Received: 4 March 2021 | Revised: 7 August 2021 | Accepted: 1 September 2021

DOI: 10.1002/bdr2.1954

RESEARCH ARTICLE

Birth Defects Research | Society for Birth Defects Research & Prevention | WILEY

Development and evaluation of standardized pregnancy identification and trimester distribution algorithms in U.S. IBM MarketScan® Commercial and Medicaid data

Kelsey M. Sumner[1,2] | **Anna Ehlinger**[3] | **Mary E. Georgiou**[1] | **Keele E. Wurst**[1]

[1]Value Evidence Outcomes Epidemiology, GlaxoSmithKline, Research Triangle Park, North Carolina, USA

[2]Department of Epidemiology, University of North Carolina Gillings School of Global Public Health, Chapel Hill, North Carolina, USA

[3]Access and Customer Engagement Strategy Pricing, GlaxoSmithKline, Research Triangle Park, North Carolina, USA

Correspondence

Keele E. Wurst, Value Evidence Outcomes Epidemiology, GlaxoSmithKline, Research Triangle Park, NC 27709.

Email: keele.e.wurst@gsk.com

Funding information

GlaxoSmithKline, Grant/Award Number: PRJ3186

Abstract

Objectives: Creation of new algorithms to identify pregnancies in automated health care claims databases is of public health importance, as it allows us to learn more about medication use and safety in a vulnerable population. Previous algorithms were largely created using international classification of disease codes, but despite the U.S. code transition in 2015, few algorithms are available with the latest ICD-10-CM codes.

Methods: Using U.S. IBM MarketScan® Commercial Claims and Encounters and Multi-State Medicaid databases for women aged 10–64 years during 2014 and 2016, two pregnancy algorithms (ICD-9-CM and ICD-10-CM) were created using expert clinical review. The algorithms were evaluated by assessing the distribution of pregnancy outcomes (live birth and pregnancy losses) within each time-based cohort and the ability of the algorithms to identify select medication use during pregnancy. Medication exposure, demographics, comorbidities, and pregnancy outcomes were compared to published literature estimates for the two time periods.

Results: For the IBM MarketScan® Commercial database, the algorithms identified 687,228 pregnancies in 2014 and 444,293 in 2016. In the IBM MarketScan® Medicaid database, 389,132 pregnancies in 2014 and 406,418 in 2016 were identified. Percentages of most pregnancy outcomes identified using the algorithms were similar to national data sources; however, percentages of preterm births and pregnancy losses were not comparable. Most medication use estimates from the algorithms were similar to or higher than published estimates.

Conclusions: By incorporating the latest ICD-10-CM codes, the new algorithms provide more complete estimates of medication use during pregnancy than algorithms using the outdated codes.

KEYWORDS

drug safety, insurance claims data, pregnancy algorithm

1 | INTRODUCTION

Real-world evidence has been increasingly utilized for assessing the safety of medications in vulnerable populations (Swift et al., 2018). Pregnant women are a key population for real-world evidence studies, because medication use during pregnancy is rarely studied in randomized-controlled trials. Automated health care insurance claims databases carry a wealth of real-world information; however, they present many challenges for examining medication use in pregnancy. In insurance claims databases, there is limited information available to estimate the timing of exposure, date of the last menstrual period (LMP) are not routinely recorded in claims, gestational age is not always available, and there can be inconsistent medical codes billed (Margulis et al., 2015). Considering these challenges, there is a need for a standardized method of identification of pregnancies, pregnancy lengths, pregnancy outcomes, and pregnancy trimester ascertainment to ensure consistency across diseases and therapeutic areas.

In the last decade, many algorithms have attempted to create pregnancy algorithms for use in U.S. health insurance claims data (Ailes, Simeone, Dawson, Petersen, & Gilboa, 2016; Ellis et al., 2020; Hornbrook et al., 2007; Kharbanda et al., 2020; Li et al., 2013; MacDonald et al., 2019; Maric et al., 2019; Matcho et al., 2018; Palmsten et al., 2013; Palmsten, Huybrechts, Kowal, Mogun, & Hernandez-Diaz, 2014; Sarayani et al., 2020; Taylor, Thelus Jean, Gordon, Fram, & Coster, 2015; Zhu et al., 2020). Some have assessed pregnancy losses in addition to live births (Ailes et al., 2016; Kharbanda et al., 2020; MacDonald et al., 2019; Matcho et al., 2018; Sarayani et al., 2020), enhanced capture of gestational age (MacDonald et al., 2019; Maric et al., 2019; Matcho et al., 2018; Zhu et al., 2020) or updated the algorithms to include the latest International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) codes (Ellis et al., 2020; Kharbanda et al., 2020; Maric et al., 2019; Sarayani et al., 2020); however, only three of these studies created algorithms with ICD-10-CM codes specifically for IBM MarketScan® Commercial Claims and Encounters data (Ellis et al., 2020; Maric et al., 2019; Sarayani et al., 2020) and none have applied algorithms with the latest classification codes to IBM MarketScan® Multi-State Medicaid data. IBM MarketScan® Multi-State Medicaid data is representative of a more diverse population than Commercial data, which is a convenience sample of Americans with employer-provided health insurance (IBM, 2021; Kaiser Family Foundation, 2021) and having an algorithm that can be applied to both databases provides a more representative picture of pregnancy and pregnancy outcomes in the United States. Thus, there is still a need for improved exposure and outcome measurement in pregnancy algorithms applied to U.S. health insurance claims data.

This study improved upon previous work by creating and evaluating two new pregnancy algorithms. The first objective was to enhance available pregnancy algorithms by creating algorithms that evaluate pregnancies that ended in live births or pregnancy losses using an expanded list of ICD-9-CM and ICD-10-CM codes in the U.S. IBM MarketScan® Commercial Claims and Encounters and Multi-State Medicaid data. The second objective was to evaluate the new pregnancy identification algorithm's performance in assessing the distribution of select pregnancy outcomes and select medication use in this population and comparing those estimates to the published literature and national surveillance data.

2 | METHODS

2.1 | Study population and design

Retrospective cohorts of women 10–64 years of age in 2014 and 2016 were identified within the U.S. IBM MarketScan® Commercial Claims and Encounters and Multi-State Medicaid databases. Women's age was defined as the age at which the pregnancy code occurred. The IBM MarketScan® Commercial claims database is a collection of health insurance claims for individuals with employer-sponsored health insurance. The database contains fully paid and adjudicated claims generated by over 40 million commercially insured individuals annually.

