Developing phenotypes across pregnant persons and infants: Utilizing pregnancy episode identification and mother-infant linkage algorithms to define outcomes

Rupa Makadia[1,2], Jill Hardin[1,2], Kevin Haynes[1], Dave Kern[1], Amir Sarayani[1], Melanie Jacobson[1]

1. Johnson and Johnson, Raritan, NJ
2. Observational Health Data Sciences and Informatics, New York, NY

Background

Maternal and infant outcomes play a crucial role in obstetric care, ensuring the well-being and health of both the pregnant person and the developing fetus. Identification of adverse pregnancy outcomes including preterm birth, small for gestational age, and congenital malformations, is challenged by the ability to accurately phenotype a pregnancy episode and capture coded diagnoses which may appear on the maternal or infant record. Several groups have developed pregnancy episode identification algorithms that impute the start of pregnancy (i.e., last menstrual period), before a marker indicating the end of pregnancy is observed (1). In addition, mother-infant linkage algorithms allow people who gave birth to be linked to their infant(s) in insurance claims databases (2). By combining both pregnancy episode and linkage algorithms, phenotypes can be developed for both pregnant people and infants. In recent years, there has been a growing body of literature utilizing observational data to develop phenotypes for outcomes such as: preterm birth, small for gestational age, and congenital malformations. However, few studies have evaluated whether preterm birth can be empirically estimated or how mother-infant linkage may affect this phenotype development process, including when and on whom (mother and/or infants) diagnostic codes occur. We aimed to combine the previous work of maternal-infant linkage and determination of pregnancy episodes with the development of adverse pregnancy outcome phenotypes.

Methods

Utilizing a pregnancy episode identification algorithm and a mother-infant linkage algorithm, we created cohorts in ATLAS for preterm birth, small for gestational age, and major congenital malformations among mother-infant (<1 year old) pairs using diagnostic codes on either mother or infant. CohortDiagnostics was used to evaluate the phenotype algorithms for the frequency of codes, incidence rates, and characterization. An additional phenotype for preterm birth was developed using an empirical derivation: pregnant people with live birth deliveries that resulted from gestations of less than 37 weeks (259 days) were identified. The databases used in this study included the Optum© De-Identified Clinformatics® Data Mart Database – Socioeconomic-(SES) (Clinformatics) and the IBM MarketScan® Commercial Claims (CCAE) databases.

Results

The results presented are for preterm birth for a single database (Clinformatics®) for brevity. Table 1 shows the top 5 SNOMED codes in Clinformatics in a cohort with pregnant people and infants with ≥1 code for preterm birth. The most common code occurred on infants (32%) and the second most common occurred on mothers at delivery (24%). The total number of persons identified by the phenotype and stratified by mother vs. infant is shown in Table 2. When restricted to linked mother-infant pairs, 70% of the cohort remained (n=308,887). Among the linked pairs where ≥1 preterm code was present, ~25% occurred only on the infant, ~10% occurred only on the pregnant person and 65% occurred on both records. Figure 2 shows the overlap between mothers with live deliveries who had ≥1 preterm birth code and those that had

calculated gestations <37 weeks. Among ~1.7 pregnant people with a live birth outcome, 17.5% had gestations estimated to be less than 37 weeks, 8.1% had ≥1 preterm code within 30 days of delivery and 6.2% had both ≥1 preterm birth code and a gestation estimated to be less than 37 weeks.

Conclusions

The ability to accurately identify phenotypes that include both the infant and pregnant person allows researchers to carefully examine outcomes that could be coded on either on pregnant people or infants. Utilizing the pregnancy identification algorithm to generate an empirical estimation of preterm birth (i.e., < 37 weeks) and comparing it with the more conventional use of codes also highlights a complexity in defining the phenotype. The use of a rigorous data-driven approach applied to multiple databases provides confidence that the phenotype algorithm can correctly identify preterm birth infants. These phenotypes are critical for use in maternal-fetal research.

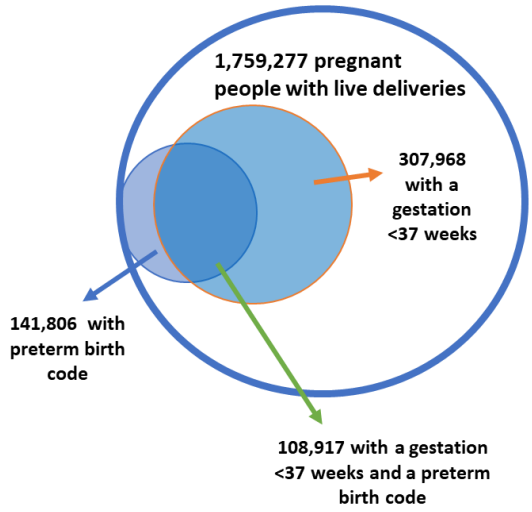
Table 1. Percentage of codes in Optum DOD with pre-term birth codes in cohort.

Database	Concept Name	Vocabulary	% of codes utilized in cohort
Optum DOD	Prematurity of infant	SNOMED	32.2%
	Premature delivery	SNOMED	24.0%
	Premature pregnancy delivered	SNOMED	21.4%
	Neonatal jaundice associated with preterm delivery	SNOMED	18.6%
	Low birth weight infant	SNOMED	9.7%

Table 2. Counts of persons with preterm birth codes stratified by pregnant people and infants.

Optum DOD	Total patients with a preterm birth code (ATLAS)	Total by subgroup	Total linked patients with ≥1 preterm birth code	Total linked by subgroup	Total linked with code in mother OR infant (only one of pair has the code)	Total linked with code on mother AND infant (both have code)
Mother	454,147	142, 976	308, 887	120, 800	29, 975	90, 825
Infant		311, 171		188, 087	75, 763	112, 324

Figure 1. Empirical estimation of preterm birth



References

1. Matcho A, Ryan P, Fife D, Gifkins D, Knoll C, Friedman A. Inferring pregnancy episodes and outcomes within a network of observational databases. PloS one. 2018;13(2):e0192033.
2. Weaver HJ, Blacketer C, Krumme A, Jacobson, M, Ryan P. Development and evaluation of an algorithm to link mothers and infants in two US commercial healthcare claims databases for pharmacoepidemiology research. medRxiv : the preprint server for health sciences. 2022.