Data and Research Center (DRC)
Committee on Access, Privacy and Security (CAPS)

All of Us Data Access and Data Privacy

Dikshya Bastakoty, PhD & Weiyi Xia, PhD
Vanderbilt University Medical Center
June 24, 2019
ELSI Workshop, NIH
Outline

● Overview of the Data types and tiers of access
● Registered Tier Access
● Data Privacy in the Registered tier and beyond
● Coming soon: New Data & New Tiers
All of Us Principles of Access and Privacy

**Broad and Open Access**
- A data resource for All of Us (Academic Researchers AND Community Scientists)
- Removal of unnecessary barriers to access

**Commitment to Protecting Privacy**
- Privacy first, Privacy always
- Strive to build a system that Participants can trust
All of Us Research Program Data

Data Collection
- Participant Surveys
- Electronic Health Records
- Physical Measurements
- Mobile/Wearable Tech
- Biospecimens

Data Curation
- Data Harmonization
- QA/QC
- Data Refinement → Privacy Methodology

Curated Data Repository
Tiers of Access

Controlled Tier
- Amended Participant-Level Data
- Other Data Types

Registered Tier
- Amended Participant Level Data

Public Tier
- Summary Statistics
- Aggregate Counts

- Registered Tier Requirements
- Institutional Sign-off
- Other Requirements

- Identity Verification
- Research Ethics Training
- Data Use Agreement

Publicly Accessible (No login required)

Researcher Workbench
Data Browser
Access to the Registered Tier

Verification of Identity
Each user will be required to undergo a process (eg: electronic ID verification) that will be used to verify the identity of the user.

Data Passport Registration
The User provides information about themselves to the AoURP. Fields include:
- Name
- Institutional Affiliation
- Role
- Demographics
- Additional affiliation

Name, Affiliation & Role displayed publicly

Research Ethics Training
Research Ethics Training, which includes a number of training modules, educating the user about the ethics of Research using Human Data, the AoU principles policies, general data safety guidelines etc.

Attestation to the Data Use Agreement
The user is required to sign a legally binding document that specifies the Codes of Conduct that the User agrees to follow.

Workspace Creation & Specification of Research Purpose
For each project for which AoU data is accessed, the user is required to create a separate Workspace, and therein, specify the research purpose for that project.

Research Purpose Description displayed publicly
Safeguards to Protect the Data

Platform/tools
- Users come to the cloud-based data platform; data does not leave the platform

Oversight of Data Use
- Any member of the public can request review if concerned about stigmatization of research participants or any other violation of the program’s Code of Conduct.
- Resource Access Board audits and monitors data usage

Data
- Data tiers and Privacy Methodology for Identity Protection
Data Privacy
Registered Tier and Beyond…
Participant Level Data and the Data Users

Survey 119
Physical Measurements 398
Medical diagnosis 85
Drug 97
Medical procedure 55
Other data from EHR 17
Hospital visits 64
Genomic data
Mobile health
...

Re-identification Attack Example

Latanya Sweeney

William Weld

Group Insurance Commission (GIC)

ZIP
Date of Birth
Sex
Ethnicity
Visit date
Diagnosis
Procedure
Medication
Total charge

Latanya Sweeney, k-anonymity: a Model for Protecting Privacy, International Journal on Uncertainty, 2002
Register Tier Protection Basic Rules Based on Standard Practice

1. Remove explicit identifiers
2. Remove free text
3. Shift dates backwards by a random number between 1 to 365 (shift is constant for each record so temporality of events is preserved)
4. Remove geolocation information smaller than a US state (include paired All of Us site, provider location, address...)
5. Remove participants aged > 89
Gross Re-identification Risk Assessment of Each Cluster of Data

- Survey
- Physical measurements
- Medical diagnosis
- Drug
- Medical procedure
- Other data from EHR
- Hospital visits
- Genomic data
- Mobile health
- 

<table>
<thead>
<tr>
<th>Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical measurements</td>
</tr>
<tr>
<td>Medical diagnosis</td>
</tr>
<tr>
<td>Drug</td>
</tr>
<tr>
<td>Medical procedure</td>
</tr>
<tr>
<td>Other data from EHR</td>
</tr>
<tr>
<td>Hospital visits</td>
</tr>
<tr>
<td>Genomic data</td>
</tr>
<tr>
<td>Mobile health</td>
</tr>
</tbody>
</table>

| Diastolic blood pressure      |
| Body temperature             |
| Body height                  |
| Alanine [Moles/volume] in Urine |
| West Nile virus IgM Ab [Presence] in Cerebral spinal fluid |
| Body weight                  |
| Globulin [Mass/volume] in Serum |

