TRUVETA

Poster: RWD2

# A hierarchical algorithm to identify pregnancy start and duration using structured EHR data

Katherine Brown, PhD, MSN, RN; Amy Sullivan, MS; Esther Kim, PhD; Nadia Tabatabaeepour, MPH; Katherine Kendrick, MPH; Jordan Swartz, MD; Sarah Platt, MS; Sunny Guin, PhD; Emily Webber, PhD

Disclosures: All authors are employees of Truveta Inc.

## Background

Pregnant individuals are routinely excluded from Phase I–III clinical trials, resulting in limited evidence on the safety, dosing, and effectiveness of most therapeutics during pregnancy.

Real-world data (RWD) studies enable evaluation of treatment exposures and pregnancy outcomes in large, heterogeneous populations, providing critical evidence to address gaps left by traditional clinical research.

Accurate identification of pregnancy start dates and durations is critical for real-world studies evaluating treatment exposures, healthcare utilization, and maternal-infant outcomes. However, estimating pregnancy start is challenging from structured electronic health record (EHR) data.

**Objective:** Adapt hierarchical algorithm to estimate last menstrual period (LMP) and pregnancy duration using structured EHR data from Truveta.

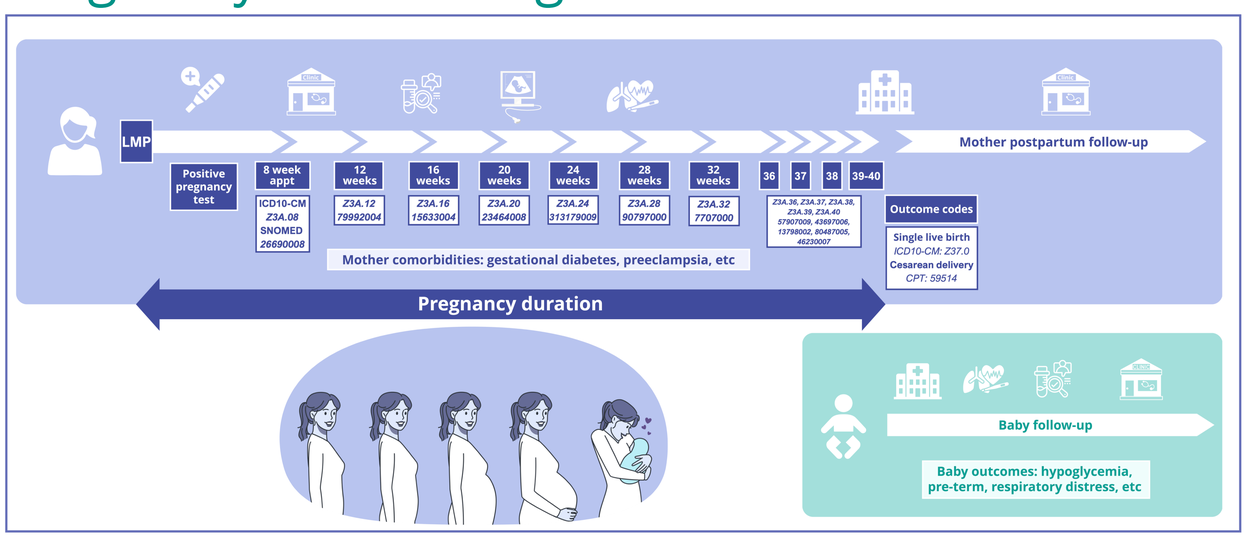
## Methods

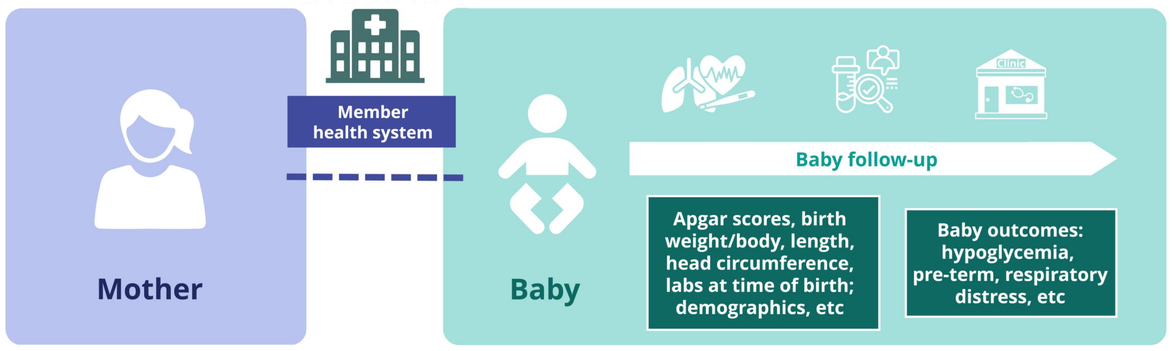
- Truveta Data is comprised of real-world US electronic health record (EHR) data, which is aggregated, normalized, and de-identified from US health care systems comprising clinics and hospitals.
- We used structured diagnosis and procedure records with deterministic mother-infant linkages.
- Women aged 12–55 with pregnancy-related codes were included. Two primary methods were used:

a. Gestational age codes—specifically ICD-10-CM Z3A.xx and SNOMED CT equivalents—to estimate weeks of gestation by counting backward from code dates[1].

