Appendix 1: Pregnancy code categories and example codes

Pregnancy code category*	Example Read code,	Description
Antenatal	62…00	Patient pregnant
Late pregnancy (≤3 weeks before delivery)	L281.00	Premature rupture of membranes
Third trimester	62N8.00	A/N 32 week examination
Delivery	L20..11	Spontaneous vaginal delivery
Stillbirth	Q4z..15	Stillbirth NEC
Ectopic	L03..00	Ectopic pregnancy
Miscarriage	L04..00	Spontaneous abortion
TOP	L052.11	Medical abortion - complete
Probable TOP	L05..12	Termination of pregnancy
Unspecified pregnancy loss	L0z..00	Pregnancy with abortive outcome NOS
Molar pregnancy	L002.00	Complete hydatidiform mole
Blighted ovum	L010.00	Blighted ovum
Postnatal (≤8 weeks after delivery)	62S7.00	Postnatal examination normal
Other postnatal**	E204.11	Postnatal depression
Preterm	L142.11	Premature delivery
Post-term	L150.00	Post-term pregnancy
Multiple	L210.00	Twin pregnancy
LMP	1513.00	Last menstrual period – 1st day
EDD	1514.12	Estimated date of delivery
EDC	Z22C500	Estimated date of conception
Pregnancy related (timing uncertain)**	L12..00	Hypertension complicating pregnancy/childbirth/puerperium

LMP=last menstrual period; EDD=estimated date of delivery; EDC=estimated date of conception; TOP=termination of pregnancy; NEC=not elsewhere classified; NOS=not otherwise specified; A/N=antenatal.

* Categories are not mutually exclusive, e.g. Late pregnancy and Third trimester codes are a subset of Antenatal codes.

**These codes are not used to determine pregnancy start and end dates due to uncertainty around which stage of pregnancy or the postnatal period they refer to.

Appendix 2: Approved ISAC application form

ISAC APPLICATION FORM

PROTOCOLS FOR RESEARCH USING THE CLINICAL PRACTICE RESEARCH DATALINK (CPRD)

For ISAC use only		
Protocol No. Submission date (DD/MM/YYYY)	....	***IMPORTANT*** *Please refer to the **guidance** for '**Completing the ISAC application form**' found on the CPRD website (www.cprd.com/isac). If you have any queries, please contact the ISAC Secretariat at isac@cprd.com.*

SECTION A: GENERAL INFORMATION ABOUT THE PROPOSED RESEARCH STUDY

- Study Title[§] (*Please state the study title below)*

Investigating pregnancies without recorded outcomes in the Clinical Practice Research Datalink / London School of Hygiene and Tropical Medicine Pregnancy Register, with the aim of improving validity.

[§]Please note: This information will be published on the CPRD's website as part of its transparency policy.

- **Has any part of this research proposal or a related proposal been previously submitted to ISAC?**
  Yes[*] ☒ No ☐

**If yes, please provide the previous protocol number/s below. Please also state in your current submission how this/these are related or relevant to this study.*
11_058 Is the original protocol for the development of the CPRD/LSHTM Pregnancy Register

- **Has this protocol been peer reviewed by another Committee? (e.g. grant award or ethics committee)**
  Yes[*] ☐ No ☒

**If Yes, please state the name of the reviewing Committee(s) below and provide an outline of the review process and outcome as an Appendix to this protocol :*

- **Type of Study** (please tick all the relevant boxes which apply)

Adverse Drug Reaction/Drug Safety	☐	Drug Effectiveness	☐
Drug Utilisation	☐	Pharmacoeconomics	☐
Disease Epidemiology	☐	Post-authorisation Safety	☐
Health care resource utilisation	☐	Methodological Research	☒
Health/Public Health Services Research	☐	Other[*]	☐

[*]If Other, please specify the type of study here and in the lay summary below:

- **Health Outcomes to be Measured[§]**

[§]Please note: This information will be published on CPRD's website as part of its transparency policy.

Please summarise below the primary/secondary health outcomes to be measured in this research protocol:

1. All end of pregnancy outcomes including live deliveries, stillbirths, early pregnancy losses and terminations.	*2.*	3.
4.	*5.*	6.
7.	*8.*	9.

[Please add more bullet points as necessary]

- **Publication: This study is intended for** (please tick all the relevant boxes which apply)**:**

Publication in peer-reviewed journals	☒	Presentation at scientific conference	☒
Presentation at company/institutional meetings	☒	Regulatory purposes	☐
Other*	☐		

**If Other, please provide further information:*

SECTION B: INFORMATION ON INVESTIGATORS AND COLLABORATORS

- **Chief Investigator§**

Please state the full name, job title, organisation name & e-mail address for correspondence - see guidance notes for eligibility. Please note that there can only be one Chief Investigator per protocol.

Jennifer Campbell, Senior Researcher, CPRD, jennifer.campbell@mhra.gov.uk

§*Please note: The name and organisation of the Chief Investigator and will be published on CPRD's website as part of its transparency policy*

CV has been previously submitted to ISAC	☒	**CV number:** 051_15CESL
A new CV is being submitted with this protocol	☐	
An updated CV is being submitted with this protocol	☐	

- **Affiliation of Chief Investigator** (full address)

The Clinical Practice Research Datalink
151 Buckingham Palace Road, London,SW1W9SZ

And

The London School of Hygiene and Tropical Medicine
Keppel St, Bloomsbury, London WC1E 7HT

- **Corresponding Applicant§**

Please state the full name, affiliation(s) and e-mail address below:

Jennifer Campbell, Clinical Practice Research Datalink and the London School of Hygiene & Tropical Medicine, Jennifer.campbell@mhra.gov.uk

§*Please note: The name and organisation of the corresponding applicant and their organisation name will be published on CPRD's website as part of its transparency policy*

Same as chief investigator	☒	
CV has been previously submitted to ISAC	☐	**CV number:**
A new CV is being submitted with this protocol	☐	
An updated CV is being submitted with this protocol	☐	

- **List of all investigators/collaborators[§]**

Please list the full name, affiliation(s) and e-mail address* of all collaborators, other than the Chief Investigator below:

[§]*Please note: The name of all investigators and their organisations/institutions will be published on CPRD's website as part of its transparency policy*

Other investigator: Professor Sara Thomas, London School of Hygiene & Tropical Medicine, sara.thomas@lshtm.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 270_15CESL
A new CV is being submitted with this protocol ☐
An updated CV is being submitted with this protocol ☐

Other investigator: Dr Caroline Minassian, London School of Hygiene & Tropical Medicine, caroline.minassian@lshtm.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 029_17
A new CV is being submitted with this protocol ☐
An updated CV is being submitted with this protocol ☐

Other investigator: Rachael Williams, Clinical Practice Research Datalink, rachael.williams@mhra.gov.uk

CV has been previously submitted to ISAC ☒ **CV number:** 130_15CESL
A new CV is being submitted with this protocol ☐
An updated CV is being submitted with this protocol ☐

Other investigator:
CV has been previously submitted to ISAC ☐ **CV number:**
A new CV is being submitted with this protocol ☐
An updated CV is being submitted with this protocol ☐

[Please add more investigators as necessary]

Please note that your ISAC application form and protocol **must be copied to all e-mail addresses listed above at the time of submission of your application to the ISAC mailbox. Failure to do so will result in delays in the processing of your application.*

- **Conflict of interest statement***

Please provide a draft of the conflict (or competing) of interest (COI) statement that you intend to include in any publication which might result from this work

Jennifer Campbell and Rachael Williams are employees of the Clinical Practice Research Datalink.

**Please refer to the International Committee of Medical Journal Editors (ICMJE) for guidance on what constitutes a COI.*

- **Experience/expertise available**

Please complete the following questions to indicate the experience/ expertise available within the team of investigators/collaborators actively involved in the proposed research, including the analysis of data and interpretation of results.

	Previous GPRD/CPRD Studies	**Publications using GPRD/CPRD data**
None	☐	☐
1-3	☐	☐
> 3	☒	☒

Experience/Expertise available	**Yes**	**No**
Is statistical expertise available within the research team? *If yes, please indicate the name(s) of the relevant investigator(s)* All investigators	☒	☐

Is experience of handling large data sets (>1 million records) available within the research team? *If yes, please indicate the name(s) of the relevant investigator(s)* All investigators	☒	☐
Is experience of practising in UK primary care available to or within the research team? *If yes, please indicate the name(s) of the relevant investigator(s)* Professor Thomas and Dr Minassian have access to advice from primary care physicians as part of the EHR working group at LSHTM.	☐	☒

- **References relating to your study**

Please list up to 3 references (most relevant) relating to your proposed study:

Margulis, A. V. *et al.* (2015) 'Beginning and duration of pregnancy in automated health care databases: Review of estimation methods and validation results', *Pharmacoepidemiology and Drug Safety*. doi: 10.1002/pds.3743

Devine, S. *et al.* (2010) 'The identification of pregnancies within the general practice research database.', *Pharmacoepidemiology and drug safety*. England, 19(1), pp. 45–50. doi: 10.1002/pds.1862.

Hardy, J. R. *et al.* (2004) 'Strategies for identifying pregnancies in the automated medical records of the General Practice Research Database.', *Pharmacoepidemiology and drug safety*, 13(11), pp. 749–759. doi: 10.1002/pds.935.

SECTION C: ACCESS TO THE DATA

- **Financial Sponsor of study[§]**

[§]*Please note: The name of the source of funding will be published on CPRD's website as part of its transparency policy*

Pharmaceutical Industry	☐	*Please specify name and country:*
Academia	☐	*Please specify name and country:*
Government / NHS	☒	*Please specify name and country:* CPRD, United Kingdom
Charity	☐	*Please specify name and country:*
Other	☐	*Please specify name and country:*
None	☐	

- **Type of Institution conducting the research**

Pharmaceutical Industry	☐	*Please specify name and country:*
Academia	☐	*Please specify name and country :*
Government Department	☒	*Please specify name and country:* CPRD, United Kingdom
Research Service Provider	☐	*Please specify name and country:*
NHS	☐	*Please specify name and country:*
Other	☐	*Please specify name and country:*

- **Data access arrangements**

The financial sponsor/ collaborator* has a licence for CPRD GOLD and will extract the data	☒
The institution carrying out the analysis has a licence for CPRD GOLD and will extract the data**	☒
A data set will be provided by the CPRD[¥€]	☐
CPRD has been commissioned to extract the data and perform the analyses[€]	☐
Other:	☐

If Other, please specify:

**Collaborators supplying data for this study must be named on the protocol as co-applicants.*
***If data sources other than CPRD GOLD are required, these will be supplied by CPRD*

[¥]*Please note that datasets provided by CPRD are limited in size; applicants should contact CPRD (enquiries@cprd.com) if a dataset of >300,000 patients is required.*

[€]*Investigators must discuss their request with a member of the CPRD Research team before submitting an ISAC application. Please contact the CPRD Research Team on +44 (20) 3080 6383 or email (enquiries@cprd.com) to discuss your requirements. Please also state*

the name of CPRD Research team with whom you have discussed this request (provide the date of discussion and any relevant reference information):

Name of CPRD Researcher Reference number (where available) Date of contact

- **Primary care data**

Please specify which primary care data set(s) are required)
Vision only (Default for CPRD studies ☒ Both Vision and EMIS®* ☐
EMIS® only* ☐

Note: Vision and EMIS are different practice management systems. CPRD has traditionally collected data from Vision practice. Data collected from EMIS is currently under evaluation prior to wider release.
Investigators requiring the use of EMIS data **must discuss the study with a member of the CPRD Research team before submitting an ISAC application*

Please state the name of the CPRD Researcher with whom you have discussed your request for EMIS data:
Name of CPRD Researcher Reference number (where available) Date of contact

SECTION D: INFORMATION ON DATA LINKAGES

- **Does this protocol seek access to linked data**

Yes* ☒ No☐ If No, please move to section E.

Research groups which have not previously accessed CPRD linked data resources **must discuss access to these resources with a member of the CPRD Research team, before submitting an ISAC application. Investigators requiring access to HES Accident and Emergency data, HES Diagnostic Imaging Dataset, PROMS data, the Pregnancy Register, Cancer Registration, SACT and CPES data and the Mental Health Services Data Set **must** also discuss this with a member of the CPRD Research team before submitting an ISAC application. Please contact the CPRD Research Team on +44 (20) 3080 6383 or email enquiries@cprd.comto discuss your requirements **before** submitting your application.*

Please state the name of the CPRD Researcher with whom you have discussed your linkage request.

Name of CPRD Researcher: Tarita Murray-Thomas Reference number (where available) Date of contact 11/10/2017

Please note that as part of the ISAC review of linkages, your protocol may be shared - in confidence - with a representative of the requested linked data set(s) and summary details may be shared - in confidence - with the Confidentiality Advisory Group of the Health Research Authority.

- **Please select the source(s) of linked data being requested§**

§Please note: This information will be published on the CPRD's website as part of its transparency policy.

