Huge cohorts, genomics, and clinical data to personalize medicine

Josh Denny, MD MS
Professor of Biomedical Informatics and Medicine
Director, Center for Precision Medicine
Vanderbilt University Medical Center
8/31/2018
Expanding Observation... the Hubble Space Telescope

Hubble’s First Image

50% More Detail
Expanding Observation… the Hubble Space Telescope

After 10 Years of Improvements and Repairs…
Observation of Medical Ailments in History

1945 Yalta Conference
People have different disease risk and variable drug response

**Challenge:** Can we use “big data” to discover and predict health variability?
Since 2005, GWAS have found >69,000 SNPs associations with hundreds of diseases and traits...

Most studies enroll thousands of people and use forms, trailers to “phenotype” people at costs of up to thousands of dollars per person
Resources for EHR-based research at Vanderbilt

The Synthetic Derivative
A de-identified and continuously-updated image of the EHR: 2,583,461 subjects
Resources for EHR-based research at Vanderbilt

The Synthetic Derivative
A de-identified and continuously-updated image of the EHR: ~2.5 million subjects

BioVU
Subjects with DNA: ~250k

- Dense (GWAS-level) genotypes: ~60,000 (soon 100k)
- Exome chip data: ~36,000 (soon >110k)
EHR data are dense and efficient for discovery: Vanderbilt’s experience (BioVU)

<table>
<thead>
<tr>
<th>EHR Data from Vanderbilt Biobank</th>
<th>BioVU start</th>
<th>248,455</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanderbilt biobank enrollment</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Data Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD Codes</td>
<td>57 million</td>
</tr>
<tr>
<td>CPT Codes</td>
<td>79 million</td>
</tr>
<tr>
<td>Clinical Documentation</td>
<td>48 million</td>
</tr>
<tr>
<td>Radiology</td>
<td>3.4 million</td>
</tr>
<tr>
<td>ECGs</td>
<td>2.9 million</td>
</tr>
<tr>
<td>Laboratory</td>
<td>193 million</td>
</tr>
<tr>
<td>Medications</td>
<td>530 million</td>
</tr>
<tr>
<td>Vital Signs</td>
<td>40 million</td>
</tr>
</tbody>
</table>
Making text documents useful for research

Clinical notes, test reports, etc

Find biomedical concepts and qualifiers; create structured data

CC: SOB
HPI: Mr. Smith is a 65yo w/ h/o CHF, ... no dm2... on atenolol 50mg daily... Mother had RA.

Deidentify: remove HIPAA identifiers + ....

Billing codes

Customized classifiers (smoking status, etc)

Structured Output
DrugName: atenolol
Strength: 50 mg
Frequency: daily

Structured Output
certainty (positive, negated)
Who experienced it? (patient or family member?)

Medication extraction

Synthetic Derivative

chief_complaint:
Shortness of Breath

history_present_illness:
Congestive Heart Failure
Type 2 diabetes, negated

mother_medical_history:
rheumatoid arthritis
Finding a “simple” disease in the EHR: Who has hypertension?
Definition: SBP > 140 or DBP > 90

Doesn’t have hypertension

Has hypertension
Our “simple” example: **Hypertension**

Multiple components are better
(and blood pressure is the worst)

Teixeira, JAMIA 2016
eMERGE Goals:
- To perform genomic studies using the EHR
- To initiate implementation of genomic medicine
What we learned - Finding phenotypes in the EHR

- Billing codes: ICD9 & CPT
- Clinical Notes (NLP - natural language processing)
- Medications & NLP (ePrescribing)
- Labs & test results NLP

True cases
The phenotyping process

Identify phenotype of interest

Common phenotype
Case & control algorithm development and refinement
Many approaches:
• Boolean logic
• Machine learning
• Regression/score

Manual review; assess precision
≥95%
Deploy in cohort
Association tests

<95%
This can take many iterations

Rare phenotype
Simple algorithm to find possible cases (controls are easy)

Manual review/deploy

Requires access to notes, image report (images?) labs, etc

Association tests
Algorithms can be deployed across multiple EHRs

Analyses can be performed using extant data

Table 1. Evaluation of Primary Hypothyroidism Algorithm at the Five eMERGE Sites

<table>
<thead>
<tr>
<th>Site</th>
<th>Primary Phenotype</th>
<th>Total Genotyped Subjects</th>
<th>Primary Hypothyroidism</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cases</td>
</tr>
<tr>
<td>Group Health</td>
<td>dementia</td>
<td>2532</td>
<td>397</td>
</tr>
<tr>
<td>Marshfield</td>
<td>cataracts</td>
<td>4113</td>
<td>514</td>
</tr>
<tr>
<td>Mayo Clinic</td>
<td>peripheral arterial disease</td>
<td>3043</td>
<td>233</td>
</tr>
<tr>
<td>Northwestern</td>
<td>type 2 diabetes</td>
<td>1217</td>
<td>92</td>
</tr>
<tr>
<td>Vanderbilt</td>
<td>normal cardiac conduction</td>
<td>2712</td>
<td>81</td>
</tr>
<tr>
<td>All sites</td>
<td></td>
<td>13,617</td>
<td>1317</td>
</tr>
</tbody>
</table>

Genotype counts represent all subjects who were found by the hypothyroidism algorithms at each site and who were genotyped. Counts are limited to those classified as “white” in the electronic medical record of each site. PPV = positive predictive value.

