

Data and Research Center (DRC)

Committee on Access, Privacy and Security (CAPS)

All of Us Data Access and Data Privacy

All of Us
RESEARCH PROGRAM | The
Future of
Health Begins
With You



National Institutes
of Health

Dikshya Bastakoty, PhD & Weiyi Xia, PhD
Vanderbilt University Medical Center

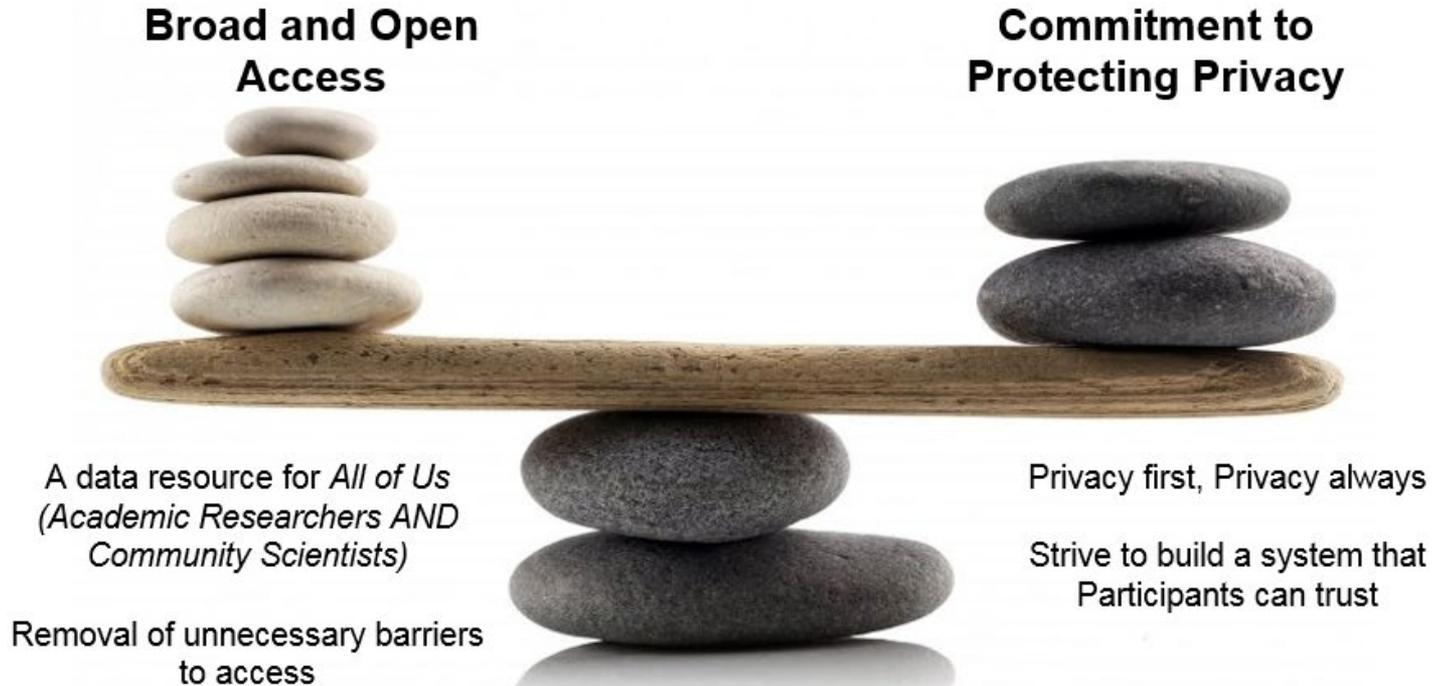
June 24, 2019

ELSI Workshop, NIH

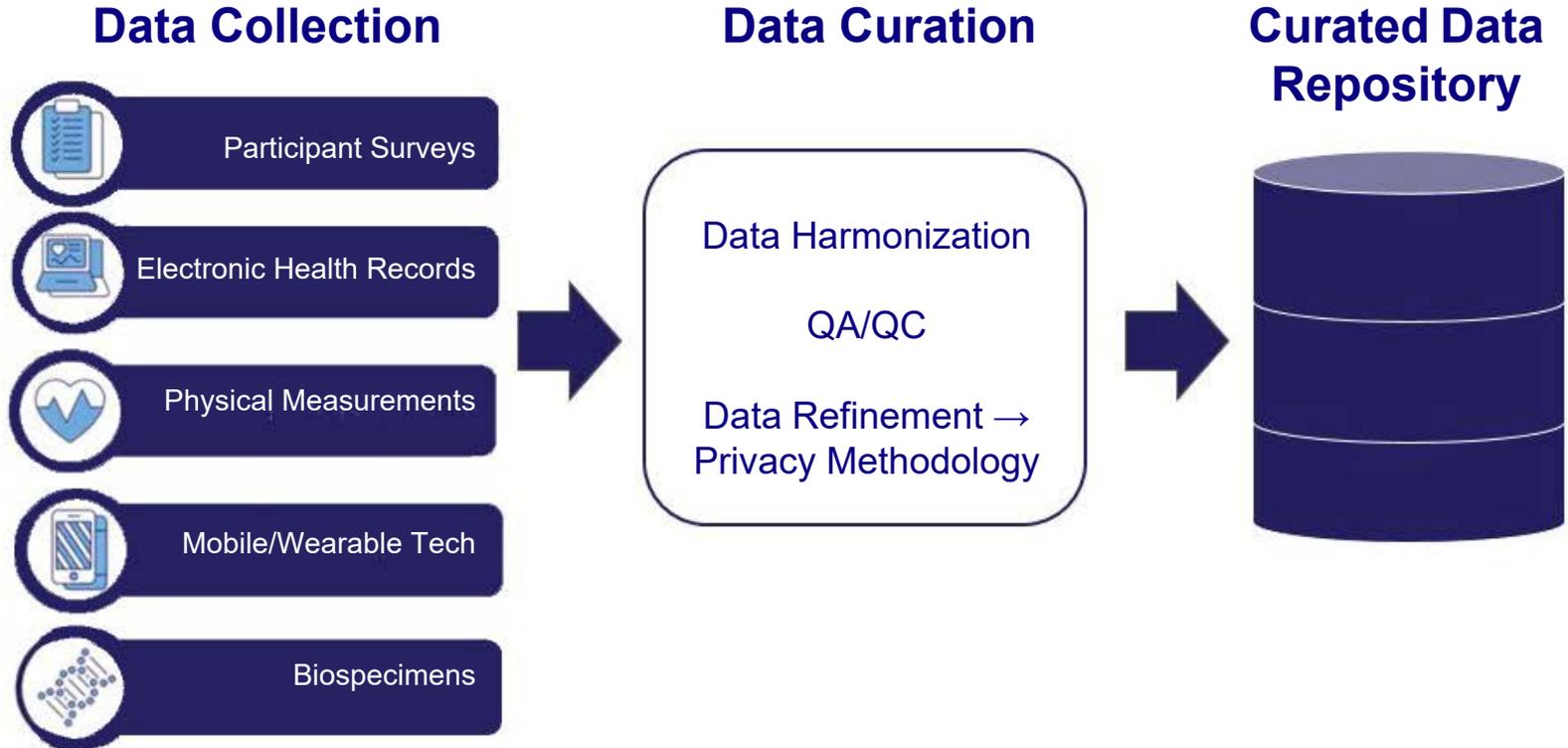
Outline

