Atlas2AoU: Enabling Comparison of OHDSI Phenotype Library Phenomic Profiles in All of Us and the UK Biobank

Abigail Newbury[1,2], Xinzhuo Jiang[1], Karthik Natarajan[1,*], Gamze Gürsoy[1,2,3,*]
1 Department of Biomedical Informatics, Columbia University, New York City
2 New York Genome Center, New York City
3 Department of Computer Science, Columbia University, New York City

Background

Precision medicine integrates data from medical history, genomics, environmental exposures, and more, to enhance our understanding of disease etiology.(1) Facilitating these efforts, two prominent biobanks are the UK Biobank (UKBB) and the All of Us Research Program (AoU). Previous work has shown that the UKBB population is healthier than the general UK population, while the AoU population has a higher disease burden compared with the US population, with the exception of psychiatric diagnoses.(2–4) Furthermore, a recent study by Zeng et al. found that the majority of diseases have significantly higher prevalence in AoU than in the UKBB.(3) Across each of these studies, disease cohort identification was performed via Phecode phenotyping or by participant self-report.(2–4)

The richness of Electronic Health Record (EHR) information contained in each of these biobanks allows researchers to define more expressive phenotype definitions to build cohorts for downstream analysis.(5,6) Here, we contend that more precisely defined phenotypes will enable a more accurate inclusion of patients in cohorts and facilitate better comparisons across biobanks. The OHDSI Phenotype Library (PL) is a repository for high-quality phenotype definitions, including those generated by domain experts and subjected to peer review, and thus is a rich resource for high-throughput computational phenotyping.(7) However, due to the technical limitations on AoU's Research Workbench, OHDSI's software ATLAS that allows cohort creation cannot be directly used in AoU.(8) This is in part because the AoU curated data repository combines data from multiple sources such as surveys, physical measurements, and EHR data.

We developed a tool to create ATLAS phenotypes within AoU's Research Workbench. We demonstrate this tool's effectiveness by creating OHDSI Phenotype Library (PL) phenotypes in AoU and comparing them with those in the UKBB.

Methods

To enable correct cohort construction in AoU, Atlas2AoU alters the original ATLAS query in two ways. First, a temporary observation period table is created and queried as opposed to the original. The temporary table is constructed using OHDSI's suggestion of taking the minimum and maximum dates of recorded clinical events for each individual across nine OMOP CDM tables.(9) Second, due to permissions on the AoU workbench, queries do not write to a cohort table, but instead return the final table directly as a Pandas Data Frame. Atlas2AoU is available at https://github.com/G2Lab/Atlas2AoU.

To illustrate the use of this tool, we present a prevalence comparison of 423 cohorts in OHDSI's PL v3.1.6 between UKBB and AoU. Furthermore, for a Chronic Obstructive Pulmonary Disease (COPD) cohort (cohortId: 28), we perform detailed analyses of geographic disease prevalence, demographics and social determinants of health (SDOH), and identify variants significantly associated with the disease through genome-wide association studies (GWAS).

Results

Table 1 demonstrates the impact of deploying an ATLAS query on the AoU database without proper configuration of the observation period table for three phenotypes. False positives are defined as individuals who were found by the original query but not by the Atlas2AoU modified query, while false negatives were missing from the original query but discovered by the Atlas2AoU query. As shown, across all three phenotypes, the Atlas2AoU query results in a larger cohort size, increasing power of downstream analysis.

