

# Pregnancy episodes in *All of Us*: Harnessing multi-source data for pregnancy-related research

Louisa H. Smith, PhD<sup>1,2</sup>

Wanjiang Wang, MS<sup>1</sup>

Brianna Keefe-Oates, PhD<sup>2</sup>

<sup>1</sup>Department of Public Health and Health Sciences, Bouvé College of Health Sciences, Northeastern University, Boston MA

<sup>2</sup>Roux Institute, Northeastern University, Portland, ME

Corresponding Author:

Louisa H. Smith

100 Fore St, Portland, ME 04101

[l.smith@northeastern.edu](mailto:l.smith@northeastern.edu)

(617) 373-7008

Key Words: pregnancy, pregnancy outcomes, gestational age, electronic health records, real-world data

Word Count: 4900

Data Availability: All analytic code is available on GitHub

((<https://github.com/louisahsmith/allofus-pregnancy>). *All of Us* Research Program data is available to registered researchers at <https://workbench.researchallofus.org/>).

Funding Statement: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author Contributions:

Louisa Smith: Conceptualization, Methodology, Software, Visualization, Writing

Wanjiang Wang: Visualization, Writing

Brianna Keefe-Oates: Visualization, Writing

Competing interests: The authors have no competing interests to declare.

## **ABSTRACT**

### **Objective**

The National Institutes of Health's *All of Us* Research Program addresses gaps in biomedical research by collecting health data from diverse populations. Pregnant individuals have historically been underrepresented in biomedical research, and pregnancy-related research is often limited by data availability, sample size, and inadequate representation of the diversity of pregnant people. *All of Us* integrates a wealth of health-related data, providing a unique opportunity to conduct comprehensive pregnancy-related research. We aimed to identify pregnancy episodes with high-quality electronic health record (EHR) data in *All of Us* Research Program data and evaluate the program's utility for pregnancy-related research.

### **Materials and Methods**

We used a previously published algorithm to identify pregnancy episodes in *All of Us* EHR data. We described these pregnancies, validated them with *All of Us* survey data, and compared them to national statistics.

### **Results**

Our study identified 18,970 pregnancy episodes from 14,234 participants; other possible pregnancy episodes had low-quality or insufficient data. Validation against people who reported a current pregnancy on an *All of Us* survey found low false positive and negative rates. Demographics were similar in some respects to national data;

however, Asian-Americans were underrepresented, and older, highly educated pregnant people were overrepresented.

## **Discussion**

Our approach demonstrates the capacity of *All of Us* to support pregnancy research and reveals the diversity of the pregnancy cohort. However, we noted an underrepresentation among some demographics. Other limitations include measurement error in gestational age and limited data on non-live births.

## **Conclusion**

The wide variety of data in the *All of Us* program, encompassing EHR, survey, genomic, and fitness tracker data, offers a valuable resource for studying pregnancy, yet care must be taken to avoid biases.

## INTRODUCTION

Despite the critical role of pregnancy in human health and development, it remains an understudied area in biomedical research. Women's health research, in general, has been historically neglected, leading to significant gaps in our understanding of conditions uniquely or predominantly affecting women.[1] This neglect extends to pregnancy, where the complexities and ethical considerations of studying the pregnant population have led to the widespread exclusion of pregnant people from clinical trials and efficacy studies. This not only leads to a gap in understanding of many health conditions, treatments, and events among pregnant people, but it also limits pregnancy-related research. Observational data is therefore necessary to study the impacts of various exposures, as well as to gain insights into the broader spectrum of pregnancy-related health conditions, behaviors, and outcomes. Nevertheless, research into pregnancy and the postpartum period remains challenging due to data and study design limitations.

Historically, studies about pregnancy and postpartum health outcomes have relied on costly cohort studies or surveillance mechanisms that do not capture the entire pregnancy period. For example, state and national governments provide representative but cross-sectional birth surveillance data that also fails to capture early outcomes or sufficient data on pregnancy exposures such as medications.[2,3] Birth certificates, while useful for documenting basic demographic information and birth outcomes, lack detailed and accurate information on maternal health, prenatal care, and pregnancy complications.[4] Additionally, birth registry data do not capture longitudinal data on maternal and child health beyond the immediate postpartum period.

Birth cohorts, observational studies that recruit pregnant or recently postpartum people and their infants, often recruit participants in the later stages of pregnancy, missing the earliest pregnancy exposures and outcomes such as miscarriages.[5–7] More recently, preconception cohort studies have been designed to prospectively collect data on fertility, pregnancy, and postpartum health outcomes.[8–11] While cohort studies are rich in longitudinal data, they are also inherently resource-intensive and often rely primarily on self-report of pregnancy timing and outcomes.

Real-world data, including electronic health records (EHR) and insurance claims, contain thorough information about diagnoses, medications, and healthcare procedures and their costs. While the longitudinal nature of the data, coupled with very large sample sizes, is promising for studying key biomedical, social, and health services outcomes, these datasets usually lack important demographic information.[12] Additionally, identifying pregnancies and measuring gestational age in these complicated records is challenging.

The *All of Us* Research Program provides an opportunity for pregnancy-related research that overcomes some of the limitations of other data sources. The program is funded by the US National Institutes of Health to improve precision health research and began collecting data on a planned 1 million Americans in 2018.[13] The study collects both EHR and survey data, as well as biospecimens and data from activity trackers, allowing for longitudinal studies examining both social and biomedical exposures and outcomes. The *All of Us* Research Program’s mission is to fully represent the diversity of the US population by explicitly including groups historically underrepresented in

biomedical research.[13] Conducting pregnancy-related research using *All of Us* Research Program data aligns well with this aim, as pregnant people, and in particular pregnant people of color and sexual and gender minorities,[14–16] make up an important and understudied segment of this population.