- 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

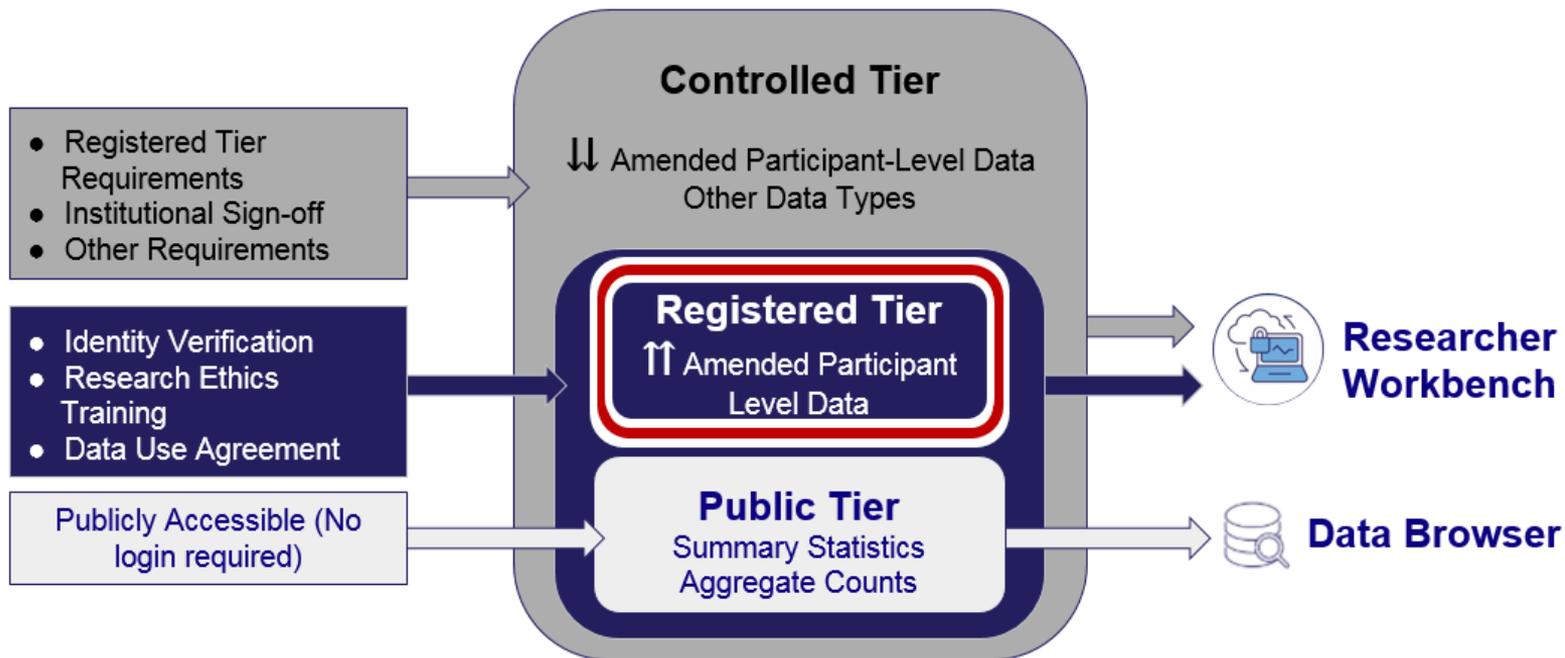
All of Us Principles of Access and Privacy



All of Us Research Program Data



Tiers of Access



Access to the Registered Tier

Become an authorized
AoU Researcher



**DATA
ACCESS**

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

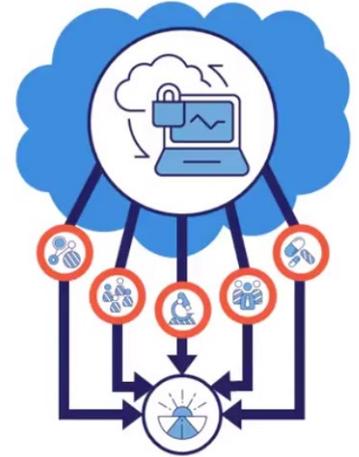
- Users come to the cloud-based data platform; data does not leave the platform

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...



All of Us
RESEARCH PROGRAM

The
Future of
Health Begins
With You

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
...

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

Sodium Chloride	241390
Acetaminophen	207950
heparin	206891
Albuterol	206390
Glucose	173920
Oxycodone	158534
POLYETHYLENE GLYCOL 3350	152404

