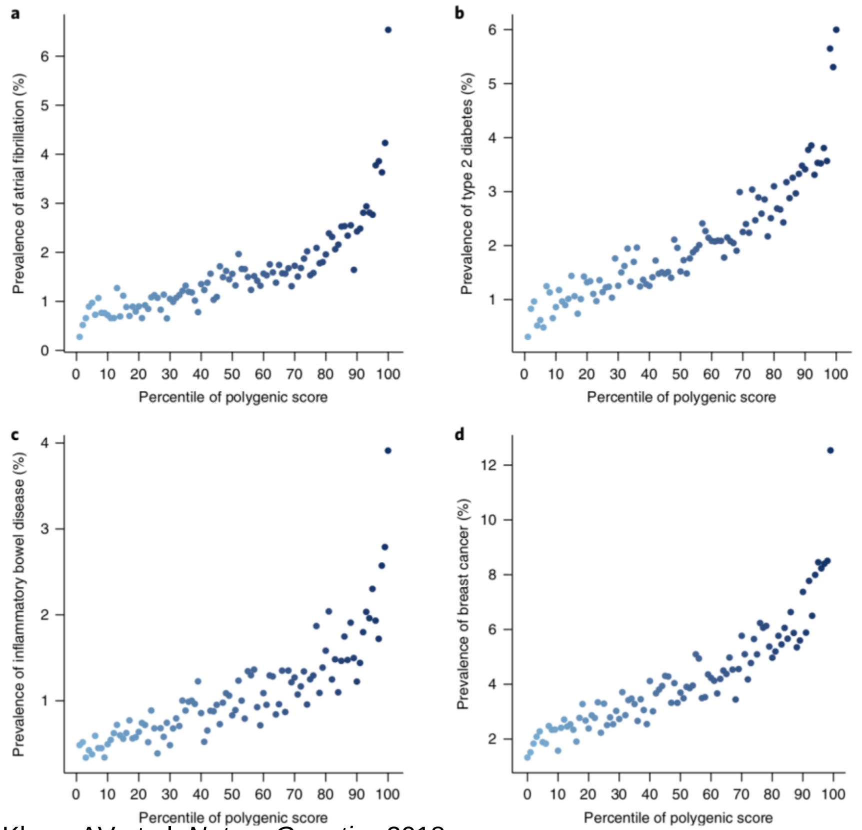


Table 3 | Prevalence and clinical impact of a high GPS

High GPS definition	Reference group	Odds ratio	95% CI	*P* value
CAD				
Top 20% of distribution	Remaining 80%	2.55	2.43–2.67	$<1\times10^{-300}$
Top 10% of distribution	Remaining 90%	2.89	2.74–3.05	$<1\times10^{-300}$
Top 5% of distribution	Remaining 95%	3.34	3.12–3.58	6.5×10^{-264}
Top 1% of distribution	Remaining 99%	4.83	4.25–5.46	1.0×10^{-132}
Top 0.5% of distribution	Remaining 99.5%	5.17	4.34–6.12	7.9×10^{-78}
Atrial fibrillation				
Top 20% of distribution	Remaining 80%	2.43	2.29–2.59	2.1×10^{-177}
Top 10% of distribution	Remaining 90%	2.74	2.55–2.94	7.0×10^{-169}
Top 5% of distribution	Remaining 95%	3.22	2.95–3.51	1.1×10^{-152}
Top 1% of distribution	Remaining 99%	4.63	3.96–5.39	2.9×10^{-84}
Top 0.5% of distribution	Remaining 99.5%	5.23	4.24–6.39	3.5×10^{-56}
Type 2 diabetes				
Top 20% of distribution	Remaining 80%	2.33	2.20–2.46	3.1×10^{-201}
Top 10% of distribution	Remaining 90%	2.49	2.34–2.66	1.2×10^{-167}
Top 5% of distribution	Remaining 95%	2.75	2.53–2.98	1.7×10^{-130}
Top 1% of distribution	Remaining 99%	3.30	2.81–3.85	1.4×10^{-49}
Top 0.5% of distribution	Remaining 99.5%	3.48	2.79–4.29	4.3×10^{-30}
Inflammatory bowel disease				
Top 20% of distribution	Remaining 80%	2.19	2.03–2.36	7.7×10^{-95}
Top 10% of distribution	Remaining 90%	2.43	2.22–2.65	8.8×10^{-88}
Top 5% of distribution	Remaining 95%	2.66	2.38–2.96	3.0×10^{-68}
Top 1% of distribution	Remaining 99%	3.87	3.18–4.66	1.4×10^{-43}
Top 0.5% of distribution	Remaining 99.5%	4.81	3.74–6.08	9.0×10^{-37}
Breast cancer				
Top 20% of distribution	Remaining 80%	2.07	1.97–2.19	3.4×10^{-159}
Top 10% of distribution	Remaining 90%	2.32	2.18–2.48	2.3×10^{-148}
Top 5% of distribution	Remaining 95%	2.55	2.35–2.76	2.1×10^{-112}
Top 1% of distribution	Remaining 99%	3.36	2.88–3.91	1.3×10^{-54}
Top 0.5% of distribution	Remaining 99.5%	3.83	3.11–4.68	8.2×10^{-38}