Maternal morbidity and mortality are of utmost concern in the US, with rising risks and severe disparities in outcomes by race and ethnicity.[17,18] The ability to link health outcome data from medical records to survey questions about health behaviors, medical history, and socioeconomic and interpersonal experiences can provide new insights into pregnancy and in the postpartum period. In particular, the survey data collected by *All of Us* provides in-depth information about social determinants of health that may help explain disparities in maternal health and, most importantly, identify interventions. Data sources beyond EHR and survey data include genetic, physical measurement, and fitness tracker data, which together have the potential to offer a comprehensive view of participants' health before, during, and after pregnancy.

While the *All of Us* Research Program provides an opportunity to ask new research questions about pregnancy, challenges to identifying and characterizing pregnancies in real-world data remain. Identifying pregnancy episodes in the *All of Us* EHR data can be difficult due to variations in coding practices, patients visiting multiple healthcare providers over the course of a pregnancy, and infrequent use of codes that identify gestational age. These challenges mean that researchers cannot rely on a single code to identify pregnancy, and complex algorithms are needed to identify pregnancies, ascertain their outcomes, and estimate gestational ages.[19]

## **OBJECTIVE**

This work aimed to identify pregnancy episodes in the most recent release of *All of Us* data using an algorithm previously published by Jones et al.[19] Objectives included determining gestational age and pregnancy outcomes, validating episodes using survey data, and characterizing the identified pregnancies. We aimed to establish how the *All of Us* Research Program could be used to answer pregnancy-related research questions important to researchers and the communities they serve.

## **METHODS**

### **Data**

The *All of Us* Research Program began enrolling adult participants across the United States in 2018; data collection is ongoing.[13] Focused recruitment occurs at an extensive network of sites nationwide; the study is also open to any volunteer. Participants must complete a baseline survey with demographic information; additional surveys, which may be completed at any time after the baseline survey, collect data on health history, social context, and more.[20] Volunteers are invited to link their EHR data (including from before joining the study) and contribute biospecimens for genomic analyses; movement, heart rate, and sleep data through wearable fitness tracking devices; clinic-based body measurements, and more. Due to the geographic diversity of the participants, EHR data is contributed by a large number of institutions with different medical record systems. The transformation of the data into the Observational Medical Outcomes Partnership common data model (OMOP CDM)[21,22] allows the EHR data to be combined across contributing

sites and analyzed on a web-based platform. After applying quality control and privacy-preserving measures, data is released to researchers who complete the required training and data use agreement. This study used data from the *All of Us* Research Program's Controlled Tier Dataset Version 7 (release C2022Q4R9), available to authorized users on the Researcher Workbench. The project followed the guidelines for ethical conduct of research put in place by *All of Us* and was determined to be exempt by the Northeastern University IRB.

### **Pregnancy identification algorithm**

We identified pregnancy episodes among *All of Us* participants who had EHR data at some point between ages 15-55 and who did not report male sex at birth. To do so we used an algorithm developed and validated in the National Covid Cohort Collaborative data (a repository of EHR data contributed from sites around the United States during the COVID-19 pandemic and harmonized according to the OMOP CDM).[19] This approach, referred to as Hierarchy and rule-based pregnancy episode Inference integrated with Pregnancy Progression Signatures (HIPPS), consists of three sub-algorithms: Hierarchy-based Inference of Pregnancy (HIP), Pregnancy Progression Signature (PPS), and Estimated Start Date (ESD). These are described here briefly; a complete description is available in the original publication.

The HIP algorithm, based on an earlier standalone algorithm for OMOP CDM data[23] classifies pregnancy episodes based on how a pregnancy ended: live birth, stillbirth, ectopic pregnancy, spontaneous or induced abortion, or delivery record only (outcome



unspecified). First, OMOP concept codes related to these outcomes are identified, prioritized in that order, and then deduplicated within patients based on minimum plausible pregnancy durations (for example, 182 days between live birth outcomes and 56 days between consecutive ectopic pregnancies) to create outcome-based episodes. Each of these episodes is additionally assigned a minimum and maximum plausible start date based on plausible gestational ages for the outcome. Next, gestation-based episodes are identified based on gestational-age-related codes (e.g., “Gestation period, 36 weeks”) by comparing differences in visit dates and between gestational age values between each such code. Progressing chronologically across all such codes within an individual, a new gestation-based episode is initiated if the time between dates is greater than 28 days more than the difference in values implied by the codes. The HIP algorithm then looks for overlap between the outcome- and gestation-based episodes.

The PPS algorithm relies on a different but overlapping set of initial pregnancy-related codes. Along with the codes for individual gestational weeks, PPS uses codes that are likely to occur only within a specific gestational age range (e.g., glucose tolerance tests most often occur within 6 and 8 months of gestation). These codes were identified by Jones et al. through a concept frequency analysis as those that are more likely to occur during HIP-identified episodes and have little variance in the observed gestational window, and were labeled by clinicians with the appropriate gestational time range in months.[19] The PPS algorithm identifies episodes with plausible progressions of these timing-specific concepts. Next, pregnancy outcome codes that overlap with the expected timing relative to the gestational age codes are used to assign outcomes to the episodes.

The HIP-identified and PPS-identified episodes are merged when they have overlapping timing. The ESD algorithm then uses both the week-specific and gestational-age range codes to identify areas of consistency in the gestational age estimates and remove outlying timing codes. These codes are used to assign an inferred pregnancy start date and a date the pregnancy outcome occurred based on the latest and most specific codes.

We translated the original code for the algorithm from PySpark SQL to work in R on the Researcher Workbench, where we primarily used dbplyr[24] and the allofus R package.[25] We conducted analyses in May 2024. Our analysis code is available on Github (<https://github.com/louisahsmith/allofus-pregnancy>).