Gross Re-identification Risk Assessment of Each Cluster of Data

Survey

Physical

measurements

Medical diagnosis

Drug

Medical procedure

Other data from EHR

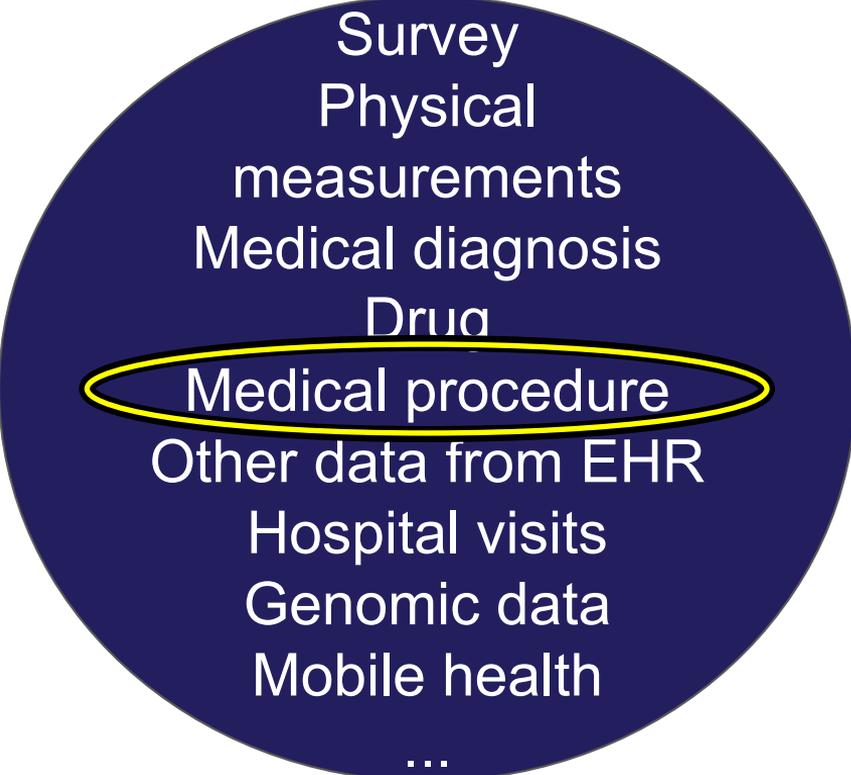
Hospital visits

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

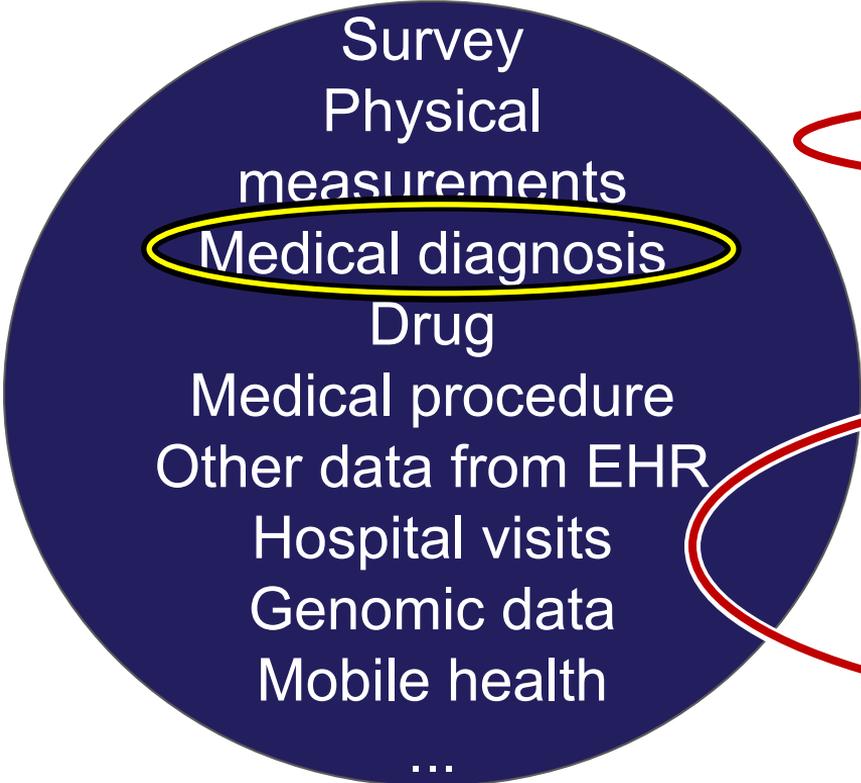
Radiologic examination, hand

Intersex surgery male to female

Lymphatics and lymph nodes imaging

Ultrasonography of digestive system

Gross Re-identification Risk Assessment of Each Cluster of Data



Ego-dystonic sexual orientation
Death cause
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

Physical

measurements

Medical diagnosis

Drug

Medical procedure

Other data from EHR

Hospital visits

Genomic data

Mobile health

...

