NCRM Research Methods e-Festival

Introduction to Using
The Clinical Practice Research Datalink (CPRD)

Introduction to Using

The Clinical Practice Research Datalink (CPRD)

An overview of available data, their structure and how they are collected

Linda Wijlaars
Arturo González-Izquierdo
Ania Zylbersztejn

NCRM Research Methods e-Festival - November 2023

Overview

Why / how is data collected?

What does the data look like?

What do I need to work with CPRD?

Why / how is data collected?

Primary care database

- Data collection by network of UK general practitioners
- Practices using Vision or EMIS practice management software
- CPRD collates into anonymised database

CPRD services not covered in this introduction course:

- patient recruitment for clinical trials (CPRD SPRINT)
- verification of coded records (CPRD PROVE)
- feasibility counts / studies
- synthetic data

Primary care databases

VAMP Medical record Management system

Free computer + software in exchange for sharing anonymised data

Linkages to hospital and other datasets

2004: QOF introduced

2017: CPRD Aurum released

1970s 1980s 1990s 2000s 2010s

5 % of GPs use computers

50 % of GPs use computers

Clinical Practice Research Datalink

Primary care database

- from ~1990 onwards
- From 2,200+ UK GP practices
- 60+ million patients

2 datasets (Gold and Aurum), 1 provider

Data Resource Profile

- CPRD gold: Herrett et al. IJE (2015)
- CPRD aurum: World et al. IJE (2019)

cprd.com

Medicines & Healthcare products Regulatory Agency

CPRD | UK data driving real-world evidence

Home ▼ News Public ▼ Data ▼ Our services ▼ General Practitioner ▼ Research ▼ Resource modules ▼

Clinical Practice Research Datalink

Clinical Practice Research Datalink (CPRD) is a real-world research service supporting retrospective and prospective public health and clinical studies. CPRD research data services are delivered by the Medicines and Healthcare products Regulatory Agency with support from the National Institute for Health and Care Research (NIHR), as part of the Department of Health and Social Care.

CPRD collects anonymised patient data from a network of GP practices across the UK. Primary care data are linked to a range of other health related data to provide a longitudinal, representative UK population health dataset. The data encompass 60 million patients, including 18 million currently registered patients.

For more than 30 years, research using CPRD data and services has informed clinical guidance and best practice, resulting in over 3,000 peer-reviewed publications investigating drug safety, use of medicines, effectiveness of health policy, health care delivery and disease risk factors.

Search

Search

GP practices - Join today

Researcher log in

Primary care data

- Most people registered with GP
  (exception: prisoners, Armed Forces, private patients, homeless persons)
- GPs opt in to contribute to CPRD, patients can use national data opt out

Data entered by GPs

- No national coding standards
- Coded data entered
- Affected by factors such as QOF (Quality Outcomes Framework)
- Specific conditions where recording was encouraged

Dynamic patient cohort

- Practices and patients join and leave at different times
- Patients can be linked over time within 1 practice

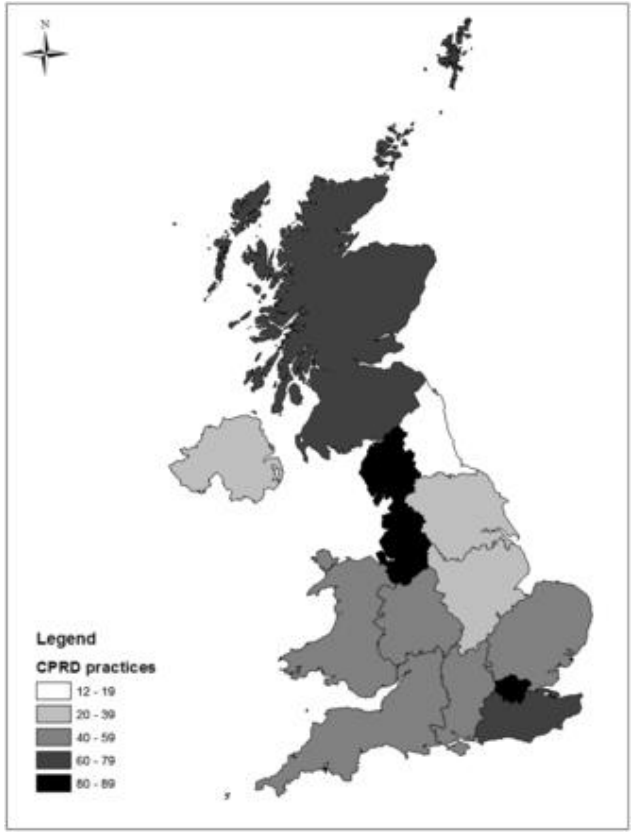
CPRD Gold and Aurum

Since 2017: two versions of CPRD:

