International Journal of Population Data Science

Journal Website: www.ijpds.org

Addressing uncertainty in identifying pregnancies in the English CPRD GOLD Pregnancy Register: a methodological study using a worked example

Yangmei Li[1*], Jennifer J. Kurinczuk[1], Fiona Alderdice[1], Maria A. Quigley[1], Oliver Rivero-Arias[1], Julia Sanders[2], Sara Kenyon[3], Dimitrios Siassakos[4,5], Nikesh Parekh[6], Suresha De Almeida[7], and Claire Carson[1]

Submission History

Submitted:	04/07/2024
Accepted:	07/01/2025
Published:	25/02/2025

[1] NIHR Policy Research Unit in Maternal and Neonatal Health and Care, National Perinatal Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Headington, Oxford, OX3 7LF, United Kingdom

[2] School of Healthcare Sciences, Cardiff University, Cardiff, United Kingdom

[3] Institute of Applied Health Research, University of Birmingham, Birmingham, United Kingdom

[4] EGA Institute for Women's Health, University College London, London, United Kingdom

[5] Wellcome/EPSRC centre for Interventional and Surgical Sciences (WEISS), London, United Kingdom

[6] Public health and wellbeing, Royal Borough of Greenwich, London, United Kingdom

[7] NHS South West London CCG, United Kingdom

Abstract

Introduction

Electronic health records are invaluable for pregnancy-related studies. The Clinical Practice Research Datalink (CPRD) Pregnancy Register (PR) identifies pregnancies in primary care records, including uncertain cases.

Objectives

This paper outlines a method to reduce uncertainty in identifying pregnancies within CPRD GOLD PR data, exemplified through a study investigating the provision of pre-pregnancy care.

Methods

We used CPRD Mother Baby Link (MBL) and Maternity Hospital Episode Statistics (HES) to clean and augment the CPRD PR data. The study included all women aged 18-48yrs, registered at an English GP practice within CPRD on 01/01/2017, with a year of prior registration and eligibility for hospital data linkage. We developed a cleaning and combining algorithm and further applied strict data quality criteria to form three populations: 'as provided', 'derived' (using our algorithm) and 'strictly derived' (with stricter data quality criteria). We compared characteristics and outcomes across these populations, examining potential biases in effect estimates using the 'as provided' population.

Results

Our algorithm added 22,270 (~7%) pregnancies from hospital data to the CPRD PR (1997-2021), eliminated conflicting pregnancies and pregnancies with unknown outcomes, and minimised potentially non-contemporaneous records of past pregnancies or partial records of pregnancies.

For all pregnancies across women's reproductive history, in the 'strictly derived' population, characterised by better data quality, a higher prevalence of pre-existing medical conditions and increased pre-pregnancy care were observed. In this dataset, recording of both exposure and outcome was better, and the magnitude of the association between exposure and outcome was reduced compared to the 'as provided' population.

Conclusion

PR data requires cleaning before use. This study presents a pragmatic and practical method to identify pregnancies using existing CPRD data and linked records, without needing additional data. Researchers should carefully consider their studies' specific requirements and may adapt our proposed methodology accordingly to align with their research questions.

Keywords

Clinical Practice Research Datalink (CPRD); Pregnancy Register; Hospital Episode Statistics (HES) Maternity; Mother Baby Link; routine National Health Service (NHS) data; pregnancies; electronic health records; electronic medical records; methodological study

*Corresponding Author:
Email Address: yangmei.li@ndph.ox.ac.uk (Yangmei Li)

https://doi.org/10.23889/ijpds.v10i1.2471

Introduction

Electronic health records are an important source of information for pregnancy-related studies, providing large datasets for epidemiological and pharmaco-epidemiological research, and an opportunity to study the provision of care. Such studies often draw on hospital admission data as, historically, the accurate identification of pregnancies and their timing in primary care records has presented a challenge [1]. The development of algorithm-based pregnancy registers, such as that released by the Clinical Practice Research Datalink (CPRD) in the UK, offers researchers the promise of better pregnancy data from routine primary care records [2].

The CPRD extracts anonymised patient record data from a network of general practitioner (GP) practices across the UK using the Vision® or EMIS® software systems. CPRD GOLD contains data contributed by practices using Vision® software. Primary care data are linked to other health-related data by a trusted third party for patients from practices that have consented to participate in the CPRD linkage scheme, providing a longitudinal, representative UK population health dataset [3]. The CPRD Pregnancy Register (PR), developed by Minassian et al, uses an algorithm to identify all pregnancies within female patient primary care records based on an extensive list of pregnancy-related codes, and consolidates the information about timing, antenatal care and pregnancy outcome in one place [2].

This approach has the advantage of using all pregnancy data in the CPRD GOLD, however, it also presents important methodological challenges. First, around 16% of all pregnancy episodes have no outcome recorded (unknown outcome); second, approximately 8.5% of pregnancy episodes conflict with another pregnancy episode for the same woman, which means there is at least one day of overlap in these episodes identified as conflicting with each other (conflicting pregnancies) [4]. CPRD policy is to give researchers all the pregnancies in the PR and let them decide how to deal with these issues, so the approaches taken will likely vary.

Researchers have explored these issues and used additional linked data, such as Hospital Episode Statistics (HES) Diagnostic Imaging Dataset (linked and provided by the CPRD and available at an additional cost), to identify several scenarios that may result in pregnancies in the PR having unknown outcomes or conflicting with one another [4]. Although approaches to dealing with these uncertain records in the PR have been suggested [4], it remains challenging for researchers to identify the best approach. For example, using some linked data as proposed previously could substantially reduce the sample size of pregnancies available. It was also unclear what the impact of including 'less certain' pregnancies in the PR data might have on real-world studies.

Given the increasing reliance on electronic health records for studies aiming to improve maternal and neonatal health outcomes, the importance of addressing data quality issues in pregnancy registers is clear. The availability of robust, accurate, and comprehensive pregnancy data is essential not only for assessing health interventions and care provision but also for conducting epidemiological studies that inform public health policies. Without adequately addressing these challenges, inaccurate pregnancy identification could lead to flawed estimates of exposure-outcome associations, ultimately undermining the reliability of research findings. Therefore, this study aims to refine current methods of identifying pregnancies in CPRD GOLD, offering improved data quality that will benefit future pregnancy-related research and healthcare evaluations.

In this study, we use the example of investigating the provision of pre-pregnancy care, where certainty in the occurrence and timing of pregnancy is particularly important, to explore novel ways to reduce uncertainty in identifying pregnancies in CPRD GOLD Pregnancy Register data. As this is the first step of a subsequent project investigating effects of pre-pregnancy care, the focus of this methodological part of the study is on ensuring pregnancies are correctly identified to provide meaningful population estimates and allow linkage to pregnancy outcomes and perinatal outcomes. Therefore, we aimed:

1. To investigate a method to identify implausible, non-contemporaneous records of a past pregnancy (hereon called 'historical' records), duplicate or overlapping pregnancies by combining all three sources of pregnancy data in CPRD GOLD and linked datasets – Pregnancy Register, Mother Baby Link (MBL) and HES Maternity dataset;

2. To describe the characteristics of the study population, number of pregnancies and births, pre-pregnancy care, health and pregnancy outcomes before and after applying the methodology developed in aim 1;

3. To explore the potential for biased estimates of effect when using the PR data as provided by CPRD, compared with the newly derived dataset.

Methods

Study population, design and setting

All women with data meeting quality standards predefined by the CPRD [5] and registered at an English GP practice participating in CPRD were included if they were aged 18-48 years on 01/01/2017, had at least one year of prior registration, and had linked hospital admissions data available.

In this methodological study, we compared three populations. The first population, termed the 'as provided' population, comprised all pregnancies identified by PR in the eligible women with minimal exclusions or data cleaning. The second population, the 'derived' population, included pregnancies from CPRD PR, augmented by the MBL and HES Maternity datasets. These pregnancies were identified using our proposed algorithm of combining and cleaning data, detailed below. Pregnancies were excluded if they ended after women transferred out of the CPRD practices, or the last collection date of data of the contributing CPRD practice, to ensure fair comparison of pregnancies added from MBL and HES. The third population, the 'strictly derived' population was further restricted to include only pregnancies that started after the woman had registered with a CPRD GP practice, and within the period when data from their GP practice were considered to be of research quality. This is in line with CPRD guidance that indicates that when patients are registered with

a CPRD GP practice, their medical records are more likely to be complete, whereas records before the current registration date may not be as complete or reliable. Therefore, the 'strictly derived' population in this study is considered to have the highest standard of research quality data.