Currently infected ZikaVirus

Race Ethnicity

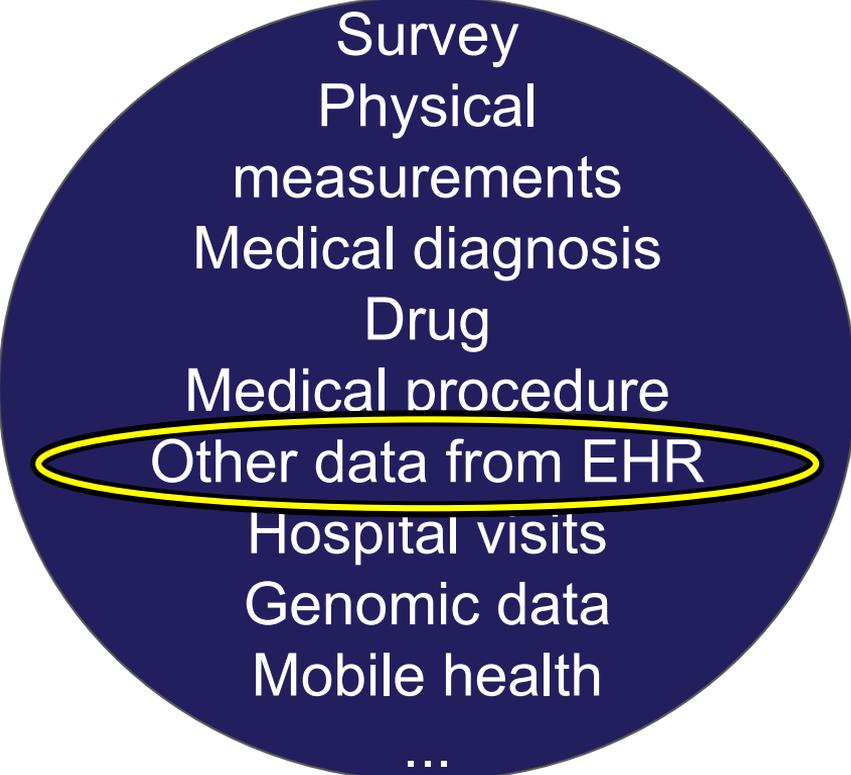
Sexual orientation

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

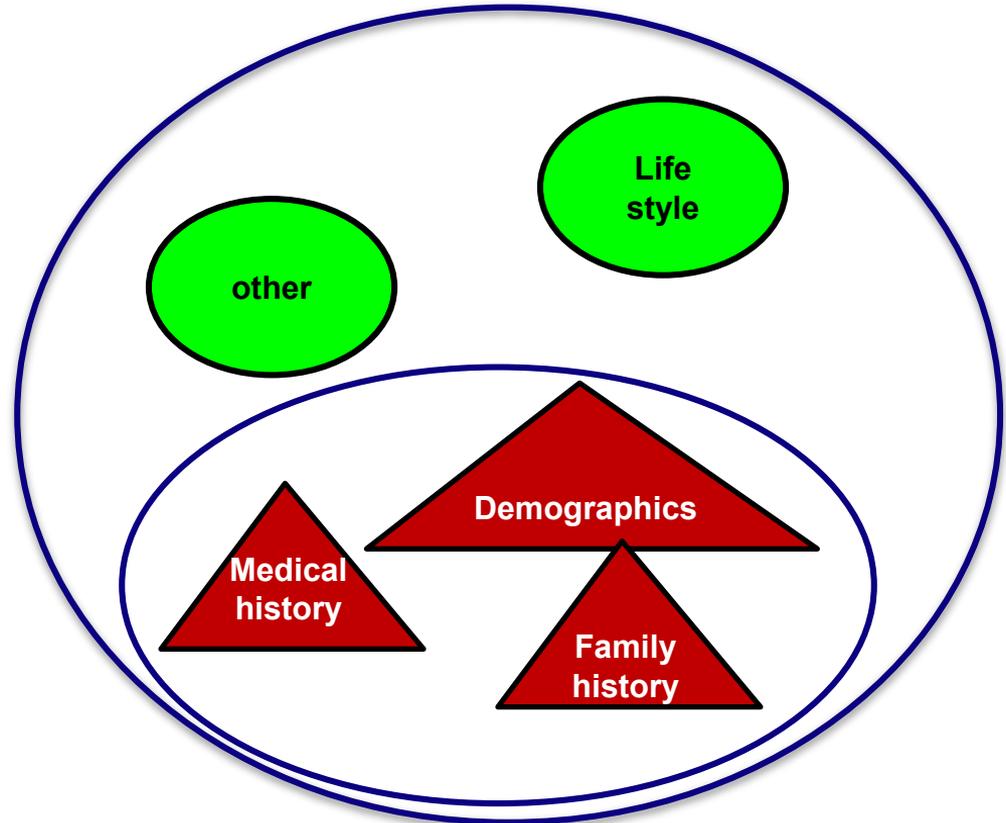
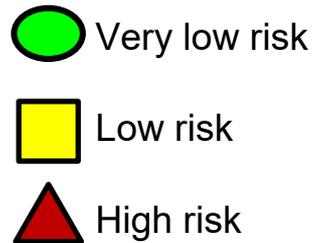
Black or African American

