All of Us RESEARCH PROGRAM

Application of a Data Quality Framework to Ductal Carcinoma In Situ Using Electronic Health Record Data From the All of Us Research Program

OHDSI Community Call

September 24, 2024

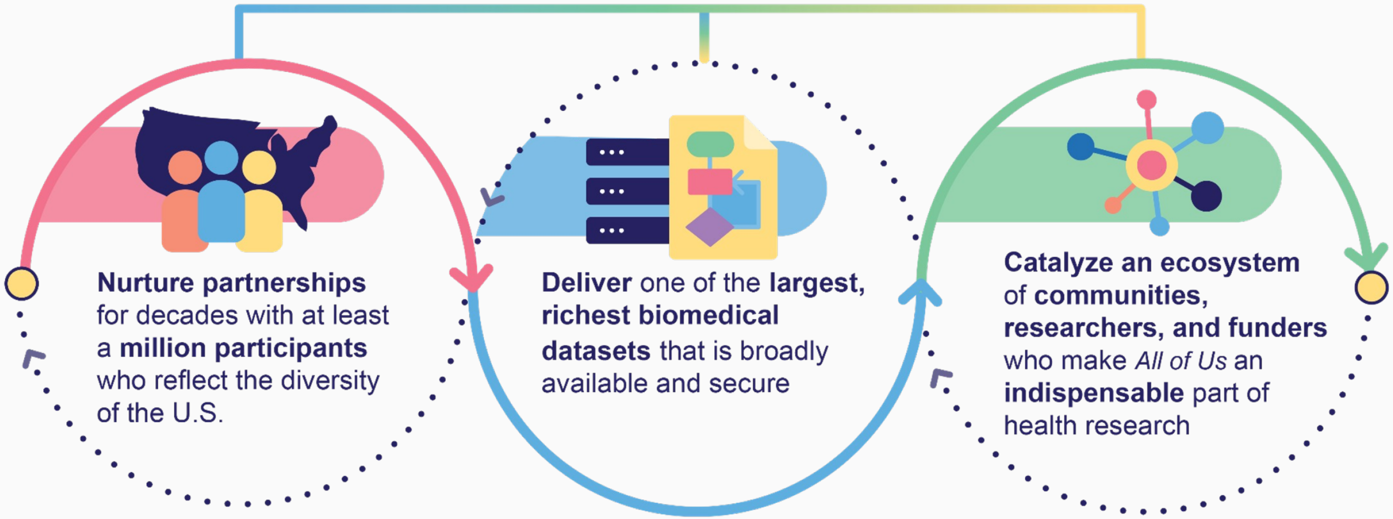
Ami Ostchega, PhD, RN, John Giannini, PhD, Lakshmi Priya Anandan, MPH, Emily Clark, MPH, Matthew Spotnitz, MD, MPH, Lina Sulieman, PhD, Michael Volynski, PhD, and Andrea Ramirez, MD, MPH, Lew Berman, PhD, MS

All of Us Research Program

NIH National Institutes of Health

The *All of Us* Research Program Mission

Accelerate health research and medical breakthroughs, enabling individualized prevention, treatment, and care for all of us



Nurture partnerships for decades with at least a **million participants** who reflect the diversity of the U.S.

Deliver one of the **largest, richest biomedical datasets** that is broadly available and secure

Catalyze an ecosystem of **communities, researchers, and funders** who make *All of Us* an **indispensable** part of health research

Made possible by a team that maintains a culture built around the program's core values

Study Objectives

Specific Aims

- Develop and operationalize an electronic health record (EHR) data quality framework
- Apply the dimensions of the framework to the phenotype and treatment pathways of ductal carcinoma in situ (DCIS) using *All of Us* Research Program data
- Propose and apply a checklist to evaluate the framework's application

Why is it significant?

- Provides insights into the fitness of using *All of Us* EHR data for specific phenotypes
- Understand the strengths & weaknesses of checklists for other phenotypes

Ductal Carcinoma In Situ Public Health Impact

- DCIS is a precancerous condition that accounts for about 25.0% of breast cancers diagnosed in the United States [1, 2]
- DCIS is a well-known and understood disease, has a clearly defined treatment protocol and, if treated, shows excellent disease-free survival without additional surgery [3]
- The incidence of DCIS has increased in recent years due to the widespread use of screening mammography [1]

[1] Allegra CJ, Aberle DR, Ganschow P, et al,. National Institutes of Health State-of-the-Science Conference Statement: Diagnosis and Management of Ductal Carcinoma In Situ September 22–24, 2009. J Natl Cancer Inst. 2010; 102(3):161–169.