The IBM MarketScan® Multi-State Medicaid claims database contains data for over 44 million individuals enrolled in state Medicaid or Medicaid managed care programs. The IBM MarketScan® Multi-State Medicaid database includes a high proportion of individuals from Hispanic and African American backgrounds(Kaiser Family Foundation, 2021). The database includes claims from 9 to 12 states in different geographic areas of the United States.

These databases consist of pseudo-anonymized member identification codes that allow individuals to be followed longitudinally as well as standard medical codes for diagnoses, procedures, and medications. Both databases include inpatient and outpatient services, outpatient pharmacy prescription claims, and healthcare plan annual enrolment information for each enrollee followed over the period of enrolment.

2.2 | Development of pregnancy algorithms

Using pregnancy outcome algorithms with ICD-9-CM codes for live birth that had been adapted from codes previously published and validated in the literature (Ailes et al., 2016; Li et al., 2013; Taylor et al., 2015), expert clinical review was utilized to expand these algorithms to include pregnancy loss and the new ICD-10-CM codes. This resulted in two new algorithms that used codes specific to those that may be contained in maternal records: (a) an extended ICD-9-CM pregnancy algorithm identifying live birth and pregnancy loss with the use of additional codes for abortion, labor, and specific birth conditions compared to previous work (Supporting information, Appendix S4) and (b) a new pregnancy algorithm including live birth and pregnancy loss using ICD-10-CM codes. The ICD-9-CM algorithm was applied to 2014 data and the ICD-10-CM algorithm was applied to 2016 data. Due to the ICD-CM code transition within the year, data from 2015 were not used except to inform 2014 pregnancies and subsequent pregnancy outcome. The LMP was estimated using codes for pregnancy, labor and delivery, and used to identify pregnancy start dates. Pregnancy episodes had to have an estimated LMPor end date in 2014 or 2016 to be attributed to that year, thus a pregnancy could have spanned 2015.

The algorithms used gestational codes (ICD-9-CM or ICD-10-CM and Current Procedural Terminology [CPT]) and clinically relevant pregnancy outcome hierarchies (codes available in the supplemental information) to identify pregnancies in each cohort. Trimester timing and pregnancy length were incorporated to create the pregnancy algorithms and demographics and selected pregnancy outcomes were evaluated (Figure 1).

2.2.1 | Algorithm creation: Trimester assignment

The identification and assignment of a pregnancy and pregnancy trimester was developed by an epidemiologist and an obstetrics and gynecology specialist using pregnancy related medical codes. First, pregnancies were identified using ICD-9-CM, ICD-10-CM, and CPT codes. Then, the pregnancy-associated ICD-CM codes were manually assigned to either first trimester, second trimester, third trimester, first or second trimester, second or third trimester, postpartum, or unspecified timing in pregnancy using the estimated LMP and end dates. Trimesters were assigned as follows: Days 0 (estimated LMP date) through 97 were classified as the first trimester, Days 98 through 195 as the second trimester, and Days 196 through the pregnancy end date as the third trimester. When multiple codes had conflicting trimester implications, the Trimester Hierarchy was applied (Appendix S1).

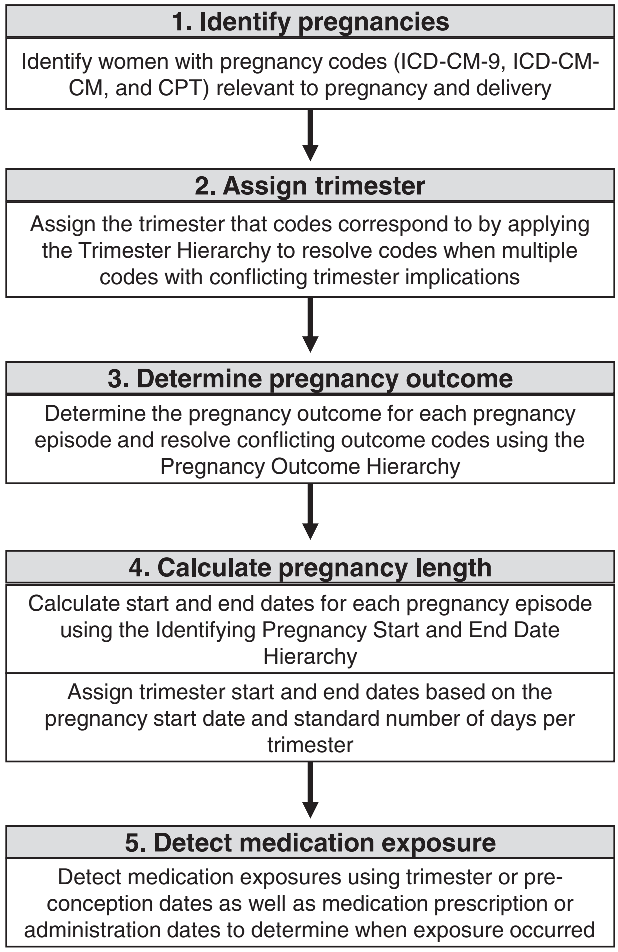

FIGURE 1 Implementation of new pregnancy algorithms in U.S. IBM MarketScan® Commercial and Medicaid databases. This figure illustrates the steps the new pregnancy algorithms take to identify pregnancies, assign trimesters, determine pregnancy outcomes and lengths, and identify medication exposures

2.2.2 | Algorithm creation: Hierarchy of pregnancy outcome assignment

Next, the pregnancy outcome was determined for each pregnancy episode. If multiple conflicting ICD-9-CM or ICD-10-CM codes occurred on the same day, a Pregnancy Outcome Hierarchy was implemented for outcome assignment (Table 1): ectopic pregnancy, spontaneous abortion, induced abortion, stillbirth, preterm birth, live birth and delivery, unclassified loss, and postpartum event. In instances where a patient had multiple pregnancy outcomes within a year, the Pregnancy Outcome

TABLE 1 Pregnancy outcome hierarchy and associated days assigned for pregnancy length identification[a]

Hierarchy	Pregnancy outcome	Days assigned
1	Ectopic pregnancy	56
2	Spontaneous abortion	70
3	Induced abortion	70
4	Stillbirth	194
5	Preterm birth	238
6	Live birth/C-section/term/post-term delivery	270
7	Unclassified loss	70
8	Postpartum event	270

[a]The table describes the hierarchy of pregnancy outcomes used if multiple conflicting outcomes occurred on the same day. When there were no ICD-9-CM or ICD-10-CM codes available to identify gestational weeks, the associated pregnancy outcome codes and outcome hierarchy were used to calculate the pregnancy's start and end dates. The days assigned for each pregnancy outcome are shown in the table and were chosen based on the literature and a data assessment (Ailes et al., 2016; Li et al., 2013; Taylor et al., 2015).

