Introduction to
The Clinical Practice Research Datalink (CPRD)

How to use CPRD for research

Ania Zylbersztejn
Linda Wijlaars
Arturo González-Izquierdo

NCRM Research Methods e-Festival - November 2023

Learning objectives

- A few examples
- How to design your study using CPRD
  things to consider for your data application
- Strengths and limitations of CPRD

Chronological map of human disease

THE LANCET
Digital Health

A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service

Valerie Kuan, Spiros Denaxas, Arturo Gonzalez-Izquierdo, Kenan Direk, Osman Bhatti, Shanaz Husain, Shailen Sutaria, Melanie Hingorani, Dorothea Nitsch, Constantinos A Parisinos, R Thomas Lumbers, Rohini Mathur, Reecha Sofat, Juan P Casas, Ian C K Wong, Harry Hemingway, Aroon D Hingorani

Summary

Background To effectively prevent, detect, and treat health conditions that affect people during their lifecourse, health-care professionals and researchers need to know which sections of the population are susceptible to which health conditions and at which ages. Hence, we aimed to map the course of human health by identifying the 50 most common health conditions in each decade of life and estimating the median age at first diagnosis.

Methods We developed phenotyping algorithms and codelists for physical and mental health conditions that involve intensive use of health-care resources. Individuals older than 1 year were included in the study if their primary-care and hospital-admission records met research standards set by the Clinical Practice Research Datalink and they had been registered in a general practice in England contributing up-to-standard data for at least 1 year during the study period. We used linked records of individuals from the CALIBER platform to calculate the sex-standardised cumulative incidence for these conditions by 10-year age groups between April 1, 2010, and March 31, 2015. We also derived the median age at diagnosis and prevalence estimates stratified by age, sex, and ethnicity (black, white, south Asian) over the study period from the primary-care and secondary-care records of patients.

Findings We developed case definitions for 308 disease phenotypes. We used records of 2784138 patients for the calculation of cumulative incidence and of 3872451 patients for the calculation of period prevalence and median age at diagnosis of these conditions. Conditions that first gained prominence at key stages of life were: atopic conditions and infections that led to hospital admission in children (<10 years); acne and menstrual disorders in the teenage years (10–19 years); mental health conditions, obesity, and migraine in individuals aged 20–29 years; soft-tissue disorders and gastro-oesophageal reflux disease in individuals aged 30–39 years; dyslipidaemia, hypertension, and erectile dysfunction in individuals aged 40–59 years; cancer, osteoarthritis, benign prostatic hyperplasia, cataract, diverticular disease, type 2 diabetes, and deafness in individuals aged 60–79 years; and atrial fibrillation, dementia, acute and chronic kidney disease, heart failure, ischaemic heart disease, anaemia, and osteoporosis in individuals aged 80 years or older. Black or

Lancet Digital Health 2019; 1: e63–77

Published Online May 20, 2019 http://dx.doi.org/10.1016/S2589-7500(19)30012-3

See Comment page e46

Institute of Cardiovascular Science (V Kuan MBBS, Prof A D Hingorani PhD), **Health Data Research UK London,** (V Kuan, S Denaxas PhD, A Gonzalez-Izquierdo PhD, K Direk PhD, R T Lumbers PhD, R Sofat PhD, Prof H Hemingway FFPH FRCP, Prof A D Hingorani), **Institute of Health Informatics** (S Denaxas, A Gonzalez-Izquierdo, K Direk, C A Parisinos MRCP, R T Lumbers, R Sofat, Prof J P Casas PhD, Prof H Hemingway), **and School of Pharmacy** (Prof I C K Wong PhD), **University College London, London, UK; Alan Turing Institute, London, UK** (S Denaxas); **Chrisp Street**

