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

AAHCI MENA Regional Conference
September 27, 2018

AKL FAHED, MD, MPH  @aklfahed
Disclosures

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

Decline in Cardiovascular Deaths

1958 Coronary arteriography developed (Sones)
1962 First beta-blocker developed (Black)
1969 First description of CABG (Favaloro)
1976 First HMG CoA reductase inhibitor described (Endo)
1980 First implantable cardioverter-defibrillator developed (Mirowski)
1972 NHBPEP
1977 Coronary angioplasty developed (Grüntzig)
1981 TIMI 1
1983 CASS
1985 NCEP
1986 GISSI and ISIS-2
1992 SAVE
1993 Superiority of primary PCI vs. fibrinolysis in acute MI noted
2002 ALLHAT
2009 Left-ventricular assist device as destination therapy in advanced heart failure shown to be effective
2009 Genomewide association in early-onset MI described
2009 Deep gene sequencing for responsiveness to cardiovascular drugs performed

Nabel E and Braunwald E. NEJM 2012
Evidence-Based Therapies (1980-2000)

SPECIAL ARTICLE

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

Translational Medicine

Bench to Bedside to Population

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

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

Precision Medicine

Artificial Intelligence

Improved Translational Medicine

Hype or Real?

Current issue
Sources of Big Data in Healthcare

- 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

Spectrum of Big Data & Machine Learning

Figure. The Axes of Machine Learning and Big Data

Deep learning
2. Google AlphaGo Zero (2017)
3. ATM check readers (1998)
5. ImageNet computer vision models (2012-2017)
7. Facebook Photo Tagger (2015)

Classic machine learning
10. EHR-based CV risk prediction (2017)

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

Risk calculators
17. CHA\textsubscript{2}-DS\textsubscript{2}-VASc Score for atrial fibrillation stroke risk (2017)
18. MELD end-stage liver disease risk score (2001)

Randomized Clinical Trials

Other
22. Clinical wisdom
# Types of Genomic Data

<table>
<thead>
<tr>
<th>Whole-Genome Genotyping</th>
<th>Whole-Exome Sequencing</th>
<th>Whole-Genome Sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Array with 100,000-1 millions SNPs Imputation: &gt;90 million SNPs</td>
<td>Coding part (1%) of the genome</td>
<td>Entirety of the genome</td>
</tr>
<tr>
<td>Common variation (allele frequency &gt;1%)</td>
<td>Rare coding variation</td>
<td>Rare and Common disease Noncoding rare variation</td>
</tr>
<tr>
<td>GWAS, Mendelian Randomization, Polygenic Risk Scores</td>
<td>Rare disease diagnosis, discovery of novel rare loss of function</td>
<td>Role of noncoding DNA</td>
</tr>
<tr>
<td>~ 50 USD</td>
<td>~ 400 USD</td>
<td>~ 1500 USD</td>
</tr>
<tr>
<td>Public data ++++</td>
<td>Public data ++++</td>
<td>Public data emerging</td>
</tr>
</tbody>
</table>
Growth and Size of Molecular Data

Wainberg et al. *Nature Biotechnology*. 2018
The Rise of the Biobanks

**UK Biobank 500,000**

UK Biobank is a national and international health resource with unparalleled research opportunities, open to all bona fide health researchers. UK Biobank aims to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses – including cancer, heart diseases, strokes, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia. It is following the health and well-being of 500,000 volunteer participants and provides health information, which does not identify them, to approved researchers in the UK and overseas, from academia and industry. Scientists, please ensure you read the background materials before registering. To our participants, we say thank you for supporting this important resource to improve health. Without you, none of the research featured on this website would be possible.