Data sources

To ensure the identification of pregnancies that can be linked to pregnancy outcomes and perinatal outcomes if they are registrable births, in addition to the CPRD Pregnancy Register (PR), we included other sources of linked pregnancy data including one primary care dataset: CPRD MBL, and one secondary care dataset: maternity data from HES Maternity as provided by CPRD. CPRD MBL identifies births in women's records and links mothers with babies born in the same family (with the same practice-specific family number primarily based on residence) within the appropriate time period. HES Maternity data is part of HES Admitted Patient Care data that is routinely linked to the CPRD [6]. It contains hospital and out-of-hospital births where care is provided by National Health Service (NHS) staff in England and includes details such as mode of birth and gestation at birth, as well as information about the baby, such as sex and birthweight [6, 7]. The CPRD clinical, referral, and therapy files, along with the CPRD PR and HES Maternity files, were used to derive exemplar exposures and outcomes for assessing potential bias introduced by the 'as provided' population.

Preparation of source files and identification of overlapping pregnancies

CPRD MBL and HES Maternity datasets were both reshaped to one record per pregnancy, before being combined with the CPRD PR data. As the data provided in HES have been pseudonymised, the HES record does not include the baby's date of birth. We used the end date of the birth episode minus the number of days of postnatal stay as a proxy. To check for duplicate pregnancies within HES Maternity, defined as pregnancies with at least one day of overlap, we subtracted the gestational age, if available, from the estimated baby's date of birth to derive the start date of each pregnancy. When the gestational age is not available, a start date of 36 weeks before the end date of a birth episode or 24 weeks before the end date of a pregnancy episode not ending in a registrable birth was assigned to each pregnancy. We selected these cut-offs to ensure they are sufficiently long to capture potentially overlapping pregnancies, based on the assumption that closely dated maternity admission records are more likely to refer to the same pregnancy. At the same time, we aimed to avoid grouping genuinely distinct pregnancies together by using a cut-off one week shorter than a typical 'term' gestation. The pregnancies were then grouped and duplicates removed using an algorithm, details of which are presented in Supplementary Appendix 1.

CPRD PR includes estimated start dates (first day of a woman's last menstrual period) and end dates for all pregnancies included in the dataset regardless of outcome. In contrast, neither the MBL nor HES has a start date recorded for pregnancies. After cleaning and reshaping the MBL and HES Maternity datasets, a start date 36/35/33 weeks before the birth date of a singleton/twin/triplet was assigned to relevant births respectively, for pregnancies where the gestational age is not available. Similar to the HES Maternity data cleaning, these approximate gestation periods were used to flag potentially overlapping pregnancies while ensuring that genuinely distinct pregnancies were not grouped together in most cases. The approach assumes that closely dated maternity admission records are more likely to reflect the same pregnancy. CPRD PR, CPRD MBL and HES Maternity data were then combined and overlapping pregnancies, identified as at least one day overlapping in dates, were identified and grouped together.

Algorithm to identify one pregnancy per record

The original PR algorithm was designed to have high sensitivity and identify all potential pregnancies, while at the same time we wanted to increase the specificity to be more certain that the pregnancies that were identified were real and had occurred at the time recorded. The aim of the combining and cleaning process was to improve the reliability of the pregnancy data, to ensure to identify true pregnancies, to reduce records to one record per pregnancy, i.e. deduplication, and to eliminate records that were likely to be historical or partial records erroneously identified as a separate unique pregnancy by the PR algorithm. Several overarching rules were developed based on the nature of the data sources to ensure priority was given to more reliable sources of data in identifying 'true' pregnancies and eliminating duplicates (Supplementary Appendix 2).

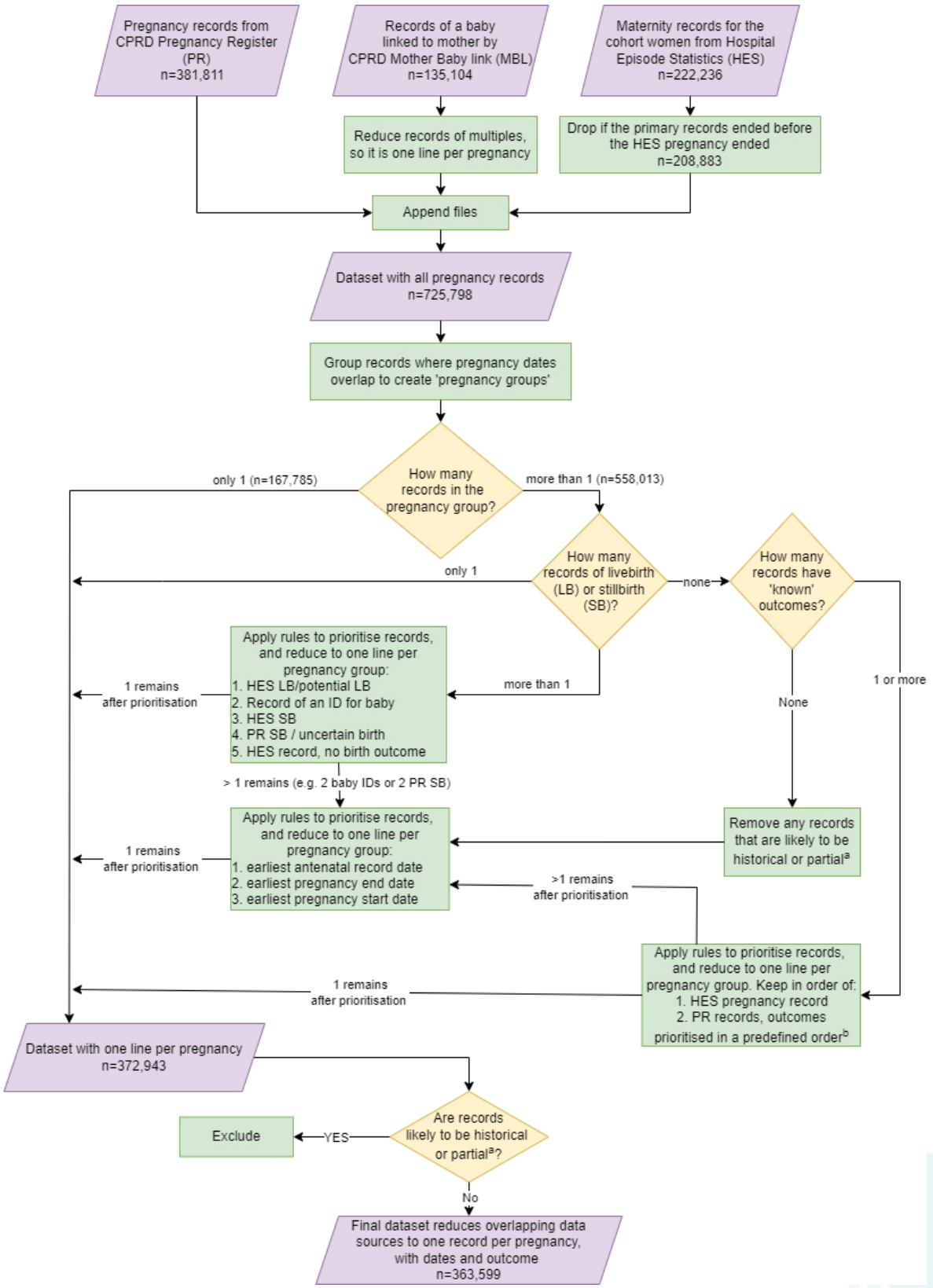
Details of the algorithm are shown in Figure 1. Records identified as belonging to the same pregnancy went through a series of combination processes until there was only one record left for the pregnancy.

After data from all sources were combined into one record per pregnancy, we further cleaned those considered potentially historical or partial records. Pregnancies with an unknown outcome that had a gestational age of exactly 28 days were removed if the date of the first antenatal record was the same as the end date of the pregnancy and if the pregnancy fell within 28 days of another pregnancy with a known outcome. This combination likely indicates historical or individual records belonging to a pregnancy that the PR algorithm did not group with other related records [2, 4].

Comparison between the 'as provided', 'derived' and 'strictly derived' populations

We calculated the number of pregnancies identified in the 'as provided', 'derived' and 'strictly derived' populations respectively. We compared the following characteristics between the three populations: maternal age, ethnicity, region, Index of Multiple Deprivation (IMD) of the GP practice and BMI, using mean and standard deviation (SD) for continuous variables and percentage for categorical variables. The characteristics were then checked against national statistics from the Office for National Statistics and other published national data for the same time period [8–13]. Additionally we compared the proportion of pregnancies with pre-pregnancy care records in the year before the pregnancy, pre-existing diseases prior to the pregnancy, and pregnancy

Figure 1: Flow chart combining Pregnancy Register (PR), Mother Baby Link (MBL), and maternity data from Hospital Episode Statistics (HES) Maternity



[a] Historical or partial records refer to those likely misidentified as separate, unique pregnancies by the PR algorithm. These include pregnancies with an unknown outcome that had a gestational age of exactly 28 days, where the date of the first antenatal record was the same as the end date of the pregnancy, and the pregnancy fell within 28 days of another pregnancy with a known outcome. This combination likely indicates historical or individual records belonging to a pregnancy that the PR algorithm did not group with other related records [2, 4].