V. Kuan et al. The Lancet Digital Health 2019; https://doi.org/10.1016/S2589-7500(19)30012-3

Research objective: to create a chronological map of human health

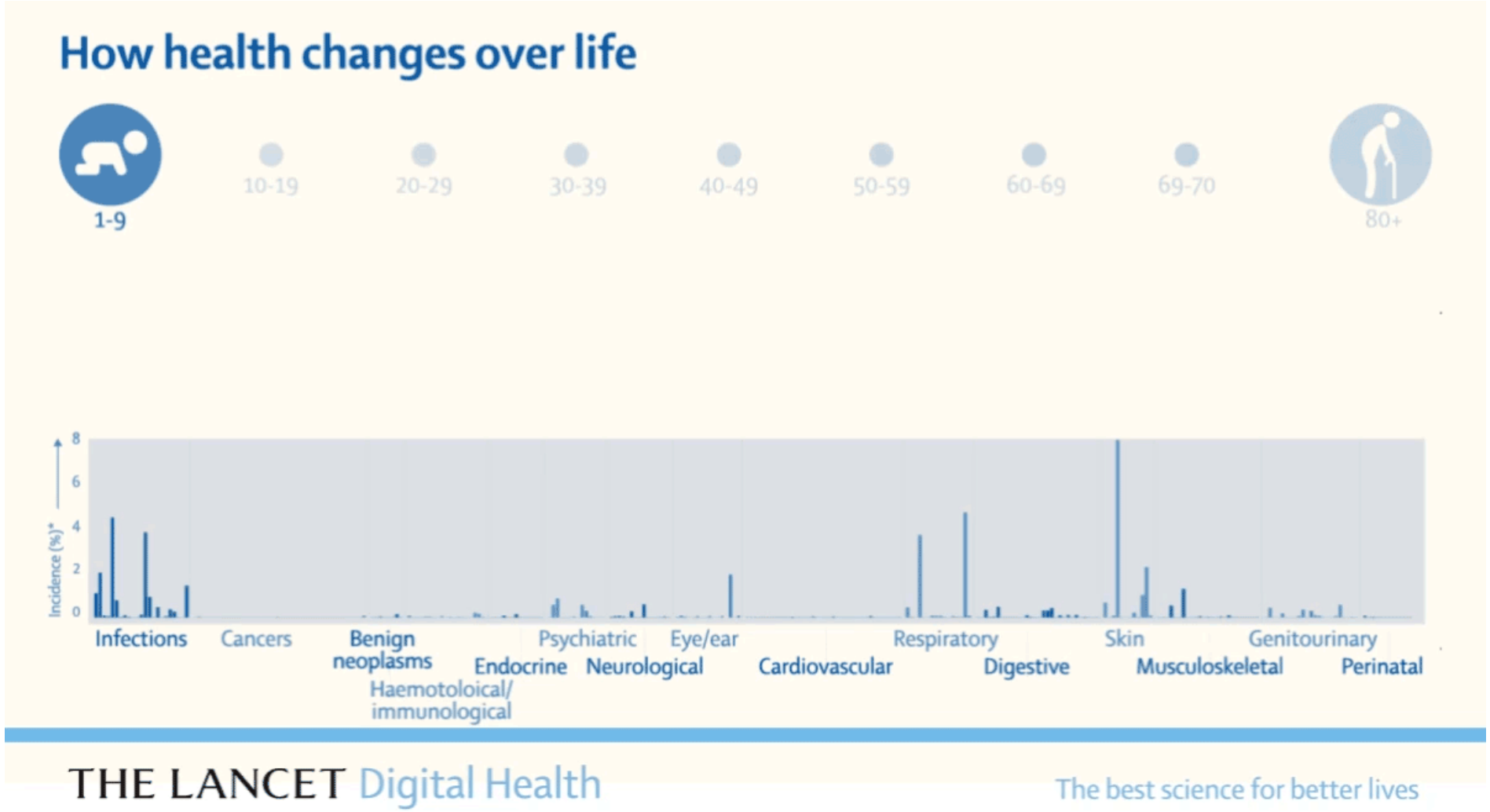
Data used: CPRD linked to HES

Study type: Cohort study (descriptive)

Study population:
- Individuals aged >1 year old
- Min 1 year follow-up from April 1, 2010, to March 31, 2015
- Registered with contributing GP practice in England (with up-to-standard data)

Outcome: diagnosis of physical and mental health conditions

Chronological map of human disease



Study population: 2,784,138 patients at the start of the study (April 1, 2010)

for calculation of cumulative incidence

V. Kuan et al. The Lancet Digital Health 2019; https://doi.org/10.1016/S2589-7500(19)30012-3

Long-term outcomes

Circulation. 2019 Sep 24;140(13):1050-1060. doi: 10.1161/CIRCULATIONAHA.118.038080. Epub 2019 Sep 23.