**USA 1,000,000**

**USA 1,000,000**

<table>
<thead>
<tr>
<th>Biobank</th>
<th>Enrollment locations</th>
<th>Initial enrollment</th>
<th>Enrollment to date</th>
<th>Target enrollment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Commercial funding</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geisinger MyCode® Community Health (Regeneron Pharmaceuticals and Others)</td>
<td>Geisinger Health System (Danville, PA)</td>
<td>2007</td>
<td>&gt;50,000</td>
<td>Unknown</td>
</tr>
<tr>
<td><strong>Government funding</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China Kadokie Biobank (<a href="http://www.ckbiobank.org/site/">http://www.ckbiobank.org/site/</a>)</td>
<td>China</td>
<td>2004</td>
<td>&gt;500,000</td>
<td>Enrollment Completed</td>
</tr>
<tr>
<td>UK Biobank (<a href="http://www.ukbiobank.ac.uk/">http://www.ukbiobank.ac.uk/</a>)</td>
<td>United Kingdom</td>
<td>2006</td>
<td>&gt;500,000</td>
<td>Enrollment Completed</td>
</tr>
<tr>
<td>Electronic Medical Records and Genomics (eMERGE) Network (<a href="https://emerge.mc.vanderbilt.edu/about-emerge/">https://emerge.mc.vanderbilt.edu/about-emerge/</a>)</td>
<td>United States Hospital Sites</td>
<td>2007</td>
<td>&gt;50,000</td>
<td>Unknown</td>
</tr>
<tr>
<td>Million Veterans Program (<a href="http://www.research.va.gov/mvp/">http://www.research.va.gov/mvp/</a>)</td>
<td>Veterans Affairs Hospital</td>
<td>2011</td>
<td>&gt;500,000</td>
<td>~1,000,000</td>
</tr>
<tr>
<td><strong>Institutional funding</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BioVu Biorepository (<a href="https://victor.vanderbilt.edu/pub/biovu/">https://victor.vanderbilt.edu/pub/biovu/</a>)</td>
<td>Vanderbilt University Medical Center (Nashville, TN)</td>
<td>2007</td>
<td>&gt;215,000</td>
<td>Unknown</td>
</tr>
<tr>
<td>Partners Healthcare Biobank (<a href="http://biobank.partners.org/">http://biobank.partners.org/</a>)</td>
<td>Partners Health Care (Boston, MA)</td>
<td>2010</td>
<td>&gt;50,000</td>
<td>~100,000</td>
</tr>
</tbody>
</table>

Khera AV and Kathiresan S. *Nature Reviews Genetics*. 2017
RAPID GWAS OF THOUSANDS OF PHENOTYPES FOR 337,000 SAMPLES IN THE UK BIOBANK

September 28, 2017

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.
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 ($\beta$)

\[ Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \]
**CAD Polygenic Risk Score**