Phenotype	Original query false positives	Original query false negatives	Original query cohort size	Atlas2AoU query cohort size
T2D	370	43,111	7,503	50,244
COPD	488	23,044	2,256	24,812
Acute MI	0	6,131	194	6,325

Table 1. Cohort size comparison between original and Atlas2AoU queries for three phenotypes

Figure 1 shows prevalence ratios for all 423 phenotypes from the OHDSI PL. A prevalence ratio of one represents equal prevalence. AoU has significantly higher prevalence compared to the UKBB for 335 of the 423 phenotypes. In contrast, the UKBB only has significantly higher prevalence for 23 phenotypes. Thus, as in Zeng et al., we find that AoU has a much higher disease burden than the UKBB.(3) All three of the diseases with the highest prevalence ratios are in the Psychiatry/Psychology category, including two phenotype definitions specific to Attention Deficit Hyperactivity Disorder. AoU has a higher prevalence for 100% of phenotypes in the Psychiatry/Psychology, Ophthalmology, Endocrinology, Sleep, Hematology, Rheumatology, Metabolism and Nutrition, Miscellanea, and Orthopedics categories. Diseases with the lowest prevalence ratios include Peripheral Ischemia, Chilblains, and Cerebrovascular accident.

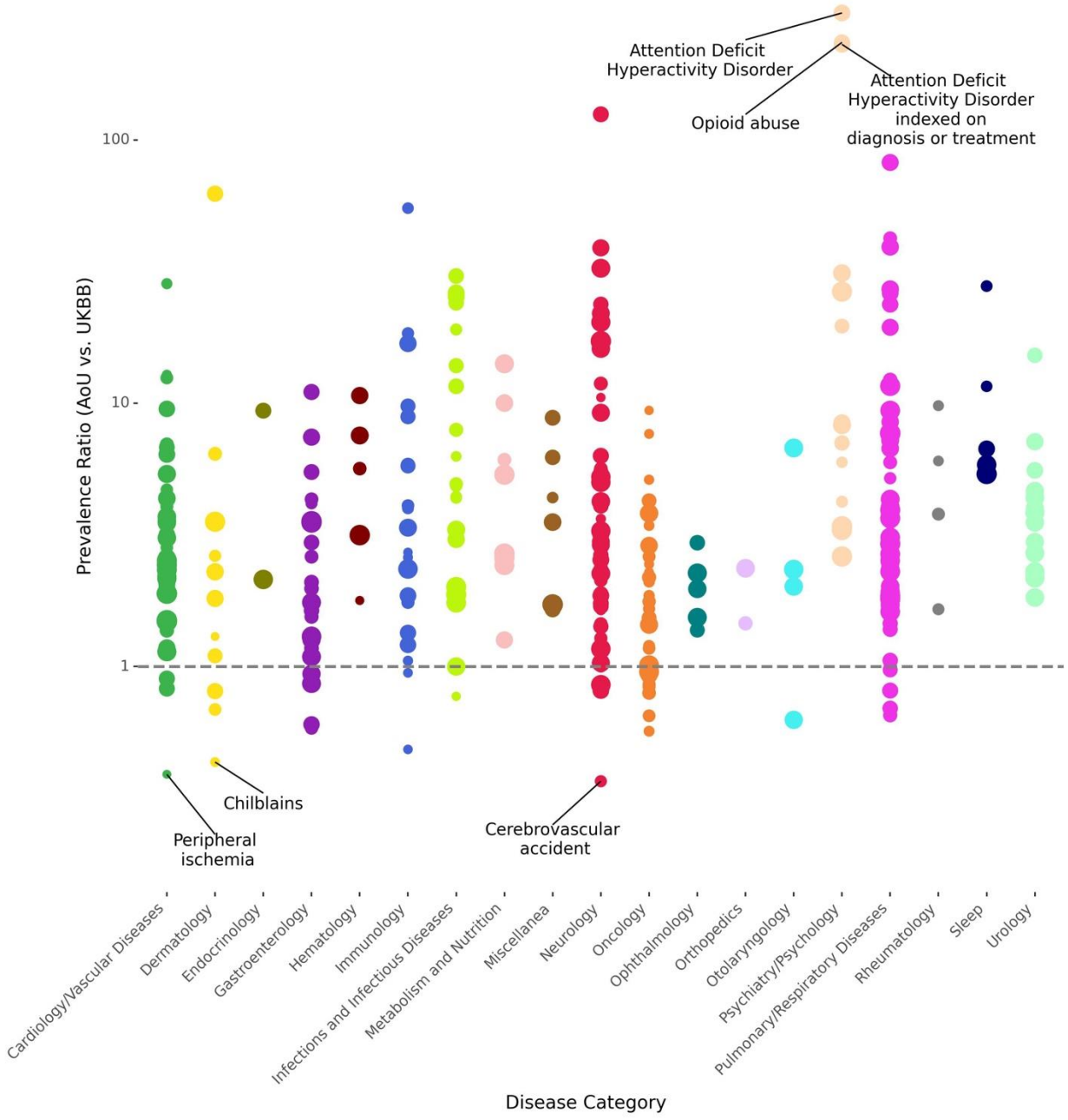
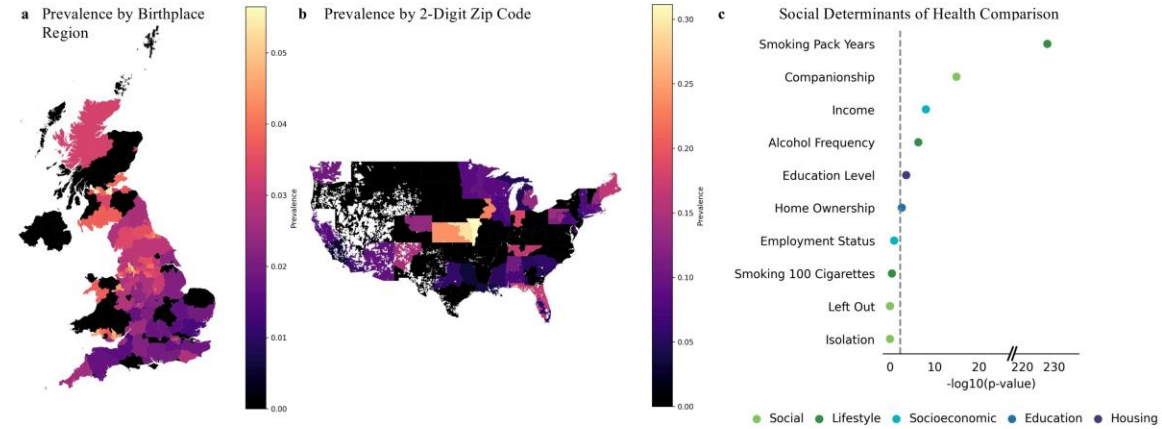