Low cervical cesarean section

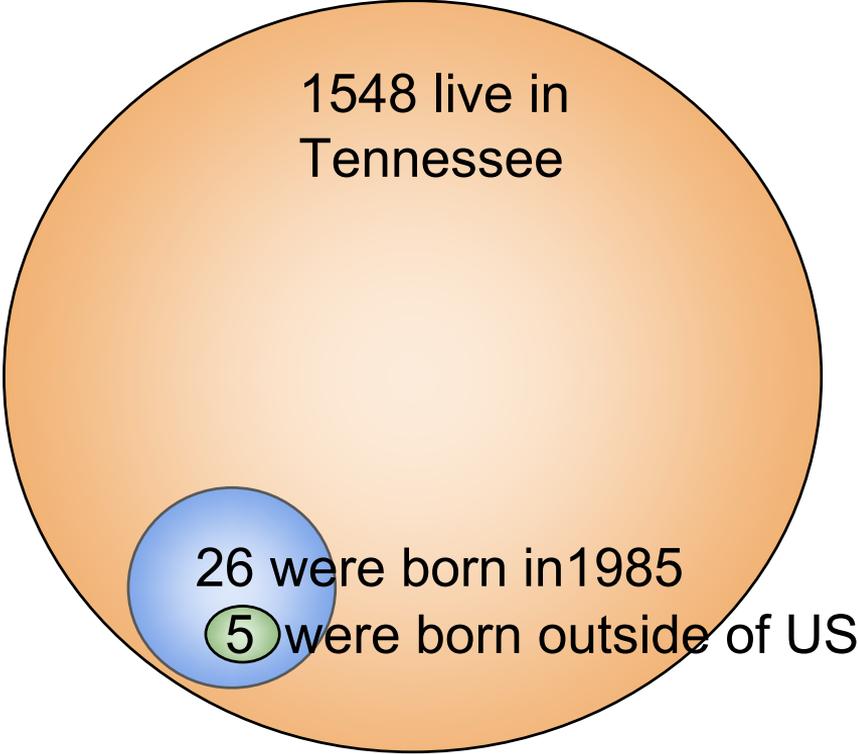
Ultrasonography of digestive system

Qualitative Analysis of the Survey Data and Data from EHR

Availability in public dataset or on the Internet

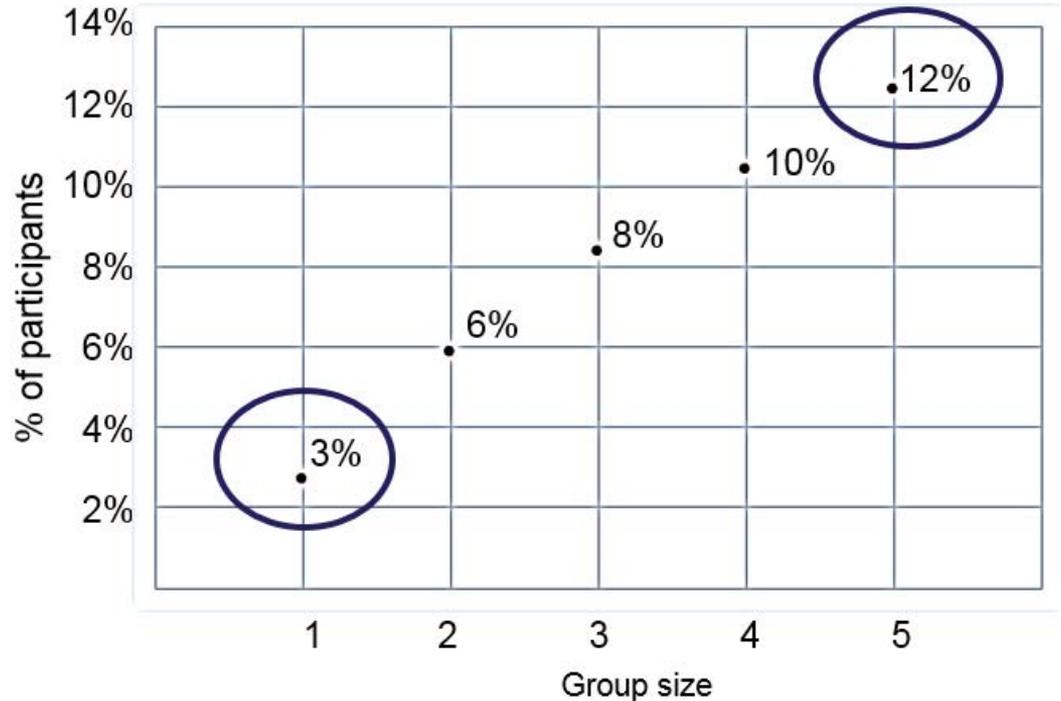
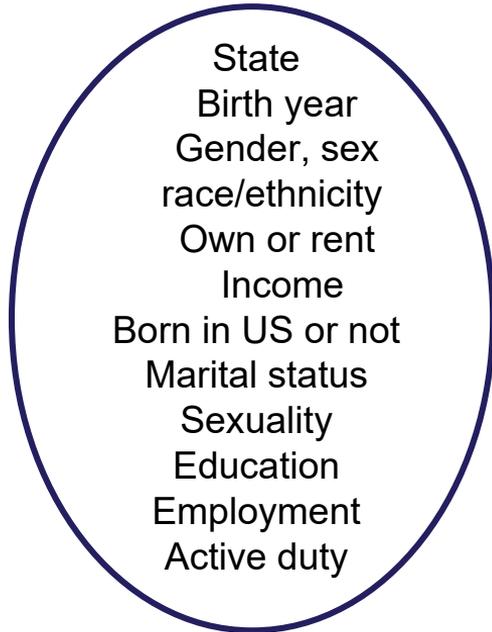


How likely can One of Us be Re-identified?



An Example of Simulated Attacks