☒ ONS Death Registration Data
☒ HES Admitted Patient Care
☒ HES Outpatient
☒ HES Accident and Emergency
☒ HES Diagnostic Imaging Dataset
☐ HES PROMS (Patient Reported Outcomes Measure)**
☐ CPRD Mother Baby Link
☒ Pregnancy Register
☐ MINAP (Myocardial Ischaemia National Audit Project)
☐ NCRAS (National Cancer Registration and Analysis Service) Cancer Registration Data *
☐ NCRAS Cancer Patient Experience Survey (CPES) data*
☐ NCRAS Systemic Anti-Cancer Treatment (SACT) data*
☐ Mental Health Services Data Set (MHDS)

☐ Practice Level Index of Multiple Deprivation (Standard)
☐ Practice Level Index of Multiple Deprivation (Bespoke)
☐ Patient Level Index of Multiple Deprivation***
☐ Patient Level Townsend Score ***
☐ Other**** *Please specify:*

**Applicants seeking access to NCRAS data must complete a Cancer Dataset Agreement form (available from CPRD). This should be submitted to the ISAC as an appendix to your protocol. Please also note that applicants seeking access to cancer registry data must provide consent for publication of their study title and study institution on the UK Cancer Registry website.*
***Assessment of the quality of care delivered to NHS patients in England undergoing four procedures: hip replacement, knee replacement, groin hernia and varicose veins. Please note that patient level PROMS data are only accessible by academics*

*** 'Patient level IMD and Townsend scores will not be supplied for the same study*
*****If "Other" is specified, please provide the name of the individual in the CPRD Research team with whom this linkage has been discussed.*

Name of CPRD Researcher | Reference number (where available) | Date of contact

- **Total number of linked datasets requested <u>including</u> CPRD GOLD**

Number of linked datasets requested *(practice/ 'patient' level Index of Multiple Deprivation, Townsend Score, the CPRD Mother Baby Link and the Pregnancy Register should **<u>not</u>** be included in this count)* 6

Please note: Where ≥5 linked datasets are requested, approval may be required from the Confidentiality Advisory Group (CAG) to access these data

- **Is linkage to a <u>local</u>¥ dataset with <1 million patients being requested?**

Yes* ☐ No ☒

**If yes, please provide further details:*
¥ *Data from defined geographical areas i.e. non-national datasets.*

- **If you have requested one or more linked data sets, please indicate whether the Chief Investigator or any of the collaborators listed in question 5 above, have access to these data in a patient identifiable form (e.g. full date of birth, NHS number, patient post code), or associated with an identifiable patient index.**

Yes* ☐ No ☒

** If yes, please provide further details:*

- **Does this study involve linking to patient *identifiable* data (e.g. hold date of birth, NHS number, patient post code) from other sources?**

Yes ☐ No ☒

SECTION E: VALIDATION/VERIFICATION

- **Does this protocol describe a purely observational study using CPRD data?**

Yes* ☒ No** ☐

** Yes: If you will be using data obtained from the CPRD Group, this study does not require separate ethics approval from an NHS Research Ethics Committee.*
*** No: You may need to seek separate ethics approval from an NHS Research Ethics Committee for this study. The ISAC will provide advice on whether this may be needed.*

- **Does this protocol involve requesting any additional information from GPs?**

Yes* ☐ No ☒

** If yes, please indicate what will be required:*

Completion of questionnaires by the GP℣ Yes ☐ No ☐
Is the questionnaire a validated instrument? Yes ☐ No ☐
If yes, has permission been obtained to use the instrument? Yes ☐ No ☐
Please provide further information:

Other (please describe)

[iv] *Any questionnaire for completion by GPs or other health care professional must be approved by ISAC before circulation for completion.*

- **Does this study require contact with patients in order for them to complete a questionnaire?**

  Yes* ☐ No ☒

**Please note that any questionnaire for completion by patients must be approved by ISAC before circulation for completion.*

- **Does this study require contact with patients in order to collect a sample?**

  Yes* ☐ No ☒

** Please state what will be collected:*

SECTION F: DECLARATION

- **Signature from the Chief Investigator**

- I have read the guidance on '***Completion of the ISAC application form***' and '***Contents of CPRD ISAC Research Protocols***' and have understood these;
- I have read the submitted version of this research protocol, including all supporting documents, and confirm that these are accurate.
- I am suitably qualified and experienced to perform and/or supervise the research study proposed.
- I agree to conduct or supervise the study described in accordance with the relevant, current protocol
- I agree to abide by all ethical, legal and scientific guidelines that relate to access and use of CPRD data for research
- I understand that the details provided in sections marked with (§) in the application form and protocol will be published on the CPRD website in line with CPRD's transparency policy.
- I agree to inform the CPRD of the final outcome of the research study: publication, prolonged delay, completion or termination of the study.

Name: Jennifer Campbell Date: 06/12/17 e-Signature (type name): J Campbell

PROTOCOL INFORMATION REQUIRED

The following sections below **must** be included in the CPRD ISAC research protocol. Please refer to the guidance on '***Contents of CPRD ISAC Research Protocols***' (www.cprd.com/isac) for more information on how to complete the sections below. Pages should be numbered. All abbreviations must be defined on first use.

Applicants must complete all sections listed below
Sections which do not apply should be completed as '*Not Applicable*'

1. Study Title[§]

[§]*Please note: This information will be published on CPRD's website as part of its transparency policy*

Investigating pregnancies without recorded outcomes in the Clinical Practice Research Datalink / London School of Hygiene and Tropical Medicine Pregnancy Register, with the aim of improving validity.

2. Lay Summary (Max. 200 words)[§]

[§]*Please note: This information will be published on CPRD's website as part of its transparency policy*

It is difficult to study the effects of medicines during pregnancy in the traditional clinical trial setting due to the potential risks for the mother and unborn child. Existing patient clinical care records represent an opportunity to answer important questions about medicines taken during pregnancy and their possible effects (for example an early pregnancy loss). To help investigate this, a register of pregnancies in the Clinical Practice Research Datalink, which includes information on the start of each pregnancy and its outcomes (live birth, still birth or early pregnancy loss), has been created. However, there are many pregnancies in the Register for which no outcome has been found. These anonymous pregnancy records are of limited use for research. If we do not know when or how the pregnancy ended, it makes studying the effects of medicines difficult. This study intends to investigate potential reasons why these pregnancies without outcome may be occurring in the register. This information will be used to improve the method by which the Register is created. Improvements to the Register will make this valuable resource more useful and enable researchers to investigate important concerns about safety.

3. Technical Summary (Max. 200 words)[§]

[§]*Please note: This information will be published on CPRD's website as part of its transparency policy*

The Pregnancy Register algorithm generates a list of all pregnancies determined in the Clinical Practice Research Datalink (CPRD). A record in the register represents a pregnancy episode and includes information on pregnancy start and outcome. However, there are approximately one million pregnancies where no outcome has been determined. Scenarios have been identified based on the algorithm's logic and how the data is structured which may explain this. The scenarios describe four problems; (i) real pregnancies where the outcome was not recorded in the, (ii) ongoing pregnancies at the end of available follow-up, (iii) the patient may not have been pregnant, or (iv) the pregnancy episode may be made up of records which are really part of another pregnancy. Descriptive analysis will use an algorithmic approach to query CPRD data and linked datasets to look for supporting evidence for each of these scenarios. Potential reasons for why a pregnancy outcome may not have been determined by the algorithm will be

tabulated. Evidence will then be used to improve the Pregnancy Register algorithm to reduce the occurrence of pregnancies without outcome and increase the usefulness of this resource.

A. Objectives, Specific Aims and Rationale

Objective

To investigate possible reasons why the pregnancy algorithm used for the CPRD/LSHTM Pregnancy Register is identifying pregnancy episodes with no associated outcome, and to use this information to attempt to reduce the occurrence of these episodes in the Pregnancy Register.

Specific Aims

- To describe the "pregnancy profile" of patients with outcome unknown pregnancies: the number of pregnancies they have (overall and by type), and the temporal relationship of their outcome unknown pregnancies to their other pregnancies.
- To use the available data to investigate identified potential scenarios explaining why pregnancy outcomes may not have been detected by the algorithm and to flag pregnancy episodes for which there is evidence that these scenarios apply.
- To use the information gathered to improve the algorithm which produces the Pregnancy Register and reduce the number of pregnancies without outcome.

Rationale

The Pregnancy Register is created by an algorithm which was developed jointly by CPRD and the London School of Hygiene and Tropical Medicine (ISAC protocol 11_058) and is now made available to CPRD data users. The Pregnancy Register lists all pregnancies identified in CPRD GOLD and includes details of each one. A single record in the Pregnancy Register represents a unique pregnancy episode. In simple terms the pregnancy algorithm works by identifying all records in CPRD GOLD representing a pregnancy delivery; these are then grouped together into delivery episodes. Based on the date of the delivery episode and other information, the algorithm then estimates the date of the woman's last menstrual period (LMP) and assigns all pregnancy records which occur between these two events to create a pregnancy episode. This process is then repeated for early pregnancy losses including terminations. Any remaining pregnancy records which are not yet associated with a pregnancy episode are grouped together sequentially, provided there are less than six weeks between them, to create pregnancy episodes without outcomes. The LMP for these episodes is estimated as four weeks before the first antenatal record in the episode.

There are approximately 950,000 pregnancy episodes identified in the Pregnancy Register which have no recorded outcome attributed to them. These pregnancies have limited use when designing a study. Both the start and end of the pregnancies are approximate estimates and when using these pregnancies for research there is uncertainty as to whether the patient was truly pregnant at any time point, which leads to a risk of exposure misclassification. Furthermore, excluding these pregnancies from a study may lead to underestimation of an outcome if outcomes such as miscarriage are less likely to be reported in the data. It is therefore important to characterise the pregnancies without outcome to attempt to understand why these episodes occur in the Register and ultimately to reduce their occurrence.

Applicants must complete all sections listed below
Sections which do not apply should be completed as *'Not Applicable'*

B. Study Background

The safety of drugs and vaccines given during pregnancy is difficult to study in the traditional trial setting. Pregnant women are excluded from many trials due to the potential risks to both the woman and her unborn child. Nevertheless, in the real-life setting pregnant women are exposed to a variety of drugs, including inadvertent exposure in the first trimester of pregnancy when the woman may not realise that she is pregnant. Exposure in early pregnancy is of particular importance to the foetus as it is the time of organogenesis and thus exposure at this time can incur the highest risk of congenital malformations(Webster and Freeman, 2003). Furthermore, vaccination of pregnant women has emerged in recent years as an increasingly important public health strategy to protect women and their infants against infection. In the UK, vaccination of pregnant women (in any trimester) against influenza was introduced in 2010 and vaccination against pertussis was introduced in 2012 (Public Health England, 2014). Post-licensure monitoring of the safety of these vaccines is essential, to continue to assess the benefits and risks of the vaccination programme.

Large datasets of electronic health records (EHR) such as CPRD GOLD have been used to assess the safety and effectiveness of vaccines and other drugs given in pregnancy (Margulis *et al.*, 2015). However, until recently there have been appreciable challenges in identifying accurately the start and end of pregnancies in these data, and thus pinpointing exposures in the first trimester. A major advance in this area was achieved in the last year, arising from a research collaboration between LSHTM and CPRD. The collaboration has resulted in the production of a Pregnancy Register, identifying in CPRD GOLD a very large number of pregnancies recorded in anonymised general practice health records, and including the start of each pregnancy and its outcomes. This important new data resource should enhance continued monitoring of the benefits and risks associated with vaccination and drugs given in pregnancy to support clinical recommendations and patient acceptance of vaccination.

Initial validation of the Pregnancy Register against linked electronic maternity records in hospitalisation data has indicated overall good agreement, suggesting that most pregnancies are well captured in the CPRD GOLD (Minassian et al). However, further methodological work is required to maximise the robustness of the Register as a research tool. This study proposes to build upon validation work which has already been carried out by attempting to investigate the large number of pregnancy episodes in the Register for which the outcome has not been identified. Whilst there have been previous algorithms which identified pregnancies in CPRD GOLD (GPRD) they have not attempted to address the situations where there is evidence of a pregnancy but no outcome detected (Hardy *et al.*, 2004; Devine *et al.*, 2010)

C. Study Type

This is a methodological study intended to further develop the algorithm used to produce the CPRD Pregnancy Register.