### **Assessment and validation**

As a preliminary assessment of the algorithm, we classified the pregnancy episodes based on whether the pregnancy outcomes from both the HIP and PPS algorithms matched, the dates of those outcomes were within 14 days of each other, and whether the estimated gestational age at the end of pregnancy was plausible. As in Jones et al.,[19] we assigned episodes a concordance score of 2 if all three criteria were met, a 1 if the outcomes did not match but an outcome was within the expected term duration, and a 0 otherwise. We also classified episodes based on how precise the gestational age concepts were: “non-specific” if there were no gestational age-related codes or if the timing could not be narrowed down below a three-month window (i.e., based on a PPS gestational range code), “4 weeks-3 months” or “1-3 weeks” if gestational age could be narrowed to one of

those windows, and “1 week (poor support)” if there was only a single gestational week concept within a pregnancy episode. In addition, we classified pregnancy episodes as “data-rich” if they occurred in 2016 or later (due to limitations of earlier EHR data) and were identified by both the HIP and PPS algorithms.

We validated the pregnancy episodes using *All of Us* survey data. First, we identified *All of Us* participants who met the eligibility criteria (had EHR data at some point between ages 15-55 and did not report male sex at birth) and who responded to the Overall Health survey. We compared those participants’ responses to the question “Are you currently pregnant?” and the identified pregnancy episodes in the EHR data to assess misclassification. Figure 1 depicts examples of possible correctly classified (Figure 1A-C) and misclassified scenarios (Figure 1 D-G). We considered true positives to occur when someone reported being pregnant at a time overlapping with an identified pregnancy episode or a two-week buffer interval to account for EHR data delays (Figure 1A-B) and false negatives other reports of pregnancy that did not overlap with an identified episode (Figure 1D-E). True negative pregnancy episodes were those in which a negative response to the survey question did not coincide with any identified episode (Figure 1C). A false positive pregnancy episode was one in which a participant reported not being pregnant on a date they were identified as greater than 12 weeks pregnant by the algorithm (Figure 1F-G). Participants estimated to be at less than 12 weeks gestation were excluded from this definition in an attempt to include as many participants who would have known they were pregnant as possible (Figure 1H); as a sensitivity analysis, we considered a cutoff of 20 weeks. We calculated sensitivity (proportion of true positives among the “Yes” survey

responses), specificity (proportion of true negatives among the “No” responses), positive predictive value (proportion of true positives among respondents surveyed at greater than 12 estimated weeks gestation), and negative predictive value (proportion of true negatives among respondents not surveyed during an identified episode).

In addition, we followed Jones et al. in quantifying occurrences of a set of 25 clinician-curated pregnancy-related codes not used in identifying the pregnancies. We compared their frequency during the expected pregnancy timing among identified episodes to the overall population frequency.

### **Characteristics of pregnancies and pregnant *All of Us* participants**

We characterized the pregnancy episodes by their outcomes, gestational lengths, year of pregnancy, and demographic characteristics of the pregnant people. We fit exploratory log-linear regression models to describe characteristics associated with a higher probability of having more than one pregnancy episode captured in the data, having a live birth vs. another pregnancy outcome, and delivering preterm (among live births). The reference level for predictor variables was the largest group in the sample. We also characterized the extent to which pregnant participants contributed additional types of data to *All of Us*, including survey, fitness tracker, and genomic data.

We used public vital statistics data to compare the demographics of *All of Us* pregnancies ending in live births to national statistics.[26] Specifically, we computed age, education, racial/ethnic, and state breakdowns of US live births from 2016-2022 and standardized those to the distribution of pregnancies by calendar year in *All of Us*.

## RESULTS

### Pregnancy episodes

There were 134,566 individuals in the Controlled Tier C2022Q4R9 release of *All of Us* (participant data cutoff date of July 1, 2022) who did not report male sex at birth and who had contributed EHR data at some point between the ages of 15 and 55 (Figure 2). Duration of retrospective EHR data varies by participant and contributing data site; we used data from as early as 1979. Overall, we identified 59,645 pregnancy episodes among 31,726 unique *All of Us* participants. Of these episodes 30,175 were identified by both the HIP and PPS algorithms, 31,516 occurred since 2016, and 18,968 were classified as “data-rich episodes” (i.e., met both criteria). Concordance differed over time, with chronologically earlier pregnancies less likely to be identified by both algorithms or result in matching outcomes and dates (Figure 3). Among the data-rich episodes, concordance was higher, with 83.5% (n = 15,836) with a concordance score of 2 (fully concordant) and an additional 9.5% (n = 1,800) with a score of 1 (plausible gestational age but mismatched outcomes, e.g., live birth vs. delivery record only). To focus on the most reliable episodes, we present our main results for the data-rich episodes only, except when otherwise specified.

### Validation

There were 63,419 *All of Us* participants who answered “yes” (n = 4,680) or “no” (n = 58,739) to the pregnancy question (“Are you currently pregnant?”) on the “Overall Health” survey whose potential pregnancies we could capture in EHR data. Of those reporting current pregnancy, 3,832 (sensitivity = 81.8%) had an identified pregnancy episode that

overlapped the survey date within a buffer of two weeks. Compared to the 848 survey respondents reporting a pregnancy we did not identify as overlapping, those with pregnancies we did identify as overlapping had more EHR data both overall (mean 90.1 codes vs. 35.9 codes,  $p < 0.001$ ) and specific to pregnancy (mean 12.0 codes vs. 2.3 codes of those used in HIP algorithm,  $p < 0.001$ ). In addition, more of their EHR codes occurred post-survey (mean time post-survey 62 days vs. -7 days,  $p < 0.001$ ) (Supplementary Table 1).