[b] Outcome of pregnancy in PR were prioritised in a predefined order: termination of pregnancy (TOP) > miscarriage or TOP > miscarriage > ectopic > molar > blighted ovum > unspecified loss > outcome unknown.

and birth outcomes. Pregnancies starting in 2017 and 2018 were investigated first to explore the impact of applying the algorithm on more recent pregnancies, and then the comparisons were expanded to include all pregnancies to explore the impact of applying both the algorithm and the 'strict' criteria on pregnancies across women's reproductive history.

Assessing the extent of bias introduced by the 'as provided' Pregnancy Register data

We explored the potential bias that may be introduced by using the 'as provided' Pregnancy Register data with two relatively common exposures as examples, ever-diagnosed asthma and actively managed asthma, i.e. asthma treated in the year before the pregnancy started. For each exposure, we looked at two relatively common outcomes, gestational diabetes (GDM) and preterm birth (PTB), as evidence suggests that asthma is associated with an increase in these outcomes [14]. Records from the CPRD clinical, referral, and therapy files were used to derive asthma (ever diagnosed), actively managed asthma (defined as ever diagnosed plus treatment in the last year) and GDM variables, using published code lists [15, 16]. Gestational age and PTB records in the PR and the HES Maternity file were used to generate the PTB variable. We compared numbers and proportions of cases of the outcomes, and calculated unadjusted and adjusted odds ratios (ORs) using logistic regression models in the three populations, adjusting for maternal age, region, practice IMD and pregnancy starting year.

Results

Combining pregnancy data from three sources and cleaning

Initially 381,811 records/pregnancies (130,429 women) were supplied in the CPRD PR dataset, plus 135,104 records/pregnancies from CPRD MBL and 208,883 pregnancies from the HES Maternity data, giving a total of 725,798 pregnancy records for 137,285 women. Among them 167,785 records were identified as not conflicting with other records and retained. For the 558,013 records identified as overlapping with at least one other, our proposed algorithm was applied to clean and remove duplicates and potentially historical or partial records. After the combining and cleaning process, there was a total of 363,599 pregnancies (137,283 women) in the cohort (Figure 1). After data were further restricted to the pregnancies that started after women became registered with a CPRD GP practice and after data from the participating GP practice were deemed to be of research quality, there were 181,381 pregnancies (91,986 women) in the cohort (Table 1).

For pregnancies starting in 2017 and 2018, there were 20,221 pregnancies (15,889 women) in the 'as provided' PR data and 17,839 pregnancies (15,518 women) in the 'derived' data. For these pregnancies, the 'strictly derived' data are the same as the 'derived' data, as it was the inclusion criterion for women to be registered with a GP practice considered contributing data of research quality in those two years.

The proportion of additional pregnancies added to the PR data from MBL and HES is showed in Supplementary Appendix 3. Most additional pregnancies were identified from HES data. Over time, 0.1% of additional pregnancies were from MBL only, 6.9% from HES only and 0.4% from both MBL and HES but not identified by PR.

Comparison between the 'as provided', 'derived' and 'strictly derived' populations

When looking at the pregnancies occurring in 2017-2018 only, the 'as provided' and the 'derived' populations were similar in characteristics and the proportion of pregnancies with pre-existing health conditions and pregnancies that received pre-pregnancy care (Table 2). Compared with the 'as provided' population, the 'derived' population included around a 9% higher proportion of live births (57.1% versus 48.3%) and around a 10% lower proportion of pregnancy with unknown outcomes (23.8% versus 34.0%).

When looking at all pregnancies across women's reproductive history, the same pattern was observed for the 'as provided' and the 'derived' populations - characteristics and the proportion of pregnancies occurring in women with pre-existing conditions and pregnancies that received pre-pregnancy care were similar, and the 'derived' population had a higher proportion of live births and a lower proportion of pregnancies with unknown outcomes (Table 3). When the pregnancies were further restricted to women with active registration and up-to-standard data quality ('strictly derived' population), across women's whole reproductive history, there were fewer pregnancies to women at a younger maternal age, as some of the health records in their early life, including earlier pregnancy history, were cut off when 'strictly derived' (Table 3). Other characteristics remain similar to the 'as provided' and the 'derived' populations. The proportion of pregnancies occurring in women with pre-existing conditions and pregnancies that received pre-pregnancy care increased with the quality restriction.

Exploration of potential bias introduced by using the 'as provided' pregnancy register data

The absolute proportions of pregnancies with GDM and pregnancies that ended with PTB increased as more restrictions were applied to the study population, suggesting detection is improved (Table 3 and Supplementary Appendix 4). The association between ever-diagnosed asthma and GDM, or ever-diagnosed asthma and PTB remained similar across the three populations (Figure 2). Similarly, the association between actively managed asthma and GDM stayed largely unchanged across different populations (Figure 3(a)). However, the association between actively managed asthma and PTB was attenuated as more restrictions were applied to the study population (Figure 3(b)). The adjusted ORs (95%CI) reduced from 1.39 (1.29–1.49) in the 'as provided' population, to 1.28 (1.19–1.37) in the 'derived' population. This further decreased to 1.16 (1.06–1.26) once the population was further restricted to those currently registered with a GP practice contributing data of research quality, i.e. the 'strictly derived' population.

Table 1: Number and percentage of pregnancies in the dataset for the Pre-pregnancy study according to different inclusion criteria

Population and inclusion criteria	Pregnancies starting in 2017–2018 in eligible women	Percentage (%)	Pregnancies in eligible women	Percentage (%)
1. 'As provided'	20,221	N/A	381,811	N/A
2. 'Derived'	17,839	100.0	363,599	100.0
3. 'Strictly derived'	17,839	100.0	181,381	49.9

Note that the 'as provided' population includes only pregnancies identified in the Pregnancy Register (PR) within the primary care (Clinical Practice Research Datalink, CPRD) data.
The 'derived' population is the population where duplicate, historical, and partial records of pregnancies in the PR have been removed, and additional pregnancies from Hospital Episode Statistics and Mother Baby Link have been included.
The 'strictly derived' population is a subset of the 'derived' population, further restricted to include only pregnancies that started after the woman registered with a CPRD general practitioner (GP) practice, and during a period when data from her GP practice were considered to be of research quality.

Discussion

Use of our proposed algorithm added around 7% of pregnancies that were only identified from the hospital data over time, eliminated conflicting pregnancies and pregnancies with unknown outcomes, and reduced potentially historical or partial records of pregnancies in the CPRD Pregnancy Register.

Characteristics are similar between 'as provided', 'derived' and 'strictly derived' populations for recent pregnancies. However, for all pregnancies across women's reproductive history, when restricted to data with better research quality, i.e. 'strictly derived' population, there are more pregnant women with pre-existing medical conditions and more who received pre-pregnancy care. In the strictest dataset, where there is better recording of both exposure and outcome, the magnitude of the association between exposure and outcome is reduced compared to the 'as provided' population, although the reduction was only observed for the preterm birth outcome and the actively managed asthma exposure.

CPRD Pregnancy Register is a valuable data source for research related to pregnancy and birth. The algorithm used to generate the Pregnancy Register is sensitive because it identifies any pregnancy related records and flags them as potential pregnancies. However, it is not specific as it may be that these codes do not denote a pregnancy at the time of recording, or a separate unique pregnancy event. It picks up most pregnancies in the record but does not necessarily time them correctly or identify when records refer to the same pregnancy. As a result, the large proportion of unknown outcomes and pregnancies conflicting with one another presents methodological challenges for researchers using the Pregnancy Register. Simply including or excluding all of these uncertain pregnancy episodes may both introduce bias for studies with a particular focus [4].

Another issue with using CPRD PR to identify pregnancies is that it does not capture all pregnancies. This may be related to recording issues, or changes in provision of maternity services. From 2007, women have been able to self-refer to midwives to directly access antenatal midwifery services without the need for a GP referral [17]. The proportion of women taking this approach has increased over time, with 59% multiparous women and 45% primiparous women going first to a midwife in 2019 without seeing their GP [18]. Data from HES additionally identified about 8.3% of pregnancies that had no PR match between 1987 and Feb 2018 [2]. This is a particularly important issue when a study is related to the prevalence of pregnancy, or pregnancy-related variables.

There is not a recommended 'standard' methodological approach to address uncertain pregnancy episodes, hence the onus is on the researchers to seek out the best approach. The data cleaning process is therefore dependent upon individual researchers and the research questions. In this worked example, we are interested in preconception care, so it is imperative that we identify all pregnancies. We therefore used augmented data from HES and MBL. In this example, we also need to ensure pregnancies are only counted once so the denominator is correct, and the timing of the start of pregnancy is accurate. Therefore, we were cautious to ensure that early pregnancy records were not counted as pre-conception care. However, the approach required may vary by research question. For example, if ever having a past pregnancy is important, then researchers may choose to keep some of the pregnancies we removed.