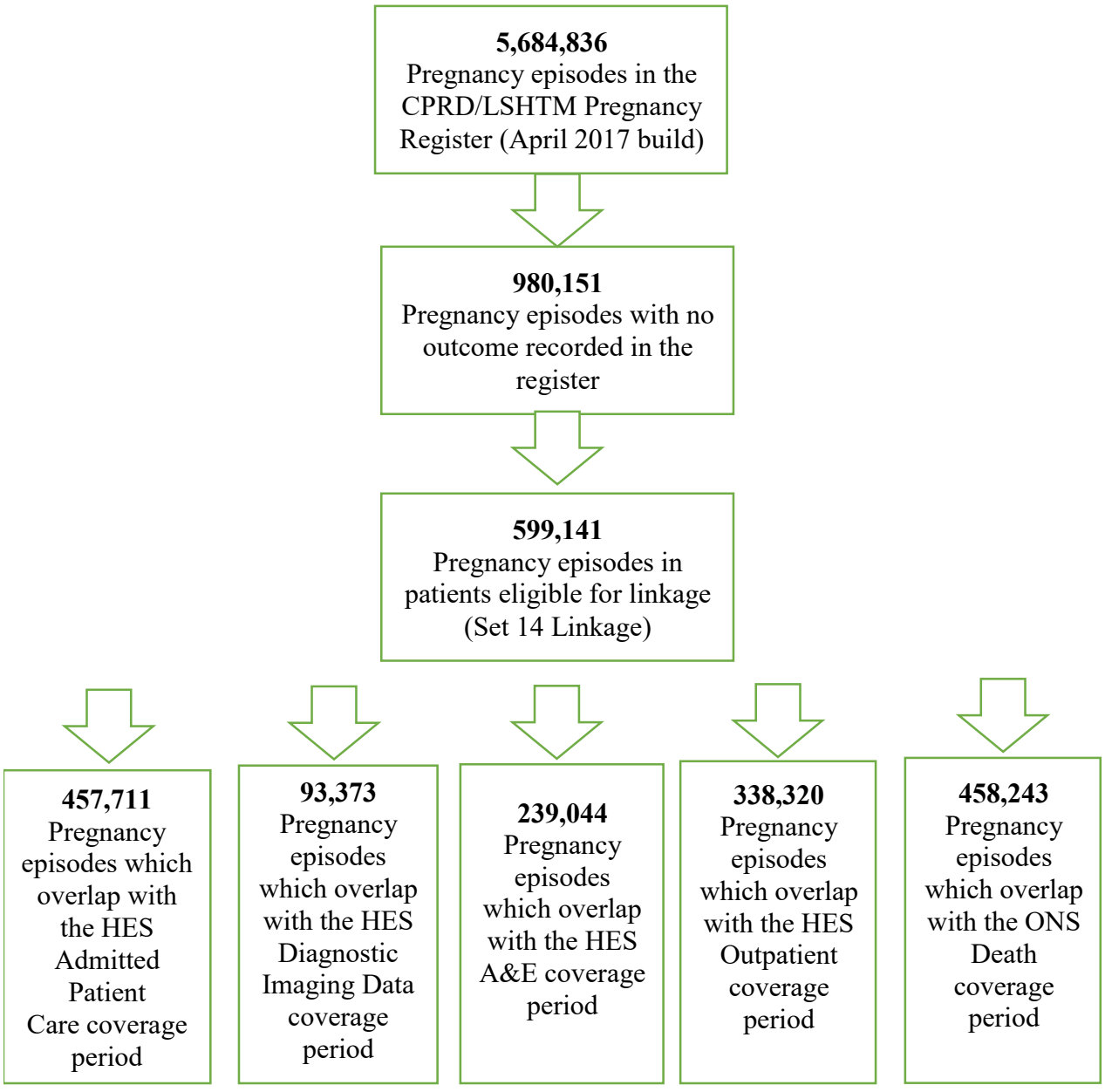
D. Study Design

This study is not a classic epidemiological study design. The study intends to explore the data in order to inform the methodology used to produce the Pregnancy Register.

Applicants must complete all sections listed below
Sections which do not apply should be completed as *'Not Applicable'*

E. Feasibility counts

For some analyses we will use all pregnancies without outcome in the Register (n=980,151, as in the Figure below); for another analyses we will restrict to those of patients eligible for HES/ONS linkage (n=93,373-458,243).



Applicants must complete all sections listed below
Sections which do not apply should be completed as *'Not Applicable'*

F. Sample size considerations

We will include all outcome unknown pregnancies in the Pregnancy Register (see section H above)

G. Data Linkage Required (if applicable):[§]

[§]*Please note that the data linkage/s requested in research protocols will be published by the CPRD as part of its transparency policy*

Approval was given for protocol 11_058 to use linked HES and ONS data to validate the pregnancy algorithm, this is nearing completion. Here we expand on how the data, including newly available linked data sources, will be further used to achieve the study aims.

We are requesting access to linked HES APC, HES Outpatient, HES A&E and HES DID. These are all data sources which may contain evidence of a patient's pregnancy or pregnancy outcome. We propose to use these data sources to look for:

- Records related to pregnancy in order to validate that a woman was actually pregnant for those pregnancy episodes without outcome in the Register which have limited supporting evidence for pregnancy in CPRD.
- Records of pregnancy outcomes (Delivery or Early pregnancy loss either spontaneous or induced).
- Dates of pregnancy outcomes in order to ascertain the end of the pregnancy episode.

In HES APC we propose to look at the Diagnosis and Procedures files for ICD10 and OPCS codes which indicate a pregnancy or its outcome. We also intend to use the HES Maternity file, looking at combinations of different data fields in the file to try to ascertain if a delivery took place. Whilst we recognise that HES Outpatient data contains limited information on diagnoses and procedures, it does contain information on the specialty which the patient visited including Maternity and Obstetrics. For HES A&E data we again intend to look at which department the patient was moved to (if applicable) as an indicator of pregnancy. In the HES DID data we intend to look for records of foetal scans.

We are also requesting access to the linked ONS Mortality data in order to check whether mothers who have a pregnancy episode without outcome died before the expected end of pregnancy. This will be in addition to using the CPRD date of death field.

We have carried out a full risk assessment regarding the use of multiple linked datasets and this is included in the appendices.

H. Study population

All woman with at least one pregnancy episode in the Pregnancy Register which has no outcome attributed to it, these pregnancies are coded as 13 in the outcome variable of the Pregnancy Register. The latest version of the Pregnancy Register will be used. We will include all patients and we will flag patients who did not have a CPRD patient acceptability flag or whose pregnancy was not during UTS follow-up as part of our analysis.

Applicants must complete all sections listed below
Sections which do not apply should be completed as *'Not Applicable'*

I. Selection of comparison group(s) or controls

Not applicable- we do not intend to use a comparison group in this analysis.

J. Exposures, Health Outcomes§ and Covariates

§Please note: Summary information on health outcomes (as included on the ISAC application form above) will be published on CPRD's website as part of its transparency policy

The aim of this study is to further develop methodology which allows researchers to study exposure during pregnancy and as such it is not a traditional study in which exposures, covariates and outcomes apply. However, as the aim of the study is to address the problem of pregnancy episodes without outcome occurring in the Pregnancy Register we are looking for: 1.all end of pregnancy outcomes including live deliveries, stillbirths, early pregnancy losses (spontaneous and induced). 2. evidence of pregnancy, as defined by Read and Gemscript and ICD codes (see appendices). In order to further understand why pregnancies without outcome are occurring we will generate the following variables from the available data for use in the proposed analysis outlined in section N:

Descriptive Variables to be generated	**Data Source**
Count of the number of pregnancies without outcome (PWO's) per woman	Pregnancy Register
Count of the number of early pregnancy losses per woman	Pregnancy Register
Count of the number of deliveries episodes (as generated by the algorithm) per woman	Pregnancy Register
Count of the number of pregnancy records within each pregnancy episode	CPRD primary care data and Pregnancy Register
Number of days since the end of the previous pregnancy episode for that woman relative to the PWO.	Pregnancy Register
Number of days until the start of the next pregnancy episode for that woman relative to the PWO.	Pregnancy Register
Flag to indicate whether the pregnancy before the PWO ended with delivery, early pregnancy loss or is a PWO.	Pregnancy Register
Flag to indicate whether the pregnancy after the PWO ended with delivery, early pregnancy loss or is a PWO.	Pregnancy Register
Flag to indicate if one of the antenatal codes in the PWO episode has an event date of the first of January.	CPRD primary care data and Pregnancy Register

Flag to indicate if the PWO is in first year of current registration.	CPRD primary care data and Pregnancy Register
Flag to indicate whether the PWO is before the start of current registration	CPRD primary care data and Pregnancy Register
Flag to indicate whether the PWO is during UTS registration	CPRD primary care data and Pregnancy Register
Flag to indicate ifr the woman's data was acceptable according to the CPRD acceptability flag	CPRD primary care data.
Flags to indicate whether each PWO episode is in a woman eligible to be linked to each of the different data sets proposed AND overlaps with coverage period for that data source.	Pregnancy Register and CPRD linkage eligibility file
Flag to indicate that there is supporting evidence of pregnancy in the linked or primary care data within appropriate time limits of the PWO in the Pregnancy Register. This will be repeated for all linked datasets separately.	HES OP, HES A&E, HES DID, HES APC, Ancillary CPRD codes.
Flag to indicate that there is supporting evidence of a pregnancy outcome in the linked data within appropriate time limits of the pregnancy episode in the Pregnancy Register. This will be repeated for all linked datasets separately.	HES OP, HES A&E, HES DID, HES APC.
Flag to indicate there is a death record for the woman within appropriate time limits following the PWO in the Pregnancy Register.	ONS Mortality data.

K. Data/ Statistical Analysis

We will begin by extracting the list of pregnancies without outcome from the Pregnancy Register. We will then generate a list of the patients who have one or more of these pregnancies without outcome, who will then form the cohort for this study. All pregnancy records for these patients will be extracted from the CPRD GOLD database using the pregnancy code list upon which the pregnancy algorithm is based to create a dataset which includes all the pregnancy records and the summary Pregnancy Register information for these woman.

Descriptive variables will be generated to further characterise the women and their pregnancies in the dataset. At the patient level these will include variables such as the number or pregnancies, pregnancies without outcome and early pregnancy losses recorded for each woman in the cohort. The full list of variables we will generate is described in section M above. For each pregnancy episode we will look at its proximity to other pregnancy episodes and whether other pregnancies in proximity were pregnancies without outcome, deliveries or pregnancy losses. We will look for information about the pregnancy without outcome: the number of codes which make up the episode and the use of Jan 1st as the recording date (as a potential default date for a past pregnancy). We will also look at the timing of the pregnancies in relation to the start and end of patient follow up in CPRD. We will ascertain whether the timing of the pregnancy without outcome falls within the coverage period of each of the linked datasets requested for those patients who are eligible for linkage. For pregnancy episodes which fall within linkage coverage we will look for evidence of pregnancies and/or pregnancy outcomes recorded within a sensible timeframe of the pregnancy record in the Register and will flag those pregnancies as having linked data evidence.

Applicants must complete all sections listed below
Sections which do not apply should be completed as *'Not Applicable'*

Based on the logic of the algorithm and how the data are structured we have identified twelve scenarios which may explain why pregnancies without outcome are occurring in the Pregnancy Register and for which we may be able to find evidence within the primary and linked data. In this analysis we intend to systematically query the data to look for evidence of these scenarios. Each pregnancy episode will be flagged for all scenarios which could potentially apply (diagrams of the scenarios are included in the appendices). The scenarios we intend to consider are as follows:

Scenario	**How does this appear in the data?** [1]
Problem 1 ***Real pregnancy outcome not captured in CPRD primary care data:***	
- The woman had a delivery, miscarriage or termination in hospital and information either wasn't fed back to the practice or wasn't recorded by the practice.	There will be no evidence of an outcome up to 38 weeks (for delivery) or 20 weeks (for miscarriage of termination) after the first antenatal record for that pregnancy in CPRD. However, there may be evidence of delivery/miscarriage/termination in one of the linked HES datasets.
- The outcome of the pregnancy is recorded in the data but has no event date associated with it and it therefore not picked up by the algorithm.	There will be an outcome of pregnancy code with missing eventdate recorded in the practice software system within 38 weeks after the first antenatal record of the pregnancy without outcome. These codes will be identified using the system date.
- The pregnancy occurred before the patient was registered at the current practice or before the start of the practices UTS follow-up. Information was recorded about the pregnancy but not the outcome.	The pregnancy without outcome episode will occur before the start of current registration. There will be no evidence of an outcome up to 38 weeks (for delivery) or 20 weeks (for miscarriage of termination) after the first antenatal record for that pregnancy in CPRD.
Problem 2 Real pregnancy ongoing at the end of available data:	
- The woman moved practices before the outcome. If a patient transfers out of a CPRD practice, then follow up is lost. OR The woman died before the outcome.	There will be a transfer out date or death date less than 38 weeks after the earliest antenatal record for the pregnancy without outcome.
- The last collection date of the practice was before the outcome.	There will be a last collection date less than 38 weeks after the earliest antenatal record for the pregnancy without outcome.
Problem 3 The patient is not pregnant at the time of the record:	
- Historical pregnancies, recorded retrospectively in the first few months after patient joins the practice.	The pregnancy without outcome occurs less than one year after the woman's current registration date.

- The woman is not pregnant but is planning a pregnancy and discusses this with the GP due to other medical conditions which may complicate pregnancy (e.g. epilepsy)	The pregnancy without outcome episode will probably be based on one code. This code is likely to be a counselling code such as "67AF.00 Pregnancy advice for patients with epilepsy"
Problem 4: The pregnancy record is really part of another pregnancy which is already captured.	
- There was a delay in the recording of the pregnancy outcome by the practice. The algorithm then calculates the LMP as being later than it was. Records which occurred earlier in the pregnancy appear as if belonging to a separate pregnancy, which is then assigned as a pregnancy without outcome.	The pregnancy without outcome will be followed by another pregnancy that starts >= 6 weeks after the end of the PWO.
- Pregnancies where the LMP is derived from records in the data that are incorrect leads to a pregnancy that is too short and uncovers codes before the pregnancy.	The pregnancy without outcome will be followed within another pregnancy ending within 42 weeks which itself will be less than 40 weeks long.
- The GP records a code relating to the patient's pregnancy outcome history during a pregnancy which is then incorrectly identified by the algorithm as the current pregnancy outcome. Records later in the pregnancy are not assigned to the pregnancy	The pregnancy without outcome must not be the patient's first pregnancy for this to apply. The pregnancy without outcome would be within 25 weeks after the previous outcome.
- In the algorithm, if there are pregnancy records within 4 weeks before the estimated LMP, the identified pregnancy episode is shifted backwards (within plausible limits) to encompass those records. This may leave unassigned pregnancy records which occurred shortly after the new estimated delivery date.	The pregnancy without outcome must not be the only pregnancy for this to apply. There will be another pregnancy which ends <8 weeks before the first antenatal record of pregnancy without outcome.
- The outcome of the pregnancy episode has been misclassified as antenatal (e.g. intrauterine death) or codes which are flagged as both antenatal and early pregnancy loss (e.g. failed abortion).	There will be a code which we identified as potentially misclassified within 38 weeks after the first antenatal record of the pregnancy without outcome.