b. Outcome-based methods used codes for live births and pregnancy losses. The algorithm prioritized four LMP estimation approaches from the two methods[1,2].

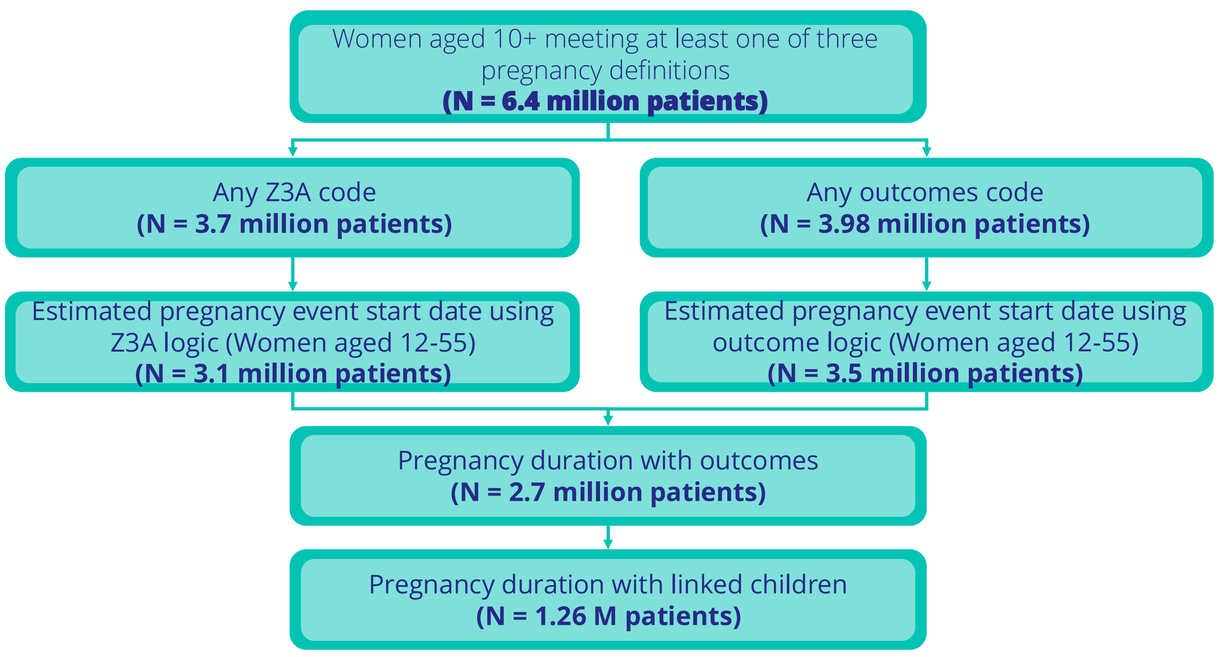
## Pregnancy duration algorithm




**Pregnancy duration algorithm and mother–infant linkage:** *Pregnancy duration is estimated using sequential Z3A codes and outcome codes to define delivery. Maternal comorbidities are captured across gestation, and deterministic linkage connects mother and infant records, enabling analysis of birth characteristics and infant outcomes.*