* Average weighted for number of samples contributed to the total.
Finding Primary Hypothyroidism in the EHR

**Case medications**
- levothyroxine, synthroid, levoxyl, unitroid,
- armour thyroid, desicated thyroid, cytomeil, triostat, liothyronine,
- synthetic triiodothyronine, liotrix, thyrolar

**ICD-9 codes for hypothyroidism**
- 244, 244.8, 244.9, 245, 245.2, 245.8, 245.9

**Abnormal lab values**
- TSH > 5 OR FT4 < 0.5

**ICD-9 codes for secondary causes of hypothyroidism**
- 244.0, 244.1, 244.2, 244.3

**ICD-9 codes for post surgical or post radiation hypothyroidism**
- 193*, 242.0, 242.1, 242.2,
- 242.3, 242.9, 244.0, 244.1, 244.2, 244.3, 258*

**CPT codes for post radiation hypothyroidism**
- 77261, 77262, 77263, 77280, 77285,
- 77290, 77295, 77299, 77300, 77301, 77305, 77310, etc.

**Exclusion keywords**
- multiple endocrine neoplasia, MEN I, MEN II, thyroid cancer, thyroid carcinoma

**Pregnancy exclusion ICD 9 codes**
- Any pregnancy billing code or lab test if all Case Definition codes, labs, or medications fall within 6 months before pregnancy to one year after pregnancy.

**Exclusion keywords**
- optiray, radiocontrast, iodine, omnipaque, visipaque, hypaque, ioversol, diatrizoate, iodixanol, isovue, iopamidol, conray, iothalamate, renografin, sinografin, cystografin, conray, iodipamide

**Case Definition**
- All three conditions required:
  - ICD-9 code for hypothyroidism OR abnormal TSH/FT4
  - Thyroid replacement medication use
  - Require at least 2 instances of either medication or lab with at least 3 months between the first and last instance of medication and lab

**Case Exclusions**
- Exclude if the following information occurs at any time in the record:
  - Secondary causes of hypothyroidism
  - Post surgical or post radiation hypothyroidism
  - Other thyroid diseases
  - Thyroid altering medication

**Case Exclusions**
- Time dependent case exclusions:
  - Recent pregnancy TSH/FT4
  - Recent contrast exposure

**Thyroid-altering medications**
- Phenytoin, Dilantin, Infatabs, Dilantin Kapseals, Dilantin-125,
  - Phenytorsk, Amiodarone Pacerone, Cordarone, Lithium, Eskalith,
  - Lithobid, Methimazole, Tapazole, Northyx, Propylthiouracil, PTU
Sharing EHR algorithms: PheKB.org

>150 phenotypes, 45 public; versioning, data dictionaries & data validation, file sharing, etc.

<table>
<thead>
<tr>
<th>Type of Phenotype</th>
<th>Public (n = 44)</th>
<th>Non-public (n = 110)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-9 or -10 codes</td>
<td>39</td>
<td>73</td>
<td>73%</td>
</tr>
<tr>
<td>Medications</td>
<td>31</td>
<td>51</td>
<td>53%</td>
</tr>
<tr>
<td>CPT codes</td>
<td>23</td>
<td>44</td>
<td>44%</td>
</tr>
<tr>
<td>NLP</td>
<td>28</td>
<td>36</td>
<td>42%</td>
</tr>
<tr>
<td>Laboratory test results</td>
<td>21</td>
<td>37</td>
<td>38%</td>
</tr>
<tr>
<td>Vital signs</td>
<td>5</td>
<td>14</td>
<td>12%</td>
</tr>
</tbody>
</table>
EHRs for drug response:

Clopidogrel adverse events associated with *CYP2C19* status

clopidogrel (inactive) $\xrightarrow{CYP2C19}$ 2-oxoclopidogrel (active)

From clinical trials

- Carriers: 12.1%
- Non-carriers: 8.0%

$N=1459$, $P=0.01$

From the EHR

- Normal metabolizers
- Carriers

$N=807$, $P=0.005$

Mega et al., *NEJM* 2009

Delaney et al. *Clin Pharm Ther.* 2012
...and with this knowledge, we can improve outcomes

clopidogrel (inactive) $\xrightarrow{CYP2C19}$ 2-oxoclopidogrel (active)

...and implementation reduces adverse outcomes

Pulley et al, CPT 2012

Cavallari et al. JACC 2017
Using Machine Learning for Phenotyping – Rheumatoid arthritis

AUCs:
- All features: 0.97
- ICD: 0.95
- NLP: 0.90
- Medications: 0.83
Deep learning for Diabetic Retinopathy

Train a machine learning algorithm over >128k images

Gulshan et al. JAMA 2016
Phenome scanning (PheWAS) in the EHR

A phenotype

Dense genomic information

Associated genotypes

A genetic variant

The curated EHR-based phenome (codes, NLP, labs)

Associated phenotypes
GWAS of normal cardiac conduction

- Find individuals with normal cardiac conduction
- Find genetic variants associated with QRS duration
- Use NLP, billing codes, lab tests, medication records to get highly accurate phenotype