Figure 1. Prevalence comparison for OHDSI PL cohorts in AoU and the UKBB

Figure 2 illustrates detailed cohort comparisons for COPD. As shown in Figure 2e, we find the prevalence of COPD to be significantly higher in AoU than in the UKBB. We find a similar estimate to that of Zeng et al. for COPD in AoU (OHDSI PL: 0.086; Phecode: 0.087), but a much lower estimate in the UKBB (OHDSI PL: 0.03; Phecode: 0.06).(3) Additionally, we see that the AoU cohort has a higher proportion of Black participants and is generally older than the UKBB cohort. Areas of high prevalence for COPD, shown in Figures 2a and 2b, are West Dunbartonshire, Glasgow City, and West Lothian, and 2-digit zip codes 64, 66, and 67 corresponding to areas of Kansas and Missouri. These regions of high prevalence coincide with

regions characterized by high smoking rates, a known risk factor for COPD.(10,11) We see that the SDOH variables of smoking pack years, companionship, income, alcohol frequency and education level are significantly different among the AoU and UKBB cohorts. While smoking pack years variable is significantly different, there is no difference in the proportion of individuals who have smoked 100 cigarettes in their lifetime. Lastly, while the AoU GWAS did not detect any significant variants within exons, the UKBB GWAS detected variants in the exons of HYKK, CHRNA3, CHRNA4, and CHRNA5, all of which are previously known to harbor variants associated with COPD.(12,13)