...
Gross Re-identification Risk Assessment of Each Cluster of Data

- Survey
- Physical measurements
- Medical diagnosis
- Drug
- Medical procedure
- Other data from EHR
- Hospital visits
- Genomic data
- Mobile health

<table>
<thead>
<tr>
<th>Drug</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sodium Chloride</td>
<td>241390</td>
</tr>
<tr>
<td>Acetaminophen</td>
<td>207950</td>
</tr>
<tr>
<td>heparin</td>
<td>206891</td>
</tr>
<tr>
<td>Albuterol</td>
<td>206390</td>
</tr>
<tr>
<td>Glucose</td>
<td>173920</td>
</tr>
<tr>
<td>Oxycodone</td>
<td>158534</td>
</tr>
<tr>
<td>POLYETHYLENE GLYCOL 3350</td>
<td>152404</td>
</tr>
</tbody>
</table>
Gross Re-identification Risk Assessment of Each Cluster of Data

- Survey
- Physical measurements
- Medical diagnosis
- Drug
- Medical procedure
- Other data from EHR
- Hospital visits
- Genomic data
- Mobile health
- ...
Gross Re-identification Risk Assessment of Each Cluster of Data

Survey
Physical measurements
Medical diagnosis
Drug
Medical procedure
Other data from EHR
Hospital visits
Genomic data
Mobile health
...

- Pregnancy detection examination
- Radiologic examination, hand
- Intersex surgery male to female
- Lymphatics and lymph nodes imaging
- Ultrasonography of digestive system
Gross Re-identification Risk Assessment of Each Cluster of Data

Survey
Physical measurements
Medical diagnosis
Drug
Medical procedure
Other data from EHR
Hospital visits
Genomic data
Mobile health

<table>
<thead>
<tr>
<th>Ego-dystonic sexual orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death cause</td>
</tr>
<tr>
<td>Toxic effect of gas, fumes AND/OR vapors</td>
</tr>
<tr>
<td>Retinal hemorrhage, left eye</td>
</tr>
<tr>
<td>Terrorism</td>
</tr>
<tr>
<td>Vehicle accident</td>
</tr>
<tr>
<td>Operations of War, Military operations</td>
</tr>
<tr>
<td>Prolonged stay in weightlessness</td>
</tr>
<tr>
<td>Assault/Homicide, Suicide</td>
</tr>
</tbody>
</table>
Gross Re-identification Risk Assessment of Each Cluster of Data

Survey
Physical measurements
Medical diagnosis
Drug
Medical procedure
Other data from EHR
Hospital visits
Genomic data
Mobile health
...

- Currently infected Zika Virus
- Race Ethnicity
- Sexual orientation
- Current ovarian cancer
- Health insurance type
- Mother diagnosed circulatory condition
Gross Re-identification Risk Assessment of Each Cluster of Data

Survey
Physical measurements
Medical diagnosis
Drug
Medical procedure
Other data from EHR
Hospital visits
Genomic data
Mobile health
...

- Non-Hispanic
- Black or African American
- Low cervical cesarean section
- Ultrasonography of digestive system
Qualitative Analysis of the Survey Data and Data from EHR

Availability in public dataset or on the Internet

- Very low risk
- Low risk
- High risk

Diagram:
- Life style
- other
- Demographics
- Medical history
- Family history
How likely can One of Us be Re-identified?

1548 live in Tennessee
26 were born in 1985
5 were born outside of US
An Example of Simulated Attacks

This is a little example of our quantitative analysis. We simulated an attacker that knows this set of information. The plot shows how many participants fall into a group with size less than or equal to the value on the x axis. We ran many simulations of “reasonable” adversaries. It turns out adding medical history and family history to it, the risk level does not increase significantly. So we kept the medical history and family history.

State
Birth year
Gender, sex
race/ethnicity
Own or rent
Income
Born in US or not
Marital status
Sexuality
Education
Employment
Active duty

\[ \text{Group size} \]

\[ \% \text{ of participants} \]