Preeclampsia and Cardiovascular Disease in a Large UK Pregnancy Cohort of Linked Electronic Health Records: A CALIBER Study

Lydia J Leon [1 2], Fergus P McCarthy [1 3], Kenan Direk [2], Arturo Gonzalez-Izquierdo [2], David Prieto-Merino [2 4], Juan P Casas [5], Lucy Chappell [1]

Affiliations + expand

PMID: 31545680 DOI: 10.1161/CIRCULATIONAHA.118.038080

Abstract

Background: The associations between pregnancy hypertensive disorders and common cardiovascular disorders have not been investigated at scale in a contemporaneous population. We aimed to investigate the association between preeclampsia, hypertensive disorders of pregnancy, and subsequent diagnosis of 12 different cardiovascular disorders.

Research objective: to examine the association between preeclampsia & hypertensive disorders of pregnancy, and risk of cardiovascular disorders

Data used: CPRD linked to HES

Study type: Cohort study (hypothesis-testing)

Study population:
- Women who were pregnant between 1997-2016 aged 11-49 years old
- Consented to linkage to HES → English GP practices

Outcome: diagnosis of 12 cardiovascular disorders *(recorded between first completed pregnancy record & 31st Dec 2016, death, first cardiovascular disorder)*

Exposure: diagnosis of preeclampsia or hypertensive disorders in primary or secondary healthcare
Comparison: unaffected pregnancies

L. Leon et al. Circulation. 2019. DOI: 10.1161/CIRCULATIONAHA.118.038080

Long-term outcomes

Circulation. 2019 Sep 24;140(13):1050-1060. doi: 10.1161/CIRCULATIONAHA.118.038080. Epub 2019 Sep 23.

Preeclampsia and Cardiovascular Disease in a Large UK Pregnancy Cohort of Linked Electronic Health Records: A CALIBER Study

Lydia J Leon [1 2], Fergus P McCarthy [1 3], Kenan Direk [2], Arturo Gonzalez-Izquierdo [2], David Prieto-Merino [2 4], Juan P Casas [5], Lucy Chappell [1]