[1] The time windows used to look for evidence for each scenario relate to specific rules in the pregnancy algorithm for establishing episodes of pregnancy and amending dates for the start and end of pregnancies

The flags and variables created (as listed in section M) will be used to produce summary tables to begin to explore the results of this characterisation. These will include the number pregnancies without outcome which fall into each scenario or combination of scenarios, how many times these scenarios occur, the mean length of pregnancies in each scenario. For pregnancies which fit scenario 1 we will summarise which of the linked data sources identified the pregnancy outcome

Applicants must complete all sections listed below Sections which do not apply should be completed as *'Not Applicable'*
Information gathered will be used to formulate suggestions for modifications to the pregnancy algorithm in order to attempt to reduce the number of pregnancies without outcome in the Register. It is hypothesised that this will either be by identifying situations where an outcome unknown pregnancy corresponds to another identified pregnancy with known outcome, where the pregnancy is not a true pregnancy episode or where the pregnancy existed but the outcome could not be identified in the primary care data. Where pregnancies fall into more than one scenario we will develop a hierarchical approach to decide the best way to attempt to resolve them.
L. Plan for addressing confounding Not applicable to this methodological study.
M. Plans for addressing missing data The objective of this study is to look for pregnancy outcome data where the outcome has not previously been identified in order to strengthen certainty around validity, type and timing of pregnancy episodes in the Pregnancy Register. In this situation missing data will not lead to bias but just limits the usefulness of the analysis. However, any evidence found is useful.
N. Patient or user group involvement (if applicable) Not applicable- there are no plans for patient or user group involvement in this methodological work.
O. Plans for disseminating and communicating study results, including the presence or absence of any restrictions on the extent and timing of publication It is our intention to submit this work to the 2018 International Conference of Pharmacoepidemiology as well as to publish the work in a suitable scientific journal. We will also report our findings to the CPRD Observational Research Team to help to inform decisions about the best way to further develop the Pregnancy Register as a useful tool for CPRD data users.
P. Limitations of the study design, data sources, and analytic methods Not all patients in the CPRD/LSHTM Pregnancy Register are eligible for linkage and not all pregnancies identified are within the coverage periods of the data sources we propose to use and so will be excluded from analyses using linked data. Furthermore, some of the HES data we intend to utilise is of limited quality, for example much of the HES A&E data are missing diagnoses. Therefore, for some pregnancies it may not be possible to gather further evidence. As there is no clear gold standard among the data

sources in this analysis, when conflicting evidence exists across data sources this will need to be noted as part of the study findings.

Q. References

Devine, S. *et al.* (2010) 'The identification of pregnancies within the general practice research database.', *Pharmacoepidemiology and drug safety*. England, 19(1), pp. 45–50. doi: 10.1002/pds.1862.

Hardy, J. R. *et al.* (2004) 'Strategies for identifying pregnancies in the automated medical records of the General Practice Research Database.', *Pharmacoepidemiology and drug safety*, 13(11), pp. 749–759. doi: 10.1002/pds.935.

Margulis, A. V. *et al.* (2015) 'Beginning and duration of pregnancy in automated health care databases: Review of estimation methods and validation results', *Pharmacoepidemiology and Drug Safety*. doi: 10.1002/pds.3743.

Minassian C, Williams R, Meeraus W, Campbell O, Thomas SL. A new data algorithm to identify pregnancies in the UK Clinical Practice Research Datalink (manuscript in preparation)

Public Health England (2014) 'Pertussis Vaccination Programme for Pregnant Women: vaccine coverage estimates in England, April to August 2014 - GOV.UK', 11(34). Available at: https://www.gov.uk/government/publications/pertussis-immunisation-in-pregnancy-vaccine-coverage-estimates-in-england-october-2013-to-march-2014/pertussis-vaccination-programme-for-pregnant-women-vaccine-coverage-estimates-in-england-april-to-august-2014.

Webster, W. S. and Freeman, J. A. D. (2003) 'Prescription drugs and pregnancy.', *Expert opinion on pharmacotherapy*. England, 4(6), pp. 949–961. doi: 10.1517/14656566.4.6.949.

List of Appendices *(Submit all appendices as separate documents to this application)*

- Risk assessment for the use of multiple linked data sets.
- Diagrams of proposed scenarios
- Read code list for identifying pregnancies and outcomes
- ICD code list for identifying early pregnancy loss
- ICD code list for evidence of pregnancy excluding early loss
- OPCS code list for identifying pregnancy outcomes
- Gemscript code for identifying pregnancy outcomes

Appendix 3: Approved ISAC application form conflicting pregnancy episodes study

INDEPENDENT SCIENTIFIC ADVISORY COMMITTEE (ISAC) PROTOCOL APPLICATION FORM

PART 1: APPLICATION FORM

IMPORTANT

Both parts of this application must be completed in accordance with the guidance note 'Completion of the ISAC Protocol Application Form', which can be found on the CPRD website (https://cprd.com/research-applications).

FOR ISAC USE ONLY	
Protocol No. -	**Submission date -**

GENERAL INFORMATION ABOUT THE PROPOSED RESEARCH STUDY

1. Study Title (Max. 255 characters including spaces)

Investigating overlapping pregnancy episodes in the Clinical Practice Research Datalink / London School of Hygiene and Tropical Medicine Pregnancy Register, with the aim of identifying and categorising validity issues.

2. **Research Area** (place '**X**' in all boxes that apply)

Drug Safety		Economics	
Drug Utilisation		Pharmacoeconomics	
Drug Effectiveness		Pharmacoepidemiology	
Disease Epidemiology		Methodological	x
Health Services Delivery			

3. Chief Investigator

Title:	Mrs
Full name:	Jennifer Campbell
Job title:	Senior Researcher
Affiliation/organisation:	CPRD
Email address:	Jennifer.campbell@mhra.gov.uk
CV Number (if applicable):	051_15CESL
Will this person be analysing the data? (Y/N)	Y

4. Corresponding Applicant

Title:	Mrs
Full name:	Jennifer Campbell
Job title:	Senior Researcher
Affiliation/organisation:	CPRD
Email address:	Jennifer.campbell@mhra.gov.uk
CV Number (if applicable):	051_15CESL
Will this person be analysing the data? (Y/N)	Y

5. List of all investigators/collaborators

Title:	Dr
Full name:	Caroline Minassian
Job title:	Assistant Professor of Epidemiology
Affiliation/organisation:	London School of Hygiene & Tropical Medicine,
Email address:	caroline.minassian@lshtm.ac.uk
CV Number (if applicable):	029_17
Will this person be analysing the data? (Y/N)	N

Title:	Professor
Full name:	Krishnan Bhaskaran
Job title:	Professor of Statistical Epidemiology
Affiliation/organisation:	London School of Hygiene & Tropical Medicine,
Email address:	Krishnan.Bhaskaran@lshtm.ac.uk
CV Number (if applicable):	15615CESL
Will this person be analysing the data? (Y/N)	N

Title:	Dr
Full name:	Helen McDonald
Job title:	Honorary Research Fellow
Affiliation/organisation:	Imperial College Healthcare NHS Trust
Email address:	Helen.mcdonald@lshtm.ac.uk
CV Number (if applicable):	320_15CES
Will this person be analysing the data? (Y/N)	N

Title:	Dr
Full name:	Rachael Williams
Job title:	Observational Research Manager
Affiliation/organisation:	CPRD
Email address:	rachael.williams@mhra.gov.uk
CV Number (if applicable):	130_15CESL
Will this person be analysing the data? (Y/N)	N

[Add more investigators/collaborators as necessary by copy and pasting a new table for each investigator/collaborator]

6. Experience/expertise available

List below the member(s) of the research team who have experience with CPRD data.

Name(s):
Jennifer Campbell
Caroline Minassian
Krishnan Bhaskaran
Helen McDonald
Rachel Williams

List below the member(s) of the research team who have statistical expertise.

Name(s):
Krishnan Bhaskaran
Rachael Williams

List below the member(s) of the research team who have experience of handling large datasets (greater than 1 million records).

Name(s):
Jennifer Campbell
Caroline Minassian

Krishnan Bhaskaran
Helen McDonald
Rachael Williams

List below the member(s) of the research team, or supporting the research team, who have experience of practicing in UK primary care.

Name(s):
Helen McDonald

ACCESS TO THE DATA

7. Sponsor of the study

Institution/Organisation:	CPRD
Address:	10 South Colonnade, Canary Wharf, London, E144PU

8. Funding source for the study

Same as Sponsor?	Yes	x	No	
Institution/Organisation:				
Address:				

9. Institution conducting the research

Same as Sponsor?	Yes	x	No	
Institution/Organisation:				
Address:				

10. Data Access Arrangements

Indicate with an '**X**' the method that will be used to access the data for this study:

Study-specific Dataset Agreement	

Institutional Multi-study Licence	x
Institution Name	CPRD
Institution Address	10 South Colonnade, Canary Wharf, London, E144PU

Will the dataset be extracted by CPRD?

Yes		No	x

If yes, provide the reference number:

11. **Data Processor(s):**

Processing	x
Accessing	x
Storing	x
Processing area (UK/EEA/Worldwide)	UK
Organisation name	CPRD
Organisation address	10 South Colonnade, Canary Wharf, London, E144PU

Processing	
Accessing	
Storing	
Processing area (UK/EEA/Worldwide)	
Organisation name	
Organisation address	

[Add more processors as necessary by copy and pasting a new table for each processor]

INFORMATION ON DATA

12. Primary care data (place '**X**' in all boxes that apply)

CPRD GOLD	x	CPRD Aurum	

Reference number (if applicable):

13. Please select any linked data or data products being requested

Patient Level Data (place '**X**' in all boxes that apply)

ONS Death Registration Data			
HES Admitted Patient Care	x		
HES Outpatient	x		
HES Accident and Emergency		NCRAS Cancer Registration Data	
HES Diagnostic Imaging Dataset	x	NCRAS Cancer Patient Experience Survey (CPES) data	
HES PROMS (Patient Reported Outcomes Measure)		NCRAS Systemic Anti-Cancer Treatment (SACT) data	
CPRD Mother Baby Link	x	NCRAS National Radiotherapy Dataset (RTDS) data	
Pregnancy Register	x	NCRAS Quality of Life Cancer Survivors Pilot (QOLP)	
Mental Health Data Set (MHDS)		NCRAS Quality of Life Colorectal Cancer Survivors (QOLC)	

Area Level Data (place '**X**' in <u>one</u> Practice / Patient level box that may apply)

Practice level (UK)		**Patient level (England only)**	
Practice Level Index of Multiple Deprivation		Patient Level Index of Multiple Deprivation	
Practice Level Index of Multiple Deprivation (index other than the most recent)		Patient Level Index of Multiple Deprivation Domains	
Practice Level Index of Multiple Deprivation Domains		Patient Level Carstairs Index for 2011 Census	
Practice Level Carstairs Index for 2011 Census (Excluding Northern Ireland)		Patient Level Townsend Score	
2011 Rural-Urban Classification at LSOA level		2011 Rural-Urban Classification at LSOA level	

Reference / Protocol number (where applicable):

14. Are you requesting linkage to a dataset not listed above?

Yes		No	**x**

If yes, provide the Non-Standard Linkage reference number:

15. Does any person named in this application already have access to any of these data in a patient identifiable form, or associated with an identifiable patient index?

Yes		No	**x**

If yes, provide further details:

VALIDATION/VERIFICATION

16. Does this protocol describe an observational study using purely CPRD data?

Yes	**x**	No	

17. Does this protocol involve requesting any additional information from GPs, or contact with patients?

Yes		No	**x**

If yes, provide the reference number:

PART 2: PROTOCOL INFORMATION

Applicants must complete all sections listed below
Sections which do not apply should be completed as '*Not Applicable*' and justification provided

1. Study Title (Max. 255 characters)

Investigating overlapping pregnancy episodes in the Clinical Practice Research Datalink / London School of Hygiene and Tropical Medicine Pregnancy Register with the aim of identifying and categorising validity issues.

2. Lay Summary (Max. 250 words)

It is important to monitor the effects of medicines during pregnancy, to ensure they are effective and safe for the mother and unborn child in real world settings as well as trials. Existing patient clinical care records represent an opportunity to answer important questions about medicines taken during pregnancy. To help investigate this, a register of pregnancies in the Clinical Practice Research Datalink, which includes anonymised information on the start of each pregnancy and its outcome (live birth, still birth or pregnancy loss), has been created. However, some women have more than one pregnancy in the Register which appear to overlap. At least one of these records must be an error, and it is unclear how to handle these records in research studies. If we do not know which, if either, are correct and at which timepoints a woman was truly pregnant, it makes studying the effects of medicines difficult. This study intends to investigate potential reasons why these apparently overlapping pregnancies may occur in the Register. This work follows on from a similar study looking at pregnancies with missing outcomes in the Register. The results of this study will provide additional information about these types of pregnancy to researchers using the data for their own studies Information from these studies will then potentially be used to improve the methods by which the Register is created. Improvements will make this valuable resource more useful enabling researchers to investigate important issues such as the safety and effectiveness of medicine during pregnancy.