Data quality is affected by the use of data preceding women's registration with their contributing GP practices. Again, approaches taken depend on the research question. For an estimate of any past history, such as past pregnancies or a proxy for parity, pregnancies outside the registration period may need to be included. In contrast, for estimates of treatment, care, or outcomes, it is important to only include pregnancies that occur during active registration and when the records are up to research standard, to ensure that cases are not missed and prevalence is more accurately ascertained. When these data quality standards are applied, there are fewer pregnancies to women at a younger maternal age. This is likely to be mainly an artefact of restricting the data to improve the quality. Some health records, including pregnancy history from when women were younger, were cut off and excluded because they occurred before the women's current registration with their GP practice or before data from the contributing GP practices met research standards. Additionally, there is a general trend of increasing maternal age over time [19]. Therefore, when stricter data quality standards are applied and earlier medical records are excluded as a

Table 2: Comparison of the characteristics between the 'as provided' and the 'derived' population for pregnancies starting in 2017–2018

	'As provided' N = 20,221	Column % 100.0	'Derived' N = 17,839	Column % 100.0	National figures for all births for comparison (%)[a]
Sociodemographic characteristics					
Age, mean (SD)	30.5 (5.8)		30.5 (5.8)		30.5
<20 yrs	412	2.0	366	2.1	3.0
20-24 yrs	3,003	14.9	2,624	14.7	14.4
25-29 yrs	5,242	25.9	4,607	25.8	28.0
30-34 yrs	6,384	31.6	5,607	31.4	31.9
35-39 yrs	3,999	19.8	3,560	20.0	18.4
40-44 yrs	1,058	5.2	958	5.4	4.0
>=45 yrs	123	0.6	117	0.7	0.3
Ethnic group					
White British	10,244	69.3	9,030	69.4	69.2
White other	1,879	12.7	1,684	12.9	7.8
Mixed	266	1.8	233	1.8	9.7[b]
Asian or Asian British	1,307	8.8	1,121	8.6	8.6
Black or Black British	727	4.9	635	4.9	4.6
Chinese or other	362	2.5	312	2.4	-
Missing	*5,436*	*26.9*	*4,824*	*27.0*	*6.2*
Geographical region					
North East, Yorkshire & The Humber	259	1.3	256	1.4	13.8
North West	3,515	17.4	3,098	17.4	13.0
West Midlands	2,604	12.9	2,164	12.1	10.7
East of England	840	4.2	784	4.4	10.9
South West	2,395	11.8	1,990	11.2	8.6
South Central	2,168	10.7	1,868	10.5	N/A
London	3,284	16.2	3,029	17.0	19.5
South East Coast	5,156	25.5	4,650	26.1	15.3
Individual-level area deprivation (IMD)					
1 (least deprived)	4,684	23.2	4,155	23.3	15.0
2	3,588	17.8	3,197	17.9	17.0
3	3,969	19.6	3,494	19.6	19.0
4	3,987	19.7	3,494	19.6	23.0
5 (most deprived)	3,982	19.7	3,488	19.6	27.0
Missing	*11*	*0.1*	*11*	*0.1*	–
Practice-level area deprivation (IMD)					
1 (least deprived)	3,322	16.4	2,936	16.5	15.0
2	3,438	17.0	3,072	17.2	17.0
3	3,943	19.5	3,432	19.2	19.0
4	3,568	17.7	3,082	17.3	23.0
5 (most deprived)	5,950	29.4	5,317	29.8	27.0
Health status and risk behaviours					
BMI (kg/m^2)					
<18.5	772	4.2	701	4.3	4.5
18.5-24.9	8,952	48.6	7,875	48.7	46.5
25-29.9	4,632	25.1	4,055	25.1	27.4
≥30	4,068	22.1	3,540	21.9	21.6
Missing	*1,797*	*8.9*	*1,668*	*9.4*	*18.7*
Pre-existing chronic health conditions					
Diabetes mellitus	201	1.0	176	1.0	
Hypertension	213	1.1	181	1.0	

Continued

Table 2: Continued

	'As provided' N = 20,221	Column % 100.0	'Derived' N = 17,839	Column % 100.0	National figures for all births for comparison (%)[a]
Asthma (ever diagnosed)	3,507	17.3	3,086	17.3	
Actively managed asthma (ever diagnosed+treated in the last year)	1,229	6.1	1,078	6.0	
Pre-pregnancy care or advice					
Specific pre-pregnancy care and advice	1,450	7.2	1,271	7.1	
General health promotion	8,517	42.1	7,410	41.5	
Opportunities for intervention	4,149	20.5	3,776	21.2	
Outcomes					
Outcomes during pregnancy					
Gestational diabetes	531	2.6	492	2.8	7.6[c]
Hypertensive disorder of pregnancy (HDP)	53	0.3	50	0.3	4.8[c]
Pregnancy outcomes					
Live birth	9,760	48.3	10,182	57.1	99.6[c]
Stillbirth	35	0.2	39	0.2	0.4[c]
Birth (live birth or stillbirth, unspecified)	0	0.0	96	0.5	N/A
Miscarriage	2,095	10.4	1,883	10.6	6.0[d]
Termination	212	1.1	206	1.2	N/A
Miscarriage or termination of pregnancy (TOP)	914	4.5	878	4.9	N/A
Other early loss	340	1.7	306	1.7	N/A
Outcome unknown	6,865	34.0	4,249	23.8	11.5[c]
Gestational age (weeks), mean (SD) in all births	39.1 (2.8)		38.7 (3.5)		
Preterm in all births	971	9.9	1,014	9.9	8.2
Missing	–	–	*110*	*1.1*	*16.6*
Birthweight (grams), mean (SD) in all births	N/A	N/A	3344.3 (584.5)		
Low birthweight (birthweight <2500 grams) in all births	N/A	N/A	539	6.9	6.8
Missing	*N/A*	*N/A*	*2,525*	*24.5*	*10.6*
Mode of birth					
Vaginal birth	N/A	N/A	4,486	58.9	58.6
Instrumental birth	N/A	N/A	948	12.4	12.6
Caesarean section	N/A	N/A	2,184	28.7	28.8
Missing	N/A	N/A	*2,706*	*26.2*	*1.5*

[a] Comparison figures drawn from: for mean maternal age and maternal age groups in 2017, from Office for National Statistics (ONS) [8]; for births by geographical area in 2017, from ONS [9]; for births by IMD group in 2017, from ONS (2019) [10]; for births by maternal ethnicity (2006-2012), from Li et al (2018) [11]; for births by maternal BMI in 2017, from Public Health England [12]; for outcomes in 2017-2018, from NHS Digital [13].

[b] This group in the referenced literature includes both Mixed and/or other groups.

[c] National statistics on outcomes are based on reports of deliveries in NHS hospitals. Cautions should be exercised when making comparisons, as the numerator only includes deliveries in NHS hospitals, excluding pregnancies that ended in early loss or termination.

[d] Miscarriages are not recorded as delivery episodes in the source data. As such, they are not included in the total delivery count, meaning the 'rate' presented is actually a ratio, since the numerator is not included in the denominator.

Table 3: Comparison of the characteristics between the 'as provided', 'derived' and 'strictly derived' populations for all pregnancies of all women in the cohort

