




ORIGINAL ARTICLE

Identification of pregnancies and infants within a US commercial healthcare administrative claims database

Monica L. Bertoia¹  | Kelesitse Phiri² | C. Robin Clifford¹ | Michael Doherty¹ | Li Zhou¹ | Laura T. Wang³ | Natalie A. Bertoia⁴ | Florence T. Wang¹  | John D. Seeger¹ 

¹Optum Epidemiology, Boston, Massachusetts, USA

²Oncology Epidemiology, GlaxoSmithKline, Collegeville, Pennsylvania, USA

³Department of Obstetrics and Gynecology, Prisma Health/University of South Carolina School of Medicine, Columbia, South Carolina, USA

⁴Obstetrics and Gynecology, Unity Hospital, Rochester, New York, USA

Correspondence

Monica L. Bertoia, Optum Epidemiology, Boston, MA, USA.
Email: monica.bertoia@optum.com

Funding information

Optum Epidemiology

Abstract

Purpose: Health care insurance claims databases are becoming a more common data source for studies of medication safety during pregnancy. While pregnancies have historically been identified in such databases by pregnancy outcomes, International Classification of Diseases, 10th revision Clinical Modification (ICD-10-CM) Z3A codes denoting weeks of gestation provide more granular information on pregnancies and pregnancy periods (i.e., start and end dates). The purpose of this study was to develop a process that uses Z3A codes to identify pregnancies, pregnancy periods, and links infants within a commercial health insurance claims database.

Methods: We identified pregnancies, gestation periods, pregnancy outcomes, and linked infants within the US-based Optum Research Database between 2015 and 2020 via a series of algorithms utilizing diagnosis and procedure codes on claims. The diagnosis and procedure codes included ICD-10-CM codes, Current Procedural Terminology (CPT) codes, and Healthcare Common Procedure Coding System (HCPCS) codes.

Results: We identified 1 030 874 pregnancies among 841 196 women of reproductive age. Of pregnancies with livebirth outcomes, 84% were successfully linked to infants. The prevalence of pregnancy outcomes (livebirth, stillbirth, ectopic, molar, and abortion) was similar to national estimates.

Conclusions: This process provides an opportunity to study drug safety and care patterns during pregnancy and may be replicated in other claims databases containing ICD-10-CM, CPT, and HCPCS codes. Work is underway to validate and refine the various algorithms.

KEYWORDS

administrative data, claims, infant, last menstrual period, pregnancy

Key Points

- Health care insurance claims databases are becoming a common data source for studies of medication safety.

Parts of this work have been presented at the 2018, 2019, 2020, and 2021 International Society for Pharmacoepidemiology annual scientific conferences.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Pharmacoepidemiology and Drug Safety* published by John Wiley & Sons Ltd.

- International Classification of Diseases, 10th revision Z3A billing codes denoting weeks of gestation provide more granular information for estimating last menstrual period (i.e., pregnancy start).
- We developed a process for identifying pregnancy episodes, outcomes, and linked infants within a claims database.
- 84% of livebirths were linked to infants and the percentage of pregnancies with a livebirth, early pregnancy loss, ectopic pregnancy, molar pregnancy, and stillbirth were similar to US national estimates.
- This process may be replicated in similar databases and provides an opportunity to study a wide variety of pregnancy-related research questions.

Plain Language Summary

This article describes a process for identifying pregnancies and infants within a health care insurance claims database. Applying this process identified 1 030 874 pregnancies among 841 196 women of reproductive age. Of pregnancies with livebirth outcomes, 84% were successfully linked to infants. The percentage of pregnancies with a livebirth, early pregnancy loss, ectopic pregnancy, molar pregnancy, and stillbirth were similar to US national estimates. The claims data from these pregnancies and infants may be used to study drug safety and care patterns during pregnancy. While this article describes results from applying this process to the Optum Research Database, researchers may apply a similar process to other insurance claims databases to study a wide variety of pregnancy-related research questions.

1 | INTRODUCTION

Pregnant women are often excluded from clinical trials and drug safety during pregnancy is typically assessed through post-marketing studies such as pregnancy registries. In recent years, administrative claims databases are becoming a more common data source for the evaluation of drug safety during pregnancy,¹ often in parallel with a registry. While pregnancy registries can ascertain attributes such as the pregnancy start date and the pregnancy outcome directly from patients, database studies must infer this information based on claims patterns.

Claims-based studies usually use codes to identify patient characteristics and clinical events. Identifying distinct pregnancy episodes in claims data is challenging given that there are no codes indicating the start and end of a pregnancy. These dates are key for studying the effects of medications used during pregnancy. Claims-based studies have historically assigned an average gestational period to pregnancies according to the observed pregnancy outcome (e.g., 39 weeks for term and post-term livebirths and 28 weeks for stillbirths). While this approach works well for some pregnancy outcomes, it is associated with greater measurement error for others such as spontaneous abortions.² ICD-10-CM Z3A codes denoting specific weeks of gestation may aid in reducing the measurement error associated with claims-estimated pregnancy start dates. The aim of this study was to develop algorithms to identify the start of pregnancy (last menstrual period [LMP]), the end of pregnancy, the pregnancy outcome, and the linked infant.

2 | METHODS

2.1 | Optum Research Database

The Optum Research Database (ORD) contains medical and pharmacy claims along with enrollment information dating as far back as 1993. One of the largest administrative health care databases in the United States, the ORD had approximately 14.3 million health plan members with medical and pharmacy coverage in 2018. This database is often used for studies mandated by regulatory agencies including those for medication safety in pregnancy specifically.³⁻⁶

2.2 | Optum dynamic assessment of pregnancies and infants

This article describes a process for identifying pregnancies and infants in claims data, referred to as the Optum Dynamic Assessment of Pregnancies and Infants (DAPI). The DAPI process applies a set of code-based algorithms to the ORD data to identify maternal characteristics, pregnancy episodes, pregnancy outcomes, linked infants, and infant characteristics.

2.3 | Identification of potential pregnancies

Women aged 12–55 years within the ORD who had both medical and pharmacy coverage any time between October 1, 2015 and

September 30, 2020 were selected. To identify potential pregnancies, the subset of women with at least one medical claim with a pregnancy-related ICD-10-CM diagnosis or procedure code, Current Procedural Terminology (CPT) code, or Healthcare Common Procedure Coding System (HCPCS) code were selected. The code list was intended to be highly sensitive, including obstetrical (O**) ICD-10-CM diagnosis codes as well as, for example, encounters for childcare instruction and pregnancy tests. The complete list of codes is provided in Appendix 1A through 1D.

2.4 | Algorithms to identify pregnancies and estimate LMP date

From the cohort of potential pregnancies, two approaches were used to identify unique pregnancies and to estimate the LMP date. The first (Z3A) approach identified active (ongoing) pregnancies using Z3A codes. The second (non-Z3A) approach identified completed pregnancies and backdated LMP according to the type of outcome and outcome date. A pregnancy episode was defined as the duration of time from the estimated LMP date (pregnancy start) through the pregnancy outcome date (pregnancy end).