Hierarchy was also applied (Table 1). For example, if a woman had records indicating a spontaneous abortion and a live birth occurred in the same year, the Pregnancy Outcome Hierarchy would be used to identify these as two pregnancy episodes within the same year. More information about the Pregnancy Outcome Hierarchy is provided in Appendix S2.

2.2.3 | Algorithm creation: Pregnancy length assignment

Finally, the pregnancy length was determined for each pregnancy episode. ICD-10-CM codes (Z3A.x, P07.3x, P07.2x) and ICD-9-CM codes (765.2x) for gestational weeks were used to calculate the estimated LMP. Trimester-specific ICD-CM codes were then used to determine pregnancy length, with outcome codes occurring within the first trimester assigned 56 days, second trimester 135 days, and third trimester 270 days. If an outcome code occurred between 0 and 28 gestational weeks and could not be differentiated between the first and second trimesters, it was assigned a pregnancy length of 135 days. If an outcome code occurred between 14 and 40 gestational weeks and could not be differentiated between the second and third trimesters, it was assigned a pregnancy length of 270 days. When an ICD-9-CM or ICD-10-CM code for specific timing in pregnancy was not present, the Pregnancy Start and End Date Hierarchy was used (Appendix S3). Briefly, when ICD-9-CM or ICD-10-CM codes occurred on the same day as gestational week codes (Z3A.x, P07.3x, P07.2x), the trimester code was reassigned based on the gestational weeks recorded in the ICD-CM coding. For codes that occurred on the same day, the frequency of trimester codes that occurred on that day was calculated and the trimester with the most instances was assigned. A description of the ICD-9-CM codes used in this algorithm compared to previous ones (Ailes et al., 2016; Li et al., 2013; Taylor et al., 2015) is included in Appendix S4. The code for this paper was generated using SAS software. Copyright © 2019 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS, Institute Inc., Cary, NC. Some figures were produced using R version 3.6.1 (R Core Team, 2019).

2.3 | Testing algorithms: Identification of pregnancies

The pregnancy algorithms were used to identify pregnancies in 2014 and 2016, independently. To be included in the analysis for the given year, a woman was required to be enrolled for at least 1 day during that year (2014 or 2016) with pharmacy benefits. Each year was analyzed separately. Rates from the IBM MarketScan® Commercial and Medicaid databases were calculated and compared separately.

2.4 | Testing algorithms: Comparisons of pregnancy outcomes

Demographics, selected comorbidities and rates of selected pregnancy outcomes, including live births (term delivery [vaginal, C-section], preterm birth, post-term birth, twins, multiple births unspecified) and pregnancy losses (ectopic pregnancy, induced abortion, spontaneous abortion, and stillbirth), were calculated and descriptively compared to rates of Centers for Disease Control and Prevention (CDC) national vital statistics collected from birth certificates in 2014 and 2016 (Hamilton, Martin, & Osterman, 2015; Martin, Hamilton, & Osterman, 2017). An age range of 10–64 was chosen to capture the full age range of pregnancy outcomes reported by the CDC. Some specific pregnancy outcomes, such as pregnancy losses, were only reported for ages 15–44 to match reports of national estimates. Comparing the pregnancy algorithm estimates for pregnancy outcomes to national values served to indirectly validate the algorithm, as direct validation was not possible with the deidentified MarketScan® databases.

2.5 | Testing algorithms: Comparisons of medication exposure

The algorithms' ability to capture medication use during pregnancy was also assessed. An anti-epileptic, lamotrigine, was used as a test case as it is an example of a medication that women need to consistently continue using during pregnancy. The prevalence of lamotrigine use (at least one prescription dispensing per woman) was estimated during pregnancy and for each trimester and compared to published estimates of use during pregnancy from 2001 to 2013 in the Sentinel Distributed Database (Illoh et al., 2018). The prevalence of tetanus, diphtheria, and pertussis (TDAP) vaccination during pregnancy was also assessed as a test case and compared to national estimates and published estimates from 2010 to 2015 in the Sentinel Distributed Database (Kawai & Panucci, 2017), as the vaccine is recommended for pregnant women between 27 and 36 weeks of gestation.

Anti-epileptic use and TDAP vaccination comparison estimates were obtained from analyses that used the Sentinel Distributed Database modules (Illoh et al., 2018; Kawai & Panucci, 2017). The Sentinel Distributed Database is a collection of quality-checked electronic healthcare databases across the United States (Ball, Robb, Anderson, & Dal Pan, 2016; Behrman et al., 2011). The U.S. Food and Drug Administration runs modules on the Sentinel Distributed Database Administration to capture national estimates of medication use in electronic healthcare databases (Ball et al., 2016; Behrman et al., 2011). Medication use during pregnancy has been captured in this database for pregnancies that resulted in live birth outcomes among women aged 10–54 years (Illoh et al., 2018; Kawai & Panucci, 2017).

3 | RESULTS

The algorithms identified 687,228 pregnancies in 2014 and 444,293 in 2016 from women aged 10–64 years with one or more pregnancy codes in the IBM MarketScan® Commercial database (Figure 2). For IBM MarketScan® Commercial data, the pregnancy rate was 31.8 per 1,000 women aged 10–64 years for 2014 and 34.8 per 1,000 women aged 10–64 years for 2016. The algorithms identified 389,132 pregnancies in 2014 and 406,418 in 2016 from women aged 10–64 years with a pregnancy code in the IBM MarketScan® Medicaid database (Figure 2). For IBM MarketScan® Medicaid data, the pregnancy rate was 93.5 per 1,000 women aged 10–64 years for 2014 and 80.3 per 1,000 women aged 10–64 years for 2016.

