

Assessing Data Completeness of Sarcoidosis Patients in the *All of Us* Research Program

Matthew Spotnitz, MD, MPH¹, Dominic R. Forte², Dhruv P. Jani³, Maria Salgado, MD³, John Giannini, PhD¹, Jennifer Adjemian, PhD¹, Mohamed Seedahmed, MD, MPH⁴, Elizabeth Park, MD, MSc³

¹*All of Us* Research Program, Office of the Director, National Institutes of Health, Bethesda, Maryland ²Department of Dental Medicine, Columbia University Irving Medical Center, New York, New York ³Department of Medicine, Columbia University Irving Medical Center, New York, New York ⁴Department of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania

Background

Sarcoidosis is a rare, chronic, systemic, granulomatous, inflammatory disease that has a range of clinical presentations [1]. A sarcoidosis diagnosis is further complicated by the fact that the granulomas do not have distinctive characteristics. Consequently, sarcoidosis can have varying case definitions, symptoms, prevalence and public health impact estimates [2]. Furthermore, investigations of sarcoidosis can have limited generalizability outside of the original study population [3]. Therefore, generating large-scale, real-world evidence on sarcoidosis could be highly valuable for further characterization of the disease and its clinical management.

Previously, we described a data quality framework, which has been used to evaluate cohorts of *All of Us* participants who had a ductal carcinoma in situ (DCIS) diagnosis, a surgical oncology procedure, rheumatoid or psoriatic arthritis, or inflammatory bowel disease diagnoses [4-8]. In this analysis, we used the same framework to evaluate the data quality of participants diagnosed with sarcoidosis.

One of the data quality dimensions that we evaluated was completeness, which is a measure of data availability. Despite being an essential component of data quality, standardized approaches to characterizing electronic health record (EHR) data are not well established. Here, we present the results from the completeness analysis.

Methods

Cases were defined by the first occurrence of an ICD-9 or SNOMED CT diagnosis code for sarcoidosis in adult *All of Us* Research Program participants, including retrospective diagnoses prior to the program enrollment date. Controls were participants who did not have sarcoidosis diagnosis codes. We used ATHENA to guide concept selection and identify OMOP CDM codes for the phenotype and disease related variables. We calculated the concept prevalence of disease specific variables in our sarcoidosis cohort.

Results

There were 2,452 sarcoidosis participants in our cohort. Representative sociodemographic characteristics are shown in Table 1.

Table 1. Sociodemographic characteristics of participants in the sarcoidosis cohort in the *All of Us* Research Program.

Variable	Sarcoidosis (N = 2,452), N (%)
Current Age	
18-44	171 (7.0%)
45-64	912 (37.3%)
≥65	1362 (55.7%)
Sex at Birth	
Female	1622 (66.3%)
Male	797 (32.6%)
Skip/None	26 (1.0%)
Race	
White	1185 (48.5%)
Black or African American	860 (35.2%)
Asian	≤20
American Indian or Alaska Native	≤20
Middle Eastern or North African	≤20
Native Hawaiian or Other Pacific Islander	≤20
More than one population	116 (4.7%)
None of these	26 (1.1%)
Skip, none indicated, prefer not to answer	220 (9.0%)
Ethnicity	
Not Hispanic or Latino	2174 (88.9%)
Hispanic or Latino	207 (8.5%)
None of these	26 (1.1%)
Skip/Prefer not to answer	38 (1.6%)
Education Level	
≤8th grade/never attended	37 (1.5%)
Grades 9-11	113 (4.6%)
High School or GED	469 (19.2%)
Some college	760 (31.1%)
College Graduate	542 (22.2%)
Advanced Degree	478 (19.6%)
Skip/Prefer not to answer	46 (1.9%)
Annual Income	
<10k	280 (11.5%)
10k-25k	349 (14.3%)
25k-35k	186 (7.6%)
35k-50k	194 (7.9%)

50k-75k	289 (11.8%)
75k-100k	198 (8.1%)
100k-150k	247 (10.1%)
150k-200k	104 (4.3%)
More than 200k	130 (5.3%)
Skip/Prefer not to answer	468 (19.1%)
Neighborhood Socioeconomic Characteristics	(Mean ± SD)
Median Income	64982 (17342)
% High School Education	87.8 (5.2)
% Below Poverty	15.8 (6.0)
% Uninsured	8.5 (3.9)
Deprivation Index	0.32 (0.07)

The results of the concept prevalence analysis are shown in Figure 1.