	'As provided' N = 381,811	Column % 100.0	'Derived' N = 363,599	Column % 100.0	'Strictly derived' N = 181,381	Column % 100.0
Sociodemographic characteristics						
Age, mean (SD)	27.4 (6.2)		27.2 (6.2)		29.2 (6.0)	
<20 yrs	45,594	11.9	45,042	12.5	11,282	6.2
20-24 yrs	84,335	22.1	81,545	22.6	30,158	16.6
25-29 yrs	105,919	27.7	100,675	27.9	49,008	27.0
30-34 yrs	94,619	24.8	87,237	24.2	55,126	30.4
35-39 yrs	42,891	11.2	38,779	10.7	29,259	16.1
40-44 yrs	7,960	2.1	7,155	2.0	6,093	3.4
>=45 yrs	493	0.1	470	0.1	453	0.3
Missing	–	–	*2,696*	*0.7*	*2*	*0.0*
Ethnic group						
White British	194,694	72.4	186,321	72.4	85,637	73.6
White other	29,964	11.1	28,505	11.1	13,019	11.2
Mixed	4,103	1.5	3,993	1.6	1,671	1.4
Asian or Asian British	20,905	7.8	19,614	7.6	8,452	7.3
Black or Black British	13,138	4.9	13,032	5.1	5,016	4.3
Chinese or other	6,087	2.3	5,843	2.3	2,578	2.2
Missing	*112,920*	*29.6*	*106,291*	*29.2*	*65,008*	*35.8*
Geographical region						
North East, Yorkshire & The Humber	15,437	4.0	14,992	4.1	7,409	4.1
North West	62,800	16.5	59,258	16.3	32,623	18.0
West Midlands	46,782	12.3	42,386	11.7	21,623	11.9
East of England	24,287	6.4	22,598	6.2	11,714	6.5
South West	40,788	10.7	37,388	10.3	19,653	10.8
South Central	44,390	11.6	39,956	11.0	19,399	10.7
London	53,460	14.0	54,830	15.1	24,607	13.6
South East Coast	93,867	24.6	92,191	25.4	44,353	24.5
Individual-level area deprivation (IMD)						
1 (least deprived)	95,417	25.0	89,902	24.7	45,850	25.3
2	69,419	18.2	65,763	18.1	33,085	18.3
3	70,759	18.5	67,388	18.5	33,623	18.6
4	72,290	18.9	69,322	19.1	34,067	18.8
5 (most deprived)	73,780	19.3	71,067	19.6	34,679	19.1
Missing	*146*	*0.0*	*157*	*0.0*	*77*	*0.0*
Practice-level area deprivation (IMD)						
1 (least deprived)	63,502	16.6	60,443	16.6	30,802	17.0
2	67,677	17.7	64,414	17.7	30,349	16.7
3	75,774	19.9	71,190	19.6	35,405	19.5
4	58,911	15.4	57,268	15.8	28,267	15.6
5 (most deprived)	115,947	30.4	110,284	30.3	56,558	31.2
'Parity'/total number of previous pregnancies						
0	130,428	34.2	137,283	37.8	91,986	50.7
1	99,651	26.1	102,277	28.1	50,138	27.6
2∼4	125,515	32.9	108,240	29.8	36,113	19.9
5+	26,217	6.9	15,799	4.4	3,144	1.7
Health status and risk behaviours						
BMI (kg/m^2)						
<18.5	11,641	4.9	10,483	5.0	7,025	4.5
18.5-24.9	128,149	53.8	112,797	54.0	82,625	52.6
25-29.9	57,288	24.0	49,914	23.9	38,652	24.6

Continued

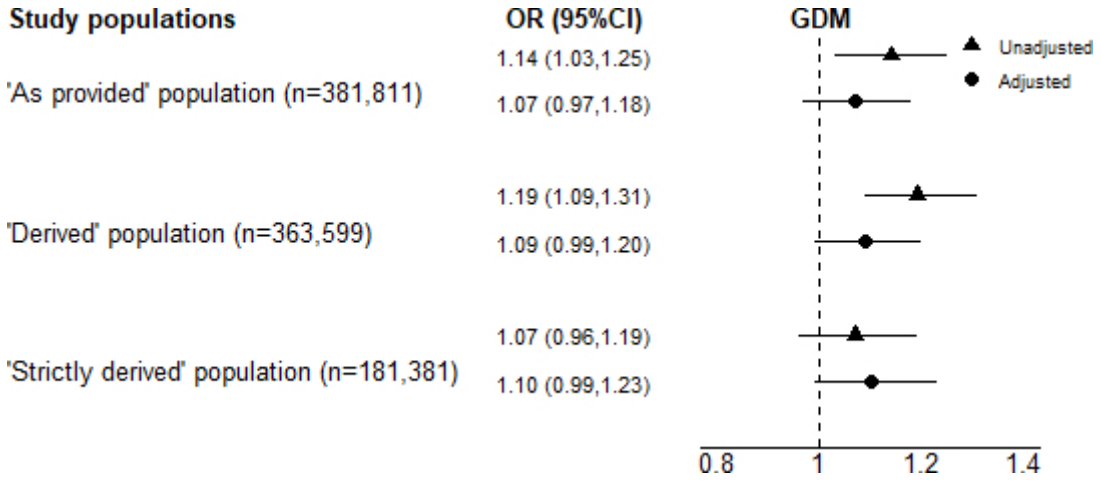
Table 3: Continued

	'As provided' N = 381,811	Column % 100.0	'Derived' N = 363,599	Column % 100.0	'Strictly derived' N = 181,381	Column % 100.0
≥30	41,209	17.3	35,546	17.0	28,842	18.4
Missing	*143,524*	*37.6*	*154,859*	*42.6*	*24,237*	*13.4*
Pre-existing chronic health conditions						
Diabetes mellitus	2,170	0.6	2,064	0.6	1,297	0.7
Hypertension	2,952	0.8	2,631	0.7	1,825	1.0
Asthma (ever diagnosed)	52,310	13.7	47,742	13.1	28,915	15.9
Actively managed asthma (ever diagnosed+treated in the last year)	16,959	4.4	14,596	4.0	11,495	6.3
Pre-pregnancy care or advice						
Specific pre-pregnancy care and advice	20,279	5.3	17,531	4.8	13,821	7.6
General health promotion	140,141	36.7	121,675	33.5	91,507	50.6
Opportunities for intervention	73,810	19.3	69,366	19.1	42,708	23.6
Outcomes						
Outcomes during pregnancy						
Gestational diabetes	4,407	1.2	4,162	1.1	2,992	1.7
Hypertensive disorder of pregnancy (HDP)	1,616	0.4	1,445	0.4	689	0.4
Pregnancy outcomes						
Live birth	223,548	58.6	237,130	65.2	119,517	65.9
Stillbirth	960	0.3	1,376	0.4	536	0.3
Birth (live birth or stillbirth, unspecified)	11	0.0	5,994	1.7	618	0.3
Miscarriage	35,153	9.2	30,570	8.4	16,795	9.3
Termination	5,895	1.5	5,700	1.6	2,293	1.3
Miscarriage or termination of pregnancy (TOP)	37,858	9.9	36,679	10.1	13,275	7.3
Other early loss	6,555	1.7	5,657	1.6	2,719	1.5
Outcome unknown	71,831	18.8	40,493	11.1	25,628	14.1
Gestational age (weeks), mean (SD) in all births	39.2 (3.7)		38.9 (3.9)		38.7 (4.1)	
Preterm in all births	15,304	6.8	17,533	7.3	10,221	8.5
Missing	–	–	*4,793*	*2.0*	*1,017*	*0.8*
Birthweight (grams), mean (SD) in all births	N/A	N/A	3359.3 (589.5)		3382.4 (584.4)	
Low birthweight (birthweight <2500 grams) in all births	N/A	N/A	11,011	6.5	5,751	6.1
Missing	*N/A*	*N/A*	*74,162*	*30.3*	*25,589*	*21.2*
Mode of birth						
Vaginal birth	N/A	N/A	107,648	63.7	59,489	63.1
Instrumental birth	N/A	N/A	21,568	12.8	11,332	12.0
Caesarean section	N/A	N/A	39,921	23.6	23,391	24.8
Missing	N/A	N/A	*75,363*	*30.8*	*26,459*	*21.9*

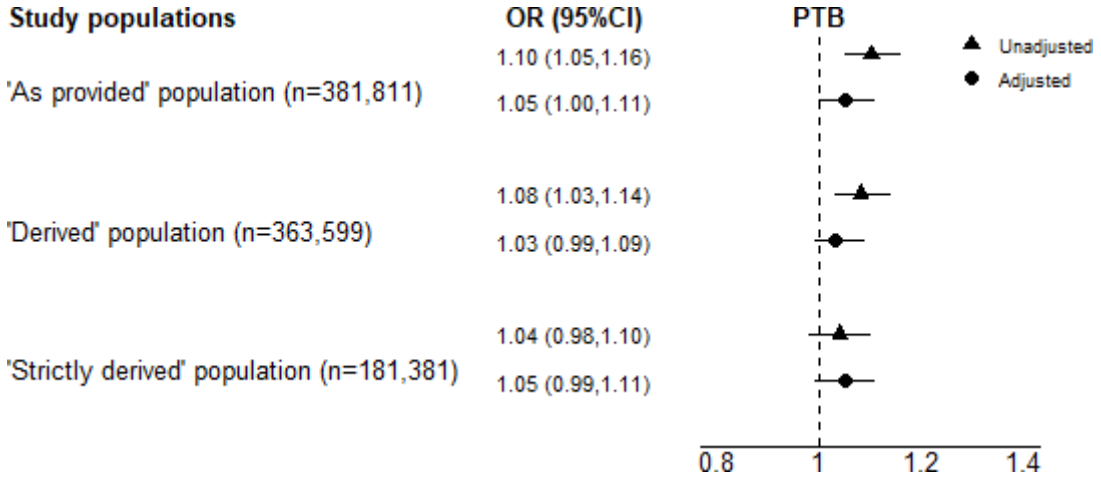
result, the average maternal age increases. We found that the magnitude of the association between exposure and outcome is reduced comparing the 'strictly derived' population to the 'as provided' population, and this reduced association is broadly in line with other estimates from routine data sources [20–22]. The most plausible explanation is that the 'as provided' and, to a lesser extent, the 'derived', artificially inflate the observed associations. Including conflicting pregnancies in the 'as provided' population means that some pregnancies are counted more than once, hence inflating the observed

Figure 2: Extent of bias introduced by the 'as provided' Pregnancy Register data (based on all pregnancies), using ever-diagnosed asthma as the exposure variable

2(a) Outcome: gestational diabetes, GDM



2(b) Outcome: preterm birth, PTB



association. Data quality affects both the detection of the exposure (in this case asthma) and the outcome (GDM or PTB). Before restriction to records of high quality, women with asthma recorded are more likely to be those with better quality of data, therefore more likely to also have health outcomes recorded and identified, which in turn may have inflated the observed association. Upon restriction to records of high quality, observed effects diminish, indicating that individuals without asthma records genuinely lack the condition, and those with adverse outcomes indeed experience them. This effect is particularly pronounced when actively managed asthma is used as the exposure, because medication records are much less reliable if women are not registered.