The percentage of live birth outcomes identified by the new pregnancy algorithms using privately insured or Medicaid insured claims data were overall similar to published nationally representative CDC estimates from birth certificates (Table 2). Percentages of preterm birth, twin birth, and multiple births were higher in the 2014 and 2016 CDC population than the IBM MarketScan® Commercial and Medicaid populations (Hamilton et al., 2015; Martin et al., 2017). In both the IBM MarketScan® Commercial and Medicaid populations, the percentages of C-section were similar to the 2014 and 2016

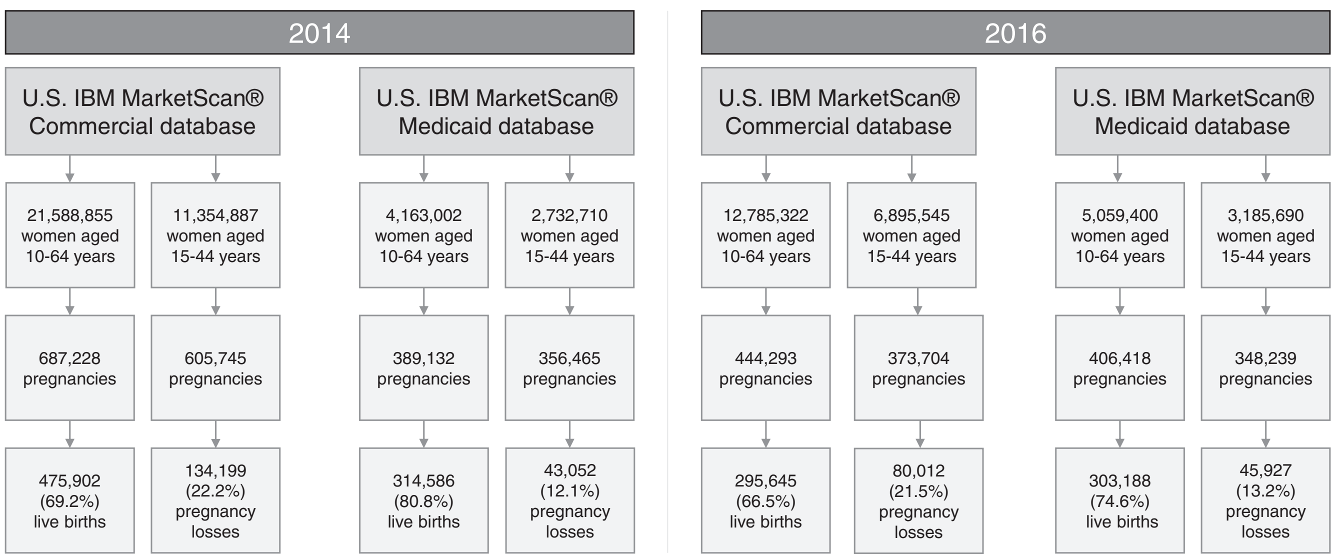

FIGURE 2 Number of pregnancies identified and associated outcomes in U.S. IBM MarketScan® Commercial and Medicaid databases. This figure details the number of women, pregnancy, and live birth or pregnancy loss outcomes within the U.S. IBM MarketScan® Commercial or Medicaid databases in 2014 and 2016. Live birth outcomes were reported for women 10–64 years. Pregnancy loss outcomes were reported for women 15–44 years

TABLE 2 Selected pregnancy outcomes among women with live births, 10–64 years[a]

	Commercial 2014	Commercial 2016	Medicaid 2014	Medicaid 2016	CDC 2014	CDC 2016
	N = 475,902	*N* = 295,645	*N* = 314,586	*N* = 303,188	*N* = 3,988,076	*N* = 3,945,875
	n (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)
Term delivery, vaginal	249,425 (52.41)	142,367 (48.15)	181,044 (57.55)	153,612 (50.67)	1,765,521 (44.27)	1,761,439 (44.64)
Term delivery, C-section	159,276 (33.47)	97,967 (33.14)	95,205 (30.26)	92,230 (30.42)	1,284,551 (32.21)	1,258,581 (31.90)
Preterm delivery	2,470 (0.52)	9,106 (3.08)	631 (0.20)	14,667 (4.84)	381,321 (9.57)	388,218 (9.84)
Post-term delivery	56,874 (11.95)	37,391 (12.65)	31,771 (10.10)	31,951 (10.54)	276,784 (6.95)	265,580 (6.73)
Twins	5,371 (1.13)	3,876 (1.31)	2,399 (0.76)	2,948 (0.97)	135,336 (3.39)	131,723 (3.34)
Multiple births	2,486 (0.52)	4,938 (1.67)	3,536 (1.12)	7,780 (2.57)	139,862 (3.51)	135,726 (3.44)

[a]The table displays the number and percent of pregnancies among women aged 10–64 years who had a live birth. The percent of each birth outcome is reported for six data sources: IBM MarketScan® Commercial 2014 database, Commercial 2016 database, Medicaid 2014, Medicaid 2016, CDC 2014 estimates (Hamilton et al., 2015), and CDC 2016 estimates (Martin et al., 2017). The CDC estimates were based on national birth certificate data.

TABLE 3 Pregnancy losses among women with a known pregnancy outcome, 15–44 years[a]

	Commercial 2014	Commercial 2016	Medicaid 2014	Medicaid 2016	Literature
	N = 605,745	*N* = 373,704	*N* = 356,465	*N* = 348,239	
	n (%)	*n* (%)	*n* (%)	*n* (%)	*n*/*N* (%)
Total pregnancy loss	134,199 (22.15)	80,012 (21.41)	43,052 (12.07)	45,927 (13.18)	14,364,623/70936411 (20.25)
Ectopic pregnancy	13,024 (2.15)	7,948 (2.13)	6,341 (1.78)	7,142 (2.05)	101,892/7271430 (1.40)
Induced abortion	25,341 (4.18)	12,754 (3.41)	1715 (0.48)	651 (0.19)	664,435/53154800 (1.25)
Spontaneous abortion	71,501 (11.80)	46,680 (12.49)	26,462 (7.42)	30,417 (8.73)	1,118,000/6578000 (17.00)
Stillbirth	2,236 (0.37)	1917 (0.51)	1852 (0.52)	2,378 (0.68)	23,595/3932181 (0.60)
Other pregnancy loss	22,097 (3.65)	10,713 (2.87)	6,682 (1.87)	5,339 (1.53)	—

[a]This figure displays the number and percent of pregnancies among women aged 15–44 years who had a pregnancy loss with a known pregnancy outcome. The percent of each birth outcome is reported for five data sources: IBM MarketScan® Commercial 2014 database, Commercial 2016 database, Medicaid 2014 database, Medicaid 2016 database, and the CDC and Stulberg, Cain, Dahlquist, and Lauderdale (2014) literature (Jatlaoui et al., 2016; MacDorman & Gregory, 2015; Stulberg et al., 2014; Ventura, Curtin, & Abma, 2012). Other pregnancy loss (i.e., unclassified loss) estimates were not obtained from the literature.