Specificity was over 99%, with 518 respondents reporting no pregnancy despite our algorithm identifying them as more than 12 weeks pregnant. Of these, the median time from the survey date to the identified pregnancy end date was 14 days (interquartile range 0, 20 days). One possible explanation consistent with these data is that these participants took the survey soon after pregnancy, but the dates related to the pregnancy outcome codes in their EHR data were delayed (e.g., as in Figure 1F). In other settings it appeared that pregnancy-related codes were carried forward for years after their first occurrence, leading to multiple false positive episodes. The positive predictive value was 83.1%, and the negative predictive value was 98.5%. Changing the estimated gestational age threshold at which we started including survey responses to 20 weeks reduced the positive predictive value to 71.9%, as many more true positives than false positives were excluded.

Overlap of the 25 clinician-curated concepts not used in the HIPPS algorithm to identify pregnancies was comparable to the overlap reported by Jones et al.[19], though was lower, as expected, when we only considered the data-rich episodes (Supplementary Tables 2-3). For example, 97.7% of occurrences of “Breech presentation” overlapped any

pregnancy episode with the appropriate timing, while there was 82.4% overlap with data-rich episodes; Jones et al. reported 95.2% overlap.

### ***All of Us* pregnancy and participant characteristics**

The majority of the data-rich pregnancy episodes ended in live birth (n = 11,385; 60.0%); 26.6% (n = 5,043) were ongoing or missing an outcome (Figure 4; Supplementary Table 4). Gestational age was able to be dated to within a week by multiple codes for 55.5% (n = 10,519) of pregnancy episodes; just 3.8% (n = 729) had non-specific gestational duration information (Supplementary Table 5).

Of the 14,237 *All of Us* participants with at least one data-rich pregnancy episode, most were Hispanic or Latino (43.2%) or non-Hispanic White (33.6%) (Table 1). Pregnant people represented 41 US states and territories, but over half had fewer than 10 participants; over two-thirds were from just four states: Arizona (24.1%), New York (22.3%), California (11.6%), and Massachusetts (9.1%). The vast majority identified as women (99.4%) and heterosexual (92.0%) (Table 1).

In the regression analysis, we found that people with incomes greater than \$100,000 per year were most likely to have more than one pregnancy episode captured (probability ratio 0.79 vs \$50-100,000; 95% CI 0.70, 0.89), as were those who were married/partnered compared to divorced/separated/widowed (0.84; 95% CI 0.70, 1.00) or to never married (0.75; 95% CI 0.67, 0.84) (Supplementary Table 6). After 30-34 years old, there was a decline with age in the probability that a pregnancy episode ended in a live birth, with probability ratios of 0.94 (95 % CI 0.87, 1.00) at 35-39 years and 0.84 (95 % CI 0.74, 0.94) at

40-44 years (Supplementary Table 7). Black participants were more likely to have preterm deliveries compared to participants who were Hispanic or Latino (probability ratio 1.24 ; 95% CI 1.06, 1.43), as were older compared to younger participants (Supplementary Table 8).

Almost all participants with data-rich pregnancy episodes completed the Lifestyle and Overall Health surveys along with the required Basics survey (Table 2). In addition, 36.1% completed the Personal and Family Medical survey, 35.3% the Healthcare Access survey, and 14.6% the Social Determinants of Health survey. Few have contributed fitness tracker data during their pregnancy (n = 211 with activity data; n = 176 with heart rate data; n = 195 with sleep data), but 88.9% have some genomic data available. Almost half of pregnancy episodes (49.4%) occurred before joining *All of Us*, but a substantial number joined during pregnancy (23.5%) or had prospective pregnancy episodes (26.8%).

*Table 1. Demographic characteristics of All of Us participants with data-rich pregnancy episodes and of live births, compared to the United States population distribution of live births (vital statistics data). Live births may represent more than one pregnancy from the same participant. Vital statistics data has been standardized to the distribution of delivery years in the All of Us data.*

	<b>Individuals</b>	<b>Live births</b>	<b>United States population</b>
	n = 14,237	n = 11,385	
<b>Gender identity</b>			
Woman	14,018 (99.4%)	11,239 (99.5%)	
Other/multiple	52 (0.4%)	-- <sup>a</sup>	
Man	28 (0.2%)	-- <sup>a</sup>	
Unknown	139	93	
<b>Sexual orientation</b>			
Straight	12,795 (92.0%)	10,455 (93.6%)	



	<b>Individuals</b>	<b>Live births</b>	<b>United States population</b>
	n = 14,237	n = 11,385	
Bisexual	833 (6.0%)	544 (4.9%)	
None	195 (1.4%)	119 (1.1%)	
Gay/lesbian	89 (0.6%)	49 (0.4%)	
Unknown	325	218	
<b>Race/ethnicity<sup>b</sup></b>			
Hispanic or Latino	6,044 (43.2%)	5,407 (48.2%)	(23.7%)
White	4,702 (33.6%)	3,422 (30.5%)	(52.7%)
Black or African-American	2,244 (16.0%)	1,582 (14.1%)	(14.7%)
Asian	465 (3.3%)	370 (3.3%)	(6.4%)
More than one race	304 (2.2%)	247 (2.2%)	(2.2%)
Other	118 (0.8%)	83 (0.7%)	--
Middle Eastern or North African	95 (0.7%)	79 (0.7%)	--
Native Hawaiian or Other Pacific Islander	25 (0.2%)	24 (0.2%)	(0.2%)
Unknown	240	171	(2.21%)
<b>Family income (\$)</b>			
< 10k	2,054 (20.7%)	1,435 (19.1%)	
10-25k	1,460 (14.7%)	1,106 (14.7%)	
25-50k	2,105 (21.2%)	1,727 (23.0%)	
50-100k	1,905 (19.2%)	1,488 (19.8%)	
> 100k	2,404 (24.2%)	1,743 (23.2%)	
Unknown	4,309	3,886	
<b>Education</b>			
Less than high school	1,566 (11.2%)	1,258 (11.2%)	(11.8%)