## Results

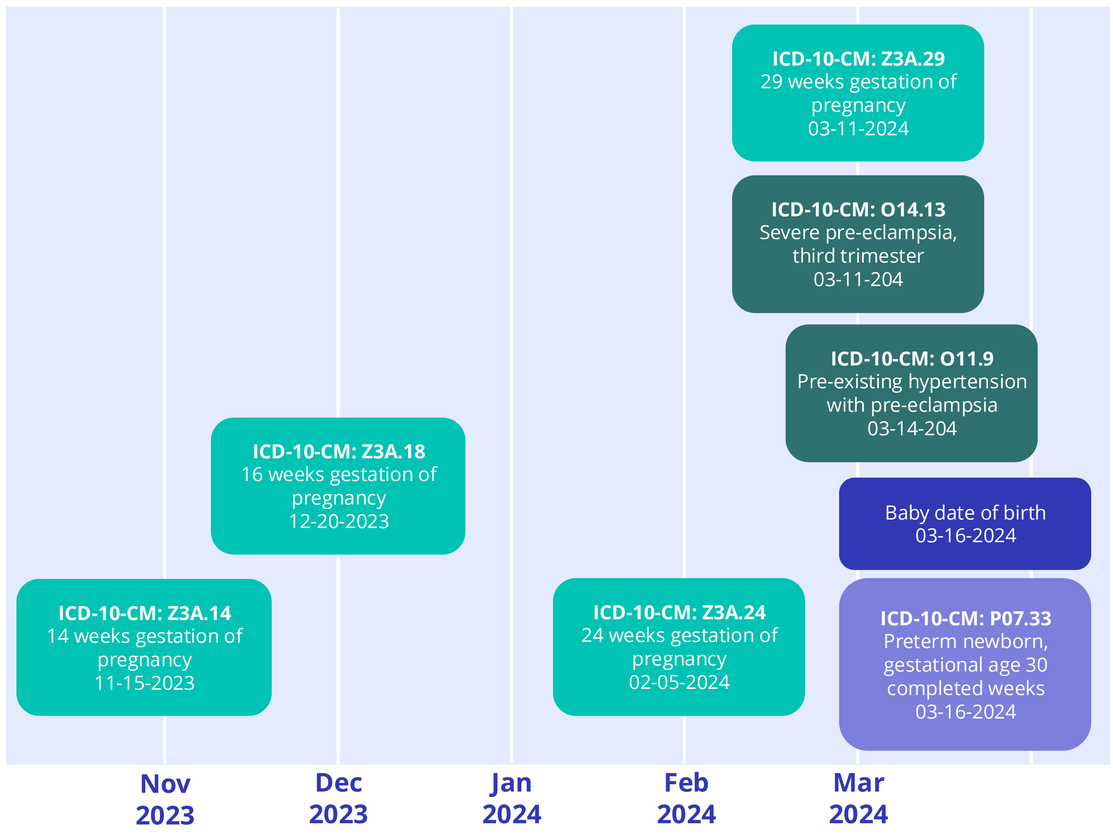



**Pregnancy episode cohort attrition:** *Stepwise attrition of women with pregnancy-related codes to final linked pregnancy episodes with outcomes and infants.*

| Pregnancy outcome | Overall | |
|---|---|---|
| | Records (N=3,488,844) | Patients (N = 2,913,469) |
| Live birth | 2,847,320 (81.6%) | 2,343,002 (80.4%) |
| Linked births | 1,525,750 (43.7%) | 1,255,868 (43.1%) |
| Pregnancy Loss | 641,524 (18.4%) | 570,467 (19.6%) |

**Pregnancy outcomes distribution:** *Among 2.9M patients, 80.4% live births and 19.6% pregnancy losses.*

**Patient timeline with pregnancy duration & mother/baby outcomes**




**Example patient timeline for pregnancy start, duration, and outcomes:** *Gestational age codes are used to estimate pregnancy start and progression, with delivery defined by the baby date of birth (or birth outcome codes). This enables linking maternal comorbidities and infant outcomes (e.g., pre-eclampsia, preterm birth, neonatal conditions, etc.) for comprehensive longitudinal analyses.*

## Results

- Among 6.4 million women with pregnancy-related codes, 3.1 million pregnancy start dates were estimated using biologically plausible durations.
- Approximately 3.5M pregnancy episodes (2.9M women) had outcome codes (2.85M live births; 0.64M non-live births), with 1.26 million linked to infant records. SNOMED codes contributed an additional 500K episodes.
- The resulting dataset includes pregnancy timing, outcome type, and infant linkage, enabling longitudinal analyses of care, treatment effects, and child outcomes.

## Conclusions

- This multi-step algorithm supports robust pregnancy episode construction using structured EHR data. Its hierarchical design improves completeness and precision, particularly when combined with mother-infant linkage.
- Future work includes clinician-reviewed validation of estimated pregnancy start dates and durations using note-based gold standards.
- This methodology enables scalable, reproducible cohort creation for regulatory-grade research, including pregnancy PASS and drug safety evaluations.

Truveta has adapted an algorithm to identify pregnancy start and duration using its mother-child linked dataset for regulatory-grade research, including post-approval pregnancy safety studies.

REFERENCES: 1. Moll K, Wong HL, Fingar K, Hobbi S, Sheng M, Burrell TA, Eckert LO, Munoz FM, Baer B, Shoaibi A, Anderson S. Validating claims-based algorithms determining pregnancy outcomes and gestational age using a linked claims-electronic medical record database. *Drug Saf*. 2021 Nov;44(11):1151-1164. doi:10.1007/s40264-021-01113-8