3. **Technical Summary (Max. 200 words**)[§]

[§]*Please note: This information will be published on CPRD's website as part of its transparency policy*

The Pregnancy Register algorithm generates a list of all pregnancies determined in the Clinical Practice Research Datalink (CPRD). A record in the register represents a pregnancy episode and includes information on pregnancy start and outcome. However, there are approximately half a million pregnancies which overlap with another pregnancy in the Register. Scenarios have been identified based on the algorithm's logic and how the data is structured which may explain this. The scenarios describe four problems; Both pregnancies are real but one episode is a historical pregnancy; Both pregnancies are historical; Both pregnancies are real but the gestation of the pregnancies applied by the algorithm is wrong; The pregnancies are really one pregnancy which has been identified as two by the algorithm. Descriptive analysis will use an algorithmic approach to query CPRD data and linked datasets to look for supporting evidence for each of these scenarios. Potential reasons for why overlapping pregnancies may have been generated by the algorithm will be tabulated. This work follows on from a previous study which assessed pregnancies without an outcome recorded in the Register (ISAC 17_285R_2) Evidence from these studies will then be used to improve the Pregnancy Register algorithm to reduce the occurrence of overlapping pregnancies and increase the usefulness of this resource.

4. Outcomes to be Measured

Pregnancy records relating to the antenatal, perinatal and postnatal period, and the pregnancy outcome in order to further understand which of the overlapping pregnancies in the Pregnancy Register are correct.

5. Objectives, Specific Aims and Rationale

Objective

To investigate possible reasons why the algorithm used to generate the Pregnancy Register identifies pregnancy episodes for the same woman which overlap with one another, and to use this information to provide advice for users wishing to utilise the Register.

Specific Aims

- To describe the "pregnancy profile" of patients with overlapping pregnancy episodes: the number of pregnancies they have (overall and by type), the temporal relationship of their overlapping pregnancies to one another, and the type of codes which make up the episodes.
- To use the available data to investigate identified potential scenarios explaining why overlapping episodes may have been created by the algorithm and to flag pregnancy episodes for which there is evidence that these scenarios apply.
- To use the information gathered to make recommendations to CPRD regarding future developments of the Pregnancy Register and to provide information to users of the Register to guide their approach to handling these episodes during analysis.

Rationale

The Pregnancy Register is created by an algorithm which was developed jointly by CPRD and the London School of Hygiene and Tropical Medicine (ISAC protocol 11_058) and is now made available to CPRD data users. The Pregnancy Register lists all pregnancies identified in CPRD GOLD and includes details of each one. A single record in the Pregnancy Register represents a unique pregnancy episode.

The Pregnancy Register described in detail elsewhere (Minassian et al., 2019) In simple terms the pregnancy algorithm works by identifying all records in CPRD GOLD representing a pregnancy delivery; these are then grouped together into delivery episodes (see appendix 1 for diagram). Based on the date of the delivery episode and other information, the algorithm then estimates the date of the woman's last menstrual period (LMP) and assigns all pregnancy records which occur between these two events to create a pregnancy episode. This process is then repeated for early pregnancy losses including terminations. Any remaining pregnancy records which are not yet associated with a pregnancy episode are grouped together sequentially, provided there are less than six weeks between them, to create pregnancy episodes without outcomes. The LMP for these episodes is estimated as four weeks before the first antenatal record in the episode.

The Pregnancy Register aims to capture all pregnancy information in CPRD regardless of completeness and thus is highly sensitive but not always specific. Some pregnancies are generated by the algorithm based on just a single antenatal record where both the start and end of the pregnancies must be treated with caution. Furthermore standard pregnancy durations are applied when no duration information is available in the data, when in reality the gestation of deliveries and of early pregnancy losses are not uniform. It is therefore unsurprising that overlapping pregnancy episodes exist in the Register. When using these pregnancies for research there is uncertainty as to whether the patient was truly pregnant at certain time points which leads to a risk of misclassification of pregnancy timing or outcome. Furthermore, excluding these pregnancies from a study may lead to underestimation of an outcome if pregnancies with outcomes such as miscarriage are more likely to overlap with other pregnancies in the Register. It is therefore important to characterise the overlapping pregnancies to attempt to understand why these episodes occur in the Register and ultimately to reduce their occurrence.

6. Study Background

The safety of drugs and vaccines given during pregnancy is difficult to study in the traditional trial setting. Pregnant women are excluded from many trials due to the potential risks to both the woman and her unborn child. Nevertheless, in the real-life setting pregnant women are exposed to a variety of drugs, including inadvertent exposure in the first trimester of pregnancy when the woman may not realise that she is pregnant. Exposure in early pregnancy is of particular importance to the foetus as it is the time of organogenesis and thus exposure at this time can incur the highest risk of congenital malformations(Webster & Freeman, 2003). Furthermore, vaccination of pregnant women has emerged in recent years as an increasingly important public health strategy to protect women and their infants against infection. In the UK, vaccination of pregnant women (in any trimester) against influenza was introduced in 2010 and vaccination against pertussis was introduced in 2012 (Public Health England, 2014). Post-licensure monitoring of the safety of these vaccines is essential, to continue to assess the benefits and risks of the vaccination programme.

Large datasets of electronic health records (EHR) such as CPRD GOLD have been used to assess the safety and effectiveness of vaccines and other drugs given in pregnancy (Margulis et al., 2015). However, until recently there have been appreciable challenges in identifying accurately the start and end of pregnancies in these data, and thus pinpointing exposures in the first trimester. A major advance in this area was achieved in the last year, arising from a research collaboration between LSHTM and CPRD. The collaboration has resulted in the production of a Pregnancy Register, identifying in CPRD GOLD a very large number of pregnancies recorded in anonymised general practice health records, and including the start of each pregnancy and its outcomes. This important new data resource should enhance continued monitoring of the benefits and risks associated with vaccination and drugs given in pregnancy to support clinical recommendations and patient acceptance of vaccination.

Initial validation of the Pregnancy Register against linked electronic maternity records in hospitalisation data has indicated overall good agreement, suggesting that most pregnancies are well captured in the CPRD GOLD (Minassian et al). However, further methodological work is required to maximise the robustness of the Register as a research tool. Some validation work has already been carried out, including a previous study which assessed pregnancies without outcomes in the Register. This study will build on this by attempting to investigate the large number of pregnancy episodes in the Register which overlap with another episode for the same woman. Whilst there have been previous algorithms which identified pregnancies in CPRD GOLD (GPRD) they have not attempted to address the situations where pregnancies apparently overlap with one another (Devine et al., 2010; Hardy, Holford, Hall, & Bracken, 2004)

7. Study Type

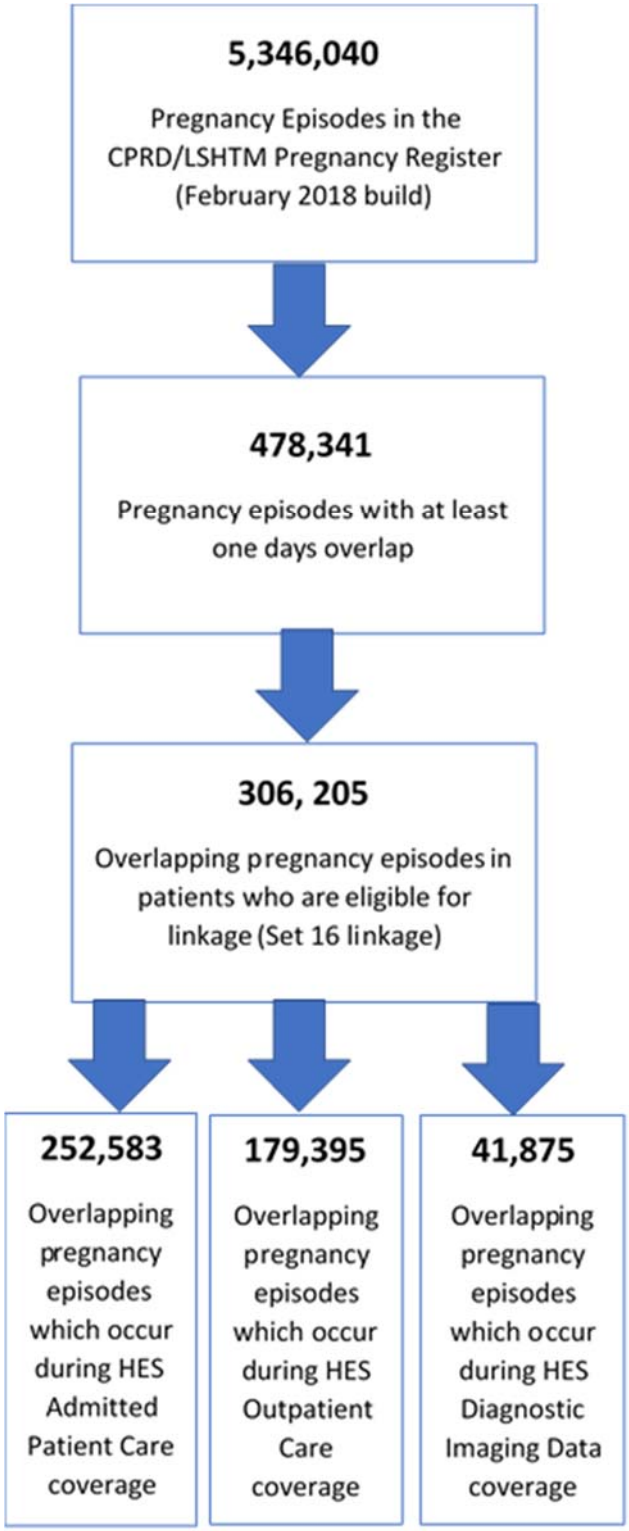
This is a methodological study intended to further develop the algorithm used to produce the CPRD Pregnancy Register.

8. Study Design

This study is not a classic epidemiological study design however, it is validating a pregnancy cohort.

9. Feasibility counts

For some analyses we will use all overlapping pregnancies in the Register (n= 478,341 as in the figure below) for other analyses we will restrict to those patients who are eligible for HES linkage.



10. Sample size considerations

We will include all overlapping pregnancies in the Pregnancy Register (see section I above)

11. Planned use of linked data (if applicable):

Approval was given for protocol 11_058 to use linked HES data to validate the pregnancy algorithm, this is now complete (Minassian C, Williams R, Meeraus W, Campbell O, n.d.). Here we expand on how the data will be further used to achieve the study aims.
We are requesting access to linked HES APC, HES Outpatient and HES DID. These are all data sources which may contain supporting evidence of a patient's pregnancy timing, gestation and outcome. We propose to use these data sources to look for:

- Records of gestation information
- Records of foetal scans or evidence suggesting a current pregnancy.
- Records of pregnancy outcomes (Delivery or other pregnancy loss either spontaneous or induced).
- Dates of pregnancy outcomes in order to ascertain the end of the pregnancy episode.

We will utilise this information to validate overlapping pregnancies in the Register and attempt to ascertain which episode is most likely to be true. Information indicating current pregnancy recorded in the linked data will be taken as evidence against an episode in the Register being historical.

In HES APC we propose to look at the Diagnosis and Procedures files for ICD10 and OPCS codes which indicate a pregnancy or its outcome. We also intend to use the HES Maternity file, looking at combinations of different data fields in the file to ascertain if and when a delivery took place. Whilst we recognise that HES Outpatient data contains limited information on diagnoses and procedures, it does contain information on the specialty which the patient visited including Maternity and Obstetrics. We intend to utilise the HES DID data to look for records of foetal scans, these will be regarded as evidence that a pregnancy recorded in CPRD is not historical.

The CPRD Mother Baby Link will also be utilised in order to obtain additional information on infant month of birth to compare with the estimated delivery dates for overlapping pregnancies.

12. Definition of the Study population

All women with at least one pregnancy episode in the Pregnancy Register which has been flagged as overlapping with another pregnancy for the same woman using the conflict field in the Pregnancy Register. Pregnancies must have at least one day of overlap to be flagged as conflicting. The February 2018 version of the Pregnancy Register will be used as this work follows on from previous work on pregnancies with no recorded outcome (ISAC 17_285R_2) which was done using this build. We will include all patients and we will flag patients who did not have a CPRD patient acceptability flag or whose pregnancy was not during UTS follow-up as part of our analysis.

13. Selection of comparison group(s) or controls

Not applicable- we do not intend to use a comparison group as part of this methodological study.