2.5 | Z3A approach

The identification of active pregnancies utilized Z3A ICD-10-CM diagnosis codes indicating weeks of gestation. Codes Z3A.08 to Z3A.42 denote exact weeks of gestation (i.e., Z3A.08 is 8 weeks gestation, Z3A.09 is 9 weeks gestation, and so on). There are 3 additional Z3A codes that do not specify a particular week of gestation. Code Z3A.00 (weeks of gestation of pregnancy not specified) was observed at various weeks of gestation among pregnancies; as such, we did not incorporate this code into the Z3A approach algorithms. Code Z3A.01 (less than 8 weeks gestation of pregnancy) was frequently observed at weeks 5–9 of gestation and was assigned to 7 weeks gestation (the median observed gestation among pregnancies with Z3A.01 and other specific Z3A codes). Code Z3A.49 (greater than 42 weeks gestation of pregnancy) was not included in the algorithms since it was observed rarely and was not associated with a specific gestational period.

For each woman, and for each observed Z3A code, the algorithm subtracted the weeks of gestation indicated by the Z3A code from the date of service indicated in the claim to estimate the LMP date. For example, if the service date on the claim was July 10, 2019 and the claim had a Z3A.10 code, the algorithm subtracted 10 weeks (70 days) from the service date, resulting in an estimated LMP date of 01 May 2019. This process resulted in multiple estimated LMP dates for many women since women often had multiple claims with Z3A codes throughout a pregnancy. We dropped the non-specific code Z3A.01 if it occurred on the same date of service as a specific Z3A code or if it occurred within 6 weeks of a specific Z3A code. This 6-week window was chosen

based on the minimum number of weeks required between pregnancies from previous publications.^{2,7}

Next, estimated LMP dates were sequentially sorted for each woman. The algorithm created LMP clusters by grouping together all LMP dates within 6 weeks of each other (starting with the earliest estimated LMP and going forward 6 weeks) (Figure 1). If any clusters contained Z3A.01 codes and specific Z3A codes, the Z3A.01 codes were dropped. In the last step of the algorithm, we estimated the LMP date for each pregnancy episode (LMP cluster) using the median LMP date. If any 2 clusters had a median LMP within 8 weeks of each other (starting with the earliest and going forward), the clusters were combined and the median LMP was re-calculated. After applying this one-time combination, no 2 clusters had a median LMP within 8 weeks of each other. With an even number of data points, it is possible to have more than one median. In these cases, if there were multiple median LMP dates (e.g. January 15, 2018 and January 16, 2018), the earliest of the dates was selected.

2.6 | Non-Z3A approach

We identified completed pregnancies using a previously established approach with some modifications.^{2,7} Our approach began by identifying pregnancy outcomes (e.g., livebirth, stillbirth, abortion). A list of codes used to identify pregnancy outcomes is provided in Appendix 2. A minimum number of weeks between successive pregnancy outcomes was applied to identify separate pregnancy episodes within each woman (Table 1). For example, a minimum of 24 weeks was required between 2 livebirth outcomes and a minimum of 10 weeks between a livebirth outcome and a spontaneous abortion outcome. All pregnancy outcomes within a woman were assessed in the following sequential order: livebirth (including livebirth and stillbirth), stillbirth, ectopic, molar, ectopic and molar, spontaneous abortion, induced abortion, other abortion, and delivery claim(s) only. “Livebirth,” “livebirth and stillbirth,” and “stillbirth” were considered three separate outcomes. For example, pregnancies with multiples could have resulted in a livebirth and stillbirth. Some codes did not distinguish between an ectopic and molar pregnancy; hence, the following were considered three separate outcomes: “molar pregnancy,” “ectopic pregnancy,” or “molar and ectopic pregnancy.”

This process began with the first (earliest) livebirth claim for a woman (where livebirth included the combined outcome livebirth and stillbirth from a multi-gestation pregnancy). If a subsequent livebirth claim date was identified within the specified period (24 weeks), it was considered the same livebirth outcome. If the next livebirth claim date was more than 24 weeks later, it was considered a separate outcome (a second livebirth). Next, stillborn claims were identified and compared to the first livebirth claim, then ectopic claims, and so on. Finally, the algorithm assigned a standard gestational period for each outcome (e.g., 39 weeks for term livebirths and 28 weeks for stillbirths) to derive the LMP date (Table 2). This period accounts for preterm birth and multiples (see Appendix 3 for a list of codes to identify preterm births and Appendix 4 for a list of codes to identify

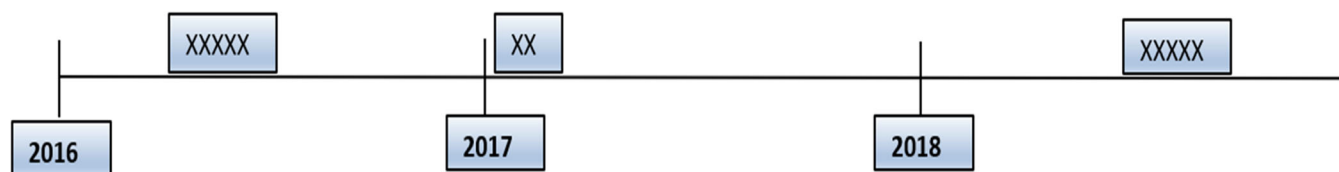


FIGURE 1 Example of LMP clusters for a woman with three potential pregnancies. The x's mark the estimated LMP dates, clustered around three different time points.

TABLE 1 Minimum number of weeks required to identify separate pregnancy outcomes.

Second outcome	First outcome								
	Livebirth, livebirth and stillbirth	Stillbirth	Ectopic	Molar	Ectopic and molar	Spontaneous abortion	Induced abortion	Other abortion, type unknown	Delivery claims only
Livebirth, livebirth and stillbirth	24	24	22	22	22	20	20	20	24
Stillbirth	24	24	22	22	22	20	20	20	24
Ectopic	10	10	8	8	8	6	6	6	10
molar	10	10	8	8	8	6	6	6	10
Ectopic and molar	10	10	8	8	8	6	6	6	10
Spontaneous abortion	10	10	8	8	8	6	6	6	10
Induced abortion	10	10	8	8	8	6	6	6	10
Other abortion, type unknown	10	10	8	8	8	6	6	6	10
Delivery claims only	24	24	22	22	22	22	22	22	24

multiples). If a woman had multiples and preterm codes, the gestational age indicated by the preterm code took preference. Preterm deliveries with unspecified gestational age were assigned 35 weeks of gestation.

The assigned lengths of gestation by pregnancy outcome (e.g., 8 weeks for ectopic pregnancy, 10 weeks for spontaneous abortion) was informed by Hornbrook et al.² with the following changes: first, we assigned singleton term and post-term livebirths and livebirths and stillbirths 39 instead of 40 weeks gestation (Table 2). We made this change because the distribution of singleton livebirths shifted to the left between 1992 and 2002, with 39 weeks becoming the most common length of gestation.⁸ In addition, previous work showed that assigning term births a gestational age of 39 weeks resulted in an estimated LMP within 2 weeks of medical record abstracted LMP in 99% of cases.⁹ Second, we assigned singleton preterm livebirths 35 instead of 34 weeks gestation because neonates 34–36 weeks gestation accounted for about 3 out of 4 of all singleton preterm births⁸ and previous work showed that assigning preterm births a gestational age of 35 weeks resulted in an estimated LMP within 2 weeks of medical record abstracted LMP in 75% of cases.⁹ Third, if the new ICD-10-CM P07 (short gestation and low birth weight) or O60 (preterm labor) codes were present, we used them to update gestational age of singleton preterm births from 35 weeks to the gestational age specified

by the code. For example, preterm newborns with code P07.31 (preterm newborn, gestational age 28 completed weeks) were assigned 28 weeks of gestation and preterm newborns with code O60.12** (preterm labor second trimester with preterm delivery second trimester) were assigned 26 weeks of gestation.