	<b>Individuals</b>	<b>Live births</b>	<b>United States population</b>
	n = 14,237	n = 11,385	
High school graduate	7,515 (53.9%)	6,243 (55.8%)	(53.3%)
College graduate	2,651 (19.0%)	2,040 (18.2%)	(21.0%)
Advanced degree	2,213 (15.9%)	1,644 (14.7%)	(12.6%)
Unknown	292	200	
<b>Maternal age <sup>c</sup></b>			
15-19 years	--	386 (3.4%)	(4.42%)
20-24 years	--	2,199 (19.3%)	(18.6%)
25-29 years	--	3,196 (28.1%)	(28.7%)
30-34 years	--	3,201 (28.1%)	(29.5%)
35-39 years	--	1,905 (16.7%)	(15.4%)
40-44 years	--	462 (4.1%)	(3.15%)
45 years and over	--	36 (0.3%)	(0.16%)

<sup>a</sup> Values omitted due to small cell counts, following *All of Us* data policy

<sup>b</sup> Race/ethnicity categories are non-exclusive; all categories apart from Hispanic or Latino are non-Hispanic

<sup>c</sup> Maternal age not presented for individuals, as age varied across pregnancies

Table 2. Additional *All of Us* data contributed by participants with identified data-rich pregnancy episodes.

<b>Timing of pregnancy episode relative to <i>All of Us</i> participation</b>	
(n = 18,968 episodes)	
Before	9368 (49.4%)
During	4456 (23.5%)
After	5076 (26.8%)
Unclear	68 (0.4%)

---

**Fitness tracking device data during pregnancy episode**

(n = 18,968 episodes)

---

Activity	211 (1.1%)
Sleep	195 (1.0%)
Heart Rate	176 (0.9%)

---

**Genomic data**

(n = 14,237 individuals)

---

Array data	12,653 (88.9%)
Whole genome variant data	10,260 (72.1%)
Long read whole genome variant	48 (0.3%)
Structural variant	506 (3.6%)

---

**Survey data**

(n = 14,237 individuals)

---

The Basics	14,237 (100%)
Lifestyle	14,233 (100%)
Overall Health	14,233 (100%)
Personal/Family Health History	5,140 (36.1%)
Healthcare Access and Utilization	5,022 (35.3%)
Social Determinants of Health	2,072 (14.6%)

---

**Live births**

Among live births, the median gestational age was 38.6 (interquartile range 37.1, 39.4), and 20.5% were inferred to have delivered before 37 weeks gestation (i.e., preterm).

Compared to vital statistics data from the same years (Table 1), live births in *All of Us* were to slightly older (21.1% vs. 18.7% 35 years or greater) and more educated (14.7% vs. 12.6%

with a graduate degree) individuals. Similar proportions of births were to Black, Native Hawaiian/Pacific Islander, or individuals with more than one race. However, compared to national data, *All of Us* had a smaller proportion of Asian (3.3% vs. 6.4%) and non-Hispanic White (30.5% vs 52.7%), and more Hispanic or Latino (48.2% vs. 23.7%) people who had given birth. Vital statistics do not capture Middle Eastern/North African ethnicity, but 95 *All of Us* participants reporting that ethnicity had data-rich pregnancy episodes (Table 1).

## **DISCUSSION**

In this study, we used *All of Us* EHR data to identify pregnancy episodes and estimate gestational age. In doing so, we validated an algorithm recently developed for use in OMOP CDM data and demonstrated the capability of the *All of Us* data to support pregnancy research with a diverse sample of pregnant people.

### **Promise and potential of *All of Us* multi-source data**

The rich combination of data sources enables a more nuanced understanding of factors affecting maternal health. By integrating EHR data with survey responses, genetic information, and physical activity data, the *All of Us* Research Program allows for a deeper examination of how social factors, health history, healthcare access, genetic predispositions, and more impact pregnancy and postpartum health.

In particular, the over 4000 people who completed *All of Us* surveys during a pregnancy for which there was high-quality electronic health record data make for a significant subset that may also have EHR or other data through the postpartum period,

allowing researchers to ask questions about relationships between social factors such as social support or experiences of discrimination in pregnancy and postpartum health and well-being. This research is crucial given the known racial disparities in maternal health in the US, including higher rates of maternal morbidity and mortality among Black and Indigenous people. By analyzing the diverse sample *All of Us* has recruited, researchers can help identify how systemic issues and social determinants of health contribute to these disparities and develop specific interventions to address their causes.

Other sources of data, including genomic and activity device data, provide additional opportunities to answer critical pregnancy-related questions. For example, preterm delivery has numerous causes but this heterogeneity makes it difficult to predict;[27,28] linking the inferred gestational ages of the *All of Us* pregnancies with these sources of big data might produce new insights. Future data types on the *All of Us* data roadmap[29] include self-reported height and weight, activity tracker data from Apple's popular platform, and data from a nutrition substudy, all of which could provide more data to study predictors and outcomes associated with weight and nutrition during preconception, pregnancy, and postpartum.

Activity tracker data shows promise not only for research on physiologic changes during pregnancy but also for inclusion in an improved pregnancy identification algorithm. Epstein and McCoy used EHR and fitness tracker data in 89 *All of Us* participants to observe the change in heart rate for pregnant people and found a peak in heart rate during the first and third trimesters and a steady increase through the second trimester.[30] Given

potential variations in daily exercise, heart rate, and sleep duration, integrating data from wearable devices could augment the algorithm's effectiveness.