[2] Sarah E Pinder and Ian O Ellis. Review:The diagnosis and management of pre-invasive breast disease Ductal carcinoma in situ (DCIS) and atypical ductal hyperplasia (ADH) — current definitions and classification. Breast Cancer Res 2003, 5:254-257

[3] About Clinical Practice Guidelines. National Comprehensive Cancer Network. 2023. Accessed October 30, 2023. https://www.nccn.org/guidelines/guidelines-process/about-nccn-clinical-practice-guidelines

Methods: Study Design

- **Phenotype**
  - ***Cases:*** Earliest occurrence of any ICD-09/10 code or SNOMED concept for DCIS mapped to OMOP concepts codes and restricting to female participants who were at least 18 years of age
  - ***Controls:*** Female participants who were at least 18 years of age and did not have a DCIS diagnosis

- **Clinical Measures and Interventions:** National Comprehensive Cancer Network (NCCN) DCIS related treatment guidelines (workup, primary treatment, and postsurgical treatment)

- **Data quality framework:** conformance, completeness, concordance, plausibility, temporality

Definitions: Data Quality Dimensions

- **Conformance:** Dataset values and elements (standards, syntax, and structure) were equivalent or represented in the same way

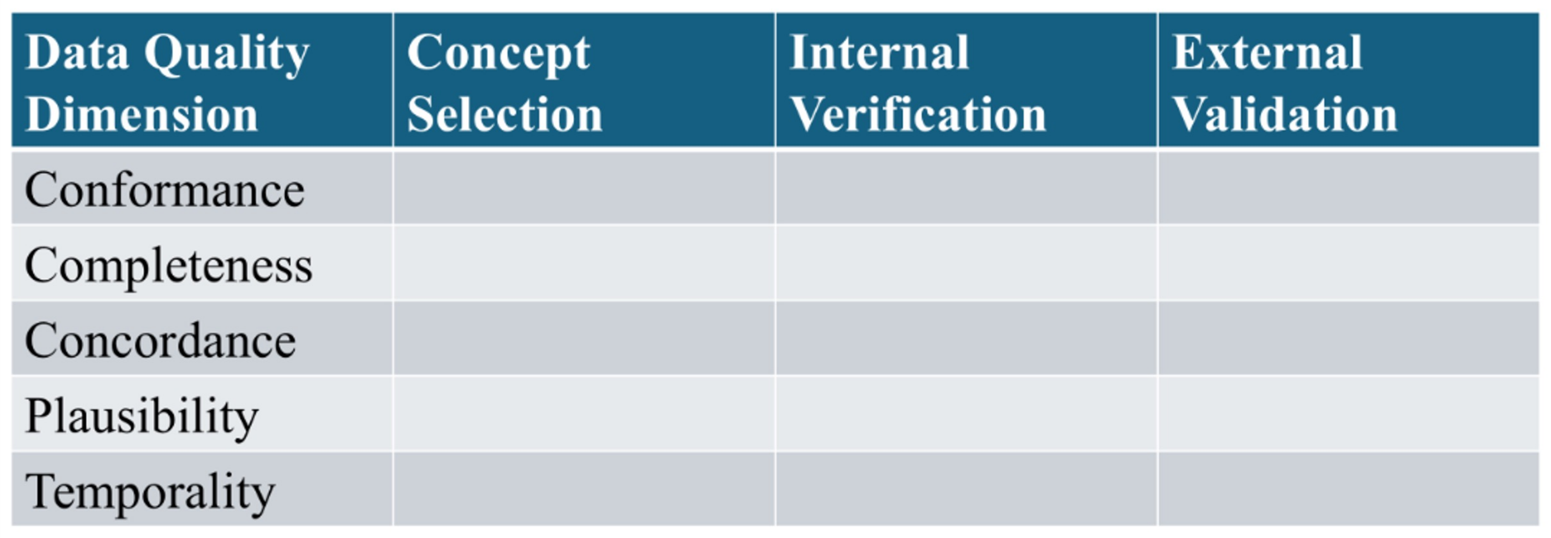
- **Completeness:** Dataset values and elements have been captured/were available

- **Concordance:** Dataset values and elements were similar or in agreement

- **Plausibility:** Dataset values and elements were believable

- **Temporality:** Dataset values and elements had valid start times, end times, and durations and followed expected order