CDC percentages, but the post-term and vaginal delivery percentages were higher than what was reported by the CDC (Hamilton et al., 2015; Martin et al., 2017).

The new algorithms produced pregnancy loss estimates that differed depending on the population compared (Table 3). Percentages of total pregnancy loss and induced abortion were higher in the 2014 and 2016 IBM MarketScan® Commercial population and the CDC literature compared to the 2014 and 2016 IBM MarketScan® Medicaid population, with the exception of ectopic pregnancy and stillbirth where rates were overall the same across databases and published literature estimates from the 2004 to 2008 Medicaid population and 1990 to 2013 CDC populations (Table 3) (Jatlaoui et al., 2016; MacDorman & Gregory, 2015; Stulberg et al., 2014; Ventura et al., 2012). The percentage of spontaneous abortion was notably higher in the nationwide CDC literature than in the IBM MarketScan® Medicaid and Commercial claims data (Jatlaoui et al., 2016).

Medication use comparisons between the literature and the algorithms' results varied overall and by trimester. Anti-epileptic medication and TDAP vaccine use widely varied by trimester across the Sentinel, IBM MarketScan® Medicaid, and IBM MarketScan® Commercial populations (Figure 3). Across all data sources, lamotrigine use was highest in the first trimester (Figure 3). Capturing use of lamotrigine during pregnancy, the algorithms produced estimates similar to the

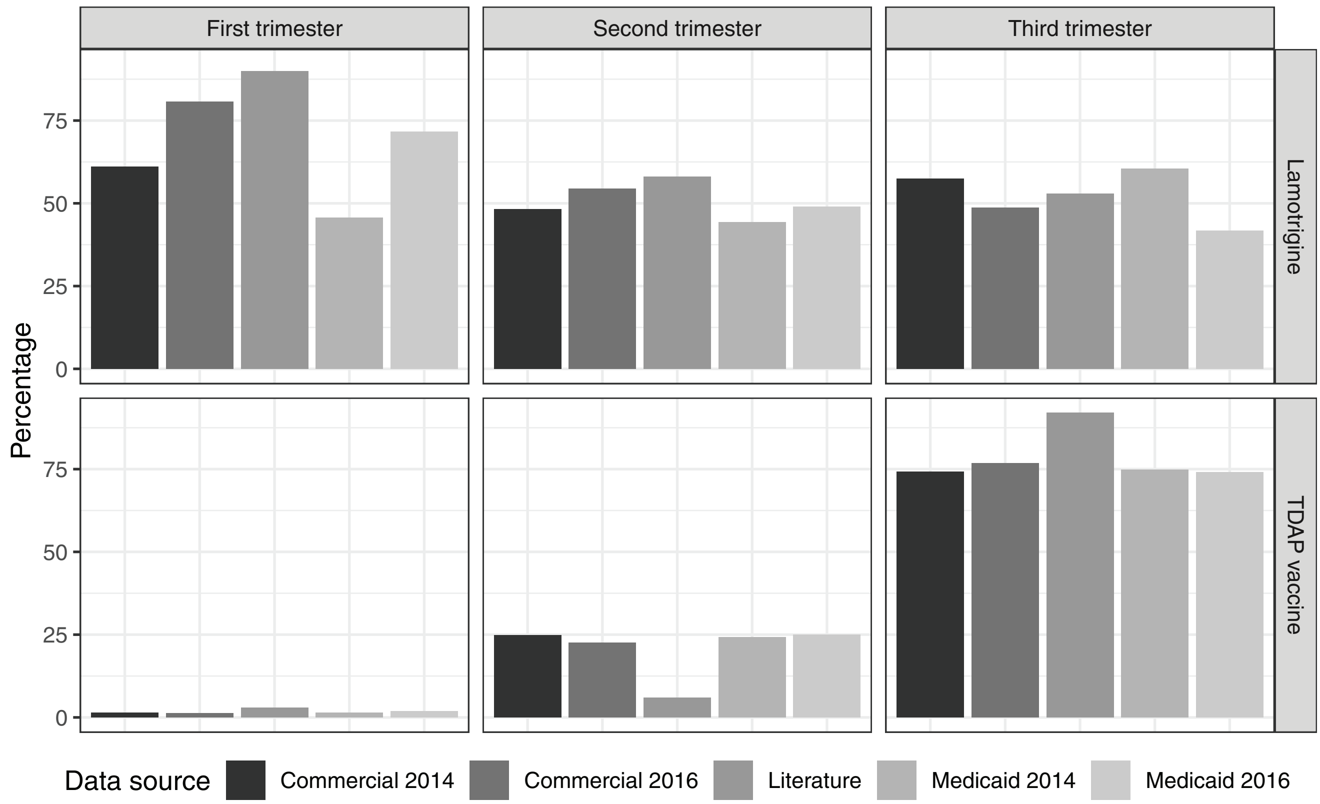

FIGURE 3 Percent of exposure in pregnancy among pregnant women, 10–64 years, by trimester. This figure displays the percentage of women exposed to lamotrigine and the TDAP vaccine in the first, second or third trimester of pregnancy. The percentage of women exposed in each trimester was obtained from the Commercial 2014 database, Commercial 2016 database, Sentinel model estimates in the literature (Illoh et al., 2018; Kawai & Panucci, 2017), Medicaid 2014 database, and Medicaid 2016 database

TABLE 4 Percent of exposure in pregnancy among pregnant women, 10–64 years, overall[a]

	Commercial 2014	Commercial 2016	Medicaid 2014	Medicaid 2016	
	N = 687,228	*N* = 444,293	*N* = 389,132	*N* = 406,418	Sentinel module
	n (%)	*n* (%)	*n* (%)	*n* (%)	*n*/*N* (%)
Lamotrigine	1,503 (0.22)	1,467 (0.33)	1,107 (0.28)	1713 (0.42)	6244/1,895,597 (0.32)
TDAP vaccine	145,090 (21.11)	155,308 (34.96)	39,657 (10.19)	80,904 (19.91)	406,321/2,633,462 (15.43)

[a]The table displays the number and percent of pregnancies among women exposed to lamotrigine and the TDAP vaccine at any point during pregnancy. The percent of women exposed was obtained from the IBM MarketScan® Commercial 2014, Commercial 2016, Medicaid 2014, Medicaid 2016, and Sentinel module estimates in the literature (Illoh et al., 2018; Kawai & Panucci, 2017). The IBM MarketScan® estimates were among women aged 10–64 years but the Sentinel module estimates were restricted to women aged 10–54 years with livebirths.