Hypothetical Record

```
<table>
<thead>
<tr>
<th>time</th>
<th>“Normal” ECG</th>
<th>Myocardial infarction</th>
<th>Atrial fibrillation</th>
</tr>
</thead>
<tbody>
<tr>
<td>No heart disease</td>
<td>No Na-blocking drugs No abnormal K, Ca, Mg</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
GWAS of QRS Duration

SCN5A/SCN10A

n=5,272

Ritchie et al., Circulation 2013
PheWAS of rs6795970 (SCN10A) (associated with longer QRS duration in “heart healthy” people)

N=13617 subjects

Ritchie et al., Circulation 2013
What happens in the “heart healthy” population?

Examined the n=5272 “heart healthy” population

Followed for development of atrial fibrillation based on genotype

Atrial fibrillation-free survival

Years since normal ECG (and no heart disease)

HR=1.49 per G allele
p=0.001
Replications of NHGRI GWAS associations via PheWAS

P-value for replication:
- All - 210/751: $2 \times 10^{-98}$
- Powered - 51/77: $3 \times 10^{-47}$
PheWAS across all HLA types
(n=37,270)

Karnes et al, Sci Trans Med, 2017
Genetics can identify new drug targets

The story of PCSK9

Nonsense mutations in PCSK9 result in very low LDL and protect against coronary disease

PCSK9 inhibitors are now on the market (alirocumab, evolocumab)

Cohen et al., NEJM 2006
Sabatine NEJM 2017
Using PheWAS to repurpose medications

Re-analysis of 3,144 SNPs previously studied by PheWAS

127 known drug-indications replicated

2,583 “new” associations with literature/clinical trial evidence

Table 2  Examples of drug-disease pairs identified from PheWAS data

<table>
<thead>
<tr>
<th>Statusa</th>
<th>Drug</th>
<th>Disease indication</th>
<th>PheWAS SNP</th>
<th>PheWAS P value</th>
<th>Associated gene</th>
<th>Medline citations/clinical trial citations (count)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known</td>
<td>Paclitaxel</td>
<td>Breast cancer</td>
<td>rs242557</td>
<td>0.041</td>
<td>MAPT</td>
<td>3,003/321</td>
</tr>
<tr>
<td>Known</td>
<td>Glyburide</td>
<td>Type II diabetes</td>
<td>rs2515629</td>
<td>0.00073</td>
<td>ABCA1</td>
<td>240/12</td>
</tr>
<tr>
<td>Strongly supported</td>
<td>Dexamethasone</td>
<td>Rheumatoid arthritis</td>
<td>rs4795067</td>
<td>0.050</td>
<td>NOS2</td>
<td>4,271/48</td>
</tr>
<tr>
<td>Strongly supported</td>
<td>Everolimus</td>
<td>Breast cancer</td>
<td>rs17036350</td>
<td>0.040</td>
<td>MTOR</td>
<td>221/46</td>
</tr>
<tr>
<td>Likely</td>
<td>Verapamil</td>
<td>Vaginal cancer</td>
<td>rs216013</td>
<td>0.011</td>
<td>CACNA1C</td>
<td>1,144/0</td>
</tr>
<tr>
<td>Likely</td>
<td>Chlorpromazine</td>
<td>Liver cancer</td>
<td>rs11214606</td>
<td>0.033</td>
<td>ARVCF</td>
<td>423/0</td>
</tr>
<tr>
<td>Novel</td>
<td>Porfimer</td>
<td>Hyper-cholesterolemia</td>
<td>rs6511720</td>
<td>2.5 × 10^{-6}</td>
<td>LDLR</td>
<td>0/0</td>
</tr>
<tr>
<td>Novel</td>
<td>Zidovudine</td>
<td>Diabetes</td>
<td>rs2736100</td>
<td>0.00029</td>
<td>TERT</td>
<td>0/0</td>
</tr>
</tbody>
</table>

aStrongly supported, some support in both the literature and clinical trial registry; likely, some support in either the literature or clinical trial registry; novel, no evidence in the databases.
Residual CHD risk on Statins

**Case definition:** Individuals on a statin with myocardial infarction or revascularization

**Discovery set:** 3,099 cases and 7,681 controls

**Replication set:** 160 cases and 1112 controls

**Alg. needs:** ICD, CPT, labs, text diagnoses, timing

- https://phekb.org/phenotype/170
- PPVs
  - Vanderbilt 96%
  - Marshfield 100%
  - Mount Sinai 96%

**Discovery OR** = 1.58

**OR, adjusted for ΔLDL** = 1.62

**OR, LDL<70 subset** = 2.43

**Replication OR** = 1.85, p=0.006

Wei et al. *Circulation* 2018
PheWAS of *LPA* locus

n=13,990

Adjusted for **Age, Sex**

Adjusted for **Age, Sex, LDL, and statin use**

Wei et al. *Circulation* 2018
Hypothesis: Mendelian genes influence complex diseases/traits

Marouli et al. Nature 2017

Variants affecting height

STC2 rs148833559, MAF = 0.1%

IHH rs142036701, MAF = 0.08%

CRISPLD2 rs148934412, MAF = 0.08%

AR rs137852591, MAF = 0.21%

80% power

Significant (discovery)

Significant after validation

Recessive variant
Knowledge-driven phenotypes in patterns

Catalog of Mendelian diseases
- CFTR
  - Cystic Fibrosis
- DGKE
  - Congenital bilateral absence of the vas deferens
  - Nephrotic Syndrome, type 7
- KIF1B
  - Charcot-Marie-Tooth disease, type 2A1
  - Pheochromocytoma

OMIM synopsis for disease
- Nephrotic Syndrome, type 7
  - Genitourinary
    - Hemolytic uremic syndrome
    - Acute kidney injury
    - Nephrotic syndrome
    - Proteinuria
    - Thickening of the glomerular basement membrane
    - Chronic kidney disease
  - Hematology
    - Hemolytic anemia
    - Thrombocytopenia

Map to phecodes & find in the EHR

Phenotype weights
- Hemolytic-uremic syndrome: 0.83
- Acute renal failure: 1.24
- Nephritis, renal sclerosis: 1.41
- Proteinuria: 0.85
- Chronic renal failure: 2.04
- Acquired hemolytic anemia: 1.06
- Thrombocytopenia: 0.00

Create a "Phenotype Risk Score (PheRS)"

PheWAS

phecodes representing OMIM traits of disease
Finding rare disease in common disease via disease patterns: 
**Phenotype Risk Scores (PheRS)**

For each record $i$, generate PheRS

$$\text{PheRS}_i = \sum_{j=1}^{k} \left\{ \begin{array}{ll} 1 & \text{phenotype } j \text{ present} \\ 0 & \text{phenotype } j \text{ absent} \end{array} \right\} \omega_j$$