Methods: Data Quality Dimensions Evaluability Checklist



Data Quality Dimension	Concept Selection	Internal Verification	External Validation
Conformance			
Completeness			
Concordance			
Plausibility			
Temporality			

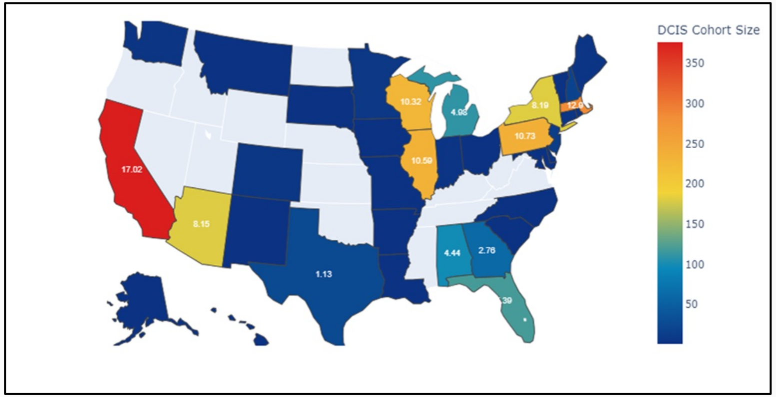
Source: Berman L, Ostchega Y, Giannini J, et. al. Application of a Data Quality Framework to Ductal Carcinoma in Situ Using Electronic Health Record Data from the All of Us Research Program. JCO Clinical Cancer Informatics.2024 Aug:8:e2400052.

DCIS Cohort Demographics and Geographic Distribution

- Out of 365,488 participants in *All of Us* more than 350,000 have shared their EHR. Among these participants, 2,209 are females with DCIS (0.6%)

- In the *All of Us* DCIS cohort, 52.5% of the females were diagnosed between the ages of 60-79, and 66.0% of the diagnosed population is non-Hispanic White

- The highest percentages of the *All of Us* DCIS cohort are in California (17.0%), Massachusetts (12.0%), Pennsylvania (10.7%), and Illinois (10.5%)

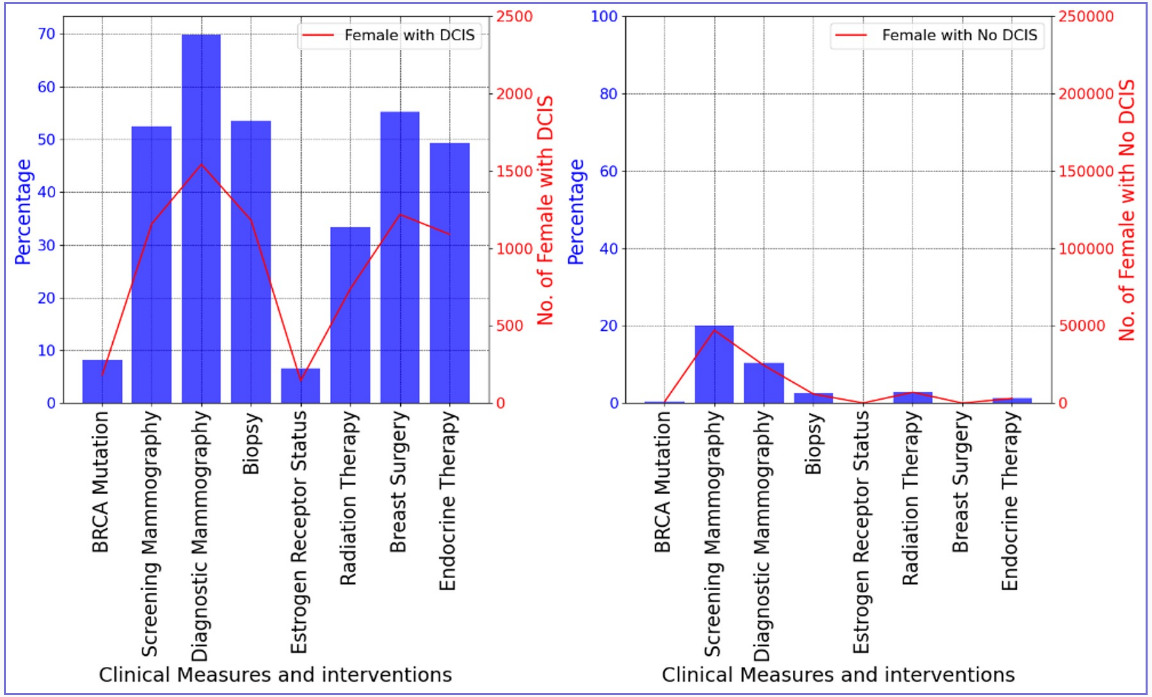


Conformance

Assessing the source distributions of the OMOP concept codes

- ICD 9/10 only: 1,924 (87.1%)
- ICD 9/10 and SNOMED: 277 (12.5%)
- SNOMED only <= 20