Sentinel modules (Table 4) (Illoh et al., 2018). The percentage of women with a TDAP vaccine during pregnancy was higher when used on the IBM MarketScan® Commercial data as compared to the IBM MarketScan® Medicaid populations; however, both were overall higher than published literature using the Sentinel modules, with the exception of Medicaid 2014 data which was lower (Table 4) (Kawai & Panucci, 2017). Given that the Sentinel modules were restricted to livebirths, conclusions regarding algorithm performance should be made with caution.

4 | DISCUSSION

The information obtained from the newly created pregnancy algorithms was overall comparable to published estimates for select outcomes when used on the 2014 and 2016 U.S. IBM MarketScan® Commercial and Multi-State Medicaid datasets. Additionally, we were able to capture lamotrigine use and TDAP vaccination during pregnancy, indicating the utility of using the new pregnancy algorithms to assess medication use in pregnancy using ICD-9-CM and ICD-10-CM codes and insurance claims data. The algorithms were indirectly validated by comparing algorithm estimates to national values.

For live birth outcomes, the percentage of births that were preterm was higher in CDC estimates from birth certificates compared to IBM MarketScan® Commercial or Medicaid estimates. This discrepancy could occur due to differences in the methods of data collection; this study utilized insurance claims, which could potentially miss cases compared to the active and passive case reporting system used by the CDC. A similar trend has been seen in previous work using IBM MarketScan® data (Ailes et al., 2016; Sarayani et al., 2020), indicating that preterm birth estimates might be underestimated within insurance claim databases compared to other live birth outcomes, such as neonatal intensive care unit admission (Andrade et al., 2013) or possibly C-section. C-section is a surgical procedure that routinely is reimbursed by a health insurance company, whereas other live birth outcomes like preterm birth do not always have to be specified in an insurance claim. Thus, outcomes like preterm birth could be mis-specified in insurance claims as a general live birth outcome. The observed IBM MarketScan® Commercial and Medicaid estimates for C-section were overall similar to the CDC estimates, indirectly validating that the algorithms accurately captured events that were routinely recorded in the insurance claims databases.

A higher pregnancy loss percentage was observed in the IBM MarketScan® Commercial Claims and Encounters population compared to the Multi-State Medicaid population. The Commercial population pregnancy loss percentage was overall similar to the CDC estimates reported in Table 3 (Jatlaoui et al., 2016; MacDorman & Gregory, 2015; Stulberg et al., 2014; Ventura et al., 2012) as well as those from previously published algorithms that used the IBM MarketScan® Commercial Claims and Encounters population (Ailes et al., 2016; MacDonald et al., 2019; Sarayani et al., 2020). In a prior study that compared pregnancy losses across both the IBM MarketScan® Commercial and Medicaid data sets, a lower pregnancy loss estimate was also observed in the Medicaid population (Matcho et al., 2018); however both estimates were a few percentage points higher than what was observed in this study. This difference in pregnancy loss estimates between the Commercial and Medicaid populations could be due to a difference in how the two populations access the healthcare system, as the Medicaid population may be less likely to attend early or regular prenatal care visits where pregnancy losses could be identified (Krieger, Connell, & LoGerfo, 1992). Alternatively, women may enroll in Medicaid after a pregnancy has reached viability. Therefore, the difference in pregnancy loss percentages between the IBM MarketScan® Commercial and Medicaid populations could be attributable to fewer pregnancy losses being captured in the Medicaid database, not necessarily less pregnancy losses occurring. Additionally, with the Hyde Amendment (Congress, 1993), induced abortions are only covered for limited reasons under Medicaid, lowering the induced abortion rate for the IBM MarketScan® Medicaid compared to Commercial populations.

The observed rates of exposure to TDAP were overall higher in the IBM MarketScan® populations than what was observed in the published Sentinel data which could be a result of inclusion of both live births and pregnancy losses in the algorithms. The inclusion of both types of pregnancy information may have increased the number of pregnancies identified in the new algorithms as compared to the Sentinel module, which focused on live births only (Illoh et al., 2018; Kawai & Panucci, 2017). The percentage of pregnancies that received a TDAP vaccination in the IBM MarketScan® and Sentinel populations were lower than the CDC percentages self-reported by women who had a live birth in 2014 (27.0%) and 2016 (48.8%) (CDC, 2017a), possibly due to some women getting TDAP outside of the claims setting.

Differences in TDAP use estimation across databases could also be due to differences in demographics. The Sentinel algorithm estimate for TDAP vaccine use was higher than the IBM MarketScan® Medicaid 2014 estimate but lower than the Medicaid 2016, Commercial 2014, and Commercial 2016 estimates. Both IBM MarketScan® Commercial, largely comprised of a working population, estimates were the highest. A previous study hypothesized that women who received the recommended TDAP vaccine during pregnancy might have received more comprehensive healthcare compared to those that did not receive the vaccine (Layton et al., 2017). The diversity in the Sentinel population could explain why its TDAP use estimate was lower than the IBM MarketScan® Commercial population, which may have greater access to care, and higher than the 2014 Medicaid population, which may have less access to care.

Lamotrigine use during pregnancy was overall similar across data sources and highest in the first trimester. Overall similarity of lamotrigine use in the Sentinel, IBM MarketScan® Commercial, and IBM MarketScan® Medicaid populations could be due to a lack of difference between antiepileptic use among women with live birth and pregnancy loss outcomes. Because the IBM MarketScan® estimates were based on prescription data, they may not reflect actual usage in the first trimester, possibly explaining high reported usage in the first trimester. Furthermore, medication exposure in the first trimester was based on the prescription date, which was not reflective of the supplied days of half-life of the medication to deduce the true level or duration of exposure during pregnancy.

