Envisioning the future: The future of data platforms for human genetics

Josh Denny, MD MS
11/28/2018
The All of Us Research Program –
Breaking Down Data Silos

Precision Medicine Initiative, PMI, All of Us, the All of Us logo, and The Future of Health Begins With You are service marks of the U.S. Department of Health and Human Services.
All of Us Research Program – Summary of Protocol and Status

Key elements:

- **Goal**: 1 million or more diverse participants
- National launch May 6, 2018
- **Current**: ~83k "full" participants, ~140k have started, >45% nonwhite, >75% underrepresented in research
- **Participants will get data back** (genetics, EHR)
- Longitudinal, **recontactable**
Current and planned in-person enrollment centers

![Map of the United States showing the locations of current and planned in-person enrollment centers, with symbols indicating different enrollment centers such as HPOs, QTC/Liedos, Quest, EMSI, Walgreens Clinics, and National Blood Collaborative.]
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
- Clinical Messaging

Much longer term
- Local Registries
- Images

Data added centrally by DRC
- Death Index
- Claims & Rx Data
- ...

From Direct Volunteers

Sync for Science

Participant provided data
- (Health surveys, activity monitors, etc)

Health data aggregators
- (PicnicHealth)

Participant exams and biospecimens

Geospatial data

Curated Data Repository
- APIs, Analysis tools, etc

Raw Data Repository
Putting the patient in charge of data mobility - Sync 4 Science (S4S)

Sync 4 Science:
- FHIR-based
- Starting with Meaningful Use Common Clinical Data set
- Piloting now

…or on your smartphone
## Scientific priorities for *All of Us*

<table>
<thead>
<tr>
<th>Scientific Priority</th>
<th>End of 2018</th>
<th>End of 2019</th>
<th>2020-2022</th>
<th>2023-2027</th>
<th>&gt;2027</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected cohort size</td>
<td>90,000</td>
<td>&gt;200,000</td>
<td>&lt;650,000</td>
<td>&gt;1,000,000</td>
<td>&gt;1,000,000</td>
</tr>
<tr>
<td>Returning data to participants</td>
<td>+</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>++++</td>
</tr>
<tr>
<td>Discover disease risk factors across genes and environment</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>++++</td>
</tr>
<tr>
<td>Better predict therapeutic safety and efficacy</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Discover disease biomarkers</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Connect mHealth with clinical outcomes</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Evaluate genetic variants</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Develop new disease classifications</td>
<td>+</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>++++</td>
</tr>
<tr>
<td>Support clinical trials</td>
<td>+</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Enable machine learning applications</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>++++</td>
</tr>
<tr>
<td>Better understand health disparities</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Enable new therapeutics</td>
<td>++</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
UK Biobank in a nutshell: an open access prospective study with large size and extensive breadth and depth of data

- 500,000 UK men and women aged 40-69 years when recruited during 2006-2010
- Consent for all types of health research by both academic and commercial researchers, and for long term follow-up through all health-related records
- Extensive baseline questions and measurements, with biological samples stored for future assays
- Open access for approved research for the benefit of the public’s health: see www.ukbiobank.ac.uk

Courtesy of Dr. Cathie Sudlow
Assays

• Complete:
  – GWAS
  – Standard panel of biochemical assays (e.g. lipids; hormones; metabolic)
• Underway:
  – Whole exome sequencing
  – Whole genome sequencing
  – Infectious disease assays
  – Nightingale metabolomics (lipidomics)
• Other:
  – Whole-body Imaging (100k)
  – Activity monitoring (100k)

• Planned to start in 2019:
  – GWAS
  – Whole genome sequencing
  – Assay pilot (a few clinically relevant labs)
• Other planned
  – 10k FitBit pilot
  – “BYOD” data donation for FitBit, Apple Watch, etc
  – GIS data linkages
• Many more under consideration
WE'VE MAPPED THE WORLD. NOW LET'S MAP HUMAN HEALTH.

Project Baseline
- Verily, Duke, Stanford, Google
- ~10,000 people
- 1-2d annual visits, quarterly assessments
- Sensors and wearables
- Surveys
- Biospecimens
Passive data follow-up via EHRs provides dense resource for efficient discovery: Vanderbilt DNA Biobank example
EHRs, surveys, wearables, etc. linked to DNA

Phenotyping (codes, labs, NLP, etc)

GWAS identifying new disease targets

Pharmacogenomic discovery

PheWAS

AI, clustering

potential ADEs

potential indications

New disease clusters, subtypes, and learned phenotypes
Genetics of rheumatoid arthritis contributes to biology and drug discovery
Okada et al., *Nature* 2014

- >100,000 cases and controls from RA studies and biobanks, multiple ethnicities
- 101 RA loci → 98 candidate genes
• Encode known targets for RA drugs, and drugs used in other diseases
• ~2/3 display associations with diseases beyond RA

Genetics of rheumatoid arthritis contributes to biology and drug discovery
Okada et al., *Nature* 2014
Finding drug response in cohorts: Clopidogrel adverse events associated with \textit{CYP2C19} status

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

\begin{itemize}
  \item From clinical trials:
    \begin{itemize}
      \item Carriers: 12.1%
      \item Non-carriers: 8.0%
      \item N=1459, P=0.01
    \end{itemize}

  \item From the EHR:
    \begin{itemize}
      \item Carriers: N=807, P=0.005
      \item Normal metabolizers
    \end{itemize}
\end{itemize}

Mega et al., \textit{NEJM} 2009

Delaney et al. \textit{Clin Pharm Ther.} 2012
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
What do cohorts tell us about potential risk of PCSK9 inhibitors? (a Mendelian randomization study)

Summary: how large cohorts may accelerate clinical trials

- Basic discovery
  - E.g., new genetic loci predicting cholesterol levels (PCSK9) or drug effects (CYP2C19 and clopidogrel efficacy)

- Direct recruitment and facilitate assessment via extant infrastructure
  - Some cohorts are **recontactable**
  - More cost effective recruitment
  - More cost effective follow-up (EHRs, basic assessment infrastructure and extant molecular data, digital devices)
  - Faster recruitment of the best targeted clinical trials (e.g., specific CFTR mutations)

- More intelligent trial design
  - Use large biobanks to generate hypotheses of possible indications and adverse effects for trials
  - Pursue the hypotheses that appear most promising, search for side effects most suggested
The cohort landscape is expanding...

- **eMERGE**
  - EHR: ~0.5M

- **pcornet**
  - EHR: >100M

- **23&me**
  - EHR: ~6M
  - >500,000
  - 52,000

- **biobankuk**
  - EHR, survey: >600k (1M)
  - >500,000

- **estonian genome center**

- **KAIser PERMANente.**
  - EHR: ~250k

- **MVP**
  - EHR, survey: >600k (1M)
  - 10,000
  - very densely phenotyped

- **CHINA KADODRE BIOBANK**
  - >500,000

- **All of Us**
  - RESEARCH PROGRAM
  - 83k (1M)

- **Discovery**
  - Recontactable for CTs
All of Us Data and Research Center Team

First All of Us Network meeting June 2016
Data from portable wearable devices

Accelerometry data: 100,000 participants

Continuous ECG monitoring: 20,000 + participants

Prospective design and large size enable well-powered studies of (causal) associations between accelerometry and cardiac rhythm measures and later onset disease.

Need scalable methods of analysing complex data to derive measures for large scale analyses.

Courtesy of Dr. Cathie Sudlow
Prospective design and large size enable well-powered studies of (causal) associations between structure and function of organs and later onset disease. Need scalable methods of analysing complex data to derive measures for large scale analyses.