2.7 | Pregnancy outcome and date

The list of diagnosis and procedure codes used to identify each pregnancy outcome is provided in Appendix 2 and the pregnancy outcome algorithms are summarized in Figures 2–8. These algorithms were applied to each pregnancy episode to identify the pregnancy outcome type (e.g., livebirth, stillbirth, abortion) and date. If diagnosis codes and procedure codes were observed on claims with the same date, the outcome date was typically assigned to the earliest date with both types of codes. Otherwise, when both types of codes were not observed on the same date, the outcome date was the earliest date with a diagnosis code (livebirth, stillbirth, livebirth, and stillbirth) or the earliest date with a procedure code (ectopic pregnancy, molar pregnancy, ectopic and molar pregnancy, abortion).

The same logic was applied to both the Z3A and non-Z3A pregnancy episodes, with some distinctions. For Z3A pregnancy episodes,

TABLE 2 Fixed number of weeks to subtract from the estimated pregnancy outcome date to estimate the date of the last menstrual period.

Livebirths																																																											
Singleton	Term or post-term 39 weeks (273 days)	Preterm (search time window ± 14 days from the outcome date) Use gestational weeks below (last column) based on the preterm flag; otherwise use 35 weeks (245 days) for preterm deliveries with unspecified gestational age																																																									
		<table border="1"> <thead> <tr> <th>ICD-10-CM code</th> <th>Code description</th> <th>Gestational weeks (days)</th> </tr> </thead> <tbody> <tr> <td>P07.21</td> <td>...gestational age less than 23 completed weeks</td> <td>22 (154)</td> </tr> <tr> <td>P07.22</td> <td>...gestational age 23 completed weeks</td> <td>23 (161)</td> </tr> <tr> <td>P07.23</td> <td>...gestational age 24 completed weeks</td> <td>24 (168)</td> </tr> <tr> <td>P07.24</td> <td>...gestational age 25 completed weeks</td> <td>25 (175)</td> </tr> <tr> <td>P07.25</td> <td>...gestational age 26 completed weeks</td> <td>26 (182)</td> </tr> <tr> <td>P07.26</td> <td>...gestational age 27 completed weeks</td> <td>27 (189)</td> </tr> <tr> <td>P07.31</td> <td>Preterm newborn, gestational age 28 completed weeks</td> <td>28 (196)</td> </tr> <tr> <td>P07.32</td> <td>Preterm newborn, gestational age 29 completed weeks</td> <td>29 (203)</td> </tr> <tr> <td>P07.33</td> <td>Preterm newborn, gestational age 30 completed weeks</td> <td>30 (210)</td> </tr> <tr> <td>P07.34</td> <td>Preterm newborn, gestational age 31 completed weeks</td> <td>31 (217)</td> </tr> <tr> <td>P07.35</td> <td>Preterm newborn, gestational age 32 completed weeks</td> <td>32 (224)</td> </tr> <tr> <td>P07.36</td> <td>Preterm newborn, gestational age 33 completed weeks</td> <td>33 (231)</td> </tr> <tr> <td>P07.37</td> <td>Preterm newborn, gestational age 34 completed weeks</td> <td>34 (238)</td> </tr> <tr> <td>P07.38</td> <td>Preterm newborn, gestational age 35 completed weeks</td> <td>35 (245)</td> </tr> <tr> <td>P07.39</td> <td>Preterm newborn, gestational age 36 completed weeks</td> <td>36 (252)</td> </tr> <tr> <td>O60.12^a</td> <td>Preterm labor second trimester with preterm delivery second trimester</td> <td>26 (182)</td> </tr> <tr> <td>O60.13^a</td> <td>Preterm labor second trimester with preterm delivery third trimester</td> <td>34 (238)</td> </tr> <tr> <td>O60.14^a</td> <td>Preterm labor third trimester with preterm delivery third trimester</td> <td>35 (245)</td> </tr> </tbody> </table>	ICD-10-CM code	Code description	Gestational weeks (days)	P07.21	...gestational age less than 23 completed weeks	22 (154)	P07.22	...gestational age 23 completed weeks	23 (161)	P07.23	...gestational age 24 completed weeks	24 (168)	P07.24	...gestational age 25 completed weeks	25 (175)	P07.25	...gestational age 26 completed weeks	26 (182)	P07.26	...gestational age 27 completed weeks	27 (189)	P07.31	Preterm newborn, gestational age 28 completed weeks	28 (196)	P07.32	Preterm newborn, gestational age 29 completed weeks	29 (203)	P07.33	Preterm newborn, gestational age 30 completed weeks	30 (210)	P07.34	Preterm newborn, gestational age 31 completed weeks	31 (217)	P07.35	Preterm newborn, gestational age 32 completed weeks	32 (224)	P07.36	Preterm newborn, gestational age 33 completed weeks	33 (231)	P07.37	Preterm newborn, gestational age 34 completed weeks	34 (238)	P07.38	Preterm newborn, gestational age 35 completed weeks	35 (245)	P07.39	Preterm newborn, gestational age 36 completed weeks	36 (252)	O60.12 ^a	Preterm labor second trimester with preterm delivery second trimester	26 (182)	O60.13 ^a	Preterm labor second trimester with preterm delivery third trimester	34 (238)	O60.14 ^a	Preterm labor third trimester with preterm delivery third trimester	35 (245)
ICD-10-CM code	Code description	Gestational weeks (days)																																																									
P07.21	...gestational age less than 23 completed weeks	22 (154)																																																									
P07.22	...gestational age 23 completed weeks	23 (161)																																																									
P07.23	...gestational age 24 completed weeks	24 (168)																																																									
P07.24	...gestational age 25 completed weeks	25 (175)																																																									
P07.25	...gestational age 26 completed weeks	26 (182)																																																									
P07.26	...gestational age 27 completed weeks	27 (189)																																																									
P07.31	Preterm newborn, gestational age 28 completed weeks	28 (196)																																																									
P07.32	Preterm newborn, gestational age 29 completed weeks	29 (203)																																																									
P07.33	Preterm newborn, gestational age 30 completed weeks	30 (210)																																																									
P07.34	Preterm newborn, gestational age 31 completed weeks	31 (217)																																																									
P07.35	Preterm newborn, gestational age 32 completed weeks	32 (224)																																																									
P07.36	Preterm newborn, gestational age 33 completed weeks	33 (231)																																																									
P07.37	Preterm newborn, gestational age 34 completed weeks	34 (238)																																																									
P07.38	Preterm newborn, gestational age 35 completed weeks	35 (245)																																																									
P07.39	Preterm newborn, gestational age 36 completed weeks	36 (252)																																																									
O60.12 ^a	Preterm labor second trimester with preterm delivery second trimester	26 (182)																																																									
O60.13 ^a	Preterm labor second trimester with preterm delivery third trimester	34 (238)																																																									
O60.14 ^a	Preterm labor third trimester with preterm delivery third trimester	35 (245)																																																									
Multiples (search time window ± 14 days from the outcome date)	36 weeks (252 days) – Twins 33 weeks (231 days) – Triplets, other multiples 31 weeks (217 days) – Quadruplets																																																										
Livebirth and stillbirth	39 weeks (273 days)																																																										
Non-livebirths																																																											
Stillbirths		28 weeks (196 days)																																																									
Abortions (all types)		10 weeks (70 days)																																																									
Molar		8 weeks (56 days)																																																									
Ectopic		8 weeks (56 days)																																																									
Ectopic and Molar		8 weeks (56 days)																																																									

Abbreviations: ICD-10-CM, International Classification of Diseases, 10th revision Clinical Modification.