- 12%
- 10%
- 8%
- 6%
- 3%

- 14%
- 12%
- 10%
- 8%
- 6%
- 4%
- 2%

- 3%
- 6%
- 8%
- 12%
# Registered Tier Race/Ethnicity

## From Survey

<table>
<thead>
<tr>
<th>Actual survey options:</th>
<th>Generalized responses:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. White</td>
<td>● White alone <em>(if only White selected)</em></td>
</tr>
<tr>
<td>2. Black, African American/African</td>
<td>● Black alone <em>(if only Black/AA selected)</em></td>
</tr>
<tr>
<td>3. Asian</td>
<td>● Asian alone <em>(if only Asian selected)</em></td>
</tr>
<tr>
<td>4. American Indian or Alaska Native*</td>
<td>● Other alone <em>(if one but NOT multiple of the options 4-7 or None selected)</em></td>
</tr>
<tr>
<td>5. Middle Eastern or North African*</td>
<td></td>
</tr>
<tr>
<td>6. Native Hawaiian or other Pacific Islander*</td>
<td></td>
</tr>
<tr>
<td>7. None of these fully describe me- Free text branching logic*</td>
<td>● Two or more races <em>(if multiple options from 1-7 selected)</em></td>
</tr>
<tr>
<td>8. Hispanic, Latino, or Spanish</td>
<td>● Hispanic, Latino or Spanish <em>(Not generalized)</em></td>
</tr>
</tbody>
</table>

*From EHR: removed during the data harmonization*
Registered Tier Gender Identity and Biological Sex

sex at birth from the survey is generalized to three groups: **male, female, other**

gender identity from the survey is generalized to three groups: **man only, woman only, other**

sexual orientation from the survey is generalized to two groups: **straight, other**

sex/gender from EHR is removed during the data harmonization
# Gross Re-identification Risk Assessment of Each Cluster of Data

<table>
<thead>
<tr>
<th>Survey 400</th>
<th>Physical measurements 119</th>
<th>Medical diagnosis 85</th>
<th>Drug 97</th>
<th>Medical procedure</th>
<th>Other data from EHR</th>
<th>Hospital visits</th>
<th>Genomic data</th>
<th>Mobile health</th>
</tr>
</thead>
</table>

- Ego-dystonic sexual orientation
- Death cause
- Type 2 diabetes mellitus
- Depression disorder
- Mood disorder
- Toxic effect of gas, fumes AND/OR vapors
- Retinal hemorrhage, left eye
- Terrorism
- Vehicle accident
- Operations of War, Military operations
- Prolonged stay in weightlessness
- Assault/Homicide, Suicide

...
Gross Re-identification Risk Assessment of Each Cluster of Data

Survey 400
Physical measurements 119
Medical diagnosis 85
Drug 97
Medical procedure
Other data from EHR
Hospital visits
Genomic data
Mobile health

- Pregnancy detection examination
- Radiologic examination, hand
- Intersex surgery male to female
- Lymphatics and lymph nodes imaging
- Ultrasonography of digestive system
- Regional lymph node excision
- Transplantation of liver
- Biopsy of lip
- Fitting and adjustment of other cardiac device
- Hepatitis A and hepatitis B vaccine
Other Transforming Rules

<table>
<thead>
<tr>
<th>Category</th>
<th>Action</th>
<th>Risk Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marital status</td>
<td>keep</td>
<td>Very low</td>
</tr>
<tr>
<td>Own or rent</td>
<td>keep</td>
<td>Very low</td>
</tr>
<tr>
<td>Born in US or not</td>
<td>keep</td>
<td>Very low</td>
</tr>
<tr>
<td>Annual household income</td>
<td>keep</td>
<td>Very low</td>
</tr>
<tr>
<td>Education</td>
<td>generalize</td>
<td>Low risk</td>
</tr>
<tr>
<td>Employment status</td>
<td>generalize</td>
<td>Low risk</td>
</tr>
<tr>
<td>Living situation</td>
<td>remove</td>
<td>High risk</td>
</tr>
<tr>
<td>Active duty</td>
<td>remove</td>
<td>High risk</td>
</tr>
</tbody>
</table>
Looking ahead...
Controlled Tier

- Registered tier Requirements
- Institutional sign-off

Controlled

- Amended Participant-Level
- Additional data types

Registered

- Amended Participant-Level

Public

- Summary Statistics

- Identity Verification
- Research Ethics Training
- Attestation to the Data Use Agreement

Publicly Accessible (No login requirements)
Controlled Tier

Data types under consideration

- Registered tier data with not generalized Demographic fields
- Unshifted Dates
- Geolocation data (more detailed than US State)
- Clinical Documents and Free Text response
- Genomic Data
New Data Types Roadmap

**Winter 2019**
- Registered Tier Launch
- 6 Survey Modules
- EHR data
- Physical Measurements

**Q2 2020**
- Genomic Data

**Q3 2020**
- Digital Health
  - Fitbit
  - Apple Health Kit
  - Data Linkage
  - National Death Index

**Q4 2020**
- New Surveys
- Data Linkage
  - Cancer Registry

**2020**
- **Controlled Tier Launch**
  - Unaltered Registered Tier Data
  - Dates
  - Genomic Data

**Winter 2019**
- Registered Tier Launch
- 6 Survey Modules
- EHR data
- Physical Measurements
Thank you...