Score for subject $i$

Add up terms for k phenotypes

0=phenotype $j$ absent

1=phenotype $j$ present

weight for phenotype $j$ derived from entire EHR

Repeat this for “all” Mendelian diseases

Bastarache et al, Science 2018
Clinical Synopsis of Cystic Fibrosis from OMIM

<table>
<thead>
<tr>
<th>System</th>
<th>Human Phenotype Ontology (HPO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growth</td>
<td>Failure to thrive</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>Cor pulmonale</td>
</tr>
<tr>
<td>Respiratory</td>
<td>Recurrent bronchopulmonary infections</td>
</tr>
<tr>
<td></td>
<td>Bronchiectasis</td>
</tr>
<tr>
<td></td>
<td>Asthma</td>
</tr>
<tr>
<td></td>
<td>Chronic lung disease</td>
</tr>
<tr>
<td></td>
<td>Recurrent pneumonia</td>
</tr>
<tr>
<td>Abdomen</td>
<td>Exocrine pancreatic insufficiency</td>
</tr>
<tr>
<td></td>
<td>Rectal prolapse</td>
</tr>
<tr>
<td></td>
<td>Biliary cirrhosis</td>
</tr>
<tr>
<td></td>
<td>Meconium ileus</td>
</tr>
<tr>
<td>Genitourinary</td>
<td>Male infertility</td>
</tr>
<tr>
<td>Metabolism/homoeostasis</td>
<td>Hypercalciuria</td>
</tr>
<tr>
<td></td>
<td>Dehydration</td>
</tr>
<tr>
<td></td>
<td>Elevated sweat chloride</td>
</tr>
</tbody>
</table>

Can we use the EHR to represent these?
## Clinical Synopsis of Cystic Fibrosis from OMIM

<table>
<thead>
<tr>
<th>System</th>
<th>Human Phenotype Ontology (HPO)</th>
<th>PheWAS code</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Growth</strong></td>
<td>Failure to thrive</td>
<td>264.2 Failure to thrive (childhood)</td>
</tr>
<tr>
<td><strong>Cardiovascular</strong></td>
<td>Cor pulmonale</td>
<td>415.1 Acute pulmonary heart disease</td>
</tr>
<tr>
<td><strong>Respiratory</strong></td>
<td>Recurrent bronchopulmonary infections</td>
<td>483 Acute bronchitis &amp; bronchiolitis</td>
</tr>
<tr>
<td></td>
<td>Bronchiectasis</td>
<td>496.3 Bronchiectasis</td>
</tr>
<tr>
<td></td>
<td>Asthma</td>
<td>495 Asthma</td>
</tr>
<tr>
<td></td>
<td>Chronic lung disease</td>
<td>496 Chronic airway obstruction</td>
</tr>
<tr>
<td></td>
<td>Recurrent pneumonia</td>
<td>480 Pneumonia</td>
</tr>
<tr>
<td><strong>Abdomen</strong></td>
<td>Exocrine pancreatic insufficiency</td>
<td>577 Diseases of pancreas</td>
</tr>
<tr>
<td></td>
<td>Rectal prolapse</td>
<td>565 Anal and rectal conditions</td>
</tr>
<tr>
<td></td>
<td>Biliary cirrhosis</td>
<td>571.6 Primary biliary cirrhosis</td>
</tr>
<tr>
<td></td>
<td>Meconium ileus</td>
<td>656.6 Perinatal disorders of digestive system</td>
</tr>
<tr>
<td><strong>Genitourinary</strong></td>
<td>Male infertility</td>
<td>609 Male infertility &amp; abnormal spermatozoa</td>
</tr>
<tr>
<td><strong>Metabolism/homeostasis</strong></td>
<td>Hypercalciuria</td>
<td>275.5 Disorders of calcium/phosphorus metab</td>
</tr>
<tr>
<td></td>
<td>Dehydration</td>
<td>276.5 Hypovolemia</td>
</tr>
<tr>
<td></td>
<td>Elevated sweat chloride</td>
<td>-</td>
</tr>
</tbody>
</table>
# Proof of concept: Cystic Fibrosis

<table>
<thead>
<tr>
<th>Age/Sex</th>
<th>CF cases</th>
<th>CF controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>18F</td>
<td>9.8</td>
<td>2.5</td>
</tr>
<tr>
<td>26M</td>
<td>4.4</td>
<td>0.7</td>
</tr>
<tr>
<td>29F</td>
<td>6.3</td>
<td>0.0</td>
</tr>
<tr>
<td>29M</td>
<td>7.8</td>
<td>0.7</td>
</tr>
</tbody>
</table>

**Chronic airway obstruction**

- **Pneumonia**
- **Diseases of pancreas**
- **Hypovolemia**
- **Acute upper respiratory infections**
- **Asthma**
- **Bronchiectasis**
- **Intestinal malabsorption**
- **Hepatomegaly**
- **Acute pulmonary heart disease**

**PRS**

- **Proof of concept:** Cystic Fibrosis
  
  Bastarache et al, Science 2018
Validating PRS on 6 clinically-diagnosed diseases in the EHR

Bastarache et al, Science 2018
Testing 6,188 rare variants (<1%) for 1200 Mendelian diseases in 21k people with Exome array genotyping

<table>
<thead>
<tr>
<th>Gene</th>
<th>Variant</th>
<th>HOM/HET</th>
<th>Associated Mendelian Disease</th>
<th>OMIM Inher.</th>
<th>Phenotype categories in PRS</th>
<th>Beta</th>
<th>P</th>
<th>ClinVar</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFTR</td>
<td>Gly542Ter</td>
<td>1/27</td>
<td>Cystic fibrosis</td>
<td>Rec</td>
<td></td>
<td>1.39</td>
<td>2.9×10⁻⁸</td>
<td>P</td>
</tr>
<tr>
<td>CHRNA4</td>
<td>Arg483Gln</td>
<td>1/21</td>
<td>Nocturnal frontal lobe epilepsy, 1</td>
<td>Dom</td>
<td></td>
<td>0.58</td>
<td>9.0×10⁻⁸</td>
<td>U</td>
</tr>
<tr>
<td>DGKE</td>
<td>Trp322Ter</td>
<td>1/14</td>
<td>Nephrotic syndrome, type 7</td>
<td>Rec</td>
<td></td>
<td>1.31</td>
<td>2.8×10⁻⁷</td>
<td>LP</td>
</tr>
<tr>
<td>SUOX</td>
<td>Arg76Ser</td>
<td>0/24</td>
<td>Sulfocysteinuria</td>
<td>Rec</td>
<td></td>
<td>0.82</td>
<td>1.7×10⁻⁶</td>
<td>U</td>
</tr>
<tr>
<td>CFTR</td>
<td>Arg553Ter</td>
<td>0/12</td>
<td>Cystic fibrosis</td>
<td>Rec</td>
<td></td>
<td>1.81</td>
<td>2.1×10⁻⁶</td>
<td>P</td>
</tr>
<tr>
<td>KIF1B</td>
<td>Thr674Ile</td>
<td>0/21</td>
<td>Charcot-Marie-Tooth disease, 2A1</td>
<td>Dom</td>
<td></td>
<td>0.79</td>
<td>5.3×10⁻⁵</td>
<td></td>
</tr>
<tr>
<td>VWF</td>
<td>Thr1951Ala</td>
<td>0/21</td>
<td>Von Willebrand disease</td>
<td>Rec/Dom</td>
<td></td>
<td>0.53</td>
<td>8.6×10⁻⁶</td>
<td></td>
</tr>
<tr>
<td>KIF1A</td>
<td>Ala993=</td>
<td>1/25</td>
<td>Spastic paraplegia-30</td>
<td>Rec</td>
<td></td>
<td>0.84</td>
<td>1.3×10⁻⁵</td>
<td>LB</td>
</tr>
<tr>