^aWildcard.

Time window for Z3A pregnancies: $LMP_median + 23*7 \leq service\ date \leq LMP_median + 42*7$

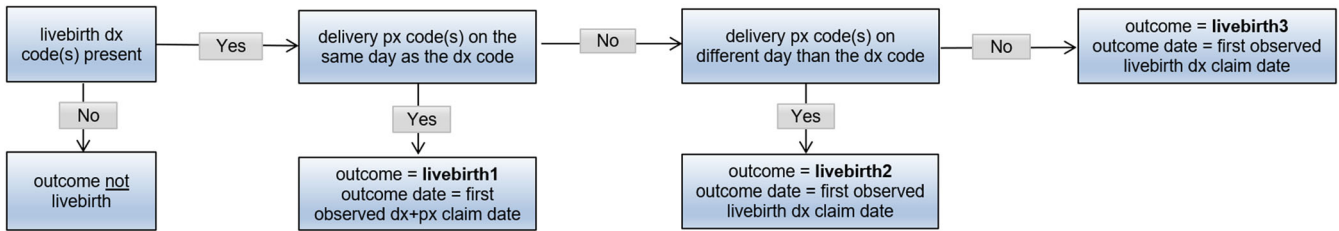


FIGURE 2 Livebirth outcome algorithm. DX, diagnosis; LMP, last menstrual period; PX, procedure.

Time window for Z3A pregnancies: $LMP_median + 20*7 \leq service\ date \leq LMP_median + 42*7$

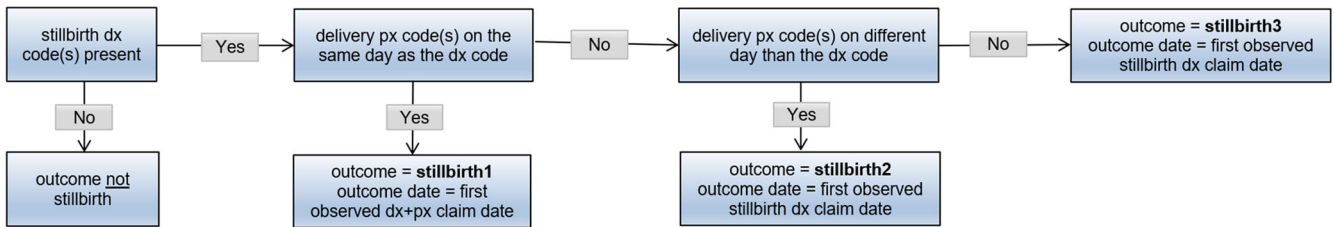


FIGURE 3 Stillbirth outcome algorithm. DX, diagnosis; LMP, last menstrual period; PX, procedure.

Time window for Z3A pregnancies: $LMP_median + 20*7 \leq service\ date \leq LMP_median + 42*7$

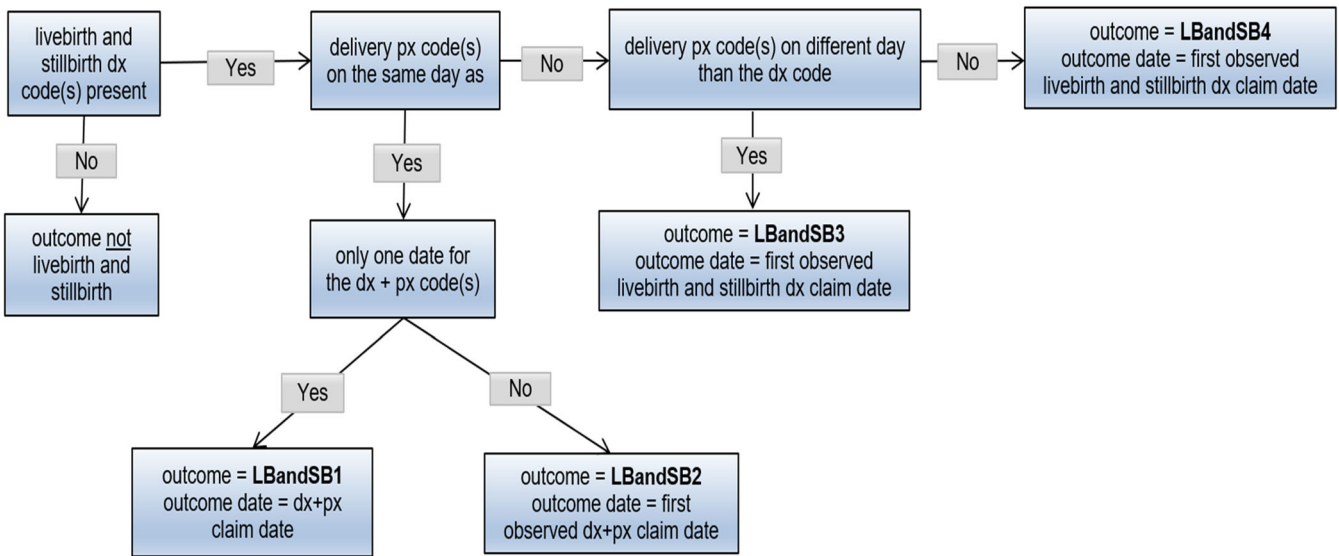


FIGURE 4 Live birth and stillbirth outcome algorithm. DX, diagnosis; LMP, last menstrual period; PX, procedure.

since information on the pregnancy outcomes was not used to identify the pregnancy period, specific time windows were assessed surrounding the estimated LMP date to search for and identify the pregnancy outcome (noted in italicized text in the figures). For example, a time window between 23 and 42 weeks after the LMP date was used to search for livebirth outcomes. For non-Z3A pregnancy episodes, a fixed time window of ± 14 days from the pregnancy outcome claim date that defined the pregnancy episode was assessed to estimate the pregnancy outcome date.

In rare instances, applying this logic assigned some Z3A and non-Z3A pregnancy episodes more than one outcome. These pregnancy episodes were removed since we were uncertain which outcome was correct.

We expected that some Z3A pregnancy episodes would have missing outcomes given the prospective identification process. For example, pregnancies that began less than 9 months prior to the end of the study period and pregnancies where women disenrolled from their health plan during the pregnancy.