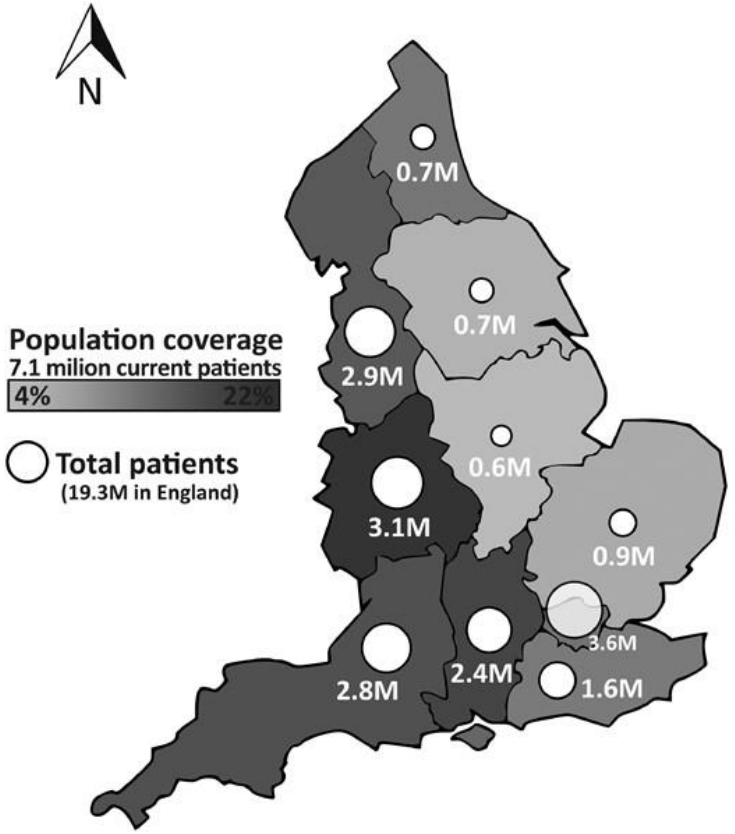
- Gold: original CPRD -> based on practices using Vision software
- Aurum: ‘new’ CPRD -> based on practices using EMIS software