<td>F10</td>
<td>Arg291Gln</td>
<td>0/15</td>
<td>Factor X deficiency</td>
<td>Rec/Dom</td>
<td></td>
<td>0.62</td>
<td>1.9×10⁻⁵</td>
<td></td>
</tr>
<tr>
<td>HFE</td>
<td>Glu168Gln</td>
<td>0/40</td>
<td>Hemochromatosis</td>
<td>Rec</td>
<td></td>
<td>1.08</td>
<td>4.0×10⁻⁵</td>
<td>U</td>
</tr>
<tr>
<td>TG</td>
<td>Gly77Ser</td>
<td>0/69</td>
<td>Thyroid dyshormonogenesis</td>
<td>Rec</td>
<td></td>
<td>0.26</td>
<td>6.0×10⁻⁵</td>
<td></td>
</tr>
<tr>
<td>SH2B3</td>
<td>Glu395Lys</td>
<td>0/22</td>
<td>Familial erythrocytosis, 1</td>
<td>Dom</td>
<td></td>
<td>1.48</td>
<td>6.1×10⁻⁵</td>
<td></td>
</tr>
<tr>
<td>SPTBN2</td>
<td>Arg2370His</td>
<td>0/11</td>
<td>Spinocerebellar ataxia</td>
<td>Rec/Dom</td>
<td></td>
<td>0.75</td>
<td>9.0×10⁻⁵</td>
<td></td>
</tr>
<tr>
<td>FAN1</td>
<td>Arg507His</td>
<td>0/434</td>
<td>Interstitial nephritis, karyomegalic</td>
<td>Rec</td>
<td></td>
<td>0.15</td>
<td>9.9×10⁻⁵</td>
<td></td>
</tr>
<tr>
<td>PANK2</td>
<td>Gly521Arg</td>
<td>0/26</td>
<td>HARP syndrome</td>
<td>Rec</td>
<td></td>
<td>0.58</td>
<td>1.1×10⁻⁴</td>
<td>P</td>
</tr>
<tr>
<td>SH2B3</td>
<td>Glu395Lys</td>
<td>0/26</td>
<td>Essential thrombocythemia</td>
<td>Dom</td>
<td></td>
<td>0.33</td>
<td>1.4×10⁻⁴</td>
<td></td>
</tr>
<tr>
<td>AGXT</td>
<td>Ala295Thr</td>
<td>1/35</td>
<td>Primary hyperoxaluria, type I</td>
<td>Rec</td>
<td></td>
<td>0.82</td>
<td>1.7×10⁻⁴</td>
<td>U/LB</td>
</tr>
<tr>
<td>PLCG2</td>
<td>Ile251Leu</td>
<td>0/10</td>
<td>Familial cold autoinflammatory syn. 3</td>
<td>Dom</td>
<td></td>
<td>0.70</td>
<td>1.9×10⁻⁴</td>
<td></td>
</tr>
</tbody>
</table>
PheWAS of a CF variant

- log_{10}(p)

- methicillin-sensitive S. aureus
- mycoses
- Varicella infection
- drug-resistant infection
- disorders of metabolism, other
- nutritional marasmus
- other nutritional deficiency
- pervasive developmental disorders

PheRS for CF
- mrsa pneumonia
- pseudomonal pneumonia
- cystic fibrosis
- bronchopneumonia and lung abscess
- intestinal malabsorption
- bronchiectasis
- hemopysis
- fungal pneumonia
- diseases of pancreas
- abx poisoning
- urticaria

Hypovolemia

- acute pulmonary heart disease
- acute bronchitis and bronchiolitis
- primary biliary cirrhosis
- anal and rectal conditions
- male infertility and abnormal spermatozoa

References:
Bastarache et al, Science 2018
PheRS identified potentially pathogenic SNVs

4/69 with TG variants with thyroidectomies
4/40 with HFE variants with liver transplants
PheRS can help us interpret undiagnosed patients

Patient with undiagnosed diseases applies to UDN

Patient info reviewed by UDN team

EHR, site-visit & in person phenotyping, lab testing

Whole exome/genome sequencing

Multidisciplinary team analysis for diagnosis
PheRS Application: Novel variant interpretation for the UDN

A UDN patient...

Clinical symptoms and physical findings

GROWTH PARAMETERS
- Failure to thrive

CARDIOVASCULAR
- Patent ductus arteriosus

GASTROINTESTINAL
- Elevated hepatic transaminase
- Gastroesophageal reflux

GENITOURINARY
- Hydrocele testis

BEHAVIOR, COGNITION AND DEVELOPMENT
- Global developmental delay
- Delayed speech and language development

DIGESTIVE SYSTEM
- Hepatomegaly

METABOLISM/HOMEOSTASIS
- Recurrent hypoglycemia
- Neonatal hypoglycemia

...with many candidate genetic variants

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chr Position rs#</th>
<th>Change</th>
<th>Effect</th>
<th>Proband</th>
<th>Mother (Unaff)</th>
<th>Father (Unaff)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COL9A1 NM_001851.4</td>
<td>chr6</td>
<td>A → T</td>
<td>splice donor</td>
<td>10.9→2.7</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td>70991091</td>
<td>c.876+2T&gt;A</td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td>rs149830493</td>
<td></td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>ELN NM_000501</td>
<td>chr7</td>
<td>G → A</td>
<td>missense</td>
<td></td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td>73470684</td>
<td>c.1234G&gt;A</td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td>rs375116795</td>
<td>p.Gly412Arg</td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>PIGN NM_012327</td>
<td>chr18</td>
<td>T → C</td>
<td>missense</td>
<td></td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td>59757754</td>
<td>c.2238A&gt;G</td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>POLG NM_002693.2</td>
<td>chr15</td>
<td>G → C</td>
<td>missense</td>
<td></td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td>89872002</td>
<td>c.1084C&gt;G</td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td>rs763248358</td>
<td>p.Leu362Val</td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>RFT1 NM_052859.3</td>
<td>chr3</td>
<td>C → T</td>
<td>missense</td>
<td></td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td>53140879</td>
<td>c.782G&gt;A</td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td>rs374781452</td>
<td>p.Arg261Gln</td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>
...and potentially we could diagnose individuals sooner

Bastarache et al, Science 2018
Finding solutions on a national scale: the All of Us Research Program

Funding:
$130M in FY2016
$230M in FY2017

21st Century Cures Act provides additional $1.5B over 10 years

“I want the country that eliminated polio and mapped the human genome to lead a new era of medicine…”

- PRESIDENT BARACK OBAMA