Time window for Z3A pregnancies: $LMP_median \leq service\ date \leq LMP_median + 16*7$

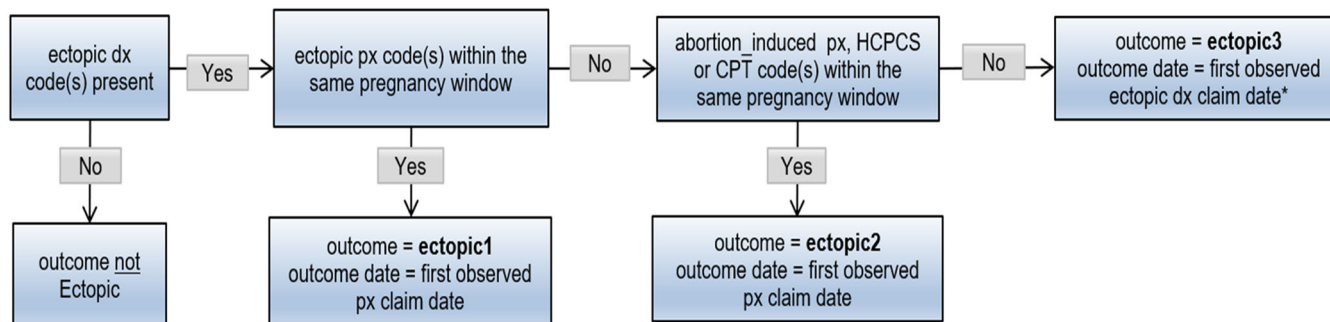


FIGURE 5 Ectopic pregnancy outcome algorithm. DX, diagnosis; LMP, last menstrual period; PX, procedure. Episodes <28 days were excluded.

Time window for Z3A pregnancies: $LMP_median \leq service\ date \leq LMP_median + 16*7$

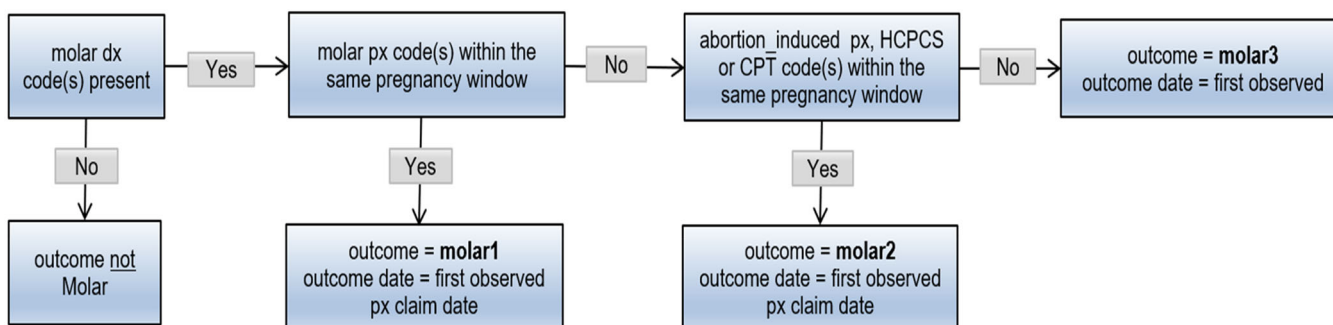


FIGURE 6 Molar pregnancy outcome algorithm. DX, diagnosis; LMP, last menstrual period; PX, procedure.

Time window for Z3A pregnancies: $LMP_median \leq service\ date \leq LMP_median + 16*7$

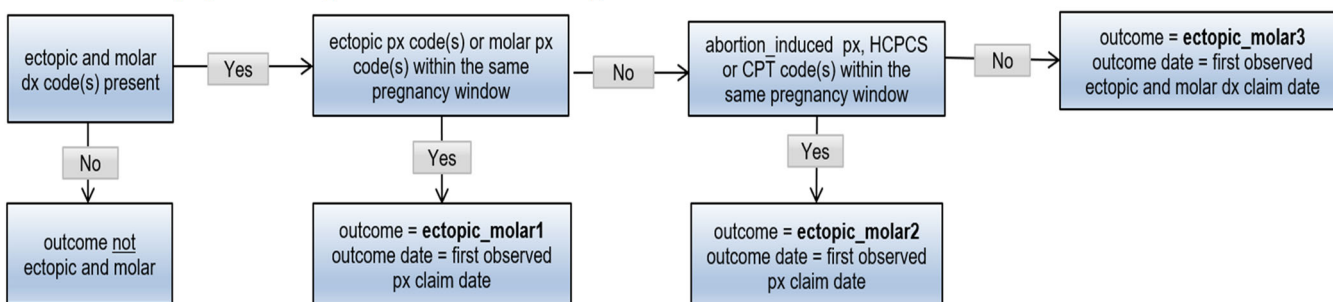


FIGURE 7 Ectopic and molar pregnancy outcome algorithm. DX, diagnosis; LMP, last menstrual period; PX, procedure.

2.8 | Z3A and non-Z3A pregnancy episode overlap

Since the non-Z3A approach was based on pregnancy outcomes, we compared all non-Z3A pregnancies to the subset of Z3A pregnancies with an identified outcome. First, we compared pregnancy outcome type and date and if any 2 dates were within 2 weeks of each other, we considered them the same pregnancy episode. For these duplicates, we retained the Z3A pregnancy episode (and Z3A estimated LMP) and removed the non-Z3A pregnancy episode.

As a final cleaning step, we combined all pregnancy episodes (those identified by Z3A and non-Z3A logic) and re-applied the Hornbrook et al.² logic (Table 1) to ensure adequate time between pregnancy episodes within a woman. Any nested pregnancy episodes were removed, as described above under “Non-Z3A approach”. For example, if a woman had two pregnancies, both resulting in a livebirth, and the livebirths were at least 24 weeks apart, the two pregnancy episodes were not considered nested and both pregnancy episodes were retained. On the other hand, if the second livebirth was less than 24 weeks later, the second livebirth was removed.

Time window for Z3A pregnancies: $LMP_median \leq service\ date < LMP_median + 20^*7$

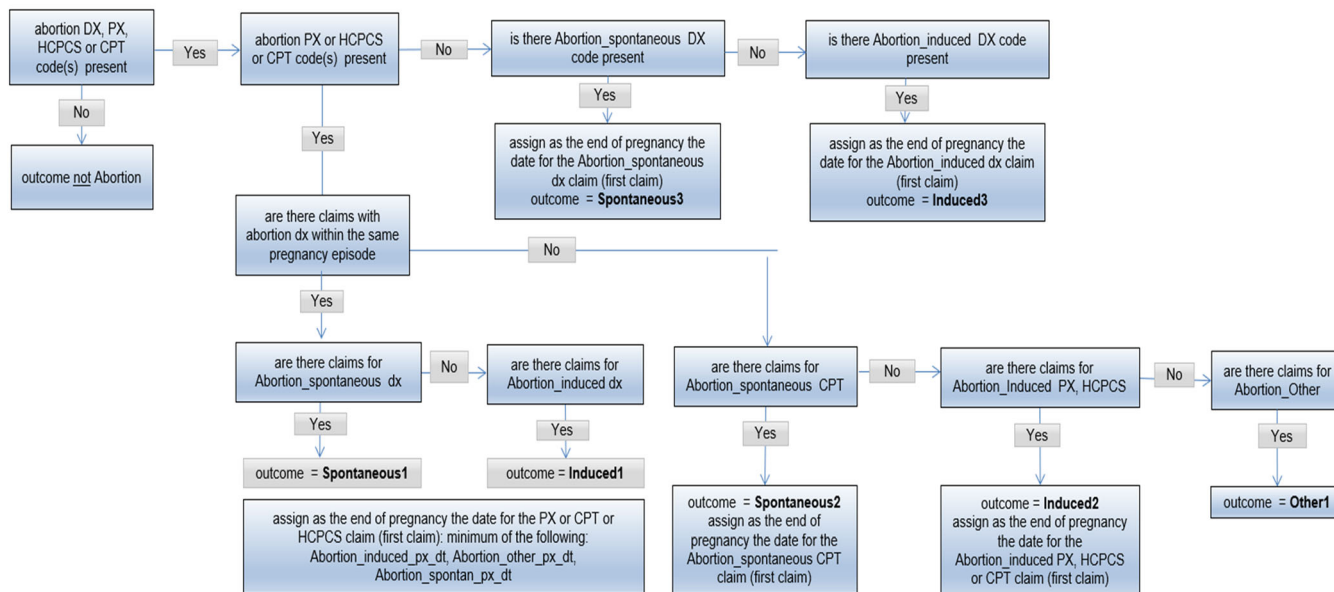


FIGURE 8 Abortion outcome algorithm. DX, diagnosis; LMP, last menstrual period; PX, procedure.