14. Exposures, Outcomes and Covariates

The aim of this study is to further develop methodology which allows researchers to study exposure during pregnancy and as such it is not a traditional study in which exposures, covariates and outcomes apply. However, as the aim of the study is to further understand why overlapping pregnancies are occurring in the Pregnancy Register and to help us untangle which, if either, of the overlapping episodes are true and correctly timed, we are looking for evidence for or against a record as a true, current pregnancy. Our outcomes of interest are therefore 1.all end of pregnancy outcomes including live deliveries, stillbirths, early pregnancy losses (spontaneous and induced); and 2. evidence of current pregnancy, as defined by Read and ICD codes (see appendices) 3. evidence against current pregnancy, for example hysterectomy codes. We will generate the following variables from the available data for use in the proposed analysis outlined in section O:

Descriptive Variables to be generated	**Data Source**
Count of the number of overlapping pregnancies per woman	Pregnancy Register
Count of the number of other episodes each overlapping episode overlaps with.	Pregnancy Register
Variable to indicate pairs/groups of overlapping episodes	Pregnancy Register
Flag to indicate the combination of outcomes for the overlapping episodes (miscarriage/delivery, delivery/delivery, etc)	Pregnancy Register
Count of the number of overlapping early pregnancy Losses (as generated by the algorithm) per woman	Pregnancy Register
Count of the number of deliveries episodes (as generated by the algorithm) per woman	Pregnancy Register
Count of the number of antenatal pregnancy records within each pregnancy episode	CPRD primary care data and Pregnancy Register
Flag to indicate if the pregnancy episode which overlaps is in first year of current registration.	CPRD primary care data and Pregnancy Register
Flag to indicate whether the pregnancy episode which overlaps is before the start of current registration	CPRD primary care data and Pregnancy Register
Flag to indicate whether the pregnancy episode which overlaps is during UTS registration	CPRD primary care data and Pregnancy Register
Flag to indicate if the woman's data was acceptable according to the CPRD acceptability flag	CPRD primary care data.
Flags to indicate whether each overlapping episode is eligible to be linked to each of the different data sets proposed AND overlaps with coverage period for that data source.	Pregnancy Register and CPRD linkage eligibility file
Flag to indicate that there is supporting evidence of a current pregnancy in the linked data within appropriate time limits of the pregnancy episode in the Pregnancy Register. This will be repeated for all linked datasets separately.	HES OP, HES APC, HES DID.
Flag to indicate there is supporting evidence of a pregnancy outcome in the linked data within appropriate time limits of the pregnancy episode in the Pregnancy Register. This will be repeated for all linked datasets separately.	HES OP, HES APC.
Flag to indicate if the pregnancy has a corresponding linked infant in the mother-baby link.	Pregnancy Register and MBL
Flag for each group of overlapping episodes to indicate whether any pregnancies in the group have a corresponding linked infant in the mother-baby link.	Pregnancy Register and MBL
Flag for each pregnancy to indicate whether there is supporting evidence of it being a current pregnancy (using a code list of antenatal flags such as foetal scan, fundus measurement etc) in the primary care data e.g. a code to indicate a foetal scan.	CPRD primary care data

Flag to indicate the patient has a record of hysterectomy prior to the pregnancy records.	CPRD primary care data, HES APC
Flag to indicate a pregnancy loss and delivery recorded on the same day for that group of overlapping episodes.	Pregnancy Register

15. Data/ Statistical Analysis

This study follows on from ISAC 17_285R_2 where we investigated pregnancies without outcome in the Pregnancy Register and a very similar methodology will be used.

We will begin by extracting the list of overlapping pregnancies from the Pregnancy Register. We will then generate a list of the patients who have two or more of these overlapping pregnancies, who will then form the cohort for this study. All pregnancy records for these patients will be extracted from the CPRD GOLD database using the pregnancy code list (Read and entity codes) upon which the pregnancy algorithm is based to create a dataset which includes all the pregnancy records and the summary Pregnancy Register information for these women.

Descriptive variables will be generated to further characterise the women and their pregnancies in the dataset. The full list of variables we will generate is described in section N above. For each overlapping pregnancy episode we will describe its temporal proximity to other pregnancy episodes and what the outcome of other pregnancies in proximity were, deliveries or pregnancy losses. We will characterise the overlapping pregnancies, specifically: the number of antenatal codes which make up the episode and the use of Jan 1st as the outcome recording date (as a potential default date for a past pregnancy). We will also investigate how the timing of the pregnancies relates to the start and end of patient follow up in CPRD. We will ascertain whether the timing of the overlapping pregnancies falls within the coverage period of each of the linked datasets requested for those patients who are eligible for linkage. For pregnancy episodes which fall within linkage coverage we will look for evidence of pregnancy recorded within the timeframe of the pregnancy record in the Register and will flag those pregnancies as having linked data evidence. We will also look for records of codes relating to events unlikely to be recorded historically such as foetal scans in the primary care data.
Based on the logic of the algorithm and how the data are structured we have identified seven scenarios which may explain why overlapping pregnancies are occurring in the Pregnancy Register and for which we may be able to find evidence within the primary and linked data. In this analysis we intend to systematically query the data to look for evidence of these scenarios. Each pregnancy episode will be flagged for all scenarios which could potentially apply (diagrams of the scenarios are included in the appendices). Evidence of these scenarios will then be used to develop a framework of advice for users of the Register. The scenarios we intend to consider are as follows (scenarios presented relate to pairs of overlapping pregnancies which are likely to account for the large majority):

Scenario	How does this appear in the data?	Overlapping Pregnancy Status	Potential use of linked data
Problem 1: Both pregnancies are real but one episode is a historical pregnancy			
1a. The GP records a past delivery during a current pregnancy > 25weeks before the true delivery of that pregnancy. OR a past pregnancy loss > 12 weeks before the actual loss of that pregnancy	Both pregnancies will have the same outcome type. Evidence of current pregnancy codes would be expected to fall within the second pregnancy.	First pregnancy in the register is past and the second pregnancy is current.	HES data will be searched for suitably timed outcomes to attempt to validate which of the deliveries is correct.
1b. If a patient has a record relating to a previous loss recorded during a pregnancy ending in delivery or vice-versa then overlapping episodes will be created by the algorithm. The algorithm first generates episodes for consecutive deliveries; it then does the same thing for pregnancy losses. There is no step in the algorithm to check that the loss episodes	The overlapping pregnancies must consist of one loss and one delivery. Evidence of current pregnancy codes would be expected to fall within the first pregnancy.	One episode is past and the other is current.	HES data will be searched for suitably timed outcomes to attempt to validate which of the outcomes identified by the algorithm is correct.

	do not coincide with the delivery episodes.			
	Problem 2: Both pregnancies are historical			
	2. A patient joins a new practice (or has another reason for a full obstetric history to be taken) and has information on historical pregnancies recorded with the current date rather than the actual date of the event. If one pregnancy ended in a loss and one in delivery they will be generated as separate overlapping pregnancies ending on the same day by the algorithm (as in scenario 3)	The overlapping pregnancies must consist of one loss and one delivery. The pregnancy end dates will be the same for both pregnancies. Both pregnancies are likely to be <1 year after the patient's current registration date. We would not expect to find codes indicating current pregnancy.	Both episodes are past pregnancies.	HES data will be searched for suitably timed pregnancy data to attempt to validate whether either of the outcomes are recorded at the correct time.
	Problem 3: Both pregnancy episodes are real but the gestation of the second pregnancy applied by the algorithm is too long.			
	3a. The woman has two pregnancy losses which are >8 weeks and <12 weeks apart. The second pregnancy has no information about gestation recorded so the algorithm applies a default of 12 weeks and the episodes overlap.	Both overlapping pregnancies must be losses. The maximum overlap between the two pregnancies must be 4 weeks. Evidence of current pregnancy codes could be found in either pregnancy.	First and second pregnancy are current, but the timings are wrong.	HES data will be searched for suitably timed outcomes to attempt to validate the timing of the loss outcomes recorded in the Register.
	3b. The woman has two pregnancies close together and the second pregnancy ends in delivery. If the information on the LMP in the data of the second pregnancy is wrong then the algorithm may generate the start too early resulting in an overlap.	The second pregnancy must be a delivery and have no information about gestation in the data. The overlap must be <15 weeks. There may be evidence of current pregnancy codes in either pregnancy	First and second pregnancy are current but the timings are wrong	HES data will be searched for suitably timed outcomes to support the second pregnancy.
	Problem 4: The pregnancy is real but is split into separate episodes by the rules of the algorithm			

4a. The GP records further information about a pregnancy > 25 weeks after the delivery date for pregnancies ending in delivery OR >8 weeks but <12 weeks for pregnancies ending in loss. The algorithm assumes this is a different pregnancy and generates a new episode, this may overlap with the “true” episode.	Both pregnancies must be of the same outcome type. Evidence of current pregnancy codes would be expected to fall within the first pregnancy.	First pregnancy in the register is current the second pregnancy refers to the same pregnancy.	HES data will be searched for pregnancy outcomes to look for evidence that these are really two separate pregnancies
4b The GP records information about a pregnancy but no outcome, with gaps in recording of more than 6 weeks between successive records. The algorithm splits the pregnancy into two separate episodes due to antenatal records being > 6 weeks apart. If there is gestational information included in the second episode the start of this episode will be assigned before the start of the previous episode resulting in a nested pregnancy episode.	Both pregnancies must be pregnancies without outcome in the register. The start of the second episode must be > 6 weeks after the end of the first episode. The start of the first pregnancy must have been generated from information in the data. There may be evidence of current pregnancy in either episode.	Both episodes are true and are part of the same pregnancy	HES data will be searched for pregnancy outcomes to look for evidence that these are really two separate pregnancies

16. Plan for addressing confounding

Not applicable to this methodological study.

17. Plans for addressing missing data

The objective of this study is to look for information on pregnancies which has not previously been identified in order to strengthen certainty around validity, type and timing of pregnancy episodes in the Pregnancy Register. In this situation missing data will not lead to bias but just limits the usefulness of the analysis. However, any evidence found is useful.

18. Patient or user group involvement (if applicable)

Not applicable- there are no plans for patient or user group involvement in this methodological work.

19. Plans for disseminating and communicating study results, including the presence or absence of any restrictions on the extent and timing of publication

It is our intention to submit this work to the 2020 International Conference of Pharmacoepidemiology as well as to publish the work in a suitable scientific journal. We will also report our findings to the CPRD Observational Research Team to help to inform decisions about the best way to further develop the Pregnancy Register as a useful tool for CPRD data users.

Conflict of interest statement: All investigators declare no conflict of interest.

20. Limitations of the study design, data sources, and analytic methods

Not all patients in the CPRD/LSHTM Pregnancy Register are eligible for linkage and not all pregnancies identified are within the coverage periods of the data sources we propose to use and so will be excluded from analyses using linked data. Furthermore, for some pregnancies it may not be possible to gather further evidence from HES. As there is no clear gold standard among the data sources in this analysis, when conflicting evidence exists across data sources this will need to be noted as part of the study findings.

21. References:

Devine, S., West, S., Andrews, E., Tennis, P., Hammad, T. A., Eaton, S., … Olshan, A. (2010). The identification of pregnancies within the general practice research database. *Pharmacoepidemiology and Drug Safety*, *19*(1), 45–50. https://doi.org/10.1002/pds.1862

Hardy, J. R., Holford, T. R., Hall, G. C., & Bracken, M. B. (2004). Strategies for identifying pregnancies in the automated medical records of the General Practice Research Database. *Pharmacoepidemiology and Drug Safety*, *13*(11), 749–759. https://doi.org/10.1002/pds.935

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

Minassian, C., Williams, R., Meeraus, W. H., Smeeth, L., Campbell, O. M. R., & Thomas, S. L. (2019). Methods to generate and validate a Pregnancy Register in the UK Clinical Practice Research Datalink primary care database. *Pharmacoepidemiology and Drug Safety*, (April), 1–11. https://doi.org/10.1002/pds.4811

Public Health England. (2014). Pertussis Vaccination Programme for Pregnant Women: vaccine coverage estimates in England, April to August 2014 - GOV.UK, *11*(34). Retrieved from https://www.gov.uk/government/publications/pertussis-immunisation-in-pregnancy-vaccine-coverage-estimates-in-england-october-2013-to-march-2014/pertussis-vaccination-programme-for-pregnant-women-vaccine-coverage-estimates-in-england-april-to-august-2014

Webster, W. S., & Freeman, J. A. D. (2003). Prescription drugs and pregnancy. *Expert Opinion on Pharmacotherapy*, *4*(6), 949–961. https://doi.org/10.1517/14656566.4.6.949

List of Appendices:

- Diagram of pregnancy register methodology
- Diagrams of proposed scenarios
- Read code list for identifying pregnancies and outcomes
- ICD code list for identifying early pregnancy loss
- ICD code list for evidence of pregnancy excluding early loss
- OPCS code list for identifying pregnancy outcomes
- Gemscript code for identifying pregnancy outcomes

Appendix 4: Approved RDG application for COVID-19 and pregnancy loss study

CPRD Research Data Governance (RDG) Application Template

ALL APPLICATIONS MUST BE COMPLETED AND SUBMITTED VIA THE CPRD ELECTRONIC RESEARCH APPLICATION PORTAL (eRAP) www.erap.cprd.com

Applicants may use this template offline to prepare their research application, prior to submission on eRAP. Applicants must also read CPRD's Research Data Governance (RDG) Guidance on how to complete their application found on the eRAP landing page under Related resources (https://www.erap.cprd.com/)

PART 1: APPLICATION FORM

GENERAL INFORMATION ABOUT THE PROPOSED RESEARCH STUDY

4. Study Title (Max. 255 characters including spaces)

5. Research Area (place '**X**' in all boxes that apply)

Drug Safety		Economics	
Drug Utilisation		Pharmacoeconomics	
Drug Effectiveness		Pharmacoepidemiology	
Disease Epidemiology	x	Methodological	x
Health Services Delivery			

6. Does this protocol describe an observational study using purely CPRD data?

Yes	**x**	No	

7. Does this protocol involve requesting any additional information from GPs, or contact with patients?

Yes		No	**x**

If yes, provide the reference number:

8. Chief Investigator

Title:	Mrs
Full name:	Jennifer Campbell
Job title:	Senior Researcher
Affiliation/organisation:	CPRD and LSHTM
Email address:	Jennifer.campbell@mhra.gov.uk
CV Number (if applicable):	
Will this person be analysing the data? (Y/N)	y

9. Corresponding Applicant

Title:	Mrs

Full name:	Jennifer Campbell
Job title:	Senior Researcher
Affiliation/organisation:	CPRD and LSHTM
Email address:	Jennifer.campbell@mhra.gov.uk
CV Number (if applicable):	
Will this person be analysing the data? (Y/N)	y

10. List of all investigators/collaborators

Title:	
Full name:	
Job title:	
Affiliation/organisation:	
Email address:	
CV Number (if applicable):	
Will this person be analysing the data? (Y/N)	

[Add more investigators/collaborators as necessary by copy and pasting a new table for each investigator/collaborator]

ACCESS TO THE DATA

11. Sponsor of the study

Institution/Organisation:	
Address:	

12. Funding source for the study

Same as Sponsor?	Yes		No	
Institution/Organisation:				
Address:				

13. Institution conducting the research

Same as Sponsor?	Yes		No	
Institution/Organisation:				
Address:				

14. Data Access Arrangements

Indicate with an '**X**' the method that will be used to access the data for this study:

Study-specific Dataset Agreement	

Institutional Multi-study Licence	
Institution Name	
Institution Address	

Will the dataset be extracted by CPRD?