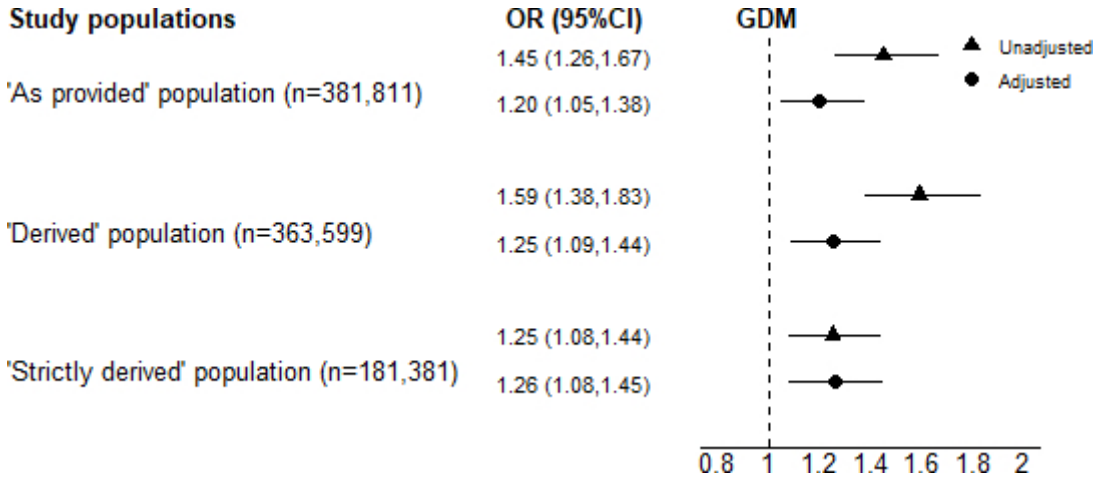
Using HES Maternity and MBL data to augment the PR data has the benefit of identifying some of the pregnancies with unknown outcomes in the PR as live births or stillbirths, without the need to understand reasons for them being unknown in the PR. Priority was given to the HES Maternity data, as HES data was used to validate the PR initially [2], and hospital data are generally more reliable than the algorithm-based primary care data and a good source to identify missing outcomes [23, 24]. Evidence of a baby registered at the GP by the parent/guardian is also good evidence of a birth, therefore this was given a priority over pregnancies with unknown outcomes in the cleaning and augmentation process. This approach also solves the problems of conflicting pregnancies in the PR. Potentially conflicting pregnancies were grouped together with only one record kept for each pregnancy following our proposed rules and algorithm, so any duplicates were removed without losing information for the pregnancy. With the augmentation from HES, it is also possible to study birth characteristics that are only available from the hospital data, such as birthweight and mode of birth.

The proposed algorithm does not require extensive use of additional data which may cost more and take time to obtain, for example, hospital image data is not usually requested by researchers to identify additional pregnancies. It provides a relatively quick, practical and cost effective way to identify real pregnancies and remove conflicting pregnancies using routine health data. The algorithm proposed also has the strength of being clear and rule-based, therefore can be replicated.

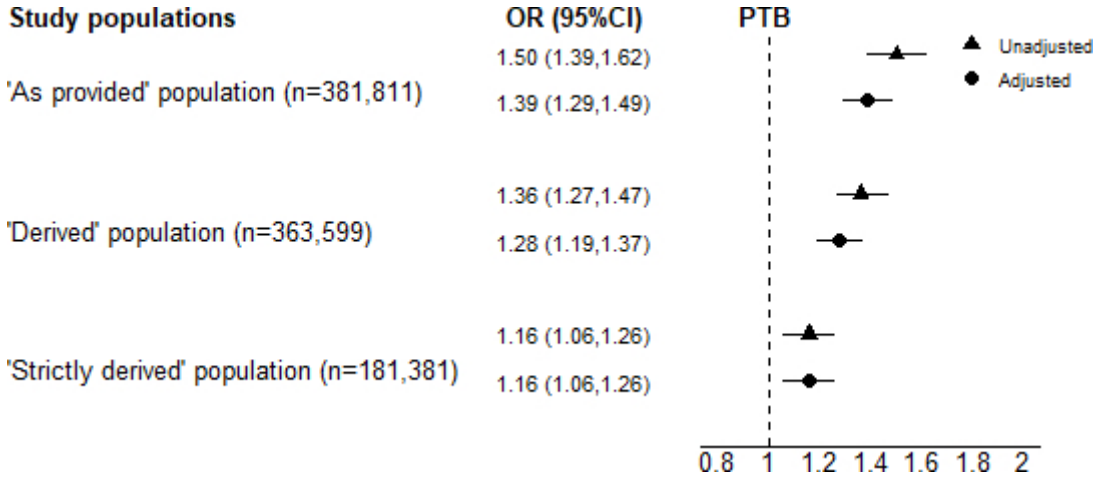
It is imperative to exercise caution and prudence when using the suggested algorithm, as there are certain caveats and limitations that require consideration. With the data used in this study, we will likely have lost some of the pregnancies that have the least reliable information recorded – so improves

Figure 3: Extent of bias introduced by the 'as provided' Pregnancy Register data (based on all pregnancies), using actively managed asthma as the exposure variable

3(a) Outcome: gestational diabetes, GDM



3(b) Outcome: preterm birth, PTB



the overall quality at the expense of the sensitivity of all pregnancies. Abortions performed by the British Pregnancy Advisory Service (BPAS) or other private providers would only be identified if the GP was informed. For the similar reason, miscarriages may also be underestimated. Not all patients in the CPRD have linked hospital data due to various reasons [25]. The completeness and quality of HES Maternity data can vary between service providers [26, 27]. HES Maternity data also have some known quality issues [25]. However, these may not have an impact on studies identifying pregnancies [25]. Validation of the proposed cleaning algorithm is not possible, as participants cannot be identified. Nonetheless our data agrees well with national statistics for those variables for which national data are available. Some of the decisions made in the proposed algorithm were driven by the example research question and the data needed, and others may decide to employ different rules as discussed above. Cut-offs chosen in the cleaning process to identify duplicating records of pregnancies can be arbitrary and assumption-based. Caution will be needed when adopting or adapting the proposed approach to clean the PR for other purposes, as other research has shown that assumptions made during data preparation can influence the outcomes of analyses [28].

Conclusion

While the CPRD Pregnancy Register is a useful resource for researchers, it has recognised limitations and needs careful and thoughtful cleaning before being used to resolve the uncertainty in identifying pregnancies. Using a worked example of investigating pre-pregnancy care, this study presents a pragmatic and practical way to identify more accurately pregnancies using data from three main CPRD and linked data sources, CPRD PR, MBL and HES Maternity, without the need for additional costly data. Researchers using the CPRD PR data need to consider carefully how inherent variability in data quality may influence study findings. Subsequently, they can align or modify the proposed approach based on their specific research questions.

Acknowledgements

We would like to thank our members of the Patient and Public Involvement (PPI) group for their valuable contributions to this research.

Statement of conflicts of interest

The authors declare that they have no competing interests.

Ethics statement

This analysis is part of a larger study approved by the CPRD Independent Scientific Advisory Committee (ISAC, protocol number: 20_000220).

Data availability statement

This study is based on data from the Clinical Practice Research Datalink (CPRD) obtained under licence from the UK Medicines and Healthcare products Regulatory Agency. The data is provided by patients and collected by the NHS as part of their care and support. The interpretation and conclusions contained in this study are those of the authors alone.

The datasets generated and/or analysed during the current study are not publicly available, as the data were provided by the CPRD under a contractual agreement that does not permit the sharing of data. Study documentation is available on request from the corresponding author.