**LDpred method (>6 million alleles)**

**Graph a:**
- Odds ratio versus remainder of population
- > threefold (8.0%)
- > fourfold (2.3%)
- > fivefold (0.5%)

**Graph c:**
- Prevalence of CAD (%)
- Percentile of polygenic score

---

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**

<table>
<thead>
<tr>
<th>High GPS definition</th>
<th>Reference group</th>
<th>Odds ratio</th>
<th>95% CI</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CAD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 20% of distribution</td>
<td>Remaining 80%</td>
<td>2.55</td>
<td>2.43-2.67</td>
<td>&lt;1x10^-30</td>
</tr>
<tr>
<td>Top 10% of distribution</td>
<td>Remaining 90%</td>
<td>2.89</td>
<td>2.74-3.05</td>
<td>&lt;1x10^-30</td>
</tr>
<tr>
<td>Top 5% of distribution</td>
<td>Remaining 95%</td>
<td>3.34</td>
<td>3.12-3.58</td>
<td>6.5x10^-24</td>
</tr>
<tr>
<td>Top 1% of distribution</td>
<td>Remaining 99%</td>
<td>4.83</td>
<td>4.25-5.46</td>
<td>1.0x10^-12</td>
</tr>
<tr>
<td>Top 0.5% of distribution</td>
<td>Remaining 99.5%</td>
<td>5.17</td>
<td>4.34-6.12</td>
<td>7.9x10^-18</td>
</tr>
<tr>
<td><strong>Atrial fibrillation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 20% of distribution</td>
<td>Remaining 80%</td>
<td>2.43</td>
<td>2.29-2.59</td>
<td>2.1x10^-27</td>
</tr>
<tr>
<td>Top 10% of distribution</td>
<td>Remaining 90%</td>
<td>2.74</td>
<td>2.55-2.94</td>
<td>7.0x10^-28</td>
</tr>
<tr>
<td>Top 5% of distribution</td>
<td>Remaining 95%</td>
<td>3.22</td>
<td>2.95-3.51</td>
<td>1.1x10^-40</td>
</tr>
<tr>
<td>Top 1% of distribution</td>
<td>Remaining 99%</td>
<td>4.63</td>
<td>3.96-5.39</td>
<td>2.9x10^-28</td>
</tr>
<tr>
<td>Top 0.5% of distribution</td>
<td>Remaining 99.5%</td>
<td>5.23</td>
<td>4.24-6.39</td>
<td>3.5x10^-16</td>
</tr>
<tr>
<td><strong>Type 2 diabetes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 20% of distribution</td>
<td>Remaining 80%</td>
<td>2.33</td>
<td>2.20-2.46</td>
<td>3.1x10^-20</td>
</tr>
<tr>
<td>Top 10% of distribution</td>
<td>Remaining 90%</td>
<td>2.49</td>
<td>2.34-2.66</td>
<td>1.2x10^-17</td>
</tr>
<tr>
<td>Top 5% of distribution</td>
<td>Remaining 95%</td>
<td>2.75</td>
<td>2.53-2.98</td>
<td>1.7x10^-40</td>
</tr>
<tr>
<td>Top 1% of distribution</td>
<td>Remaining 99%</td>
<td>3.30</td>
<td>2.81-3.85</td>
<td>1.4x10^-30</td>
</tr>
<tr>
<td>Top 0.5% of distribution</td>
<td>Remaining 99.5%</td>
<td>3.48</td>
<td>2.79-4.29</td>
<td>4.3x10^-20</td>
</tr>
<tr>
<td><strong>Inflammatory bowel disease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 20% of distribution</td>
<td>Remaining 80%</td>
<td>2.19</td>
<td>2.03-2.36</td>
<td>7.7x10^-46</td>
</tr>
<tr>
<td>Top 10% of distribution</td>
<td>Remaining 90%</td>
<td>2.43</td>
<td>2.22-2.65</td>
<td>8.8x10^-46</td>
</tr>
<tr>
<td>Top 5% of distribution</td>
<td>Remaining 95%</td>
<td>2.66</td>
<td>2.38-2.96</td>
<td>3.0x10^-48</td>
</tr>
<tr>
<td>Top 1% of distribution</td>
<td>Remaining 99%</td>
<td>3.87</td>
<td>3.18-4.66</td>
<td>1.4x10^-30</td>
</tr>
<tr>
<td>Top 0.5% of distribution</td>
<td>Remaining 99.5%</td>
<td>4.81</td>
<td>3.74-6.08</td>
<td>9.0x10^-37</td>
</tr>
<tr>
<td><strong>Breast cancer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 20% of distribution</td>
<td>Remaining 80%</td>
<td>2.07</td>
<td>1.97-2.19</td>
<td>3.4x10^-09</td>
</tr>
<tr>
<td>Top 10% of distribution</td>
<td>Remaining 90%</td>
<td>2.32</td>
<td>2.18-2.48</td>
<td>2.3x10^-10</td>
</tr>
<tr>
<td>Top 5% of distribution</td>
<td>Remaining 95%</td>
<td>2.55</td>
<td>2.35-2.76</td>
<td>2.1x10^-12</td>
</tr>
<tr>
<td>Top 1% of distribution</td>
<td>Remaining 99%</td>
<td>3.36</td>
<td>2.88-3.91</td>
<td>1.3x10^-14</td>
</tr>
<tr>
<td>Top 0.5% of distribution</td>
<td>Remaining 99.5%</td>
<td>3.83</td>
<td>3.11-4.68</td>
<td>8.2x10^-18</td>
</tr>
</tbody>
</table>

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.
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