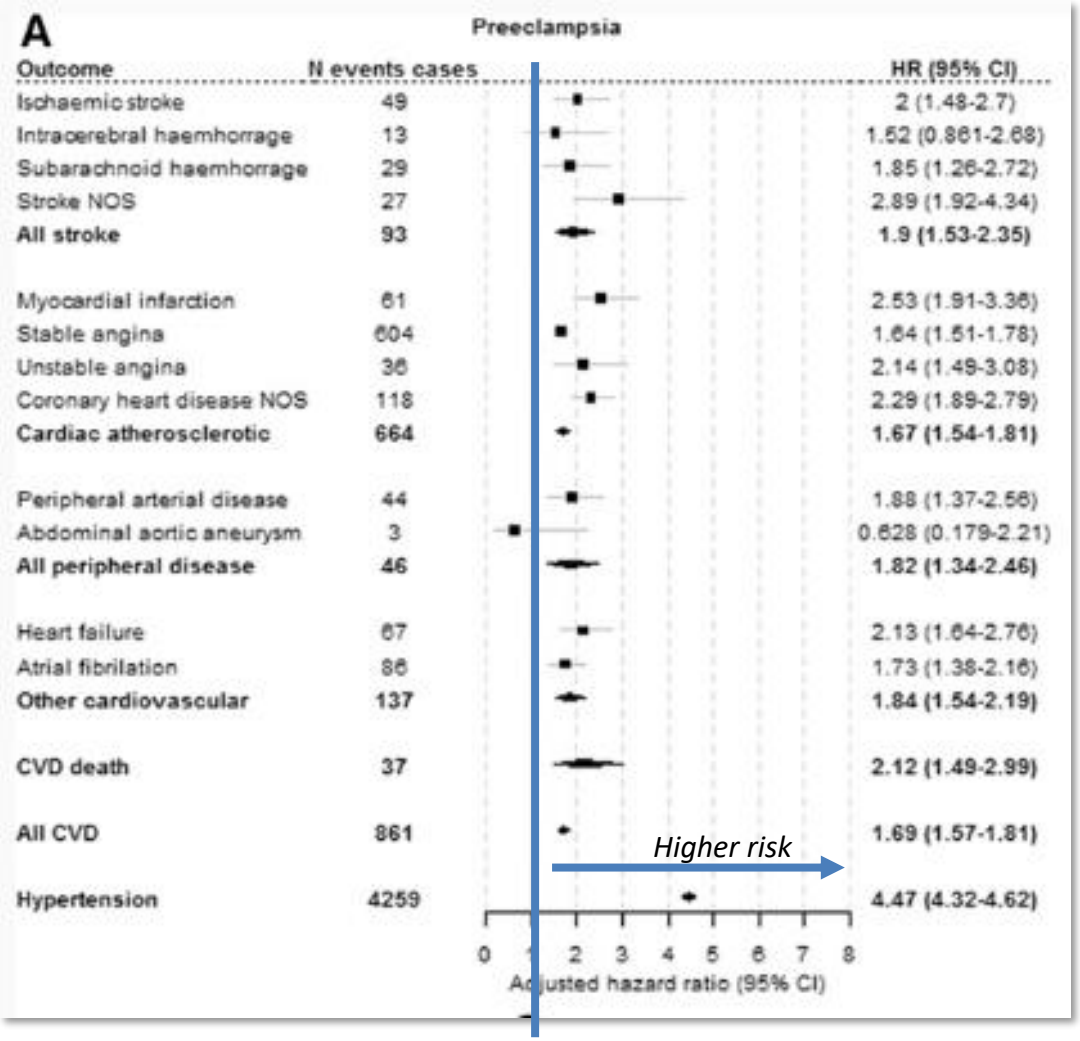
Affiliations + expand

PMID: 31545680 DOI: 10.1161/CIRCULATIONAHA.118.038080

Abstract

Background: The associations between pregnancy hypertensive disorders and common cardiovascular disorders have not been investigated at scale in a contemporaneous population. We aimed to investigate the association between preeclampsia, hypertensive disorders of pregnancy, and subsequent diagnosis of 12 different cardiovascular disorders.

Women with any hypertensive disorders of pregnancy are at increased risk of all cardiovascular disorders



No difference

L. Leon et al. Circulation. 2019. DOI: 10.1161/CIRCULATIONAHA.118.038080

Real-world evidence

thebmj covid-19 Research Education News & Views Campaigns Jobs

Research

Safety of pertussis vaccination in pregnant women in UK: observational study

BMJ 2014 ; 349 doi: https://doi.org/10.1136/bmj.g4219 (Published 11 July 2014)
Cite this as: *BMJ* 2014;349:g4219

Article | Related content | Metrics | Responses | Peer review

Katherine Donegan, pharmacoepidemiologist, Bridget King, scientific assessor, Phil Bryan, scientific assessor

Author affiliations

Correspondence to: K Donegan katherine.donegan@mhra.gsi.gov.uk

Accepted 18

Abstra

Objectiv

Design

Setting

Participa nd a matched historical unvaccinated control group.

Findings: There is no evidence of an increased risk of adverse pregnancy outcomes in women vaccinated in 3rd trimester

Research objective: to examine the safety of pertussis vaccination in pregnancy
(Vaccination programme started began on 1 October 2012)

- **Study population**: Pregnant women
- **Exposure:** a record of pertussis vaccination during pregnancy after 1 October 2012
- **Comparison:** historical cohort of unvaccinated women (giving birth before 1 October 2012)

Study type: Matched cohort study

Outcome: adverse pregnancy outcomes (including stillbirth, maternal or neonatal mortality)

Donegan et al. BMJ. 2014. doi: https://doi.org/10.1136/bmj.g4219

Pharmacoepidemiology

thebmj

PEN ACCESS

for updates

Associations between macrolide antibiotics prescribing during pregnancy and adverse child outcomes in the UK: population based cohort study

Heng Fan,[1] Ruth Gilbert,[1] Finbar O'Callaghan,[2] Leah Li[1]

olicy and Practice
eat Ormond
of Child Health,
ge London,
1EH, UK
al Neurosciences
eat Ormond
of Child Health,
ge London,

ce to:Heng Fan
cl.ac.uk
002-4203-2516)

erial is published
view please visit
ne.
/ 2020;368:m331
10.1136/bmj.m331
nuary 2020