State of the Union Address, Jan. 20, 2015
Factors of Risk in the Development of Coronary Heart Disease—Six-Year Follow-up Experience

The Framingham Study

WILLIAM B. KANNEL, M.D., THOMAS R. DAWBER, M.D., F.A.C.P., ABRAHAM KAGAN, M.D., F.A.C.P., NICHOLAS REVOTSIE, M.D., AND JOSEPH STORES, III, M.D.
Framingham, Massachusetts

Increasingly reliable estimates of the prevalence and incidence of coronary heart disease (CHD) emphasize the importance of this disease as a contemporary health hazard. Cardiovascular disease is now the leading cause of death, with coronary heart disease accounting for two-thirds of all heart disease deaths. While advances in the diagnosis and therapeutic manage-

Since it has been established that coronary atherosclerosis is present for many years prior to the development of symptomatic CHD, it seems evident that efforts at prevention must begin many years before the appearance of clinical CHD. A knowledge of the epidemiology of the disease is highly desirable if a program of prevention is to be developed. From a study of the character-

Enrolled 5209 men and women in 1948

Some discoveries:
• 1960 – Cigarettes increase heart disease
• 1961 – cholesterol, blood pressure increase heart disease
• 1967 – exercise decreases risk of heart disease; obesity increases it
• 1970 – high blood pressure and atrial fibrillation cause stroke

The impact of Framingham (and similar cohorts) has been dramatic

https://www.cdc.gov/Mmwr/preview/mmwrhtml/mm4830a1.htm
All of Us Research Program - Mission and Objectives

Nurture relationships
with one million or more participant partners, from all walks of life, for decades

Our mission
To accelerate health research and medical breakthroughs, enabling individualized prevention, treatment, and care for all of us

Deliver the largest, richest biomedical dataset ever that is easy, safe, and free to access

Catalyze a robust ecosystem of researchers and funders hungry to use and support it
Why Diversity?

**PERSISTENT BIAS**

Over the past seven years, the proportion of participants in genome-wide association studies (GWAS) that are of Asian ancestry has increased. Groups of other ancestries continue to be very poorly represented.

- **2009**: 373 studies, 1.7 million samples
  - 96% European ancestry
  - 4% Non-European ancestry

- **2016**: 2,511 studies, 35 million samples
  - 81% European ancestry
  - 19% Non-European ancestry

4% GWAS represents >33% US population

All of Us Research Program – Summary of Protocol and Status

Key elements:
- 1 million or more engaged participants
- Participants get data back (genetics, EHR)
- Longitudinal, recontactable

Key dates:
- Network launched **July 2016**
- Protocol developed with 50-state pilot testers, 76 engagement studios in 17 cities
- Beta testing began **May 31, 2017**
- National launch **May 6, 2018**

Current stats:
- >100k have signed up
- >54k have completed all steps
- >200 recruitment sites, web sites, mobile van, recruitment materials and videos, initial EHR uploads
Major Building Blocks of *All of Us*

**DATA AND RESEARCH CENTER (DRC)**
Big data capture, cleaning, curation, & sharing in secure environment
*Vanderbilt, Verily, Broad Institute*

**BIOBANK**
Repository for processing, storing, & sharing biosamples
*Mayo Clinic*

**PARTICIPANT CENTER**
Direct volunteer participant enrollment, digital engagement innovation, & consumer health technologies
*Scripps Research Institute (with multiple partners)*

**PARTICIPANT TECHNOLOGY SYSTEMS CENTER**
Web & phone-based platforms for participants
*Vibrent Health*

**HEALTH CARE PROVIDER ORGS (HPOs)**
Clinical & scientific expertise network, enrollment & retention of participants
*20+ regional med centers, FQHCs, VA, future awards to grow network*

**COMMUNICATIONS & ENGAGEMENT**
Comms, marketing, & design expertise; Engagement coordination & community partners network
*Wondros, HCM, future awards to grow network of community partners*
Current Consortium Members

### Data and Research Center
- Vanderbilt University Medical Center
- Broad Institute
- verily

### Other Platform Development
- Scripps Translational Science Institute
- Sage
- WONDROS

### Communication & Engagement
- WONDROS
- HSM
- BlueCross BlueShield

### DV Network (Direct Volunteers)
- Walgreens
- patientslikeme
- WebMD

### HPO Network (Health Care Provider Organizations)

#### RMCs
- California Precision Medicine Consortium
- Illinois Precision Medicine Consortium
- New England Precision Medicine Consortium
- Trans-American Consortium for the Health Care Systems Research Network
- New York City Precision Medicine Consortium
- Southern All of Us Network
- SouthEast Enrollment Center

#### FQHCs (Federally Qualified Health Centers)
- UC San Diego Health
- Keck Medical Center of USC
- UC Irvine Health
- UCHealth
- UC Davis Health
- Rush University Medical Center
- NorthShore University Health System
- Partners Healthcare
- Massachusetts General Hospital
- Brigham and Women’s Hospital
- Beth Israel Deaconess Medical Center
- Baystate Health
- Beth Israel Deaconess Medical Center
- Columbia University Medical Center
- NewYork-Presbyterian
- University of Chicago Medicine
- University of Illinois Hospital
- Spectrum Health
- Weill Cornell Medicine
- University of Kentucky Medical Center
- University of Louisville Medical Center
- University of Texas Medical Branch