Odds ratios were calculated by comparing those with high GPS with the remainder of the population in a logistic regression model adjusted for age, sex, genotyping array, and the first four principal components of ancestry. The breast cancer analysis was restricted to female participants. CI, confidence interval.

Khera AV et al. *Nature Genetics* 2018

Polygenic Risk Prediction

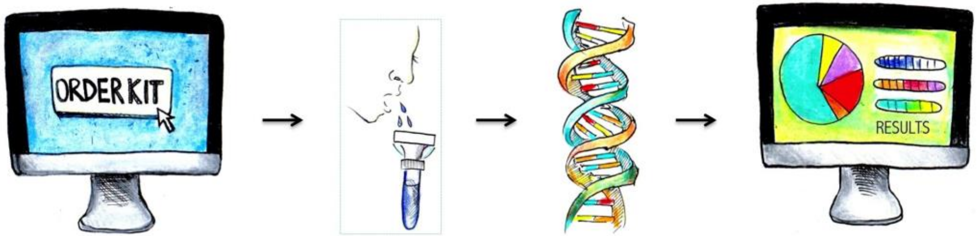
20% of the study population are at ≥ threefold

increased risk for at least 1 of the 5 diseases studied !

"The First" risk factor

~100 USD

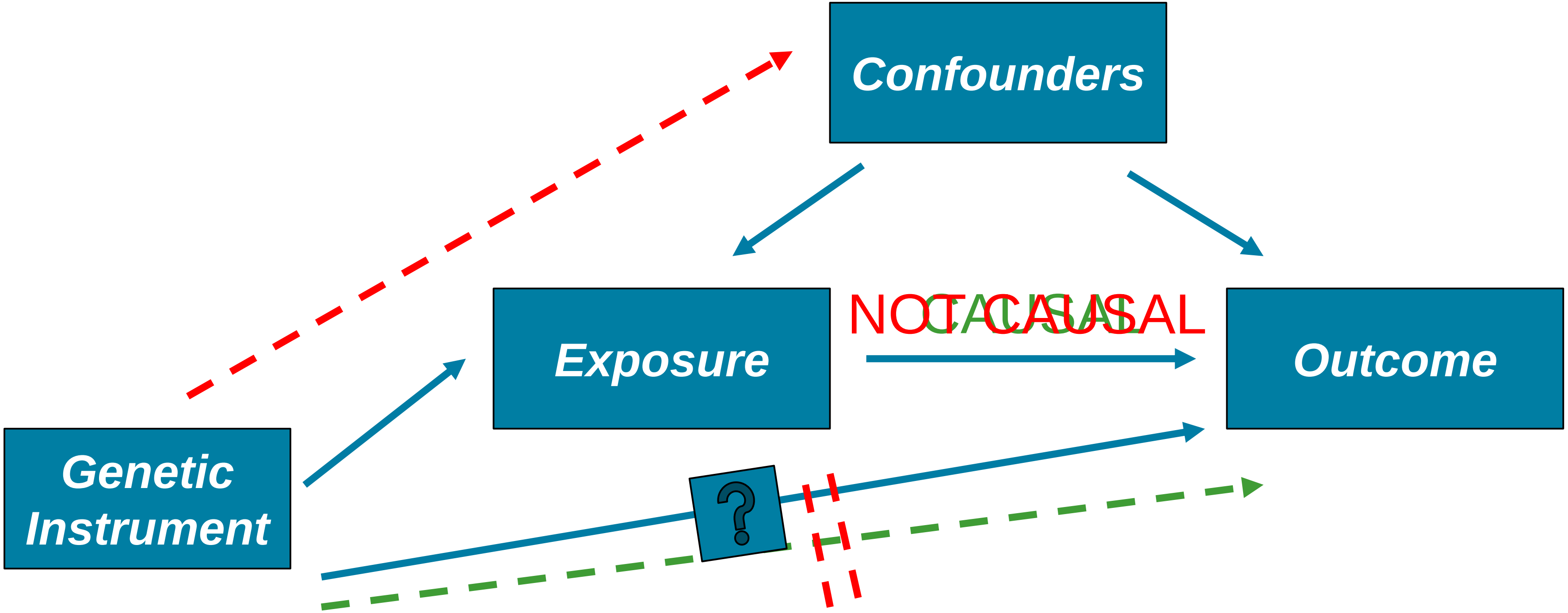
Direct to Consumer Genetics



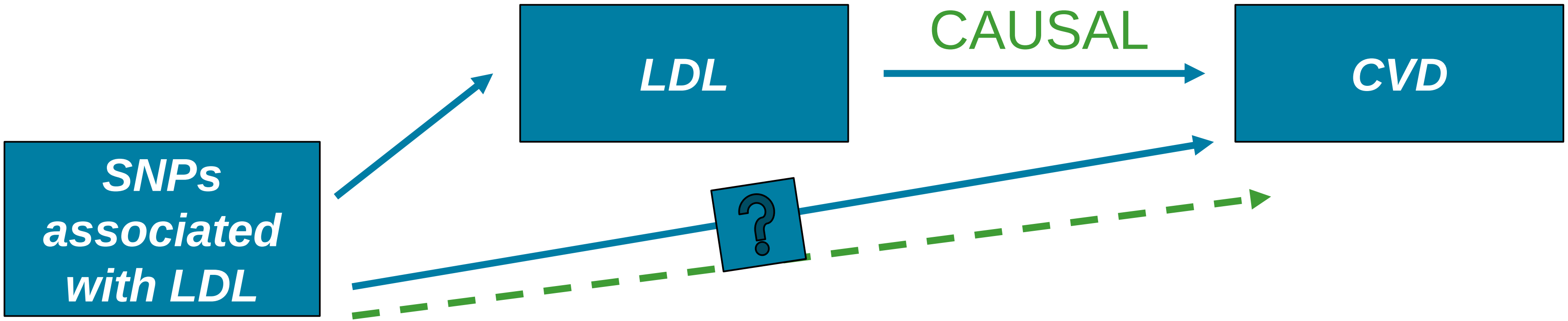
Khera AV et al. *Nature Genetics* 2018
https://pged.org/direct-to-consumer-genetic-testing/

Mendelian Randomization

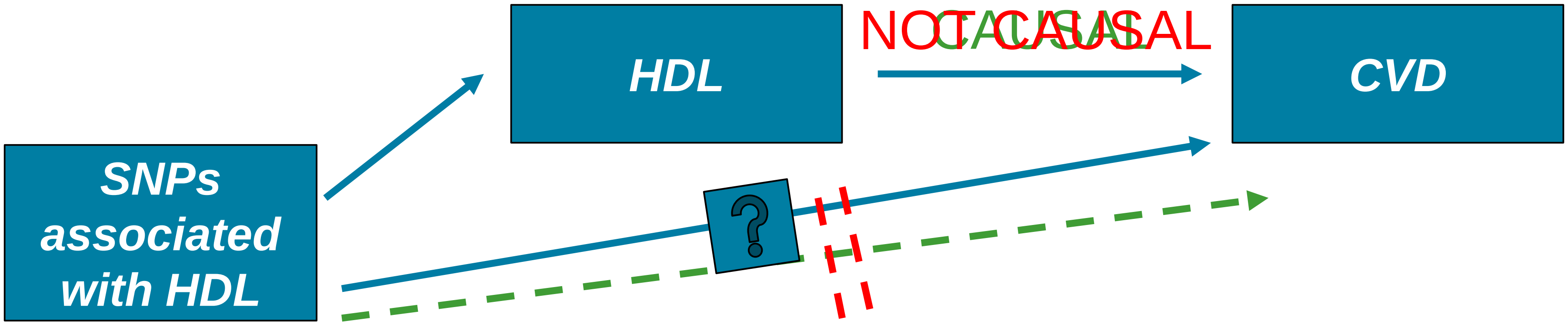
MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER



Mendelian Randomization



Mendelian Randomization



Human Knockout Project

LETTER

doi:10.1038/nature22034

Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity

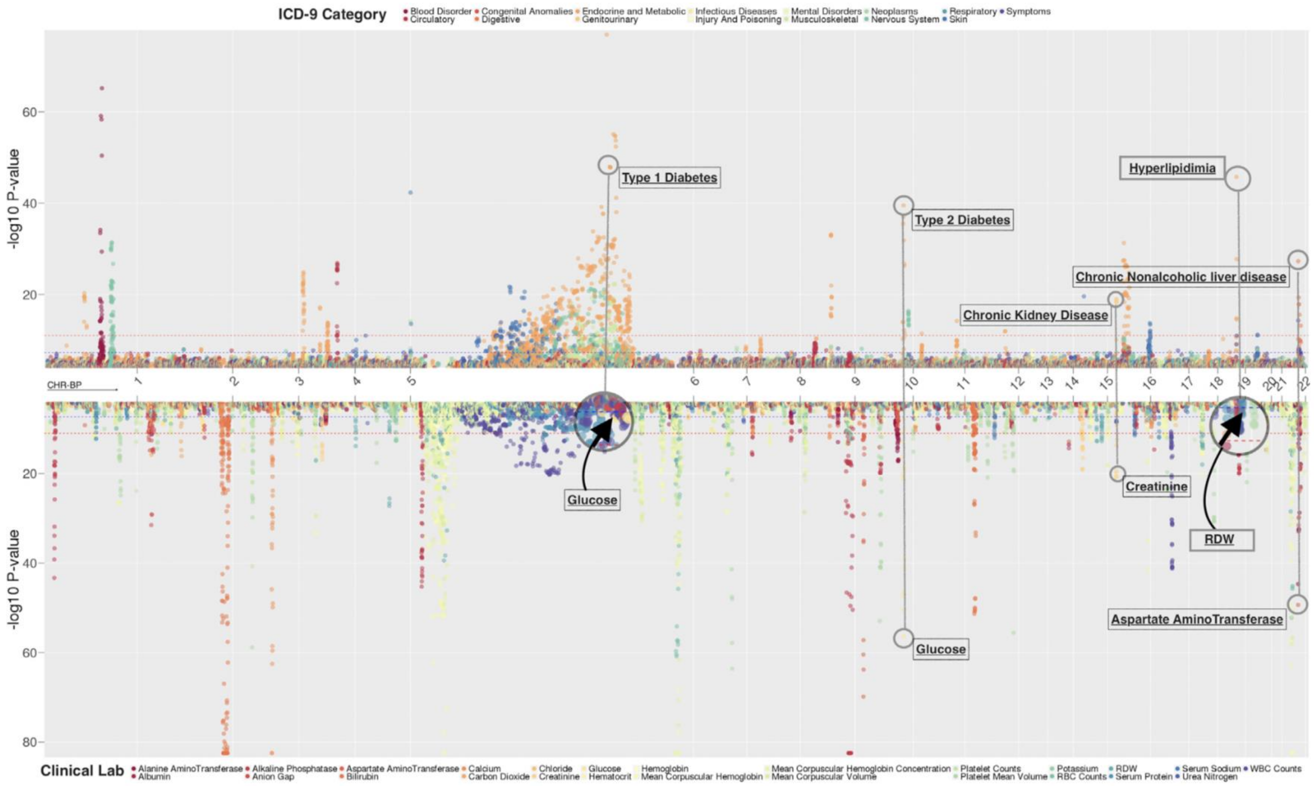
Danish Saleheen[1,2*], Pradeep Natarajan[3,4*], Irina M. Armean[4,5], Wei Zhao[1], Asif Rasheed[2], Sumeet A. Khetarpal[6], Hong-Hee Won[7], Konrad J. Karczewski[4,5], Anne H. O'Donnell-Luria[4,5,8], Kaitlin E. Samocha[4,5], Benjamin Weisburd[4,5], Namrata Gupta[4], Mozzam Zaidi[2], Maria Samuel[2], Atif Imran[2], Shahid Abbas[9], Faisal Majeed[2], Madiha Ishaq[2], Saba Akhtar[2], Kevin Trindade[6], Megan Mucksavage[6], Nadeem Qamar[10], Khan Shah Zaman[10], Zia Yaqoob[10], Tahir Saghir[10], Syed Nadeem Hasan Rizvi[10], Anis Memon[10], Nadeem Hayyat Mallick[11], Mohammad Ishaq[12], Syed Zahed Rasheed[12], Fazal-ur-Rehman Memon[13], Khalid Mahmood[14], Naveeduddin Ahmed[15], Ron Do[16,17], Ronald M. Krauss[18], Daniel G. MacArthur[4,5], Stacey Gabriel[4], Eric S. Lander[4], Mark J. Daly[4,5], Philippe Frossard[2]§, John Danesh[19,20]§, Daniel J. Rader[6,21]§ & Sekar Kathiresan[3,4]§