Funding statement

This research is funded by the National Institute for Health and Care Research (NIHR) Policy Research Programme, conducted through the Policy Research Unit in Maternal and Neonatal Health and Care, PR-PRU-1217-21202. SK is part funded by NIHR grant 970014 through the Applied Research Collaborative (ARC) West Midlands (Maternity Theme). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Authors' contributions

C.C. and J.K. developed the protocol with input from all the other authors. Y.L. and C.C. developed the analysis plan with input from all the other authors. Y.L. compiled the code lists for pre-pregnancy care with input from C.C., N.P. and S.D.A. Y.L. cleaned, prepared and managed the data and conducted the statistical analysis with input from all the other authors. Y.L. and C.C. drafted the article with input from all the other authors. All authors were involved in interpretation of the findings, revised the manuscript critically for important intellectual content, and approved the final version.

References

1. Margulis AV, Palmsten K, Andrade SE, Charlton RA, Hardy JR, Cooper WO, 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):335-42. https://doi.org/10.1002/pds.3743

2. Minassian C, Williams R, Meeraus WH, Smeeth L, Campbell OMR, Thomas SL. Methods to generate and validate a Pregnancy Register in the UK Clinical Practice Research Datalink primary care database. Pharmacoepidemiol Drug Saf. 2019;28(7):923-33. https://doi.org/10.1002/pds.4811

3. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827-36. https://doi.org/10.1093/ije/dyv098

4. Campbell J, Bhaskaran K, Thomas S, Williams R, McDonald HI, Minassian C. Investigating the optimal handling of uncertain pregnancy episodes in the CPRD GOLD Pregnancy Register: a methodological study using UK primary care data. BMJ Open. 2022;12(2):e055773. https://doi.org/10.1136/bmjopen-2021-055773

5. Medicines & Healthcare products Regulatory Agency. CPRD GOLD Glossary of terms/Data definitions. 2023. Available from: https://www.cprd.com/sites/default/files/2023-02/CPRD%20GOLD%20Glossary%20Terms%20v2.pdf.

6. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). Int J Epidemiol. 2017;46(4):1093–i. https://doi.org/10.1093/ije/dyx015

7. Dattani N, Datta-Nemdharry P, Macfarlane A. Linking maternity data for England, 2005-06: methods and data quality. Health Stat Q. 2011(49):53–79. https://doi.org/10.1057/hsq.2011.3

8. Office for National Statistics. Birth characteristics in England and Wales: 2017. 2019. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthcharacteristicsinenglandandwales/2017#average-ages-of-mothers-and-fathers-of-all-babies-have-continued-to-rise.

9. Office for National Statistics. Live births in England and Wales by sex and characteristics of mother: national/regional. 2017. Available from: https://www.nomisweb.co.uk/query/construct/summary.asp?mode=construct&version=0&dataset=203.

10. Office for National Statistics. Figures on births by gestation, ethnic group, Index of Multiple Deprivation and area of usual residence 2017. 2019. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/adhocs/009477figuresonbirthsbygestationethnicgroupindexofmultipledeprivationandareaofusualresidence.

11. Li Y, Quigley MA, Dattani N, Gray R, Jayaweera H, Kurinczuk JJ, et al. The contribution of gestational age, area deprivation and mother's country of birth

to ethnic variations in infant mortality in England and Wales: A national cohort study using routinely collected data. PLoS One. 2018;13(4):e0195146. https://doi.org/10.1371/journal.pone.0195146

12. Public Health England. Health of women before and during pregnancy: health behaviours, risk factors and inequalities. 2019. Available from: https://assets.publishing.service.gov.uk/media/5dc00b22e5274a4a9a465013/Health_of_women_before_and_during_pregnancy_2019.pdf.

13. NHS Digital. NHS Maternity Statistics, England 2017-18. 2018. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/nhs-maternity-statistics/2017-18#.

14. Lao TT, Annie Hui SY. The obstetric aspects of maternal asthma. Best Pract Res Clin Obstet Gynaecol. 2022;85(Pt A):57–69. https://doi.org/10.1016/j.bpobgyn.2022.08.005

15. Nissen F, Morales DR, Mullerova H, Smeeth L, Douglas IJ, Quint JK. Validation of asthma recording in the Clinical Practice Research Datalink (CPRD). BMJ Open. 2017;7(8):e017474. https://doi.org/10.1136/bmjopen-2017-017474

16. Vounzoulaki E, Khunti K, Miksza JK, Tan BK, Davies MJ, Gillies CL. Screening for type 2 diabetes after a diagnosis of gestational diabetes by ethnicity: A retrospective cohort study. Prim Care Diabetes. 2022;16(3):445–51. https://doi.org/10.1016/j.pcd.2022.03.008

17. Department of Health. Maternity Matters: Choice, access and continuity of care in a safe service. 2007. Available from: https://webarchive.nationalarchives.gov.uk/ukgwa/20130103035958/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_074199.pdf.

18. NHS Care Quality Commission. 2019 survey of women's experiences of maternity care statistical release. 2019. Available from: https://www.cqc.org.uk/sites/default/files/20200128_mat19_statisticalrelease.pdf.

19. Office for National Statistics. Birth characteristics in England and Wales: 2021. 2023. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthcharacteristicsinenglandandwales/2021#age-of-parents.

20. Kemppainen M, Lahesmaa-Korpinen AM, Kauppi P, Virtanen M, Virtanen SM, Karikoski R, et al. Maternal asthma is associated with increased risk of perinatal mortality. PLoS One. 2018;13(5):e0197593. https://doi.org/10.1371/journal.pone.0197593

21. Shaked E, Wainstock T, Sheiner E, Walfisch A. Maternal asthma: pregnancy course and outcome. J Matern Fetal Neonatal Med. 2019;32(1):103-8. https://doi.org/10.1080/14767058.2017.1372414

22. Tronnes H, Wilcox AJ, Markestad T, Tollanes MC, Lie RT, Moster D. Associations of maternal atopic diseases with adverse pregnancy outcomes: a national cohort study. Paediatr Perinat Epidemiol. 2014;28(6):489-97. https://doi.org/10.1111/ppe.12154

23. Harron K, Gilbert R, Cromwell D, van der Meulen J. Linking Data for Mothers and Babies in De-Identified Electronic Health Data. PLoS One. 2016;11(10):e0164667. https://doi.org/10.1371/journal.pone.0164667

24. Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, Strongman H. Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. Eur J Epidemiol. 2019;34(1):91-9. https://doi.org/10.1007/s10654-018-0442-4

25. Medicines & Healthcare products Regulatory Agency. Hospital Episode Statistics (HES) Admitted Patient Care and CPRD primary care data Documentation (set 22/January 2022). 2021. Available from: https://cprd.com/sites/default/files/2022-02/Documentation_HES_APC_set22.pdf.

26. Knight HE, Gurol-Urganci I, Mahmood TA, Templeton A, Richmond D, van der Meulen JH, et al. Evaluating maternity care using national administrative health datasets: how are statistics affected by the quality of data on method of delivery? BMC Health Serv Res. 2013;13:200. https://doi.org/10.1186/1472-6963-13-200

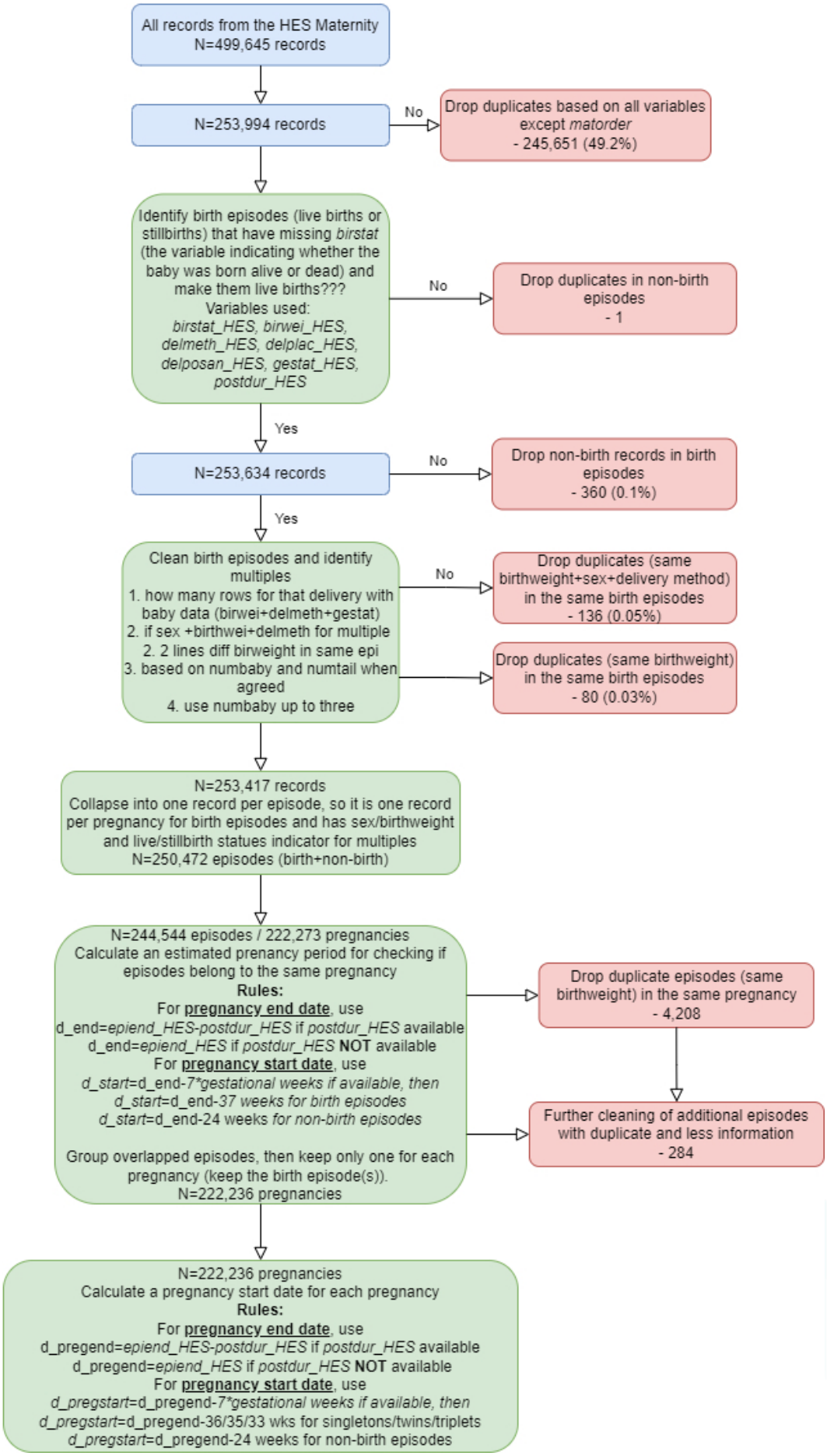
27. Murray J, Saxena S, Modi N, Majeed A, Aylin P, Bottle A, et al. Quality of routine hospital birth records and the feasibility of their use for creating birth cohorts. J Public Health (Oxf). 2013;35(2):298-307. https://doi.org/10.1093/pubmed/fds077