- University of Texas Health Science Center
- University of Virginia Health System
- University of Washington Medical Center
- UAB Medicine
- VA Medical Centers

#### VA Medical Centers
- VA
- U.S. Department of Veterans Affairs

#### Other Platform Development
- DXC
- Quest Diagnostics
- Emsi Health
- QTC
- A Galileo Company
Current and planned in-person enrollment centers
Summary of our Current Protocol

Enroll, Consent & EHR
- Recruit 18+ years old initially
- Plan to include children in next iteration
- eConsent, paper coming
- EHRs collected periodically

Surveys
- Current modules: The Basics, Overall Health, Lifestyle, Family history, Healthcare access
- More in development (Meds, Diet, Personal Medical History)
- Linked to EHR data elements

Physical Measurements
- Blood pressure
- Heart rate
- Weight
- Height
- BMI
- Hip circumference
- Waist circumference

Biosamples
- Blood (or saliva, if blood draw is unsuccessful)
- Urine

Full protocol published at allofus.nih.gov. Planning for new releases every 2.5 to 3.5 years.
PPI/Survey Modules

Enrollment Surveys
1. The Basics
2. Overall Health
3. Lifestyle

Later:
4. Personal Health History
5. Family History

In development:
6. Medications
7. Health Care Access and Utilization
8. Sleep
9. Environment/exposures

Elements being coordinated with EHR analogs in our Common Data Model
All of Us will aggregate data from many sources

From Healthcare Provider Orgs

- Version 1 (2018)
  - Visits
  - Billing codes
  - Meds
  - Labs

- Version 2
  - Clinical Notes & Reports
  - Much longer term
  - Local Registries
  - Images

Data added centrally by DRC

- Death Index
- Claims & Rx
- GIS
- ...?

From Direct Volunteers

Sync for Science

- Participant provided info
  (Health surveys – linked to EHR)
- Digital health tech.
  (activity monitors, BP watches, etc)
- Physical measures and biospecimens

Raw Data Repository

Curated Data Repository

APIs, Analysis tools, etc
Sync 4 Science (S4S) – a new technology to share health data

- FHIR-based
- Starting with MU Common Clinical Data set
- Notes in the future?
Data Access is centralized in *All of Us*

**Traditional Approach:** Bring data to researchers

**Problems**
- Data sharing = data copying
- Security (data handoffs)
- Huge infrastructure needed
- Siloed compute

**AoU Approach:** Bring researchers to the data

**Advantages**
- Cost
- Threat detection and auditing
- Increased Accessibility
- Shared compute
1. **Public:** Data that poses minimal risks to the privacy of research participants. It can be accessed without logging into the All of Us Research Platform. *(e.g. aggregate statistics)*

1. **Registered:** Data that has some risk to the privacy of research participants. It can only be accessed after logging into the *All of Us* Research Platform; all access will be logged and may be audited. *(e.g. PPI responses, EHR structured data)*

1. **Controlled:** Data that poses the most significant risks to the privacy of research participants; researchers must be approved by the RAB to access it. *(e.g. EHR clinical notes)*
Data Access Protocol (DRAFT)

Public

- no login required
- Access Data

Registered

1. Registration and Identity Verification
2. Research Ethics Training
3. Sign Code of Conduct
4. RAB Approval
5. Create Project and State Purpose
6. Access Data
7. Names and Projects posted on a public website

Controlled

1. Registration and Identity Verification
2. Research Ethics Training
3. Sign Code of Conduct
4. eRA Commons ID
5. RAB Approval
6. Create Project and State Purpose
7. Access Data

automated & instantaneous
Research Hub Public Site

http://researchallofus.org

There are thousands of research questions. Let’s find some answers.

The All of Us Research Program is building one of the largest biomedical resources of its kind to explore how lifestyle, environment, and biological makeup affect health and disease. When it’s available, researchers will be able to use the diverse data here to explore a wide range of biomedical and scientific hypotheses.

A unique participant community...

The All of Us Research Program aims to engage a community of one million or more volunteers who reflect the diversity of America, including many people who haven’t taken part in medical research before. We welcome participants both healthy and sick, of all backgrounds and walks of life, from all regions across the country.

Participants answer surveys about their health, lifestyle, and environment. Some participants will also contribute physical measurements, biosamples (blood and urine), their Electronic Health Record (EHR), and more. Taken together, these data give the program the scale and scope to enable research for most common diseases, as well as many rare conditions, and may help increase our understanding of healthy states.

Learn more about the All of Us Research Program protocol >

Sign up to receive updates

Learn About Program Data, Tools, & Access
Building tools to facilitate research (tools in development)

Section 1: Notebooks in a Biomedical context

Why are notebooks used in biomedical research?

Try to answer in the cell below by:
1. Clicking on the cell
2. Change the dropdown box to the top of the screen that says "Code" to "Markdown"
3. When you are done providing an answer click the play button above or use the shortcut (Shift+Enter)

In []:

```
Good work - by typing in your answer and executing the cell, you just used a notebook! You can see how Markdown can be used in a cell to simply type and transform text. Now we will tell you the answer:

**Answer:**
In progress [TM]. Notebooks make it easy to record and reproduce data analysis steps. (Add more detail and include link to Helen Shen paper: https://www.nature.com/ncomms/interactive-notebooks-sharing-the-code-1-16261 and https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3736611) Notebooks can be used for any industry where data analysis is required... Great example notebook: http://nbviewer.jupyter.org/github/masayamab/Zika-RNAseq-Pipeline/blob/master/Zika.pdf
```

What is the relationship between the notebook and the workspace?
Return of Information to Participants

Participants may receive, depending on their preferences:

- Individual health information
- Survey data (comparative)
- EHR data, claims data
- Research results
- Ongoing study updates
- Aggregated results
- Scientific findings
- Opportunities to be contacted for other research opportunities
Participants will learn of genetic sequence results.

3.5% of all tested had an actionable result. 86% were new diagnoses.
Frequency of actionable genotypes in the first 10,000 Vanderbilt PGx patients

- At least one high risk variant
- At least one actionable variant
- No actionable variants

91%

Van Driest, CPT 2014
National launch: Sunday, May 6!

THE NIH LAUNCHES ITS AMBITION MILLION-PERSON GENETIC STUDY.

Pay it forward: Join with All of Us Research Program to build a healthier future.

By signing up for All of Us, you will help accelerate the growth of precision medicine. We will be able to deliver better health for every American.

In 1948, more than 5,000 people in the little town of Framingham, Mass., volunteered for a study to find answers about the mysterious, growing epidemic of heart disease. Every two years for decades, they had a physical exam, gave blood and urine samples, and answered questions about their health. Their children and grandchildren joined, too. Because of them, we now know the big risk factors for cardiovascular disease and have saved millions of lives through new prevention strategies, drugs, procedures and education.

The Framingham study taught us something important: Driving the future of medicine requires the generosity of volunteers who, in the service of science and generations to come, choose to "pay it forward" and help provide the data today that produce cures tomorrow. The Department of Health and Human Services is devoted to improving the quality of health care for every American, and today, we have tools to advance that work.
Anyone can signup now! http://joinallofus.org

The future of health begins with you.

The All of Us Research Program has a simple mission. We want to speed up health research breakthroughs. To do this, we’re asking one million people to share health information. In the future, researchers can use this to conduct thousands of health studies.

JOIN NOW

WATCH INTRODUCTION
May 31, 2017: Launched Beta phase

- Version 1 protocol developed, tested, & IRB approved
- Completed security plan/tests & Authority to Operate
- Completed site visits then end-to-end “dress rehearsals”, then alpha, then beta tests at every site
- Call center & command center up & running
- Direct Volunteer capability & HPO network established
- Launched beta phase in 19 awardees, >120 locations
- Kicked off mobile exhibit, the All of Us Journey
- In development:
  - Genomics plan (GWAS & WGS starting 2019)
  - Pediatric enrollment (~ 1 year)
  - Researcher tools

Today… >54,000 full participants and >100k have registered
The paradox of personalized medicine: we will need huge populations to best understand the care for the individual.
The power of a *data biosphere* of common semantics and APIs

http://databiosphere.org

© 2009 www.outline-world-map.com
Data and Research Center Team
A small sampling of the VUMC Team +many others!