- Genetic Instrument
- Exposure
- Confounders
- Outcome

Causal relationships:
- Genetic Instrument → Exposure
- Exposure → Outcome

Non-causal relationships:
- Confounders → Exposure, Outcome

CAUSAL vs. NOT CAUSAL
Mendelian Randomization

SNPs associated with LDL → LDL → CAUSAL → CVD
Mendelian Randomization

SNPs associated with HDL

HDL

CVD

NOT CAUSAL

CAUSAL
Human Knockout Project

LETTER

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

Danish Saleheen1,2, Pradeep Natnaji3,4, Irina M. Armean4,5, Wei Zhao1, Asif Rasheed2, Sumeet A. Khetarpal6, Hong-Hee Won7, Konrad J. Karczewski1,5, Anne H. O’Donnell-Luria4,6, Kaitlin E. Samocha4,5, Benjamin Weisburd1,7, Namrata Gupta4, Mozammal Zaidi2, Maria Samuel1, Atif Imran7, Shahid Abbas8, Faisal Majeed9, Madiha Ishaq1, Saba Akhtar2, Kevin Trindade6, Megan Muckavage9, Nadeem Qamar10, Khan Shah Zaman11, Zia Yaqoob12, Tahir Saghir13, Syed Nadeem Hasan Rizvi10, Anis Memon14, Nadeem Hayyat Mallick15, Mohammad Ishaq16, Syed Zahed Rasheed17, Fazal-ur-Rehman Memon13, Khalid Mahmood18, Naveeduddin Ahmied19, Ron Do20,21, Ronald M. Krauss18, Daniel G. MacArthur14,22, Stacey Gabriel14, Eric S. Lander4, Mark J. Daly24, Philippe Frossard25, John Danesh16,26, Daniel J. Rader27,28 & Sekar Kathiresan18,28

- 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

Safety check for drug development
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

OUR FIELD HAS BEEN STRUGGLING WITH THIS PROBLEM FOR YEARS.

STRUGGLE NO MORE! I'M HERE TO SOLVE IT WITH ALGORITHMS!

SIX MONTHS LATER:
WOW, THIS PROBLEM IS REALLY HARD.
YOU DON'T SAY.

Source: Twitter @AndrewLBeam
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
Opportunity for Academic Health Centers

- 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 triple aim: care, health, and cost
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
Figure: Analogy Between a Mendelian Randomization Study and a Randomized Trial

Mendelian Randomization Study

- Eligible Population
  - SNP associated with LDL-C
    - Naturally Random Allocation of Alleles
  - Lower LDL-C Allele
    - Treatment Arm
  - Other Allele
    - Usual Care Arm
  - $\Delta$ LDL-C
  - Incident Major Cardiovascular Events

Randomized Controlled Trial

- Eligible Population
  - LDL-C Lowering Therapy
    - Random Allocation of Treatment
  - Treatment Arm
  - Usual Care Arm
  - $\Delta$ LDL-C
  - Incident Major Cardiovascular Events

Ferrence PA. ACC.org Expert Analysis
Integrating Clinical and Polygenic Risk Prediction

- Clinical risk
  - Cholesterol: per 40 mg/dl increase
  - Smoking: per 50 cigarettes/day
  - Systolic blood pressure: per 20 mmHg increase

- Polygenic risk
  - Low polygenic risk
  - Intermediate polygenic risk
  - High polygenic risk

- Combined risk
  - Clinical risk and high polygenic risk
  - Clinical risk and intermediate polygenic risk
  - Clinical risk and low polygenic risk
  - Clinical risk and unmeasured polygenic risk

- CAD relative risk
- CAD absolute risk

Torkamani A et al. Nature Reviews Genetics 2018
Timeline of Molecular Data

1957 Sanger sequencing invented

1958 First X-ray crystallography of protein

1963 Ultrasound commercially available

1969 "AI Winter" begins

1970 Perceptron invented

1971 Precursor to neural networks

1973 DNA structure discovered

1975 Sanger sequencing invented

1977 First DNA genome sequenced (viral)

1977 Protein Data Bank (PDB) launched

1980 MRI image first used in clinic

1982 Perceptron used for gene-finding

1982 Genbank database launched

1986 Backpropagation algorithm

1988 More "AI Winter"

1989 Convolutional neural networks invented

1992 DNA editing technology invented

1999 Recurrent networks used to predict protein contact maps

2001 Random forest invented

2001 Human genome sequenced

2001 Fast matrix multiply on commercial GPUs

2001 Develops into an enabling technology for deep learning.

2008 1000 Genomes project launches

2007 ChIP-seq invented

2012 CRISPR–Cas9 gene editing technology invented

2012 Deep learning wins Merck Molecular Activity Challenge

2012 Deep learning wins ImageNet challenge

2014 Generative adversarial networks (GANS) invented

2014 Deep learning improves splice prediction

2015 Human genome sequencing for $1000

2015 Deep learning improves protein binding prediction

2015 Deep learning boosts power of Alzheimer’s clinical trial by improved patient enrollment
“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

*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