Database	Genes with Significant Exon Variants
UKBB	IREB2, HYKK, PSMA4, CHRNA5, CHRNA3
AoU	None

e

Characteristic	UKBB	AoU
Cohort Size	14,901	24,812
Prevalence	0.030	0.086*
Sex at birth (%)		
Female	44.4	51.6
Male	55.6	48.4
Race (%)		
Asian	1.0	1.1
Black	0.1	25.4
White	98.9	73.5
Age (%)		
18-34	0.0	0.7
35-49	5.2	4.7
50-64	56.6	37.3
65+	38.2	57.3

Figure 2. Geographic, SDOH, genomic, and demographic profiles for COPD

Conclusion

We demonstrate that Atlas2AoU can enable downstream deployment of ATLAS queries in the AoU researcher workbench. Through application of Atlas2AoU with 423 OHDSI PL phenotypes, we show that AoU has a significantly higher disease burden than the UKBB. Furthermore, we present detailed comparisons between AoU and the UKBB for COPD. This detailed comparison demonstrates differences in prevalence, demographics, and SDOH in the two databases, as well as highlights important regions of high risk for disease susceptibility, both in terms of geography and genetics.

References

1. Johnson KB, Wei W, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision medicine, AI, and

the future of personalized health care. Clin Transl Sci. 2021 Jan;14(1):86–93.

2. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. American Journal of Epidemiology. 2017 Nov 1;186(9):1026–34.

3. Zeng C, Schlueter DJ, Tran TC, Babbar A, Cassini T, Bastarache LA, et al. Comparison of phenomic profiles in the All of Us Research Program against the US general population and the UK Biobank. Journal of the American Medical Informatics Association. 2024 Jan 23;ocad260.

4. Barr PB, Bigdeli TB, Meyers JL. Prevalence, comorbidity, and sociodemographic correlates of psychiatric diagnoses reported in the All of Us Research Program. JAMA Psychiatry. 2022 Jun 1;79(6):622–8.

5. Swerdel JN, Ramcharran D, Hardin J. Using a data-driven approach for the development and evaluation of phenotype algorithms for systemic lupus erythematosus. PLoS One. 2023 Feb 16;18(2):e0281929.

6. Hardin J, Murray G, Swerdel J. Phenotype algorithms to identify Hidradenitis Suppurativa using real-world data: Development and Validation Study. JMIR Dermatol. 2022 Nov 30;5(4):e38783.

7. Rao G. PhenotypeLibrary: The OHDSI Phenotype Library [Internet]. 2024. Available from: https://ohdsi.github.io/PhenotypeLibrary/, https://github.com/OHDSI/PhenotypeLibrary

8. Software Tools – OHDSI [Internet]. Available from: https://www.ohdsi.org/software-tools/

9. Philofsky M, EHR Working Group. OHDSI: Observational Health Data Sciences and Informatics. Observation period considerations for EHR data. Available from: https://ohdsi.github.io/CommonDataModel/ehrObsPeriods.html

10. Centers for Disease Control and Prevention. CDC State Tobacco Activities Tracking and Evaluation (STATE) System. Map of current cigarette use among adults. Available from: https://www.cdc.gov/statesystem/cigaretteuseadult.html

11. Revie L, Davies B, Mais D. Office for National Statistics. Adult smoking habits in the UK: 2021. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/bulletins/adultsmokinghabitsingreatbritain/2021

12. Thorgeirsson TE, Steinberg S, Reginsson GW, Bjornsdottir G, Rafnar T, Jonsdottir I, et al. A rare missense mutation in CHRNA4 associates with smoking behavior and its consequences. Mol Psychiatry. 2016 May;21(5):594–600.

13. Cho MH, McDonald MLN, Zhou X, Mattheisen M, Castaldi PJ, Hersh CP, et al. Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. Lancet Respir Med. 2014 Mar;2(3):214–25.