### **Strengths and limitations of EHR pregnancy data**

Real-world data like the EHR data used in this study offers valuable insights into pregnancies and health-related characteristics of pregnant people. Unlike pregnancy-specific studies, it can span years of patient health records, enabling a comprehensive understanding of medical history, including pre-pregnancy and post-pregnancy phases, and providing a holistic view of participant health. Furthermore, compared to clinical trial data, which generally has stringent inclusion/exclusion criteria, EHR data like that in *All of Us* better represents real-world populations, including underrepresented groups, fostering a more inclusive, reliable, and comprehensive approach to research.[31] In addition, the validity of gestational age diagnosis codes in characterizing pregnancies is critical to pregnancy and perinatal research, as gestational age is one of the most important factors in clinical decision-making and in neonatal prognosis.[32]

However, EHR data comes with its own limitations. Firstly, coding errors lead to incomplete or inaccurate documentation of patient records. Certain medical conditions may not be fully captured or explained by these codes, leaving important information conveyed through free text, which is not included in *All of Us* data. Secondly, patients often receive care from multiple healthcare systems, resulting in fragmented records that result in missing information. For instance, a patient receiving prenatal care at one hospital may need emergency labor services at a different hospital within another health system.

Although the *All of Us* Research Program harmonizes data from multiple systems, not all sites where patients receive care contribute data. Indeed, over one-quarter of the otherwise data-rich pregnancy episodes we identified were missing an outcome, though in some cases this was likely due to pregnancies that continued past the data cut-off point.

In our survey-based validation substudy that included people who joined *All of Us* while pregnant, we estimated sensitivity exceeding 80% and specificity approaching 100%, affirming the Jones et al approach for reliably identifying pregnancy episodes. Other studies have also reported high agreement rates and positive predictive values in line with our study, from 70% to close to 100%.<sup>[19,33–37]</sup> However, we could not specifically validate gestational age at the outcome, which may be less accurately identified than the outcome itself, particularly for non-live birth outcomes.<sup>[23]</sup> In addition, our approach to validation primarily relied on EHR information occurring around the time participants joined *All of Us* and took the surveys, when their healthcare is more likely to be occurring within systems that contribute data to *All of Us*, leading to an overestimate of the likely sensitivity of the algorithm over the entire scope of the data.

As with other pregnancy algorithms, HIPPS leverages medical codes that represent key factors such as prenatal care procedures, gestational age, and a range of pregnancy outcomes. The algorithm was developed for data that has been translated to the OMOP CDM, which is made up of a common vocabulary of concept codes representing other code sets, including ICD-9, ICD-10, and CPT codes. This makes the algorithm highly transferable to different settings and across time. Indeed, we identified pregnancies as early as the 1980s despite changes in medical coding since then. However, the early

episodes were of notably lower quality based on concordance between algorithms and estimated gestational age, reflecting improvements in electronic health record keeping and the usage of gestational age-specific ICD-10 codes.[38,39]

### **Recommendations and future directions**

Properly accounting for timing is critical in pregnancy-related research and will be even more so in *All of Us* studies, as different data components are contributed at different times relative to a given pregnancy. Half of the pregnancies we identified occurred before a participant joined *All of Us*, limiting the sample size for research questions in which exposures of interest and covariates may change over time and are drawn from survey questions. Researchers should be careful not to make analyses conditional on joining *All of Us* post-pregnancy; for example, pre-*All of Us* pregnancies are guaranteed not to have resulted in maternal mortality. Nonetheless, some of the survey responses (e.g., race/ethnicity) can be combined with EHR regardless of timing, as can genetic data. As *All of Us* grows, we can expect more prospective pregnancies to occur.

While *All of Us* aims to be inclusive, it is not necessarily representative. We found that in several respects, demographic data on pregnancies in *All of Us* did not match that from vital statistics. While not inherently a problem, researchers should consider selection as a source of bias in their studies, thinking carefully about who is joining *All of Us* and contributing each type of data. While targeted recruitment is necessary to meet the Program's commitment to outreach to underrepresented communities, a resulting lack of geographic diversity suggests a possible lack of diversity in other, unmeasured respects. In



future work, we will consider how to address possible biases due to selection and missing data and improve the generalizability of the data.

Although the algorithm we used was not perfectly accurate even according to our limited validation exercises, the use of an algorithm like this one represents an improvement compared to a simple code search for pregnancy or delivery-related codes. While live birth is a relatively straightforward outcome to recognize, other outcomes, such as ectopic pregnancy, require more supporting information before a single code should be considered indicative of an event.[40] In informal reviews of some participants' medical histories, we found the same code referring to ectopic pregnancy or miscarriage repeated for years with no other indication of pregnancy, suggesting that in some cases these codes are carried forward in the problem list without representing new events. Future research on pregnancies that do not end in live birth will involve more thorough review to assess the accuracy of the algorithm for these outcomes. Furthermore, we followed Jones et al.[19] in combining spontaneous and induced abortion in the presentation, as distinguishing the two brings additional challenges. Given the changing landscape of abortion access in the United States, addressing these challenges represents an important future research contribution.

## **CONCLUSION**

The Jones et al. pregnancy algorithm can be used by the community of researchers working on *All of Us* to identify pregnancy episodes and ask novel questions about experiences longitudinally with fertility, pregnancy, birth, and the postpartum and long-

term health of a diverse sample of pregnant people. However, limitations of electronic health record data, *All of Us* survey and measurement timing, and selection into the study should be thoughtfully considered in such research.

## **ACKNOWLEDGMENTS**

We gratefully acknowledge *All of Us* participants for their contributions, without whom this and future pregnancy-related research would not be possible. We also thank the National Institutes of Health's *All of Us* Research Program for making available the data examined in this study. The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276.