Yes		No	

If yes, provide the reference number:

15. Data Processor(s):

Processing	
Accessing	
Storing	
Processing area (UK/EEA/Worldwide)	
Organisation name	
Organisation address	

Processing	
Accessing	
Storing	
Processing area (UK/EEA/Worldwide)	
Organisation name	
Organisation address	

[Add more processors as necessary by copy and pasting a new table for each processor]

INFORMATION ON DATA

16. Primary care data (place '**X**' in all boxes that apply)

CPRD GOLD		CPRD Aurum	x

Reference number (if applicable):

17. Please select any linked data or data products being requested

Patient Level Data (place '**X**' in all boxes that apply)

ONS Death Registration Data		NCRAS Cancer Registration Data	
HES Admitted Patient Care	x	NCRAS Cancer Patient Experience Survey (CPES) data	
HES Outpatient	x	NCRAS Systemic Anti-Cancer Treatment (SACT) data	
HES Accident and Emergency		NCRAS National Radiotherapy Dataset (RTDS) data	
HES Diagnostic Imaging Dataset	x	NCRAS Quality of Life Cancer Survivors Pilot (QOLP)	
HES PROMS (Patient Reported Outcomes Measure)		NCRAS Quality of Life Colorectal Cancer Survivors (QOLC)	
CPRD Mother Baby Link		Second Generation Surveillance System (SGSS, COVID-19)	x
Pregnancy Register	x	COVID-19 Hospitalisations in England Surveillance System (CHESS)	
Mental Health Data Set (MHDS)			

Area Level Data (place '**X**' in <u>one</u> Practice / Patient level box that may apply)

Practice level (UK)		**Patient level (England only)**	
Practice Level Index of Multiple Deprivation		Patient Level Index of Multiple Deprivation	x
Practice Level Index of Multiple Deprivation (index other than the most recent)		Patient Level Index of Multiple Deprivation Domains	
Practice Level Index of Multiple Deprivation Domains		Patient Level Carstairs Index for 2011 Census	
Practice Level Carstairs Index for 2011 Census (Excluding Northern Ireland)		Patient Level Townsend Score	
2011 Rural-Urban Classification at LSOA level		2011 Rural-Urban Classification at LSOA level	

Reference / Protocol number (where applicable):

18. Are you requesting linkage to a dataset not listed above?

Yes		No	**x**

If yes, provide the Non-Standard Linkage reference number:

19. Does any person named in this application already have access to any of these data in a patient identifiable form, or associated with an identifiable patient index?

Yes		No	**x**

If yes, provide further details:

PART 2: PROTOCOL INFORMATION

Applicants must complete all sections
R. Study Title **Does having Covid-19 whilst pregnant increase the risk of pregnancy loss (miscarriage or stillbirth)? A matched cohort study**
A. Lay Summary (Max. 250 words) Covid-19 is a new disease caused by infection with the SARS-CoV2 virus, little is known so far about the impact it might have on pregnancy outcomes if a woman gets Covid-19 whilst pregnant. We are proposing to use de-identified electronic primary care records to investigate whether having Covid-19 during pregnancy increases the chances of the pregnancy ending in miscarriage or stillbirth (pregnancy loss).We will use data from the first wave of the Covid-19 pandemic in the UK (01/03/2020 - 10/06/2020) We will look at whether there is any change in risk depending on the severity of the Covid-19 disease or at which stage of pregnancy it is contracted. Furthermore, there is the potential that broader factors associated with the pandemic itself resulted in an increased chance of pregnancy loss for example reduced contact with healthcare providers, we will explore this and any interaction it may have with the relationship between Covid-19 and pregnancy loss. Finally, ascertaining when a woman is pregnant in electronic primary care records can be challenging. There may be a chance that pregnancy loss is less likely to be recorded completely than live births. Building on previous research we have done we will use this study to examine whether adjusting the way we define pregnancy in the data changes the results of our analysis. We hope to provide valuable insight into the potential risks of having Covid-19 whilst pregnant.
S. Technical Summary (Max. 300 words) Since the emergence of the SARS-COv2 virus and its associated illness Covid-19, a number of studies and case reports have hypothesised a potential increased risk of pregnancy loss (miscarriage or stillbirth) associated with Covid-19 whilst pregnant. We propose to examine this potential association with a large-scale observational cohort study in CPRD Aurum. We will use the CPRD Aurum Pregnancy Register to select women whose pregnancy began between 01/03/2020 and the 10/06/2020 with no record of Covid-19 prior to their pregnancy. We will match women with a record of Covid-19 during pregnancy with controls who are at the same gestational and maternal age. Using a Cox regression model we will estimate the hazard ratio for the risk of pregnancy loss between those who had Covid-19 whilst pregnant and those who did not. We will also select a historical comparison cohort of woman who were pregnant in 2018 in order to examine whether factors associated with the pandemic itself, such as reduced healthcare contacts, had an impact on the risk of pregnancy loss. Whilst CPRD pregnancy registers are extremely useful the nature of the data means they contain uncertain pregnancy episodes. Our previous research examined reasons these episodes might exist and the potential impact of their inclusion or exclusion from studies. We will use this study to apply recommendations from our research on the optimal way to handle uncertain episodes by changing the criteria we use to select pregnancies and examining any impact this has on the results. We will use linked secondary care to obtain additional pregnancy information missing from the primary care records. This study will not only provide valuable insight into the relationship between Covid-19 and pregnancy loss but will also act as proof of concept study for our recommended methodologies when using electronic health records to study pregnancy.

T. Outcomes to be Measured

Miscarriage and Stillbirth.

U. Objectives, Specific Aims and Rationale

Objective

To evaluate the potential impact of having Covid-19 during pregnancy on the risk of pregnancy loss (miscarriage or stillbirth).

Specific Aims

10. To evaluate whether having Covid-19 during pregnancy is associated with risk of miscarriage or stillbirth and whether this differs by trimester and severity (hospitalised vs non-hospitalised patients) using a Cox regression model.
11. To evaluate whether pandemic changes such as the reduction in the utilisation of primary care services may have increased the risk of miscarriage or stillbirth independently of Covid-19 infection.
12. To evaluate the robustness of our model to changes in the definition of the study population (i.e. including pregnancies with less detail recorded in primary care).

Rationale

Covid-19 emerged as a novel viral disease towards the end of 2019 and little is known so far about its impact on pregnancy outcomes if contracted whilst pregnant. A number of smaller studies and case reports have hypothesised a risk to the unborn foetus likely to be mediated by placental damage (1) However, to date there have been no large observational studies which have looked at the potential association between having Covid-19 whilst pregnant and miscarriage or stillbirth. We therefore propose to utilise CPRD Aurum primary care data including the Pregnancy Register to attempt to address this question in a large population-based cohort study. Furthermore, it has been shown that during the first wave of the Covid-19 pandemic there was a significant reduction in overall primary care contacts in the UK (2). We would like to investigate whether this is true amongst the pregnant population and any impact this may have on the risk of miscarriage or stillbirth.

Finally, whilst the CPRD pregnancy registers are extremely useful for identifying episodes of pregnancy within the CPRD primary care data, they were designed to be sensitive rather than specific. As a result of this there are many pregnancy episodes within the register which are uncertain either because they have no recorded outcome in the pregnancy register, or they overlap with another pregnancy episode for the same woman. In previous work on the CPRD GOLD pregnancy register we examined the potential impact of including or excluding uncertain pregnancy episodes from a study cohort (3). We concluded that excluding all uncertain pregnancy episodes may result in an underestimation of pregnancies ending in loss. We therefore wish to examine whether adjusting the inclusion criteria of our pregnancies cohorts in a second analysis has any impact on the observed relationship between having Covid-19 whilst pregnant and pregnancy loss.

V. Study Background

SARS-CoV-2 emerged as a new coronavirus at the end of 2019 spreading rapidly to cause a global pandemic of its associated illness Covid-19. Many millions of people around the world have been infected with the virus including pregnant women. In the UK the first wave of Covid-19 infections occurred between February and July 2020 (UK Government, 2022)

Since the start of the pandemic there have been numerous studies which have examined the potential effect of Covid-19 on pregnancy outcomes (5,6) Several of them have concluded an increased risk of miscarriage in mothers who tested positive for SARS-CoV-2 (6) Studies discussed a potential aetiological effect of the virus on the placenta causing inflammation, resulting in foetal growth retardation potentially inducing miscarriage (1). However, to date there have been no large-scale observational studies which have looked at any potential association between Covid-19 and risk of pregnancy loss (miscarriage or stillbirth).

Electronic Health Record databases such as CPRD Aurum represent a useful tool to study outcomes and exposures in pregnant women with the potential to easily establish a large study cohort (7). Development of an algorithm to produce a register of pregnancy episodes in CPRD data in 2016 made the CPRD primary care data even more useful (8) . The CPRD Aurum database contains electronic health records for ~ 40 million patients including over 16 million women with a pregnancy record. CPRD Aurum therefore offers the opportunity to conduct a large observational study on the impact of Covid-19 on pregnancy outcomes during the first wave of the pandemic in the UK.

The CPRD pregnancy registers are designed to be sensitive, capturing all records of pregnancy within the primary care databases. There are therefore pregnancy episodes within the registers which are uncertain either because the outcome of the pregnancy is missing or because they overlap with another episode for the same woman. In previous work we have examined potential reasons for the existence of uncertain pregnancy episodes in the CPRD Pregnancy Registers and produced recommendations for researchers on how to handle them (3). Alongside examining the association between Covid-19 and pregnancy loss we intend to use this study as a proof of concept to examine whether making changes to the way in which we define our pregnant cohort has an impact on the results of the study.

W. Study Type

Hypothesis Testing.
Our null hypothesis is that there is no association between having Covid-19 whilst pregnant and the risk of pregnancy loss (miscarriage or stillbirth).

X. Study Design

Matched cohort study

Y. Feasibility counts

There are 89,008 women in the May 2021 version of CPRD Aurum who have a pregstart date after 01/03/2020 and before 10/06/2020. Of these 14,199 had a record of Covid-19 between their pregstart and pregend (as defined by the pregnancy register).

Z. Sample size considerations

At a conservative estimate, the prevalence of miscarriage in the general population is 12% (9).

Probability of Covid-19 exposure from our feasibility count is 14199 / 89008 = 0.16 and the standard deviation is ~0.366. Therefore, with 89,000 women in our cohort and a 0.12 probability of the outcome the largest hazard ratio that can be detected with a preserved 80% power and a 95% confidence interval is roughly 1.08.

AA. Planned use of linked data (if applicable):

Pregnancy Register

The CPRD Aurum pregnancy register will be used to determine pregnancy episodes in the primary care data. The Pregnancy Register will also be used to determine the outcome of interest (outcome field = 2,3, and 4)

HES-Admitted Patient Care (APC)

HES-APC data will be used to obtain additional Covid-19 diagnoses which may not be in the primary care record. It will also be used to generate a sub cohort of patients who were hospitalised with Covid-19 in order to examine if risk differs by severity. Finally, HES-APC will be used in our second analysis to look for additional pregnancy outcomes which may be missing from the Pregnancy Register.

HES- Outpatient

HES Outpatient will be used in our second analysis to look for additional pregnancy outcomes which may be missing from the Pregnancy Register.

HES- Diagnostic Imaging Data (DID)

HES-DID will be used to look for evidence that foetal scans took place in order to assess healthcare utilisation and to look for further confirmatory evidence of current pregnancy in our second analysis.

All three of the HES datasets above will also be used to obtain additional ethnicity information which will feed into an ethnicity algorithm developed by CPRD colleagues designed to determine a patients most likely ethnicity.

SGSS

The SGSS data will be used as supplementary information to look for additional records of positive Covid-19 tests which may be missing from primary care.

Patient Level IMD

Patient level IMD will be used to adjust for socio-economic status in our model in an attempt to address any potential confounding.

Use of linked data in the study will allow us to obtain further data on covariates, outcomes and exposures which may be missing from primary care making the study more robust. This study will benefit patients in England and Wales by providing valuable further information on the potential risks Covid-19 poses to pregnant women and their babies.

BB. Definition of the Study population

Patients will be selected from CPRD Aurum based on the following criteria:

Inclusion criteria

- Women with a record of pregnancy in the CPRD Aurum pregnancy register with a pregstart between the 01/03/2020 and the 10/06/2020 (to allow for nine months of linked HES follow-up after the latest pregstart)
- Who are flagged as acceptable for research
- Who are eligible for linkage to HES secondary care data.
- Whose pregnancy start is at least 280 days before their last data collection date.

Exclusion criteria

- Women with a record of Covid-19 (in either primary care, HES-APC or SGSS) prior to their pregnancy start.