ABSTRACT

OBJECTIVE

To assess the association between macrolide antibiotics prescribing

prescribed macrolides and 1666 of 95 973 children (17.36 per 1000) whose mothers were prescribed penicillins during pregnancy. Macrolide prescribing

sociated with an
ation compared
0, adjusted
l 1.19 to 2.03)
nations (10.60
macrolide
ated with an
(4.75 *v* 3.07
ypospadias).
associated with
ation (27.39 *v*
o statistically
other system
elopmental
sitivity

ng the first
ed with an
ormation and

Findings: Macrolide antibiotics prescribing during the 1st trimester of pregnancy was associated with an increased risk of any major birth defect

macrolides should be used with caution during pregnancy

Fan et al. BMJ. 2020 doi: https://doi.org/10.1136/bmj.m331

Research question: Is macrolide antibiotics prescribing during pregnancy associated with adverse outcomes in fetuses and children?

Data used: CPRD (mother-baby link)

Study type: Cohort study (hypothesis-testing)
Emulating a clinical trial

Study population: 104,605 children born from 1990 to 2016 mother was prescribed:

- Macrolides (**exposure)**
- Penicillin (**comparison**)

after 4th week of pregnancy

Outcome: major birth defect or neurodevelopmental condition

How to use CPRD for research?

Tips on how to design your study

How to use CPRD for research?

Your study question

Can it be answered using primary care data?

	Research question definition	Can it be defined in admin data?
Population		
Intervention / exposure		
Comparison		
Outcome		

How to use CPRD for research?

Your study question

to examine the association between preeclampsia & hypertensive disorders of pregnancy, and risk of cardiovascular disorders

Can it be answered using primary care data?

	Research question definition	Can it be defined in admin data?
Population	Pregnant women	
Intervention / exposure	Preeclampsia & hypertensive disorders of pregnancy	
Comparison	Unaffected pregnant women	
Outcome	Cardiovascular disorder (after completed pregnancy)	

How to use CPRD for research?

Your study question

to examine the association between preeclampsia & hypertensive disorders of pregnancy, and risk of cardiovascular disorders

Can it be answered using primary care data?

	Research question definition	Can it be defined in admin data?
Population	Pregnant women	Yes – pregnant women frequently interact with healthcare *CPRD pregnancy register & HES delivery records*
Intervention / exposure	Preeclampsia & hypertensive disorders of pregnancy	Yes – likely to be recorded in healthcare records
Comparison	Unaffected pregnant women	Yes – likely to be recorded in healthcare records
Outcome	Cardiovascular disorder (after completed pregnancy)	Yes – likely to be recorded in healthcare records

How to use CPRD for research?

Study Design:

- Cohort (birth cohort, open cohort);
- Cross-sectional design?
- Case-control?

Population: eligibility criteria

- How will eligible patients be identified?
  - E.g.: based on patient characteristics (age, sex, presence of specific health condition)
- Minimum length of follow-up required? 1 year?
- Linked data required?
  - e.g.: mother-baby cohort, linked HES

How to use CPRD for research?

Study Design:

- Cohort (birth cohort, open cohort);
- Cross-sectional design?
- Case-control?

Population: eligibility criteria

- How will eligible patients be identified?
  - E.g.: based on patient characteristics (age, sex, presence of specific health condition)
- Minimum length of follow-up required? 1 year?
- Linked data required?
  - e.g.: mother-baby cohort, linked HES

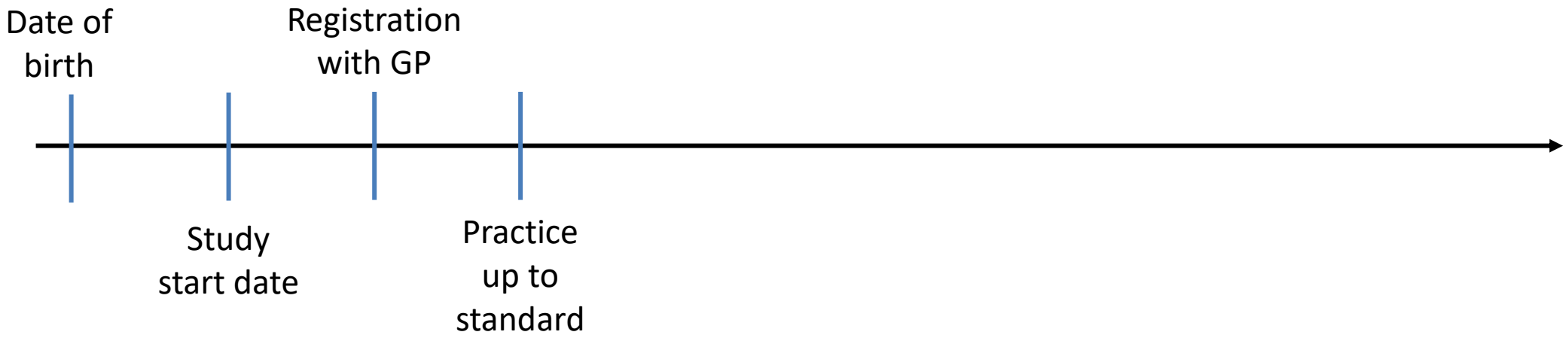
The association between preeclampsia and risk of cardiovascular disorders