This is a little example of our 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.



Registered Tier Race/Ethnicity

From Survey

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

***From EHR: removed during the data harmonization**

Registered Tier Gender Identity and Biological Sex

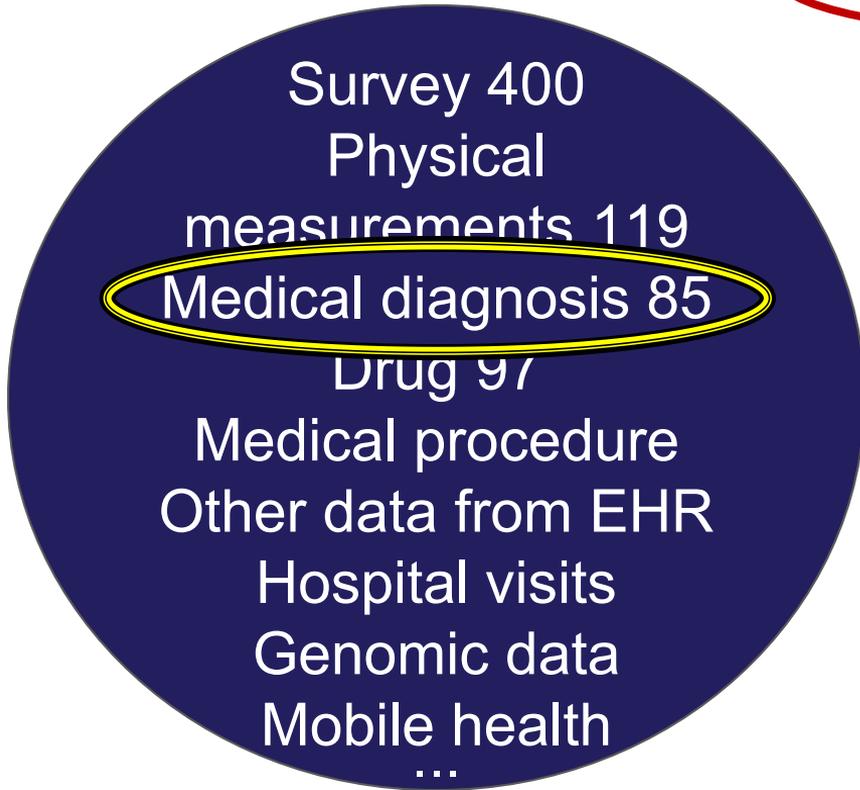
sex at birth from the survey is generalized to three groups: male, female, other