## FIGURE CAPTIONS

Figure 1. Schematic describing the validation study using survey data and scenarios resulting in errors. The participants represented in A and B reported being pregnant during, or within two weeks of, an assigned pregnancy episode, representing true positives. Participant C reported not being pregnant during a time not overlapping any identified pregnancy episode, representing a true negative. Participants D and E reported being pregnant at points in time not overlapping with any identified pregnancy episode (false negatives), either because there was no EHR data at that time (D) or because no outcome data resulted in a misaligned episode. False positives are represented by F and G, either due to no corresponding true pregnancy in E (e.g., if historical codes are carried forward in the EHR) or due to misalignment in G (e.g., because of delayed dates in the EHR data). Participant H answered the survey before 12 weeks of an assigned pregnancy episode, meaning we did not know whether they would have known they were pregnant at the time (these situations were excluded from the false-negative calculations no matter the response).

Figure 2. *All of Us* participant flow diagram: eligibility, validation substudy, and pregnancy episode identification. Boxes refer to numbers of individuals, and dashed boxes refer to numbers of pregnancy episodes.

Figure 3. Concordance between pregnancy identification algorithms by date across all pregnancies. Fully concordant pregnancies (score = 2) have matching HIP and PPS

outcomes, similar end dates (within 14 days), and a plausible gestational age. Somewhat (score = 1) and not concordant (score = 0) episodes differed on outcome category or timing. Additional episodes were identified by only one of the two algorithms. We included episodes starting after 2016 (dashed line) that were identified by both algorithms in our main analysis.

Figure 4. Distribution of pregnancy outcomes among all identified pregnancies (left panel) and the data-rich episodes (right panel). Episodes are stratified by the precision of the gestational age information used to assign pregnancy timing. Week-level pregnancy episodes were able to be dated to within less than a month; those with poor support only had a single week-specific code. Month-level episodes were dated to within 1-3 months, and non-specific episodes were less precise. In addition, shaded portions of the bars represent the highly concordant episodes, on which both HIP and PPS algorithms agreed on outcome and timing, and light-colored portions represent partial or non-concordance.

## REFERENCES

- 1 Bill & Melinda Gates Foundation, National Institutes of Health. Women's Health Innovation Opportunity Map 2023: 50 High-Return Opportunities to Advance Global Women's Health R&D. 2023. [https://orwh.od.nih.gov/sites/orwh/files/docs/womens-health-rnd-opportunity-map\\_2023\\_508.pdf](https://orwh.od.nih.gov/sites/orwh/files/docs/womens-health-rnd-opportunity-map_2023_508.pdf) (accessed 1 June 2024)
- 2 Shulman HB, D'Angelo DV, Harrison L, *et al.* The Pregnancy Risk Assessment Monitoring System (PRAMS): Overview of Design and Methodology. *Am J Public Health.* 2018;108:1305–13. doi: 10.2105/AJPH.2018.304563
- 3 Schoendorf KC, Branum AM. The use of United States vital statistics in perinatal and obstetric research. *Am J Obstet Gynecol.* 2006;194:911–5. doi: 10.1016/j.ajog.2005.11.020
- 4 Ziogas C, Hillyer J, Saftlas AF, *et al.* Validation of birth certificate and maternal recall of events in labor and delivery with medical records in the Iowa health in pregnancy study. *BMC Pregnancy Childbirth.* 2022;22:232. doi: 10.1186/s12884-022-04581-7
- 5 Kishi R, Kobayashi S, Ikeno T, *et al.* Ten years of progress in the Hokkaido birth cohort study on environment and children's health: cohort profile—updated 2013. *Environ Health Prev Med.* 2013;18:429–50. doi: 10.1007/s12199-013-0357-3
- 6 Magnus P, Irgens LM, Haug K, *et al.* Cohort profile: The Norwegian Mother and Child Cohort Study (MoBa). *Int J Epidemiol.* 2006;35:1146–50. doi: 10.1093/ije/dyl170
- 7 Niswander KR, Gordon MJ, Gordon M. *The Women and Their Pregnancies: The Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke.* National Institute of Health 1972.
- 8 Wise LA, Rothman KJ, Mikkelsen EM, *et al.* Design and conduct of an internet-based preconception cohort study in north america: pregnancy study online. *Paediatr Perinat Epidemiol.* 2015;29:360–71. doi: 10.1111/ppe.12201
- 9 Voorst SF van, Vos AA, Jong-Potjer LC de, *et al.* Effectiveness of general preconception care accompanied by a recruitment approach: protocol of a community-based cohort study (the Healthy Pregnancy 4 All study). *BMJ Open.* 2015;5:e006284. doi: 10.1136/bmjopen-2014-006284
- 10 Spry E, Olsson CA, Hearps SJC, *et al.* The Victorian Intergenerational Health Cohort Study (VIHCS): Study design of a preconception cohort from parent adolescence to offspring childhood. *Paediatr Perinat Epidemiol.* 2020;34:86–98. doi: 10.1111/ppe.12602