The primary limitation of this study was, due to the nature of using anonymized data, the inability to directly validate the pregnancy estimates and outcomes derived from the pregnancy algorithms against medical records or birth certificates. Thus, gestational age and trimester ascertainment were assigned based on available anonymized data, making it possible that some pregnancy outcomes and gestational lengths were misclassified. Comparison to published literature and estimates from previously validated algorithms indirectly validated the new algorithms, but future studies using other data sources that may be linked to primary data could directly validate these algorithms. Additionally, early pregnancy losses were unlikely to be captured in this dataset as losses can occur before a woman seeks care for a pregnancy. Although IBM MarketScan® Commercial and Medicaid databases are generally representative of the United States (IBM, 2021; Kaiser Family Foundation, 2021) there may be differences in the populations compared to those included in the calculation of the CDC estimates. The CDC reported some specific birth outcomes for only ages 15–44; however, pregnancies outside of this age range made up less than 1% of the live births in the Medicaid and Commercial data sets. Differences in the estimates could also be due to population demographic differences instead of differences in true live birth prevalence. At the time of algorithm development, 2017 data was not available and thus pregnancies that occurred in 2016 but only had a code in 2017 may not have been captured thus underestimating the pregnancy rate for 2016.

The new pregnancy identification algorithms improved upon previous work by including a broader range of ICD-9-CM codes for U.S. IBM MarketScan® data as well as incorporating first-ever use of ICD-10-CM codes in U.S. IBM MarketScan® Multi-state Medicaid data. By incorporating Medicaid data and a more sensitive pregnancy capture methodology, these algorithms provide more representative and complete estimates of medication exposure during critical periods of pregnancy compared to algorithms only based on old classification codes and Commercial populations. The algorithms performed especially well at capturing pregnancy outcomes including live births, C-sections, total pregnancy losses, ectopic pregnancy, and stillbirth. For some outcomes such as preterm birth and spontaneous abortion, additional work, such as electronic health record linkage or linkages to patient reported data, may improve the algorithms' capture of these outcomes.

ACKNOWLEDGMENTS

The authors would like to acknowledge Marcy Powell for her work on the trimester assignment, Chi Truong for her contributions to identifying and classifying relevant codes, and Marianne Cunnington for contributions to the protocol and manuscript. This study was funded by GlaxoSmithKline PRJ3186. The funding body was involved in the design of the study; data collection, analysis, and interpretation; and manuscript writing.

CONFLICT OF INTEREST

KEW, MEG, and AE are employees of GlaxoSmithKline. KEW, AE and MEG are stockholders of GlaxoSmithKline. KMS is employed by the University of North Carolina at Chapel Hill to provide research assistance to GlaxoSmithKline.

DATA AVAILABILITY STATEMENT

This analysis used de-identified secondary data collection. The data that support the findings of this study are available from IBM MarketScan® but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Code lists for the algorithms implemented in this manuscript are publicly available.

ORCID

Keele E. Wurst https://orcid.org/0000-0002-5177-1759

REFERENCES

Ailes, E. C., Simeone, R. M., Dawson, A. L., Petersen, E. E., & Gilboa, S. M. (2016). Using insurance claims data to identify and estimate critical periods in pregnancy: An application to antidepressants. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, *106*(11), 927–934. https://doi.org/10.1002/bdra.23573

Andrade, S. E., Scott, P. E., Davis, R. L., Li, D. K., Getahun, D., Cheetham, T. C., … Cooper, W. O. (2013). Validity of health plan and birth certificate data for pregnancy research. *Pharmacoepidemiology and Drug Safety*, *22*(1), 7–15. https://doi.org/10.1002/pds.3319

Ball, R., Robb, M., Anderson, S. A., & Dal Pan, G. (2016). The FDA's sentinel initiative—A comprehensive approach to medical product surveillance. *Clinical Pharmacology and Therapeutics*, *99*(3), 265–268. https://doi.org/10.1002/cpt.320

Behrman, R. E., Benner, J. S., Brown, J. S., McClellan, M., Woodcock, J., & Platt, R. (2011). Developing the sentinel system—A natioanl resources for evdience development. *The New England Journal of Medicine*, *364*(6), 498–499.

CDC. (2017a). *Tdap vaccine*. Retrieved from https://www.cdc.gov/vaccines/pregnancy/hcp-toolkit/tdap-report.html

Congress. (1993). H.R.2518 — Departments of Labor, Health and Human Services, and Education, and Related Agencies Appropriations Act, 1994.

Ellis, R. P., Hsu, H. E., Song, C., Kuo, T. C., Martins, B., Siracuse, J. J., … Ash, A. S. (2020). Diagnostic category prevalence in 3 classification systems across the transition to the international classification of diseases, tenth revision, clinical modification. *JAMA Network Open*, *3*(4), e202280. https://doi.org/10.1001/jamanetworkopen.2020.2280

Hamilton, B. E., Martin, J. A., Osterman, M. J. K., Curtin, S. C., & Matthews, T. J. (2015). Births: Final Data for 2014. *National Vital Statistics Reports*, *64*(12), 1–64.

Hornbrook, M. C., Whitlock, E. P., Berg, C. J., Callaghan, W. M., Bachman, D. J., Gold, R., … Williams, S. B. (2007). Development of an algorithm to identify pregnancy episodes in an integrated health care delivery system. *Health Services Research*, *42*(2), 908–927. https://doi.org/10.1111/j.1475-6773.2006.00635.x

IBM. (2021). *IBM MarketScan Research Databases for life sciences researchers* [White Paper]. Retrieved from https://www.ibm.com/downloads/cas/OWZWJ0QO

Illoh, O. A., Toh, S., Andrade, S. E., Hampp, C., Sahin, L., Gelperin, K., … Bird, S. T. (2018). Utilization of drugs with pregnancy exposure registries during pregnancy. *Pharmacoepidemiology and Drug Safety*, *27*(6), 604–611. https://doi.org/10.1002/pds.4409

Jatlaoui, T. C., Ewing, A., Mandel, M. G., Nguyen, A., Oduyebo, T., Petersen, E., & Whiteman, M. K. (2016). Abortion surveillance—United States, 2013. *Morbidity and mortality weekly report. Surveillance summaries*, *65*(12), 1–44. https://doi.org/10.15585/mmwr.ss6512a1

Kaiser Family Foundation. (2021). *Medicaid enrollment by race/-ethnicity*. Retrieved from Medicaid Enrollment by Race/Ethnicity|KFF

Kawai, A., & Panucci, G. (2017). Tdap vaccination during pregnancy. *Sentinel*. https://doi.org/10.1002/pds4409

Kharbanda, E. O., Vazquez-Benitez, G., DeSilva, M. B., Spaulding, A. B., Daley, M. F., Naleway, A. L., … Romitti, P. A. (2020). Developing algorithms for identifying major structural birth defects using automated electronic health data. *Pharmacoepidemiology and Drug Safety*, *30*(2), 266–274. https://doi.org/10.1002/pds.5177

Krieger, J. W., Connell, F. A., & LoGerfo, J. P. (1992). Medicaid prenatal care: A comparison of use and outcomes in fee-for-service and managed care. *American Journal of Public Health*, *82*, 185–190.