2.9 | Mother-infant linkage

The linkage algorithm used the infant's date of birth, the estimated pregnancy end date (delivery date), and a family member identifier, which is a number that uniquely identifies a family unit for insurance purposes. A mother was classified as linked to her infant(s) if the infant's date of birth was within 7 days of the estimated pregnancy end date for singleton pregnancies, or within 32 days for multi gestation pregnancies; otherwise the pregnancy episode was classified as not linked.

3 | RESULTS

Figure 9 describes the pregnancy identification process ending with 1 030 874 pregnancies. While the majority of pregnancies were identified by the Z3A and non-Z3A logic (677 526), many were identified by only one of the two methods: 213 976 (209 338 + 4638) were identified by the Z3A logic only and 145 228 by the non-Z3A logic only. Note that these numbers do not add to the final number of pregnancies due to downstream data cleaning steps, as described in the figure. Given the precision associated with using Z3A codes, we gave preference to LMP dates estimated by Z3A codes resulting in 86.5% of pregnancies with a Z3A logic LMP and 13.5% of pregnancies with a non-Z3A logic LMP.

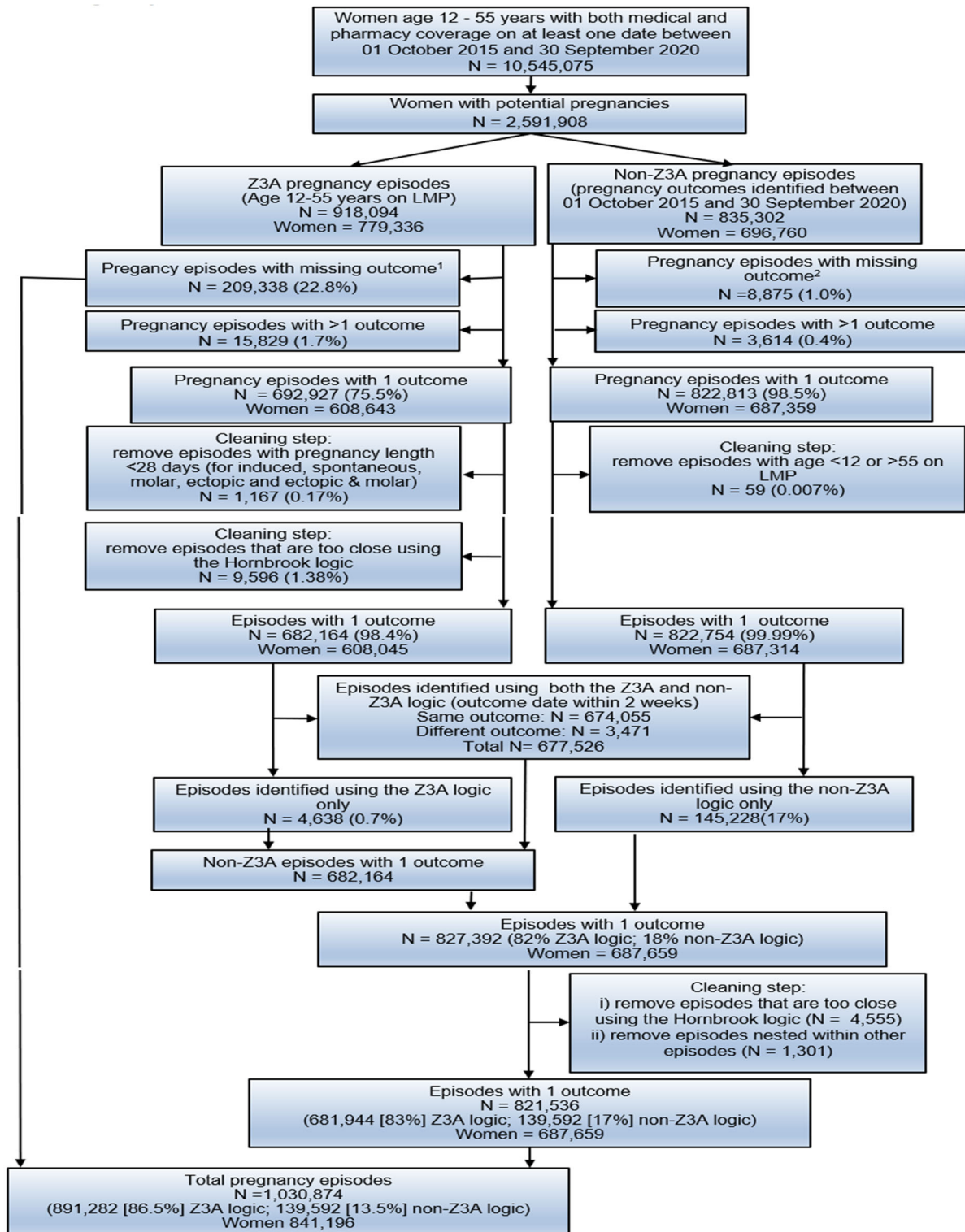
The frequency of pregnancy outcomes and linkage proportion of mothers and infants is described in Figure 10. Of the 821 536 pregnancies with a non-missing outcome, 76.2% were livebirths, 20.2% abortions (including spontaneous, induced, and unknown type), 3.0% ectopic or molar pregnancies, and 0.6% stillbirths. This is similar to national estimates including 10–28% early pregnancy losses,¹⁰ 2%

ectopic pregnancies,¹¹ <1% molar pregnancies,¹² and 0.6% stillbirths.¹³

While rare, 4368 (0.7%) pregnancies with non-missing outcomes were identified by the Z3A logic only. Differences between the Z3A and non-Z3A logic explain this small discrepancy. For example, the Z3A approach starts by identifying clusters of Z3A codes within women to identify a preliminary set of potential pregnancies. In contrast, the non-Z3A approach starts by identifying pregnancy outcomes and applying the required minimum number of weeks between sequential pregnancies (Table 1) to identify a preliminary set of potential pregnancies. Hence, we expected some differences in the initial set of potential pregnancies identified by each approach, where some pregnancies with non-missing outcomes are in the Z3A logic set only. Of the 4638 pregnancies, 2556 (55%) were abortions, 1652 (36%) were livebirths, 350 (8%) were stillbirths, 65 (1%) were ectopic pregnancies, 8 (0.2%) were livebirth and stillbirth, 6 (0.1%) were molar pregnancies, and one (0.02%) was an ectopic and molar pregnancy.