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



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

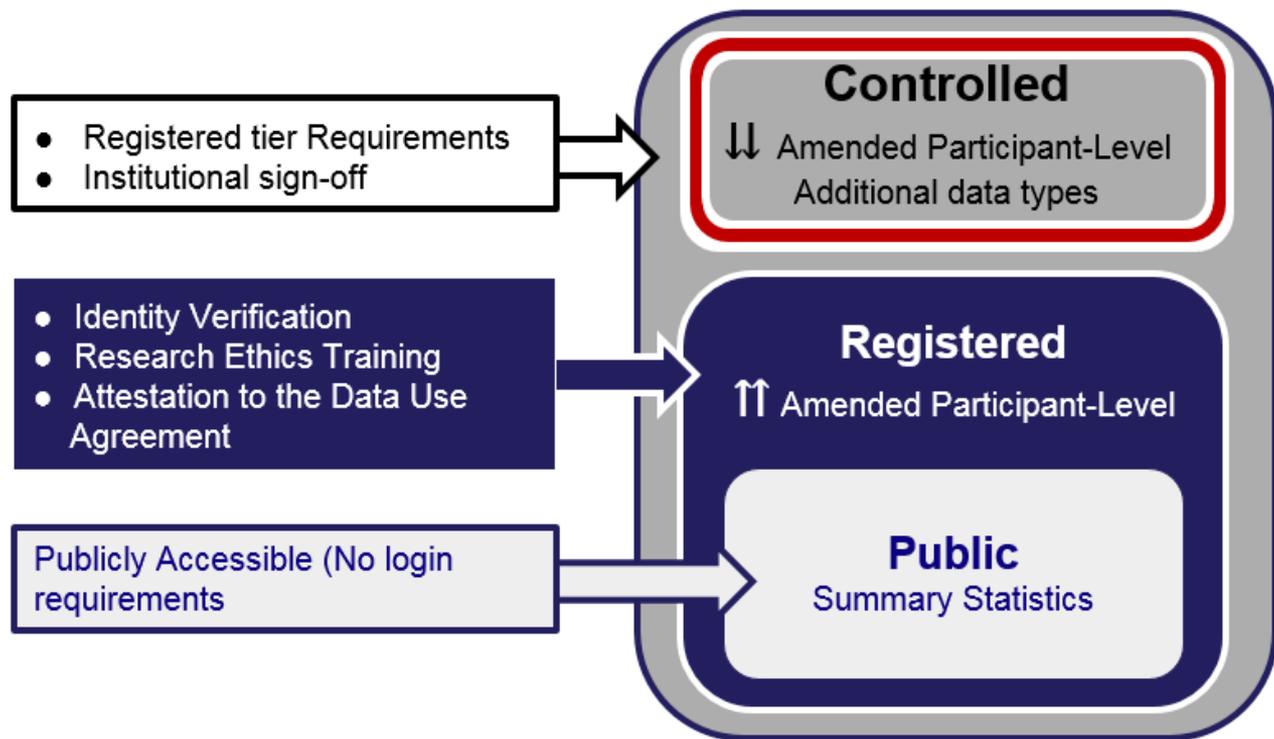
Other Transforming Rules

Marital status	keep	
Own or rent	keep	
Born in US or not	keep	
Annual household income	keep	
Education	generalize	
Employment status	generalize	
Living situation	remove	
Active duty	remove	

-  Very low risk
-  Low risk
-  High risk

Looking ahead...

Controlled Tier



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

Q4 2020

**New Surveys
Data Linkage**

Cancer Registry

**Biospecimen
Access**

Cohort Access

2020

Controlled Tier Launch

Unaltered Registered Tier Data
Dates
Genomic Data

Q2 2020

Genomic Data

Q3 2020

Digital Health

Fitbit
Apple Health Kit

Data Linkage

National Death Index

Winter 2019

Registered Tier Launch

6 Survey Modules
EHR data
Physical Measurements

Thank you...