Study population:

- Women
- Pregnant between 1997-2016
- Aged 11-49 years old
- Consented to linkage to HES → English GP practices

How to use CPRD for research?

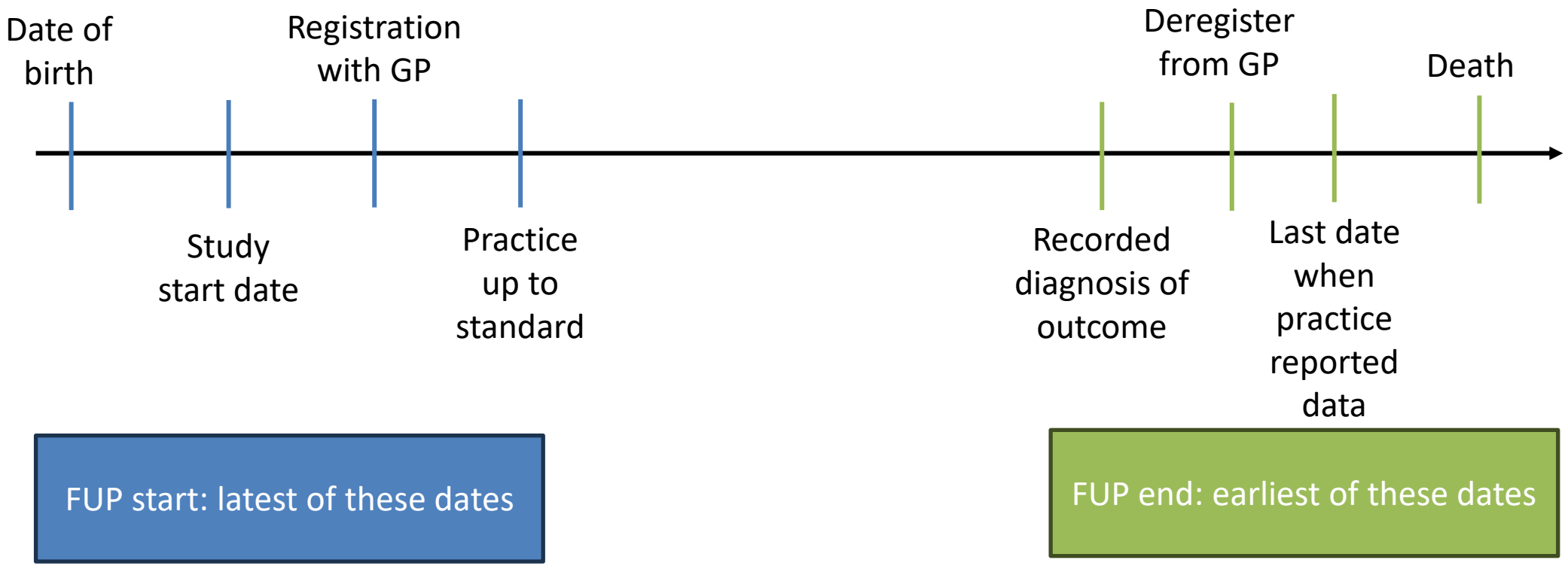
Definition of start & end of follow-up



FUP start: latest of these dates

How to use CPRD for research?

Definition of start & end of follow-up



How to use CPRD for research?