Safety check for drug development

- Exome sequencing of 10,503 Pakistani subjects
- Identify individuals carrying predicted homozygous loss-of-function mutations
- Perform phenotypic analysis of >200 biochemical disease traits
- e.g. *APOC3* hom pLoF low fasting TG and blunted post-prandial lipaemia

Phenome Wide Association Studies (PheWAS)

Association of SNPs with Medical Diagnoses and Clinical Measures in the EHR



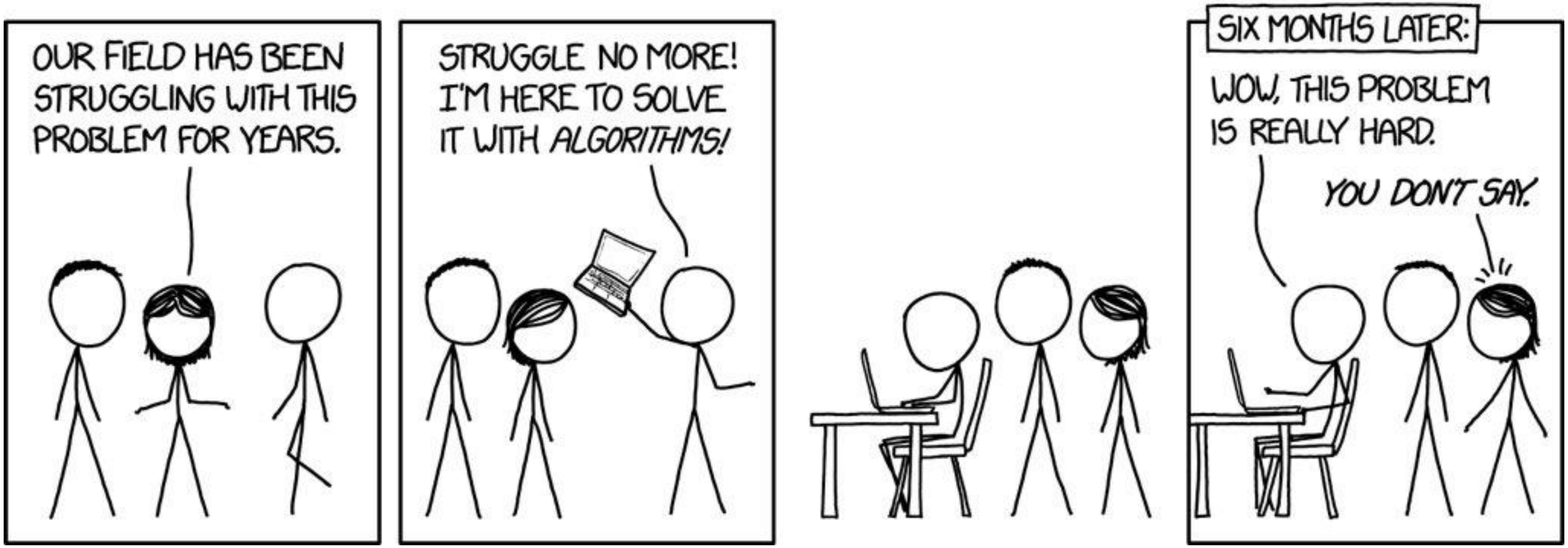
Verma A. et al. *AJHG 2018*

Pitfalls of Big Data and ML

- Improved generation of hypotheses
  - But burden of proof remains on the basic scientist
- Polygenic risk implementation in care
  - Will it change outcomes?
- Biobanks phenotypic classification (case/control definitions)
- EHR/Administrative data has inherent biases of observational data
  - Informative missing data
  - Risk of false positives and negatives (i.e. misclassification)
  - Treatment selection bias i.e. unmeasured confounding variables

Data Science in Academic Health Centers

MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER



Source: Twitter @AndrewLBeam

Doctors have a ‘hunch’ and it matters!

MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER

FULL SCREEN

A new study from MIT computer scientists suggests that human doctors provide a dimension that, as yet, artificial intelligence does not.

Image: Chelsea Turner, MIT

Doctors rely on more than just data for medical decision making

Computer scientists find that physicians’ “gut feelings” influence how many tests they order for patients.

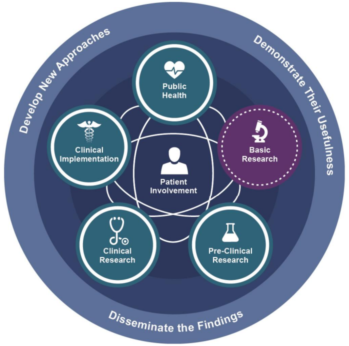
Watch Video

Anne Trafton | MIT News Office
July 20, 2018

Press Inquiries

RELATED

Opportunity for Academic Health Centers



The triple aim: care, health, and cost

- Data Science as part of the framework of translational research
- Essential basic, translational and epidemiologic research for new technologies
- Unique partnerships with industry
- Products that are cost-effective, scientifically solid, and needed to advance patient care

The new med school classroom?

- Computationally-Enabled Medicine
- “Pathways” curriculum
- Harvard Medical School 3rd year students

https://hms.harvard.edu/news/knowing-unknown

Thank you

@aklfahed

fahed@mail.harvard.edu

Acknowledgements:

NIH NHLBI
Sekar Kathiresan, MD

William S. Weintraub, MD
John S. Rumsfeld, MD, PhD

MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER

CENTER FOR GENOMIC MEDICINE

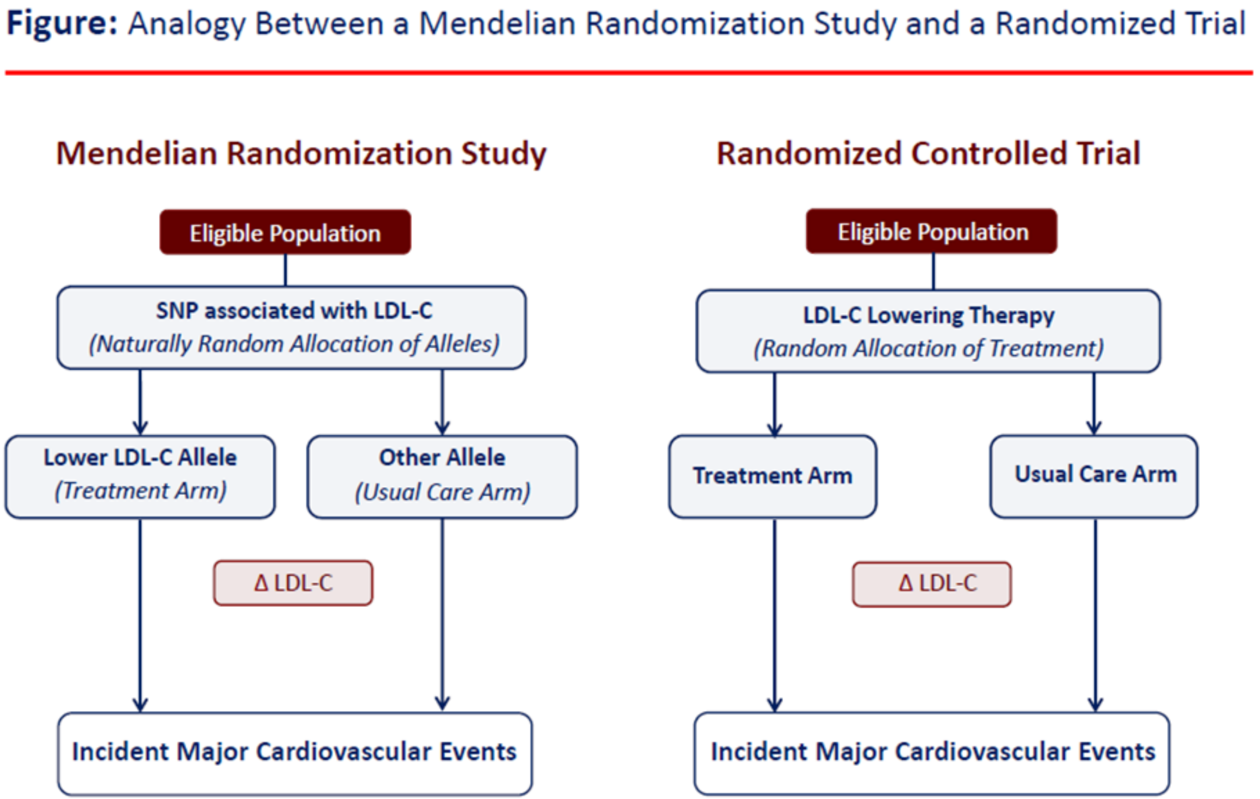
HARVARD MEDICAL SCHOOL TEACHING HOSPITAL