Layton, J. B., Butler, A. M., Li, D., Boggess, K. A., Weber, D. J., McGrath, L. J., & Becker-Dreps, S. (2017). Prenatal Tdap immunization and risk of maternal and newborn adverse events. *Vaccine*, *35*(33), 4072–4078. https://doi.org/10.1016/j.vaccine.2017.06.071

Li, Q., Andrade, S. E., Cooper, W. O., Davis, R. L., Dublin, S., Hammad, T. A., … Toh, S. (2013). Validation of an algorithm to estimate gestational age in electronic health plan databases. *Pharmacoepidemiology and Drug Safety*, *22*(5), 524–532. https://doi.org/10.1002/pds.3407

MacDonald, S. C., Cohen, J. M., Panchaud, A., McElrath, T. F., Huybrechts, K. F., & Hernandez-Diaz, S. (2019). Identifying pregnancies in insurance claims data: Methods and application to retinoid teratogenic surveillance. *Pharmacoepidemiology and Drug Safety*, *28*(9), 1211–1221. https://doi.org/10.1002/pds.4794

MacDorman, M. F., & Gregory, E. C. W. (2015). Fetal and perinatal mortality: United States, 2013. *National Vital Statistics Reports*, *64*(8), 1–24.

Margulis, A. V., Palmsten, K., Andrade, S. E., Charlton, R. A., Hardy, J. R., Cooper, W. O., & Hernandez-Diaz, S. (2015). Beginning and duration of pregnancy in automated health care databases: Review of estimation methods and validation results. *Pharmacoepidemiology and Drug Safety*, *24*(4), 335–342. https://doi.org/10.1002/pds.3743

Maric, I., Winn, V. D., Borisenko, E., Weber, K. A., Wong, R. J., Aziz, N., … Shaw, G. M. (2019). Data-driven queries between medications and spontaneous preterm birth among 2.5 million pregnancies. *Birth Defects Research*, *111*(16), 1145–1153. https://doi.org/10.1002/bdr2.1580

Martin, J. A., Hamilton, B. E., & Osterman, M. J. K. (2017). NCHS data brief. Births in the United States, 2016. *NCHS Data Brief*, (287), 1–8.

Matcho, A., Ryan, P., Fife, D., Gifkins, D., Knoll, C., & Friedman, A. (2018). Inferring pregnancy episodes and outcomes within a network of observational databases. *PLoS One*, *13*(2), e0192033. https://doi.org/10.1371/journal.pone.0192033

Palmsten, K., Huybrechts, K. F., Kowal, M. K., Mogun, H., & Hernandez-Diaz, S. (2014). Validity of maternal and infant outcomes within nationwide Medicaid data. *Pharmacoepidemiology and Drug Safety*, *23*(6), 646–655. https://doi.org/10.1002/pds.3627

Palmsten, K., Huybrechts, K. F., Mogun, H., Kowal, M. K., Williams, P. L., Michels, K. B., … Hernandez-Diaz, S. (2013). Harnessing the medicaid analytic eXtract (MAX) to evaluate medications in pregnancy: Design considerations. *PLoS One*, *8*(6), e67405. https://doi.org/10.1371/journal.pone.0067405

R Core Team. (2019). *R: A language and environment for statistical computing (Version 3.6.1)*. Vienna, Austria: Author. Retrieved from https://www.R-project.org/

Sarayani, A., Wang, X., Thai, T. N., Albogami, Y., Jeon, N., & Winterstein, A. G. (2020). Impact of the transition from ICD-9-CM to ICD-10-CM on the identification of pregnancy episodes in US health insurance claims data. *Clinical Epidemiology*, *12*, 1129–1138. https://doi.org/10.2147/CLEP.S269400

Stulberg, D. B., Cain, L. R., Dahlquist, I., & Lauderdale, D. S. (2014). Ectopic pregnancy rates and racial disparities in the Medicaid population, 2004–2008. *Fertility and Sterility*, *102*(6), 1671–1676. https://doi.org/10.1016/j.fertnstert.2014.08.031

Swift, B., Jain, L., White, C., Chandrasekaran, V., Bhandari, A., Hughes, D. A., & Jadhav, P. R. (2018). Innovation at the intersection of clinical trials and real-world data science to advance patient care. *Clinical and Translational Science*, *11*(5), 450–460. https://doi.org/10.1111/cts.12559

Taylor, L. G., Thelus Jean, R., Gordon, G., Fram, D., & Coster, T. (2015). Development of a mother-child database for drug exposure and adverse event detection in the Military Health System. *Pharmacoepidemiology and Drug Safety*, *24*(5), 510–517. https://doi.org/10.1002/pds.3759

Ventura, S. J., Curtin, S. C., & Abma, J. C. (2012). Estimated pregnancy rates and rates of pregnancy outcomes for the United States, 1990–2008. *National Vital Statistics Reports*, *60*(7), 1–21.

Zhu, Y., Hampp, C., Wang, X., Albogami, Y., Wei, Y. J., Brumback, B. A., … Winterstein, A. G. (2020). Validation of algorithms to estimate gestational age at birth in the Medicaid analytic eXtract-quantifying the misclassification of maternal drug exposure during pregnancy. *Pharmacoepidemiology and Drug Safety*, *29*(11), 1414–1422. https://doi.org/10.1002/pds.5126

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Sumner, K. M., Ehlinger, A., Georgiou, M. E., & Wurst, K. E. (2021). Development and evaluation of standardized pregnancy identification and trimester distribution algorithms in U.S. IBM MarketScan® Commercial and Medicaid data. *Birth Defects Research*, *113*(19), 1357–1367. https://doi.org/10.1002/bdr2.1954