Sarcoidosis

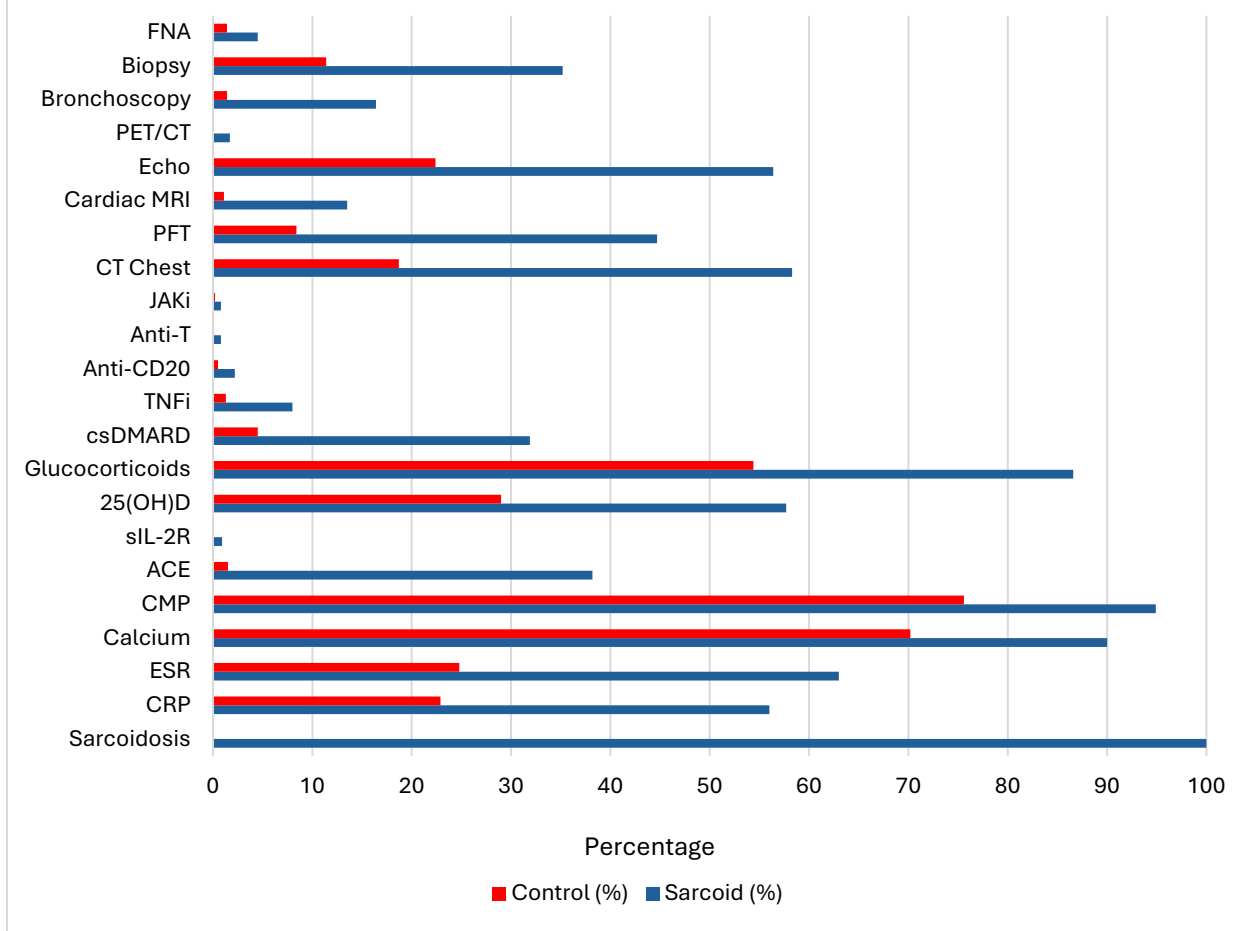


Figure 1: Bar chart of clinical concepts for participants who had sarcoidosis (bottom, blue) and controls (red, top). C-Reactive Protein = CRP; Erythrocyte Sedimentation Rate = ESR; Calcium = calcium measurement; Comprehensive Metabolic Panel = CMP, including components of the Basic Metabolic Panel and Liver Function Test panels; Angiotensin Converting Enzyme Measurement = ACE; Serum Soluble Interleukin-2 Receptor = sIL-2R; 25-hydroxyvitamin D = 25(OH)D; Conventional Synthetic Disease-Modifying Antirheumatic Drug = DMARD (csDMARD); Tumor Necrosis Factor Inhibitors = TNFi; Anti B-Cell Inhibitor = Anti-CD20; Anti T-Cell Inhibitor = Anti-T; Janus-associated Kinase Inhibitors = JAKi; Chest Computed Tomography = CT Chest; Pulmonary Function Test = PFT; Cardiac Magnetic Resonance Imaging = Cardiac MRI; Echocardiography = Echo; Positron Emission Tomography/Computed Tomography = PET/CT; Fine Needle Aspiration = FNA; Biopsy of skin, heart, liver, lymph node, brain, lung, thyroid, bone marrow, muscle, nerve, or breast = Biopsy; Data Source: The *All of Us* Research Program.

Overall, the concept prevalence was higher for cases than controls ($p < 0.001$). However, some counts were lower than expected. For example, 58.3% of the cohort had a chest CT and 44.7% had pulmonary function tests.

Conclusion

We have shown successful implementation of our data quality completeness analysis to evaluate a sarcoidosis cohort. Additionally, we have characterized the sociodemographic characteristics of that cohort. Without source validation, it is unclear to what extent the low clinical variable concept prevalence is due to data missingness or variance in practice patterns. However, non-digitized and out of network records may have contributed to the low counts. Manual concept selection was a limitation of our analysis. Also, it is unclear if alternative phenotype definitions would have yielded different results. Overall, our findings may have implications for whether sarcoidosis data from the *All of Us* program are fit for use.