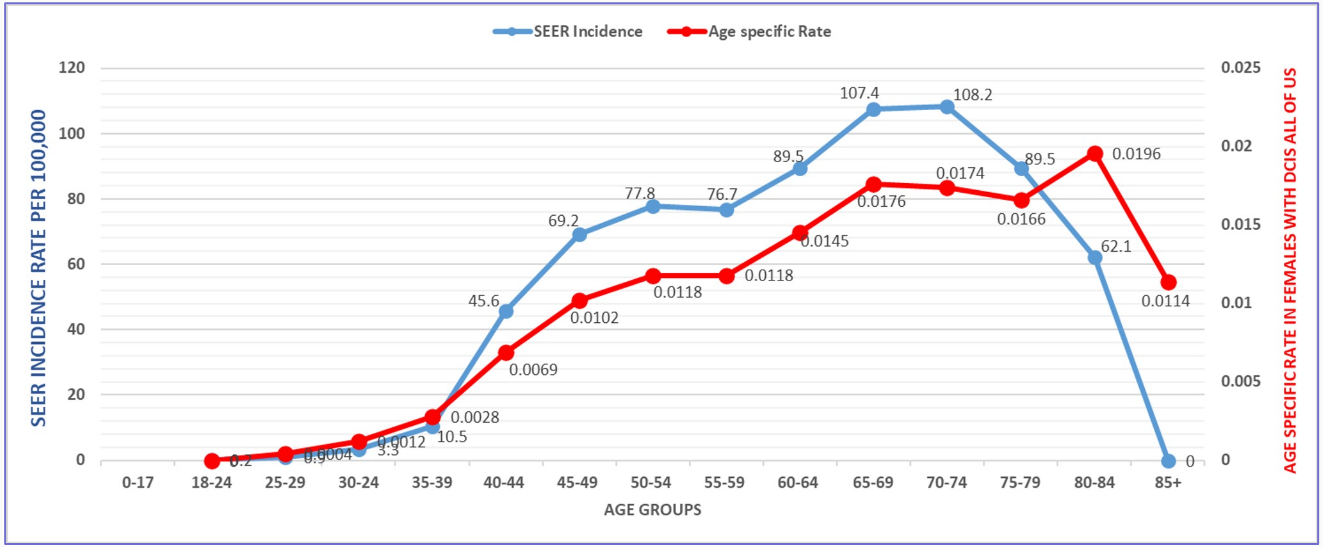
Completeness



The DCIS and non-DCIS groups showed differences in the proportions of NCCN guideline related concepts, which included diagnostic mammography (69.9% vs. 11.2%), biopsy (53.5% vs. 3.2%), surgery (55.18% vs. 1.3%), and endocrine therapy (49.4% vs. 1.9%), ($p<0.01$).