For our first analysis we will also exclude

- Women whose pregnancy episode has no recorded outcome in the pregnancy register
- Women whose pregnancy episode overlaps with another episode for the same woman in the pregnancy register.

Start of follow-up will be from week 4+1 of pregnancy as defined by the Pregnancy Register. End of follow-up will be the earliest of, end of pregnancy, last collection date, transfer out date or death date.

CC. Selection of comparison group(s) or controls

Women who do not have a record of Covid-19 before week 4 +1 of pregnancy will be eligible as controls. Women with Covid-19 (cases) will be matched (without replacement) to up to 3 controls who have not had Covid-19 by the same gestational age as the case at index date. Controls will also be matched on maternal age. Controls who get Covid-19 will be censored as controls at the earliest Covid-19 record and will then become a case and be allocated their own set of controls.

In addition we will also include historical controls who will also be matched 3:1 to cases by gestational age and maternal age at index date.

The Historical control pool will be selected as follows:
Inclusion criteria

- Women with a record of pregnancy in the CPRD Aurum Pregnancy Register with a pregstart between the 01/02/2018 and the 10/06/2018
- Who are flagged as acceptable for research.
- Who are eligible for linkage.
- Whose pregnancy start is after the practice UTS date.
- Whose pregnancy start is at least 280 days before the last data collection date.

Exclusion Criteria

For our first analysis we will also exclude

- Women whose pregnancy episode has no recorded outcome in the pregnancy register. (Analysis 1 only)
- Women whose pregnancy episode overlaps with another episode for the same woman in the pregnancy register (as defined by the conflict flag).

DD. Exposures, Outcomes and Covariates

Exposure

Exposure will be defined as a record of Covid-19 between the start and end of pregnancy. Covid-19 records may be a record of a medcode associated with Covid-19 infection in the primary care data (Appendix 1) OR a record of a Covid-19 ICD code in the linked HES-APC data (appendix 2) OR a record of a positive COVID-19 test in the SGSS data. A woman will be considered to be pregnant between the dates of the pregstart and pregend variable as defined by the CPRD Aurum pregnancy register.

We will examine symptomatic Covid-19 rather than SARS-CoV2 infection as exposure of interest in order to avoid misclassification of non-symptomatic cases. Asymptomatic testing was not available in the UK during our study period so we presume that cases recorded in CPRD Aurum during this time must have been detected through symptomatic disease.

Outcomes

The main outcome of interest will be pregnancy loss by miscarriage after week 4 (any time after and including week4 +1 day) of pregnancy or stillbirth as recorded in the CPRD Aurum Pregnancy Register. For our second analysis we will also look for miscarriage and stillbirth records in the linked HES APC data using ICD (appendix 3) within 294 days (42 weeks) of the pregstart.

Covariates

We will include the following covariates in our model (Maternal age and gestational age will be matching variables):

- Maternal Age at pregnancy start. This will be calculated by subtracting the mother's year of birth from the year of pregstart.
- Gestational Age at index. This will be calculated by subtracting the indexdate from the pregstart variable date.
- SES will be defined using patient-level IMD data
- Smoking status categorised as current, never and ex will be taken from the last smoking record prior to the pregnancy end date.
- BMI will be calculated using the latest height measurement recorded in CPRD Aurum and the last weight record prior to the pregnancy start date. BMI will be categorised as Not obese (<30) Obese class 1 (30-34.9) Obese class 2 (35-39.9) Obese class 3 (>=40)
- Ethnicity will be categorised as White, Mixed, South Asian, Black, Other using an ethnicity algorithm developed by CPRD colleagues.
- Chronic health conditions
    - Diabetes (categorised as Controlled (HbA1C <58mmol/mol) Uncontrolled (HbA1c >=58mmol/mol), Unknown HbA1c)
    - Gestational Diabetes
    - Long term kidney disease (any record prior to pregnancy start)
    - HIV infection (any record of HIV infection prior to pregnancy start)
    - Rheumatoid arthritis (any record prior to pregnancy start)
- Immunosuppressive drug use (a prescription record in the 6 months prior to pregnancy start)

EE. Data/ Statistical Analysis

Analysis will be conducted using Stata.

Descriptive baseline tables will be produced for each of the three cohorts. These will include mean patient follow-up time in days and the proportional distribution of each of the covariates included in the model

Hazard ratios will be calculated for:

- Women with a record of Covid-19 during pregnancy compared with those without in 2020
- Women with a record of Covid-19 during pregnancy compared with those without in a pre-pandemic time period (historical cohort)
- Woman without a record of Covid-19 during pregnancy in 2020 compared to those in the historical cohort.

Cox regression analysis, with stratification of the baseline hazard by matched set, will be performed to adjust for potential confounders. Adjusted hazard ratios will be produced comparing the groups above.

Cases will contribute person time to the analysis from their index date which will be their first Covid-19 record (or 4weeks plus one day of pregnancy if they have a Covid-19 record in the first 4 weeks of gestation). Cases will be censored before the event if they reach the end of CPRD follow-up (earliest of, last collection date, transfer out date or death date), when they give birth, or if their pregnancy ends in termination, ectopic or molar pregnancy. We will only include the first pregnancy each woman has within the study period.

A potential issue with presenting a single HR for the whole pregnancy is non-proportional hazards. We will use log-log survival curve and we will apply a goodness of fit test of the Schoenfeld residuals to check the correlation between the residuals and survival time. If the proportional hazards assumption holds for each trimester we partition the time axis in order to calculate hazard ratios in cumulative time periods by trimester (4-12 weeks, 4-24 weeks and 4-42 weeks). As a secondary analysis we will fit an interaction by the trimester at index date.

Codes for miscarriage recording vs elective termination can sometimes be ambiguous. We will therefore perform sensitivity analyses to investigate whether using a more restrictive code list for miscarriage has an effect on the observed association.

We will examine the feasibility of using HES- APC data to look at Covid-19 hospitalisation as a measure of severity. We will look at how hospitalisations are recorded and whether it is possible to identify when women were hospitalised with Covid-19 as the primary cause. We will assess numbers of women in the hospitalised group and if possible, we will conduct a secondary exploratory analysis with 3 way exposure (un-exposed, Covid-19, hospitalised with Covid-19).

In order to further examine any impact the pandemic itself had any effect on the risk of pregnancy loss, we will describe the patterns of healthcare utilisation in the two comparison cohorts. This will include frequency and type of GP visits and records of foetal scans recorded in both CPRD Aurum and HES DID.

We will re-run the model described above making stepwise adjustments to the way in which we define pregnancy. These adjustments will be made based on our previous research conducted around the potential reasons for uncertain pregnancy records within the pregnancy register(10). We will apply the following

Step one including pregnancies with no recorded outcome which meet the following criteria:

- We will utilise linked HES data to look for missing outcomes and include any additional pregnancies. Any pregnancies for which outcomes can be obtained will be included in the cohort.
- We will restrict pregnancies to those where the woman has at least nine months follow-up after the pregnancy start to ensure that all episodes have the potential for the outcome to be recorded.
- Exclude episodes which are likely to be derived from historical data based on our previous

research.

Step two including conflicting pregnancy episodes which meet the following criteria:

- Using criteria developed in our previous study we will ascertain conflicting pregnancy episodes which are likely to be truly one pregnancy and have been split by the algorithm which generates the pregnancy register. We will merge these episodes and adjust the start and end dates accordingly (deciding which of the outcomes is likely to be the true outcome based on the scenarios we have described and then estimating a start date. This will be based on a combination of the patient's antenatal records and default duration dependent on outcome type)

- Exclude episodes which are likely to be derived from historical data based on our previous research.

Step three including pregnancies that meet the criteria of step one AND step two.

FF. Plan for addressing confounding

We will use our model to assess the following potential confounders (see covariates section for definitions):

- Age
- SES (defined using patient-level IMD data)
- Smoking status (categorised as current, never and ex)
- BMI (categorised as >=40 or <40) taken from the last record prior to the woman's pregnancy start
- Ethnicity
- Chronic health conditions
    - Diabetes (including gestational diabetes)
    - Long term kidney disease
    - HIV infection
    - Rheumatoid arthritis
- Long term immunosuppressive drug use

GG. Plans for addressing missing data

We will utilise linked data to look for additional Covid-19 records which may be missing from CPRD Aurum.

For objective three of our study will utilise HES to try to obtain pregnancy outcomes which are missing from CPRD Aurum. We will assess the impact of this on the observed hazard ratio.

We will use complete case analysis rather than multiple imputation as data on covariates such as smoking, and BMI are unlikely to be missing at random. We will carry out sensitivity analysis using other methods such as multiple imputation and assumptions around when data is likely to be missing, for example patients who smoke are more likely to have a smoking record than those who do not.

HH. Patient or user group involvement

We have assembled a small focus group of women who were pregnant during the first wave of the Covid-19 pandemic. We will discuss their experiences of healthcare provision during that time and how they feel this compares to pre-pandemic maternity care. We will use this information to inform our analyses and discussion. We will share our results and draft publication with them in order to obtain feedback from a patient perspective.

II. Plans for disseminating and communicating study results

We plan to disseminate our results through conference presentations and a publication in a peer reviewed open access publications utilising the STROBE (Strengthening the Reports of Observational studies in Epidemiology) principles.

JJ. Conflict of interest statement

Jennifer Campbell and Rachael Williams are employees of CPRD

KK. Limitations of the study design, data sources, and analytic methods

Using electronic health records to study exposures and outcomes relies on the assumptions that events are recorded in the patient's medical record and that the dates associated with them are correct however, this may not always be the case. There is a chance that a patient's Covid-19 may not be recorded especially in the first wave of the pandemic when home testing was not available. We will therefore also conduct an analysis using a historical comparison cohort in order to examine misclassification of Covid-19 exposure in the control group.

Women who miscarry may not always report it to their GP and therefore we may under ascertain the outcome of interest. We will also examine whether using hospital data alongside primary care data allows us to obtain more miscarriage records. Furthermore, coding of miscarriage can be difficult to interpret with some ambiguous codes around pregnancy termination making it difficult to distinguish between miscarriage and termination. We will conduct sensitivity analyses to assess what impact inclusion and exclusion of these codes may have on the results.

The start and end of pregnancies can be difficult to ascertain using electronic health data. Whilst the CPRD pregnancy register helps to minimise this there is still the potential for error in the pregnancy start and end dates and therefore the misclassification of exposure status.

Due to our requirement for HES data and the time lag in the availability of linked HES data we will only be able to study the first wave of the Covid-19 pandemic prior to a vaccine becoming available. We will therefore be unable to assess whether vaccination status have any effect on the relationship between Covid-19 and pregnancy loss.

There may be further unmeasured confounding around health seeking behaviour which we are not able to adjust for. For example woman who are "health seeking" may be more likely to report both Covid-19 infections and also more likely to report their miscarriage.

A potential problem with the calculation of hazard ratios is the depletion of susceptibles over time (11). We will therefore present cumulative hazard ratios in order to examine this and if necessary conduct a secondary analysis fitting on interaction by trimester at baseline

LL. References

1. Arthurs AL, Jankovic-Karasoulos T, Roberts CT. COVID-19 in pregnancy: What we know from the first year of the pandemic. Biochim Biophys Acta - Mol Basis Dis. 2021;1867(12).
2. Mansfield KE, Mathur R, Tazare J, Henderson AD, Mulick AR, Carreira H, et al. Indirect acute effects of the COVID-19 pandemic on physical and mental health in the UK: a population-based study. Lancet Digit Heal. 2021;3(4):e217–30.
3. Campbell J, Bhaskaran K, Thomas SL, 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. Under Rev. 2022;
4. UK Government. Coronavirus (COVID-19) in the UK [Internet]. [cited 2022 Jan 24]. Available from: https://coronavirus.data.gov.uk/
5. Kazemi SN, Hajikhani B, Didar H, Hosseini SS, Haddadi S, Khalili F, et al. COVID-19 and cause of pregnancy loss during the pandemic: A systematic review. PLoS One [Internet]. 2021;16(8 August):1–10. Available from: http://dx.doi.org/10.1371/journal.pone.0255994
6. Soheili M, Ghobad M, Hamid Reza B, Soheili M, Mahidi Makhtari M, Moradi Y. Clinical manifestation and maternal complications and neonatal outcomes in pregnant women with COVID-19: a comprehensive evidence synthesis and meta-analysis. J Matern Fetal Neonatal Med. 2021;18:1–14.
7. Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, et al. Data resource profile : Clinical Practice Research Datalink (CPRD) Aurum. 2019;1–8.
8. 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 [Internet]. 2019;(April):1–11. Available from: http://doi.wiley.com/10.1002/pds.4811
9. NHS. Miscarriage Overview [Internet]. 2021. Available from: https://www.nhs.uk/conditions/miscarriage/
10. 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 [Internet]. 2022;12:55773. Available from: http://dx.doi.org/10.1136/bmjopen-2021-055773
11. Hernan MA. Paraninfo Digital. Epidemiology [Internet]. 2013;21(1):13–5. Available from: http://dx.doi.org/10.1016/j.earlhumdev.2015.09.003%5Cnhttp://dx.doi.org/10.1016/j.earlhumdev.2014.01.002%5Cnhttp://dx.doi.org/10.1016/S0378-3782(12)70006-3%5Cnhttp://www.sciencedirect.com/science/article/pii/S2341287914000763%5Cnhttp://dx.doi.org/10.1016/

List of Appendices

1. Covid 19 codelist CPRD Aurum
2. Covid codelist ICD 10
3. ICD codelist Miscarriage stillbirth

Grant ID *(optional)*