References

- [1] Drent M, Crouser ED, Grunewald J. Challenges of Sarcoidosis and Its Management. *N Engl J Med* 2021;385:1018-32. doi: 10.1056/NEJMr2101555
- [2] Gerke AK, Judson MA, Cozier YC, Culver DA, Koth LL. Disease Burden and Variability in Sarcoidosis. *Ann Am Thorac Soc* Vol 14, Supplement 6, pp S421–S428, Dec 2017. doi: 10.1513/AnnalsATS.201707-564OT
- [3] Obi ON, Polite PY, Fish KM et. al. A Snapshot of United States Sarcoidosis Patients and their Perceived Disease Impact: Results of the Sarcoidosis Research Institute Survey. *Lung* (2025) 203:31. doi: 10.1007/s00408-024-00761-8
- [4] Berman L, Osthega 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;8:e2400052. doi: 10.1200/CCI.24.00052
- [5] Spotnitz M, Giannini J, Osthega Y, et. al. Assessing the Data Quality Dimensions of Partial and Complete Mastectomy Cohorts in the *All of Us* Research Program: Cross-Sectional Study. *JMIR Cancer* 2025;11:e59298. doi: 10.2196/59298
- [6] Spotnitz M, Giannini J, Clark E et. al. Assessing the Data Quality Dimensions of Surgical Oncology Cohorts in the *All of Us* Research Program. *JCO Clinical Cancer Informatics*, 2025 9:e2500078. doi: 10.1200/CCI-25-000
- [7] Spotnitz M, Giannini J, Clark E et. al. Assessing Data Quality of Rheumatoid and Psoriatic Arthritis Patients in the *All of Us* Research Program. *JAMIA Open* 2026, 9(2), ooag028. doi: 10.1093/jamiaopen/ooag028
- [8] Spotnitz M, Faye AS, Giannini J, et. al. Assessing Data Quality of Inflammatory Bowel Disease Patients in the *All of Us* Research Program. *JAMIA Open* 2026 9(3), ooag084. doi: 10.1093/jamiaopen/ooag084