- 11 Loo EXL, Soh S-E, Loy SL, *et al.* Cohort profile: Singapore Preconception Study of Long-Term Maternal and Child Outcomes (S-PRESTO). *Eur J Epidemiol.* 2021;36:129–42. doi: 10.1007/s10654-020-00697-2
- 12 Daw JR, Auty SG, Admon LK, *et al.* Using Modernized Medicaid Data to Advance Evidence-Based Improvements in Maternal Health. *Am J Public Health.* 2023;113:805–10. doi: 10.2105/AJPH.2023.307287
- 13 The “All of Us” research program. *N Engl J Med.* 2019;381:668–76. doi: 10.1056/nejmsr1809937
- 14 Gomez SE, Sarraju A, Rodriguez F. Racial and Ethnic Group Underrepresentation in Studies of Adverse Pregnancy Outcomes and Cardiovascular Risk. *J Am Heart Assoc.* 2022;11:e024776. doi: 10.1161/JAHA.121.024776
- 15 Girardi G, Longo M, Bremer AA. Social determinants of health in pregnant individuals from underrepresented, understudied, and underreported populations in the United States. *Int J Equity Health.* 2023;22:186. doi: 10.1186/s12939-023-01963-x
- 16 Moseson H, Fix L, Hastings J, *et al.* Pregnancy intentions and outcomes among transgender, nonbinary, and gender-expansive people assigned female or intersex at birth in the United States: Results from a national, quantitative survey. *Int J Transgender Health.* 2020;22:30–41. doi: 10.1080/26895269.2020.1841058
- 17 Fleszar LG, Bryant AS, Johnson CO, *et al.* Trends in State-Level Maternal Mortality by Racial and Ethnic Group in the United States. *JAMA.* 2023;330:52–61. doi: 10.1001/jama.2023.9043
- 18 Hoyert D. Maternal mortality rates in the United States, 2021. National Center for Health Statistics (U.S.) 2023. <https://doi.org/10.15620/cdc:124678>
- 19 Jones SE, Bradwell KR, Chan LE, *et al.* Who is pregnant? Defining real-world data-based pregnancy episodes in the National COVID Cohort Collaborative (N3C). *JAMIA Open.* 2023;6:ooad067. doi: 10.1093/jamiaopen/ooad067
- 20 Survey Explorer – All of Us Research Hub. <https://www.researchallofus.org/data-tools/survey-explorer/> (accessed 27 June 2024)
- 21 Stang PE, Ryan PB, Racoosin JA, *et al.* Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Ann Intern Med.* 2010;153:600. doi: 10.7326/0003-4819-153-9-201011020-00010
- 22 Observational Health Data Sciences and Informatics. *The Book of OHDSI.* 2021. <https://ohdsi.github.io/TheBookOfOhdsi/> (accessed 4 April 2024)

- 23 Matcho A, Ryan P, Fife D, *et al.* Inferring pregnancy episodes and outcomes within a network of observational databases. *PLOS ONE*. 2018;13:e0192033. doi: 10.1371/journal.pone.0192033
- 24 Wickham H, Girlich M, Ruiz E. dbplyr: A “dplyr” back end for databases. 2024. <https://CRAN.R-project.org/package=dbplyr>
- 25 Smith LH, Cavanaugh R. allofus R package. 2023. <https://doi.org/10.5281/ZENODO.10420610>
- 26 Centers for Disease Control and Prevention, National Center for Health Statistics. National vital statistics system, natality. <http://wonder.cdc.gov/natality-expanded-current.html> (accessed 27 November 2023)
- 27 Romero R, Dey SK, Fisher SJ. Preterm labor: One syndrome, many causes. *Science*. 2014;345:760–5. doi: 10.1126/science.1251816
- 28 Mitrogiannis I, Evangelou E, Efthymiou A, *et al.* Risk factors for preterm birth: an umbrella review of meta-analyses of observational studies. *BMC Med*. 2023;21:494. doi: 10.1186/s12916-023-03171-4
- 29 Data Sources – All of Us Research Hub. <https://www.researchallofus.org/data-tools/data-sources/> (accessed 4 April 2024)
- 30 Modde Epstein C, McCoy TP. Linking Electronic Health Records With Wearable Technology From the All of Us Research Program. *J Obstet Gynecol Neonatal Nurs*. 2023;52:139–49. doi: 10.1016/j.jogn.2022.12.003
- 31 Ramirez AH, Sulieman L, Schlueter DJ, *et al.* The All of Us Research Program: Data quality, utility, and diversity. *Patterns*. 2022;3:100570. doi: 10.1016/j.patter.2022.100570
- 32 Leonard SA, Panelli DM, Gould JB, *et al.* Validation of ICD-10-CM Diagnosis Codes for Gestational Age at Birth. *Epidemiology*. 2023;34:64–8. doi: 10.1097/EDE.0000000000001557
- 33 Canelón SP, Burris HH, Levine LD, *et al.* Development and evaluation of MADDIE: method to acquire delivery date information from electronic health records. *Int J Med Inf*. 2021;145:104339. doi: 10.1016/j.ijmedinf.2020.104339
- 34 Chomistek AK, Phiri K, Doherty MC, *et al.* Development and Validation of ICD-10-CM-based Algorithms for Date of Last Menstrual Period, Pregnancy Outcomes, and Infant Outcomes. *Drug Saf*. 2023;46:209–22. doi: 10.1007/s40264-022-01261-5



- 35 Zhu Y, Bateman BT, Hernandez-Diaz S, *et al.* Validation of claims-based algorithms to identify non-live birth outcomes. *Pharmacoepidemiol Drug Saf.* 2023;32:468–74. doi: 10.1002/pds.5574
- 36 Devine S, West S, Andrews E, *et al.* The identification of pregnancies within the general practice research database. *Pharmacoepidemiol Drug Saf.* 2010;19:45–50. doi: 10.1002/pds.1862
- 37 Li Q, Andrade SE, Cooper WO, *et al.* Validation of an algorithm to estimate gestational age in electronic health plan databases. *Pharmacoepidemiol Drug Saf.* 2013;22:524–32. doi: 10.1002/pds.3407
- 38 Sarayani A, Wang X, Thai TN, *et al.* Impact of the Transition from ICD–9–CM to ICD–10–CM on the Identification of Pregnancy Episodes in US Health Insurance Claims Data. *Clin Epidemiol.* 2020;Volume 12:1129–38. doi: 10.2147/CLEP.S269400
- 39 Ailes EC, Zhu W, Clark EA, *et al.* Identification of pregnancies and their outcomes in healthcare claims data, 2008–2019: An algorithm. *PLOS ONE.* 2023;18:e0284893. doi: 10.1371/journal.pone.0284893
- 40 Scholes D, Yu O, Raebel MA, *et al.* Improving automated case finding for ectopic pregnancy using a classification algorithm. *Hum Reprod.* 2011;26:3163–8. doi: 10.1093/humrep/der299