The majority of non-Z3A only pregnancies were early losses. Comparing the 139 592 non-Z3A only pregnancies to the 891 282 Z3A pregnancies, 70% of non-Z3A only pregnancies were abortions and 15% were ectopic or molar pregnancies compared to 8% and 1% of Z3A pregnancies, respectively.

Table 3 describes the distribution of pregnancies by maternal age and calendar year. Given that some pregnancies were identified by Z3A codes denoting the third trimester or term births at the very beginning of the study period (October 1, 2015), a handful of pregnancies were identified that began at the end of 2014 (0.1%). Between 2015 and 2019, about 200 000 pregnancies were identified per year. As expected, the number of pregnancies identified in 2020 was lower given that this was a partial year: the study period ended on 30 September 2020 and with a data lag of 6 months, we expected



¹170,431 (81%) disenrolled from health plan in the time window LMP to LMP+42 weeks¹

²7,478 (84%) episodes defined by delivery codes only
Abbreviations: LMP, last menstrual period.

FIGURE 9 Pregnancy identification flow chart. ¹170 431 (81%) disenrolled from health plan in the time window last menstrual period (LMP) to LMP + 42 weeks. ²7478 (84%) episodes defined by delivery codes only.

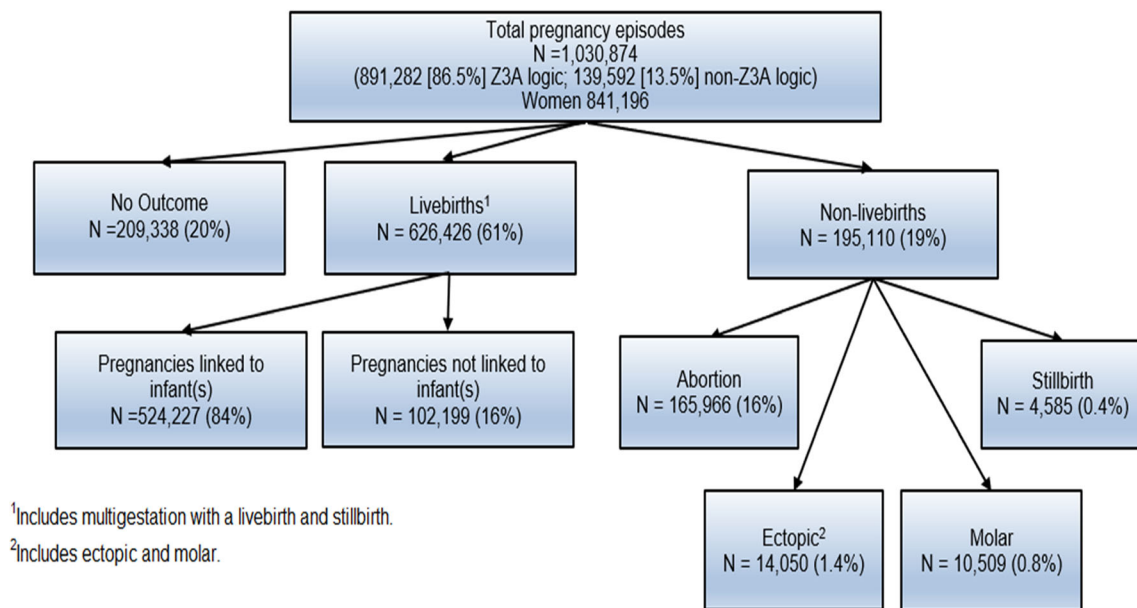


FIGURE 10 Mother-infant linkage and pregnancy outcomes. ¹Includes multigestation with a livebirth and stillbirth. ²Includes ectopic and molar.

TABLE 3 Distribution of pregnancies by maternal age and calendar year, Optum Research Database, October 1, 2015 to September 30, 2020

	N (%) of Pregnancies
Age	
12–19 years	22 553 (2.2%)
20–29 years	411 961 (40.0%)
30–39 years	538 820 (52.3%)
40+ years	57 540 (5.6%)
Year of LMP	
2014	965 (0.1%)
2015	166 185 (16.1%)
2016	206 590 (20.0%)
2017	204 067 (19.8%)
2018	201 444 (19.5%)
2019	189 186 (18.4%)
2020	62 437 (6.1%)
Total	1 030 874

Abbreviation: LMP, last menstrual period.

complete data for patients through the end of March 2020 (3 calendar months).

4 | DISCUSSION

Historically, registries have been the most common type of post-marketing pregnancy study requested by the US Food and Drug

Administration, however claims-based analyses are now requested more often.¹ While pregnancy registry studies can collect detailed patient information, they are resource intensive, recruitment is slow, and interpretation of these studies may be limited due to selection bias, recall bias, and loss to follow-up. Often registries lack an internal comparison group which further limits interpretation of results.¹⁴ On the other hand, claims-based analyses are relatively more efficient and can include very large sample sizes allowing researchers to study the effects of drugs used for rare conditions in pregnancy and the subsequent risk of rare pregnancy and infant outcomes.

This study builds on previous work^{7,9,15–17} identifying pregnancy episodes and outcomes in North American claims databases. Compared to previous work, our algorithms incorporate a key element: we used the additional granularity provided by ICD-10-CM Z3A codes to estimate LMP. Although we developed ICD-10-CM algorithms to build on previous work which used International Classification of Diseases, 9th revision Clinical Modification (ICD-9-CM) codes, others have incorporated ICD-10-CM codes into their pregnancy identification process as well.¹⁷ While we used a similar approach to Hornbrook et al.² to identify pregnancy outcomes and to estimate LMP when Z3A codes were absent, we did not replicate the Hornbrook algorithms exactly as we did not have the additional data sources available including gestational age from hospital discharge summaries and an EMR-based preterm birth prevention database. Like Wentzell et al.¹⁸ that ranked outcomes as more reliable if the outcome date was from an inpatient stay, we ranked outcomes as more reliable if both a procedure code and diagnosis code were observed on the same date.

We used a similar approach to Hornbrook et al.² given its high validity: 88% or greater agreement on outcome (livebirth, livebirth and stillbirth, therapeutic abortion, spontaneous abortion, ectopic

pregnancy, stillbirth) comparing the algorithm classification to medical chart clinical adjudication. When gestational age was estimated based on the outcome type and date, Hornbrook et al. observed the following agreement on estimated gestational age within 4 weeks: livebirth 98%, spontaneous abortion 67%, and ectopic pregnancy 70%. We aimed to achieve better estimation of gestational age using Z3A codes which is why we gave preference to Z3A-estimated LMP when a pregnancy was identified by both the Z3A and non-Z3A approaches. Given that many pregnancies were identified using the non-Z3A logic only, our suggestion is to use both approaches to identify pregnancies in claims databases. Alternatively, rather than applying both approaches to all pregnancies, the non-Z3A logic could be applied to the subset of pregnancies where no Z3A codes are observed within a specified number of weeks surrounding the pregnancy outcome.

While the algorithms described in this article are based on previous work, they have not been validated (compared to medical records) and there are no published results describing the validity of using Z3A codes to estimate LMP. When Z3A codes are not present, using traditional non-Z3A algorithms to estimate LMP (that are based on the presence of pregnancy outcomes) is expected to introduce a greater degree of measurement error for outcomes such as spontaneous abortions which had, for example, 67% agreement between actual and estimated gestational age.² In addition, given the inexact estimation of LMP, classification of trimesters and the corresponding timing of medication exposure could be inaccurate, with such non-differential misclassification of a binary exposure likely biasing study results toward the null.