Defining exposures, outcomes and covariates:

- Which datasets?
- Which time periods?
- Are there code lists or phenotyping algorithms available?
- Comparison group

The association between preeclampsia and risk of cardiovascular disorders

Exposure: pre-exlampsia

- prespecified list of Read or ICD10 Codes
- preeclampsia code recorded within 20 weeks on either side of a pregnancy end date

How to use CPRD for research?

Common challenges to consider:

- Missing data: *selective missingness?*
  - *E.g.: BMI more likely to be recorded for individuals with specific diseases?*
- Comparison group
  - *patients who have more contact with the medical system have more opportunities to receive diagnoses?*
- Confounding
  - Is required data available in CPRD?
  - Address unmeasured confounding?

Target population-based study

Protocol Component	Target population-based study specification
Eligibility criteria	
Study design	
Exposure definition	
Follow-up period	
Outcome	
Target of estimation	
Analysis plan	

With thanks to Dr Kate Lewis

Target population-based study

Protocol Component	Target population-based study specification	Emulation study using administrative records	Sources of bias & mitigation strategies
Eligibility criteria			
Study design			
Exposure definition			
Follow-up period			
Outcome			
Target of estimation			
Analysis plan			

Hernán et al. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016. doi: 10.1093/aje/kwv254

With thanks to Dr Kate Lewis

RECORD

REporting of studies Conducted using Observational Routinely-collected Data

- Reporting guidelines for studies using routinely collected data
- Extension of STROBE statement
  - RECORD-PE – for pharmacoepidemiology
- Reporting checklist required by many journals

http://record-statement.org/

The RECORD statement – checklist of items, extended from the STROBE statement, that should be reported in observational studies using routinely collected health data.

Methods					
Study Design	4	Present key elements of study design early in the paper			
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection			
Participants	6	*(a) Cohort study* - Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study* - Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study* - Give the eligibility criteria, and the sources and methods of selection of participants *(b) Cohort study* - For matched studies, give matching criteria and number of exposed and unexposed *Case-control study* - For matched studies, give matching criteria and the number of controls per case		RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided. RECORD 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided. RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage.	
Variables	7	Clearly define all outcomes		RECORD 7.1: A complete list of codes	

How to use CPRD for research?

Strengths and limitations of CPRD

Strengths of CPRD data

Large sample size

- CPRD Aurum, September 2023: https://doi.org/10.48329/6j2c-nh78
  - 45.1mln patients (35.2 mln eligible for linkage),
  - **15.6 mln current patients (23% of UK population),**
  - median FUP: 5 years (2-13 years)
- CPRD GOLD, October 2023 version: https://doi.org/10.48329/czpn-2s41
  - 21.4mln patients (9.3 mln eligible for linkage),
  - **nearly 3 mln current patients (4.5% of UK population),**
  - median FUP: 5.6 years (2-13.5 years)

Long-term follow-up

Ongoing data collection (monthly updates)

Unselected population

Rare exposures and outcomes

Representative of the UK population

Strengths of CPRD data

Rich data collection: GPs are the first port of call for care, providing a range of primary care services

Real world data:

- insight into disease epidemiology, population health, treatment, and clinical pathways
- Opportunities for target trial emulation where Randomised Control Trials might not be feasible / ethical

Data linkages

- national secondary healthcare databases,
- the national cancer registry,
- death registrations
- deprivation measures
- Derived datasets – linking mothers and babies; pregnancy register

Challenges of using CPRD data

Data not collected for research:

- Administration, not research data
- Relevant information might not be collected
- **Missing data:** some health information such as smoking status, BMI, or ethnicity data may only be recorded when this is relevant to the patient's health condition

What is <u>not</u> collected:

- Complete data on primary care **prescriptions** for medications and devices available, but not: dispensed medications, secondary care prescriptions and over-the-counter use
- Hospital discharge letters are not coded separately → linkage to HES

Challenges of using CPRD data

No standardised definitions:

- Researchers derive & validate phenotypes (Code lists or algorithms)
- New phenotypes require understanding how conditions of interest is treated in the UK (primary / secondary health care) & whether CPRD appropriate source

Changes in coding/ GP IT systems over time:

- Possible variations in coding between practices and over time
- Multiple coding systems used, e.g. Read codes and SNOMED for recording medical observations in CPRD; ICD-10 for diagnoses in HES
- Impact of QOF – increased coding depth, caution when examining time trends
- Differences between CPRD Gold and Aurum (due to differences in between EMIS Web® and Vision® software)