28. Pye SR, Sheppard T, Joseph RM, Lunt M, Girard N, Haas JS, et al. Assumptions made when preparing drug exposure data for analysis have an impact on results: An unreported step in pharmacoepidemiology studies. Pharmacoepidemiol Drug Saf. 2018;27(7):781-8. https://doi.org/10.1002/pds.4440

Abbreviations

CPRD:	Clinical Practice Research Datalink
GDM:	gestational diabetes
GP:	general practitioner
HES:	Hospital Episode Statistics
IMD:	Index of Multiple Deprivation
MBL:	Mother Baby Link
NHS:	National Health Service
OR:	odds ratio
PR:	Pregnancy Register
PTB:	preterm birth

Supplementary Appendix 1: Cleaning Hospital Episode Statistics Maternity data



Supplementary Appendix 2: Priority setting and overarching rules were developed based on the nature of the data sources to ensure priority was given to more reliable sources of data in identifying 'true' pregnancies and eliminating duplicates

1. To ensure pregnancies duplicated across data sources were only counted once, all data sources were considered at the same time.
2. The Pregnancy Register uses an algorithm to estimate pregnancy dates, and birth dates ('pregend') may be adjusted when records conflict with estimated dates or fall outside plausible periods. As a result, the estimated birth date may not be the actual birth date. The variable 'deldate' in CPRD Mother Baby Link (MBL) is an estimated date based on record of a birth in the mother's file. Hospital Episode Statistics (HES) data does not include a 'pregend' or a 'deldate' variable, so we used the end date of the birth episode. As a result, dates were not expected to match exactly.
3. Live births are always given precedence over other outcomes.
4. Linkage to a child in the MBL (evidenced by presence of a baby patient id number) was considered strong evidence of a birth.
5. A HES record of a birth or stillbirth was given precedence over the PR records.
6. When other information was the same, the record with the earliest antenatal record date, pregnancy end date, and then pregnancy start date, subsequently, was chosen to avoid pregnancy care being misclassified to pre-pregnancy care.

Supplementary Appendix 3: Proportion of pregnancies added from Mother Baby Link (MBL) and Hospital Episode Statistics (HES, 1997 onwards)

Source of data by pregnancy					
Year when pregnancy started	**Total number**	**PR (with or w/o evidence from MBL/HES)**	**MBL only**	**HES only**	**MBL+HES**
1997	10,260	71.7	0.0	28.1	0.1
1998	10,512	79.0	0.0	20.8	0.1
1999	10,500	85.5	0.1	14.3	0.2
2000	10,761	86.3	0.0	13.4	0.2
2001	11,869	88.8	0.0	11.0	0.2
2002	12,642	90.0	0.0	9.7	0.3
2003	13,759	90.7	0.0	8.9	0.3
2004	14,779	91.5	0.0	8.1	0.4
2005	15,262	91.7	0.1	7.9	0.4
2006	16,402	93.2	0.1	6.4	0.3
2007	17,629	93.2	0.3	6.4	0.2
2008	17,429	94.0	0.2	5.5	0.4
2009	18,088	94.3	0.0	5.1	0.5
2010	18,505	94.2	0.0	5.2	0.6
2011	18,329	95.2	0.1	4.3	0.4
2012	17,773	95.8	0.0	3.7	0.5
2013	17,249	96.7	0.1	2.7	0.5
2014	16,677	97.4	0.1	2.0	0.5
2015	16,132	98.3	0.1	1.1	0.6
2016	13,923	98.2	0.1	1.3	0.5
2017	10,528	98.0	0.1	1.4	0.5
2018	7,311	97.1	0.1	2.2	0.6
2019	5,116	96.5	0.2	2.9	0.4
2020	2,440	98.2	0.5	1.2	0.0
2021	355	100.0	0.0	0.0	0.0
Total	324,230	92.7	0.1	6.9	0.4

PR Pregnancy Register.
MBL Mother Baby Link.
HES Hospital Episode Statistics.

Supplementary Appendix 4: Extent of bias introduced by the 'as provided' Pregnancy Register data (based on all pregnancies)

	'As provided' N = 381,811	'Derived' N = 363,599	'Strictly derived' N = 181,381	'As provided'		'Derived'		'Strictly derived'	
				Unadjusted OR	Adjusted OR[a]	Unadjusted OR	Adjusted OR[a]	Unadjusted OR	Adjusted OR[a]
Exposure 1, Asthma	**Number of cases (row%)**	**Number of cases (row%)**	**Number of cases (row%)**						
Outcome 1, gestational diabetes (GDM)									
No exposure	3,734 (1.1)	3,527 (1.1)	2,488 (1.6)	Reference	Reference	Reference	Reference	Reference	Reference
Exposure	673 (1.3)	635 (1.3)	504 (1.7)	1.14 (1.03-1.25)	1.07 (0.97-1.18)	1.19 (1.09-1.31)	1.09 (0.99-1.20)	1.07 (0.96-1.19)	1.10 (0.99-1.23)
Total	4,407 (1.2)	4,162 (1.1)	2,992 (1.7)						
Outcome 2, preterm birth (PTB)									
No exposure	13,068 (6.7)	15,087 (7.2)	8,563 (8.5)	Reference	Reference	Reference	Reference	Reference	Reference
Exposure	2,236 (7.4)	2,446 (7.8)	1,658 (8.8)	1.10 (1.05-1.16)	1.05 (1.00-1.11)	1.08 (1.03-1.14)	1.03 (0.99-1.09)	1.04 (0.98-1.10)	1.05 (0.99-1.11)
Total	15,304 (6.8)	17,533 (7.3)	10,221 (8.5)						
Exposure 2, Actively managed asthma	**Number of cases (row%)**	**Number of cases (row%)**	**Number of cases (row%)**						
Outcome 1, gestational diabetes (GDM)									
No exposure	4,130 (1.1)	3,904 (1.1)	2,760 (1.6)	Reference	Reference	Reference	Reference	Reference	Reference
Exposure	277 (1.6)	258 (1.8)	232 (2.0)	1.45 (1.26-1.67)	1.20 (1.05-1.38)	1.59 (1.38-1.83)	1.25 (1.09-1.44)	1.25 (1.08-1.44)	1.26 (1.08-1.45)
Total	4,407 (1.2)	4,162 (1.1)	2,992 (1.7)						
Outcome 2, preterm birth (PTB)									
No exposure	14,373 (6.7)	16,609 (7.2)	9,491 (8.5)	Reference	Reference	Reference	Reference	Reference	Reference
Exposure	931 (9.7)	924 (9.6)	730 (9.7)	1.50 (1.39-1.62)	1.39 (1.29-1.49)	1.36 (1.27-1.47)	1.28 (1.19-1.37)	1.16 (1.06-1.26)	1.16 (1.06-1.26)
Total	15,304 (6.8)	17,533 (7.3)	10,221 (8.5)						