We used Z3A codes to estimate LMP, however Z3A codes may also provide information for outcome classification. For example, Andrade et al. incorporated Z3A codes denoting 20 or more gestational weeks into their ICD-10 based stillbirth algorithm.¹⁹ The value in using Z3A codes to identify other pregnancy outcomes remains to be explored.

While the mother-infant linkage remains to be validated, it reflects the process used by others. For example, Palmsten et al.²⁰ used state, Medicaid Case Number (family identifier), and delivery/birth dates to link mothers and infants in the Medicaid Analytic eXtract (MAX) database. We linked 84% of livebirths to infants whereas Palmsten et al. reported a wide variation in linkage percentage depending on state (0%–96%) with about half of deliveries in the database linked to an infant.

While claims data represent an efficient approach for the examination of drug safety and pharmacotherapy treatment patterns during pregnancy, all claims databases have certain inherent limitations. These data reflect administrative record-keeping for financial transactions and were not designed for research. We used some diagnosis codes to identify pregnancy outcomes and the presence of a diagnosis code on a medical claim does not guarantee the presence of a condition or disease, as the diagnosis code may be incorrectly coded or indicate a rule-out criterion rather than the presence of an actual condition. To counteract false positives, claims databases such as the ORD have the option to seek patient medical charts for confirmation of pregnancy outcomes, infant outcomes, and drug exposures, and to abstract information that is missing from the database (e.g., reproductive history).

In conclusion, additional research is required to assess the performance of our algorithms and to correctly classify pregnancies with missing outcomes, pregnancies with multiple outcomes, and outcomes that are less clear. However, applying the DAPI process to the ORD identified claims data from mothers and infants that meets the minimum criteria required for pregnancy drug safety research¹ and has multiple options available for supplementation and enrichment. Overall, the DAPI process has many advantages that may make it an appropriate tool for a variety of pregnancy safety studies.

FUNDING INFORMATION

The work was funded by Optum Epidemiology (Optum is part of UnitedHealth Group).

ETHICS STATEMENT

The protocol was reviewed by an Optum ethics review committee to assure the data could not be used to re-identify patients. An institutional review board review was not required because the study used de-identified data. All data access conformed to applicable Health Insurance Portability and Accountability Act policies.

ORCID

Monica L. Bertoia  <https://orcid.org/0000-0001-8790-7243>

Florence T. Wang  <https://orcid.org/0000-0001-8776-9434>

John D. Seeger  <https://orcid.org/0000-0002-8020-3144>

REFERENCES

- Andrade SE, Bérard A, Nordeng HM, Wood ME, van Gelder MM, Toh S. Administrative claims data versus augmented pregnancy data for the study of pharmaceutical treatments in pregnancy. *Curr Epidemiol Rep.* 2017;4(2):106-116.
- Hornbrook MC, Whitlock EP, Berg CJ, et al. Development of an algorithm to identify pregnancy episodes in an integrated health care delivery system. *Health Serv Res.* 2007;42(2):908-927.
- Cole JA, Ephross SA, Cosmatos IS, Walker AMJP. Safety d. paroxetine in the first trimester and the prevalence of congenital malformations. *Pharmacoepidemiol Drug Saf.* 2007;16(10):1075-1085.
- Cole JA, Modell JG, Haight BR, Cosmatos IS, Stoler JM, Walker AM. Bupropion in pregnancy and the prevalence of congenital malformations. *Pharmacoepidemiol Drug Saf.* 2007;16(5):474-484.
- Carman WJ, Accortt NA, Anthony MS, Iles J, Enger CJP. Safety d. pregnancy and infant outcomes including major congenital malformations among women with chronic inflammatory arthritis or psoriasis, with and without etanercept use. *Pharmacoepidemiol Drug Saf.* 2017;26(9):1109-1118.
- Wyszynski DF, Carman WJ, Cantor AB, et al. Pregnancy and birth outcomes among women with idiopathic thrombocytopenic purpura. *J Pregnancy.* 2016;2016:1-8.
- Matcho A, Ryan P, Fife D, Giffkins D, Knoll C, Friedman A. Inferring pregnancy episodes and outcomes within a network of observational databases. *PLoS One.* 2018;13(2):e0192033.
- Davidoff MJ, Dias T, Damus K, et al. Changes in the gestational age distribution among US singleton births: impact on rates of late preterm birth. 1992 to 2002. *Semin Perinatol.* 2006;30(1):8-15.
- Margulis AV, Setoguchi S, Mittleman MA, Glynn RJ, Dormuth CR, Hernández-Díaz S. Algorithms to estimate the beginning of pregnancy in administrative databases. *Pharmacoepidemiol Drug Saf.* 2013;22(1):16-24.

10. Rossen LM, Ahrens KA, Branum AMJP. Epidemiology p. trends in risk of pregnancy loss among US women, 1990–2011. *Paediatr Perinat Epidemiol*. 2018;32(1):19-29.
11. Centers for Disease Control and Prevention. Current trends in ectopic pregnancy - United States, 1990-1992. Accessed January 29, 2021. <https://www.cdc.gov/mmwr/preview/mmwrhtml/00035709.htm>
12. March of Dimes. Molar pregnancy. Accessed January 29, 2021. <https://www.marchofdimes.org/complications/molar-pregnancy.aspx>
13. Centers for Disease Control and Prevention. What is stillbirth? Accessed January 29, 2021. <https://www.cdc.gov/ncbddd/stillbirth/facts.html#ref>
14. Margulis AV, Andrews EB. The safety of medications in pregnant women: an opportunity to use database studies. *Pediatrics*. 2017; 140(1):e20164194.
15. Margulis AV, Palmsten K, Andrade SE, et al. Beginning and duration of pregnancy in automated health care databases: review of estimation methods and validation results. *Pharmacoepidemiol Drug Saf*. 2015;24(4):335-342.
16. Toh S, Mitchell AA, Werler MM, Hernández-Díaz S. Sensitivity and specificity of computerized algorithms to classify gestational periods in the absence of information on date of conception. *Am J Epidemiol*. 2008;167(6):633-640.
17. Naleway AL, Crane B, Irving SA, et al. Vaccine safety datalink infrastructure enhancements for evaluating the safety of maternal vaccination. *Ther Advan Drug Saf*. 2021;12:20420986211021233.
18. Wentzell N, Schink T, Haug U, Ulrich S, Niemeyer M, Mikolajczyk R. Optimizing an algorithm for the identification and classification of pregnancy outcomes in German claims data. *Pharmacoepidemiol Drug Saf*. 2018;27(9):1005-1010.
19. Andrade SE, Shinde M, Moore Simas TA, et al. Validation of an ICD-10-based algorithm to identify stillbirth in the sentinel system. *Pharmacoepidemiol Drug Saf*. 2021;30(9):1175-1183.
20. Palmsten K, Huybrechts KF, Mogun H, et al. Harnessing the Medicaid analytic eXtract (MAX) to evaluate medications in pregnancy: design considerations. *PLoS One*. 2013;8(6):e67405.

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: Bertoia ML, Phiri K, Clifford CR, et al. Identification of pregnancies and infants within a US commercial healthcare administrative claims database. *Pharmacoepidemiol Drug Saf*. 2022;31(8):863-874. doi:10.1002/pds.5483