BROAD INSTITUTE

Massachusetts Institute of Technology

Mendelian Randomization

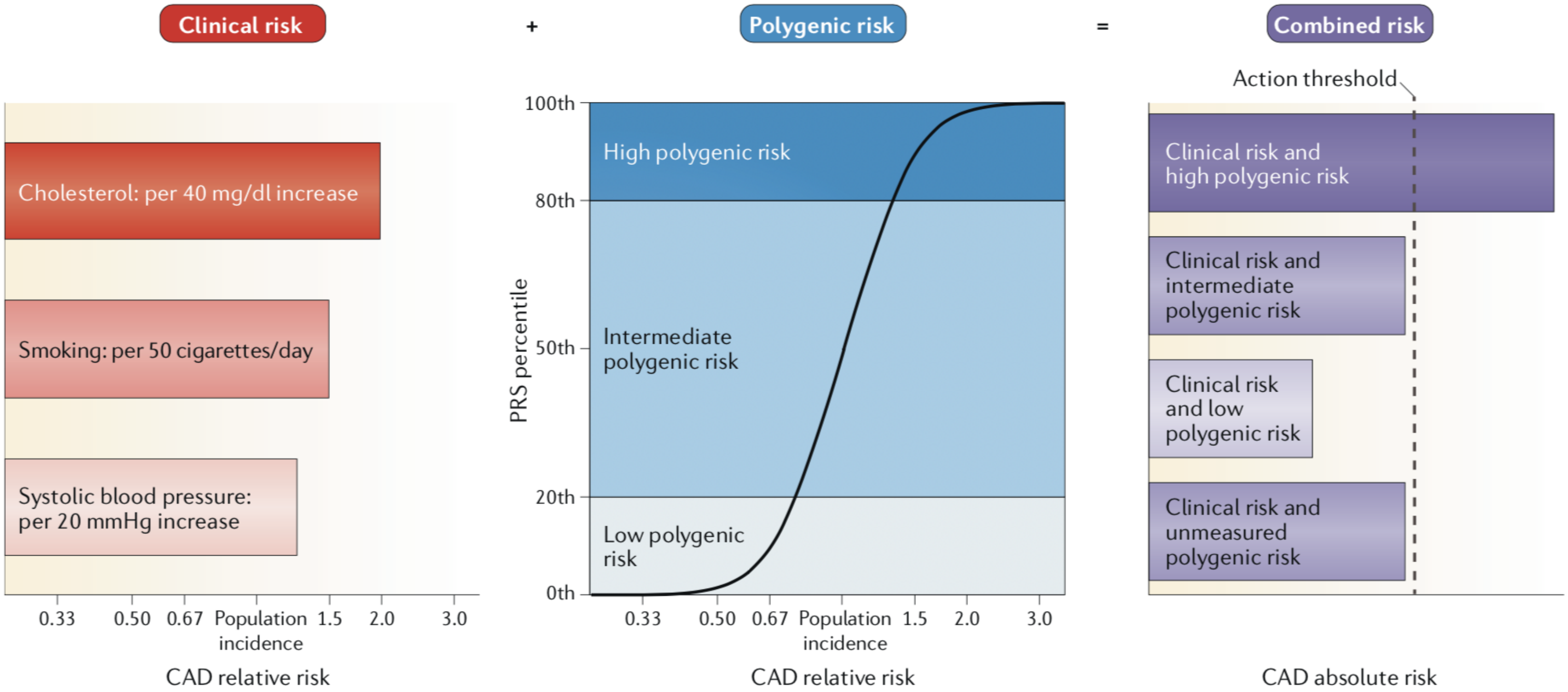
MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER



Ferrence PA. ACC.org Expert Analysis

Integrating Clinical and Polygenic Risk Prediction

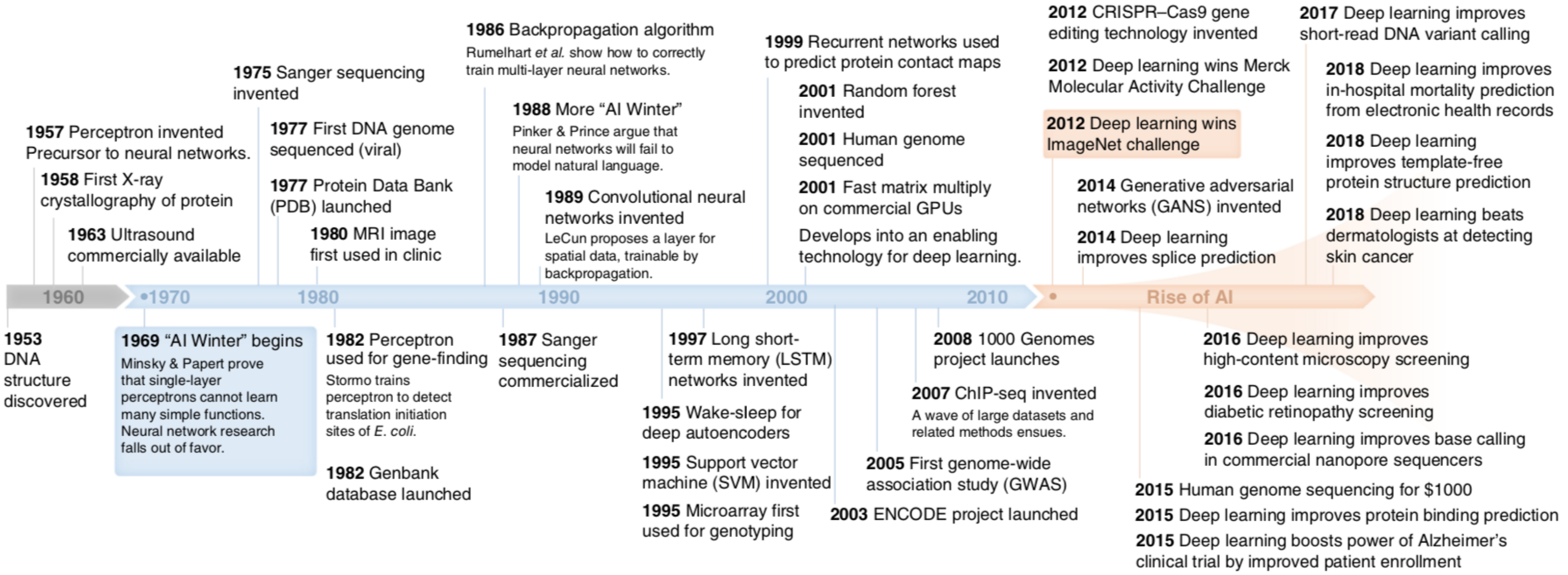
MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER



Torkamani A et al. *Nature Reviews Genetics* 2018

Timeline of Molecular Data

MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER



Wainberg et al. *Nature Biotechnology*. 2018

Take Home

MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER

“ Machine Learning should try to do:

1- What doctors cannot do

2 What doctors do NOT what to do ”

Not all Data are Created Equal

MASSACHUSETTS GENERAL HOSPITAL
CORRIGAN MINEHAN HEART CENTER

Low Quality for ML

- EHR
- Administrative Data

Good Quality for ML

- Image interpretation
  - CT
  - MRI
  - Echocardiography
- Detection of Dysrhythmias
  - Cardiac rhythm
- Wearables/Biosensors
  - HR/ Other physiological data
- Molecular data