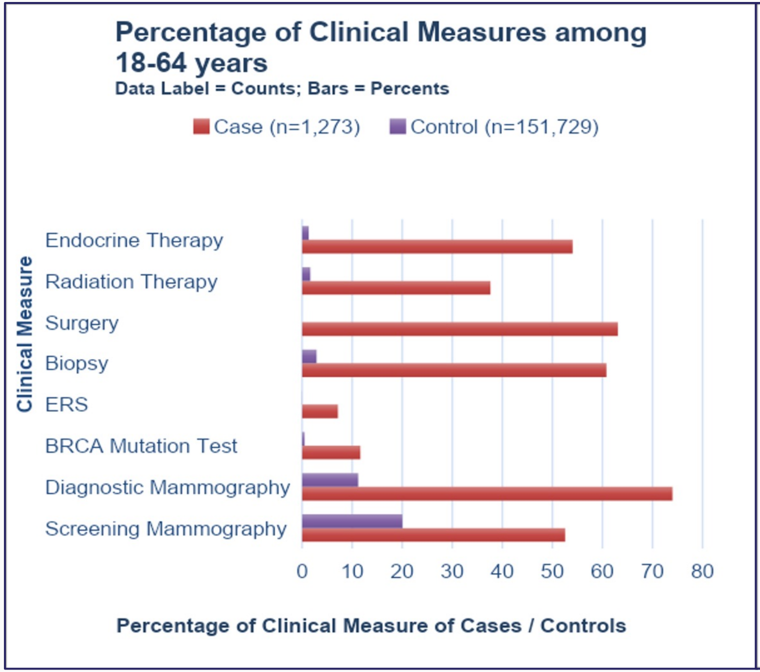
Concordance

The correlation between DCIS *All of Us* age-specific prevalence rates and SEER (1975-2017) reported incidence of DCIS by age groups



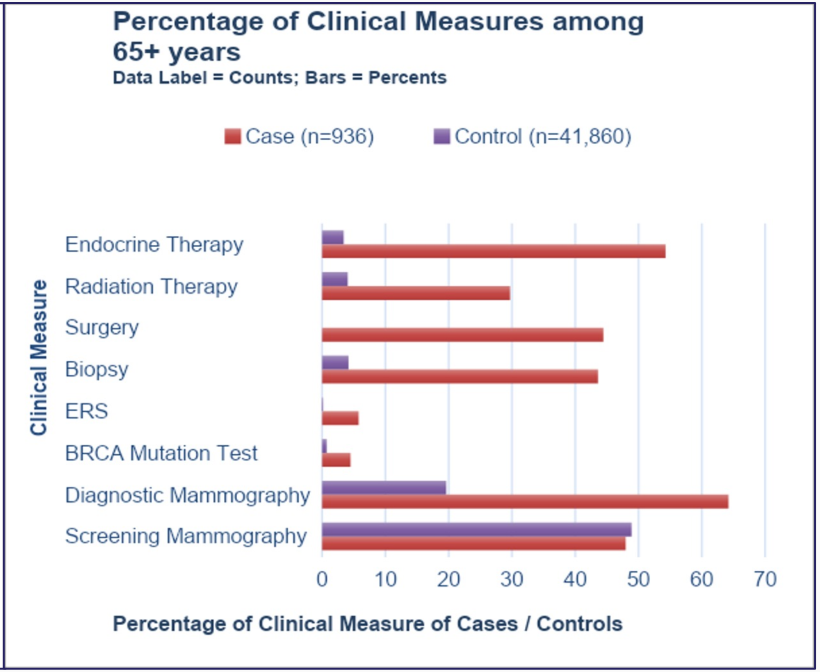
The Spearman rank correlation coefficient between these two sets of values is 0.85 (p<0.01).

Plausibility



Percentage of Clinical Measures among 18-64 years

Data Label = Counts; Bars = Percents



Percentage of Clinical Measures among 65+ years

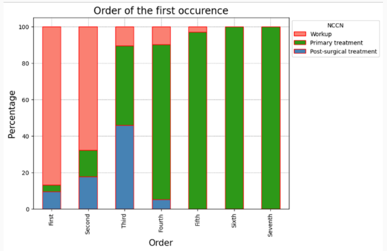
Data Label = Counts; Bars = Percents

The relative number of cases and controls who had clinical measures and interventions differed by age group p<0.01.

Temporality

- Biopsy-Diagnosis Interval
  - Antecedent biopsy data were available for 1,023 females (47%)
  - The median time from biopsy to diagnosis was 8 weeks

- NCCN Phase Analysis
  - Organized clinical measure and intervention concept sets by NCCN phase guidelines (e.g. workup, primary treatment, postsurgical treatment)
  - Sequenced the concept sets by those phases
  - Most participants progressed from workup to primary treatment



Key Findings

- Developed a systematic and generalizable approach to assessing phenotype data quality

- All five dimensions were evaluated successfully for concept selection, internal verification, and external validation. External validation was limited by external benchmarks

- Used case and control phenotype definitions to ensure there are differences between these groups

Limitations and Next Steps

- Limitations
  - Over-representation of academic medical centers / healthcare provider organizations
  - Manual selection of concepts limited by knowledge of clinical experts
  - Fragmented EHR data
  - Absence of data from unstructured sources

- Next Steps
  - Application of the framework to other phenotypes
  - Refinement of temporal analysis

Contact Information

Yechiam Ostchega, PhD,RN
National Institutes of Health
Office of the Director
All of Us Research Program
Division of Technology and Platform Development
ami.ostchega@nih.gov