CPRD coverage

CPRD Gold

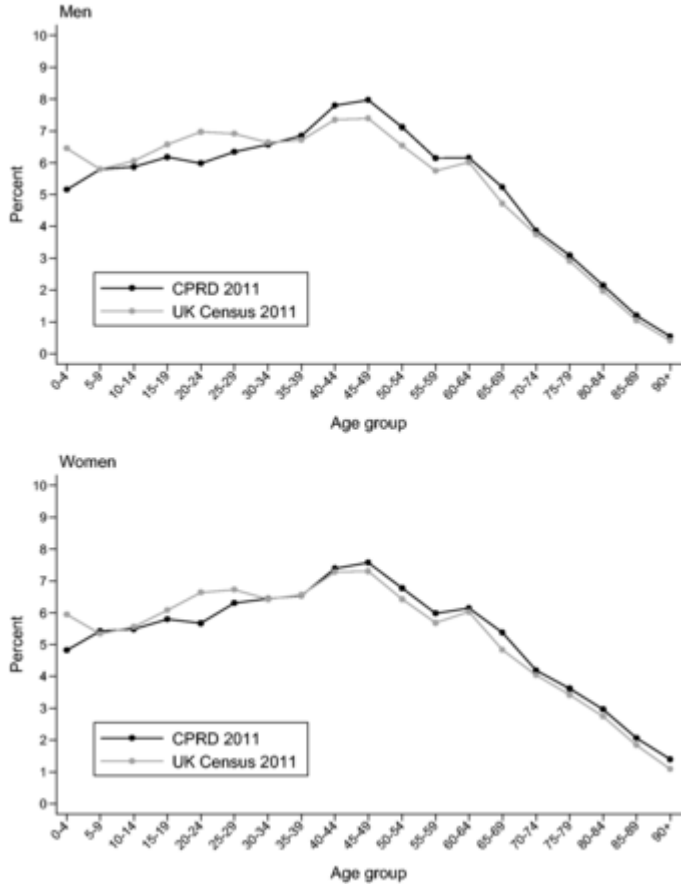


CPRD Aurum

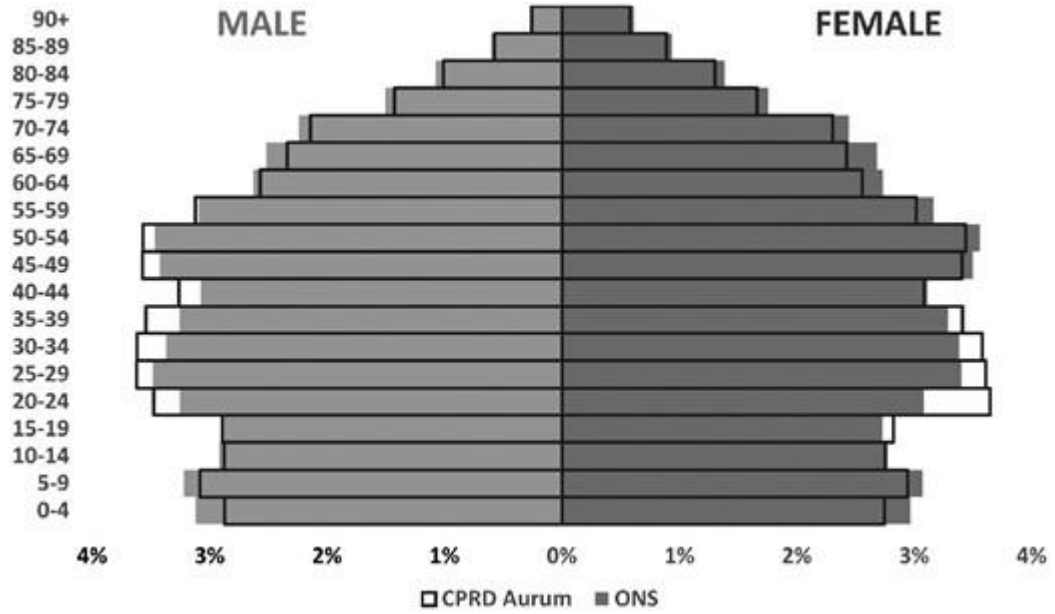


CPRD coverage

CPRD Gold

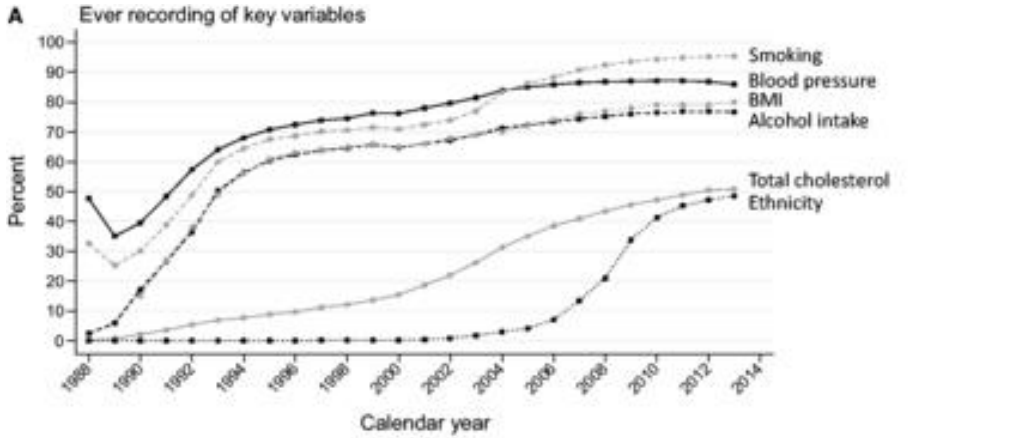


CPRD Aurum

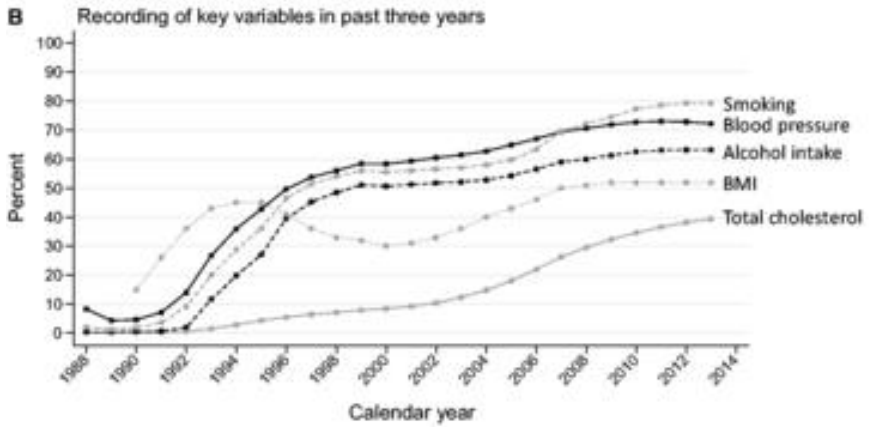


CPRD coverage

CPRD Gold



Note: Based on a random sample of one million acceptable patients. Adults aged 18+ were included. Records outside of UTS were included. Denominators are mid-year registered populations. Total cholesterol has poorer completeness as this would not be routinely recorded and would require a clinical indication. Completeness of ethnicity recording for new registrants after 2004 approached 70% in 2011.(31)



Note: based on a random sample of one million acceptable patients. Adults aged 18+ were included. Records outside of UTS were included. Denominators are mid-year registered populations. Total cholesterol has poorer completeness as this would not be routinely recorded and would require a clinical indication.

Other Primary Care Datasets

QRESEARCH

THIN
THE HEALTH IMPROVEMENT NETWORK

Research One
Transforming data into knowledge

OPEN SAFELY

Primary care database paper topics

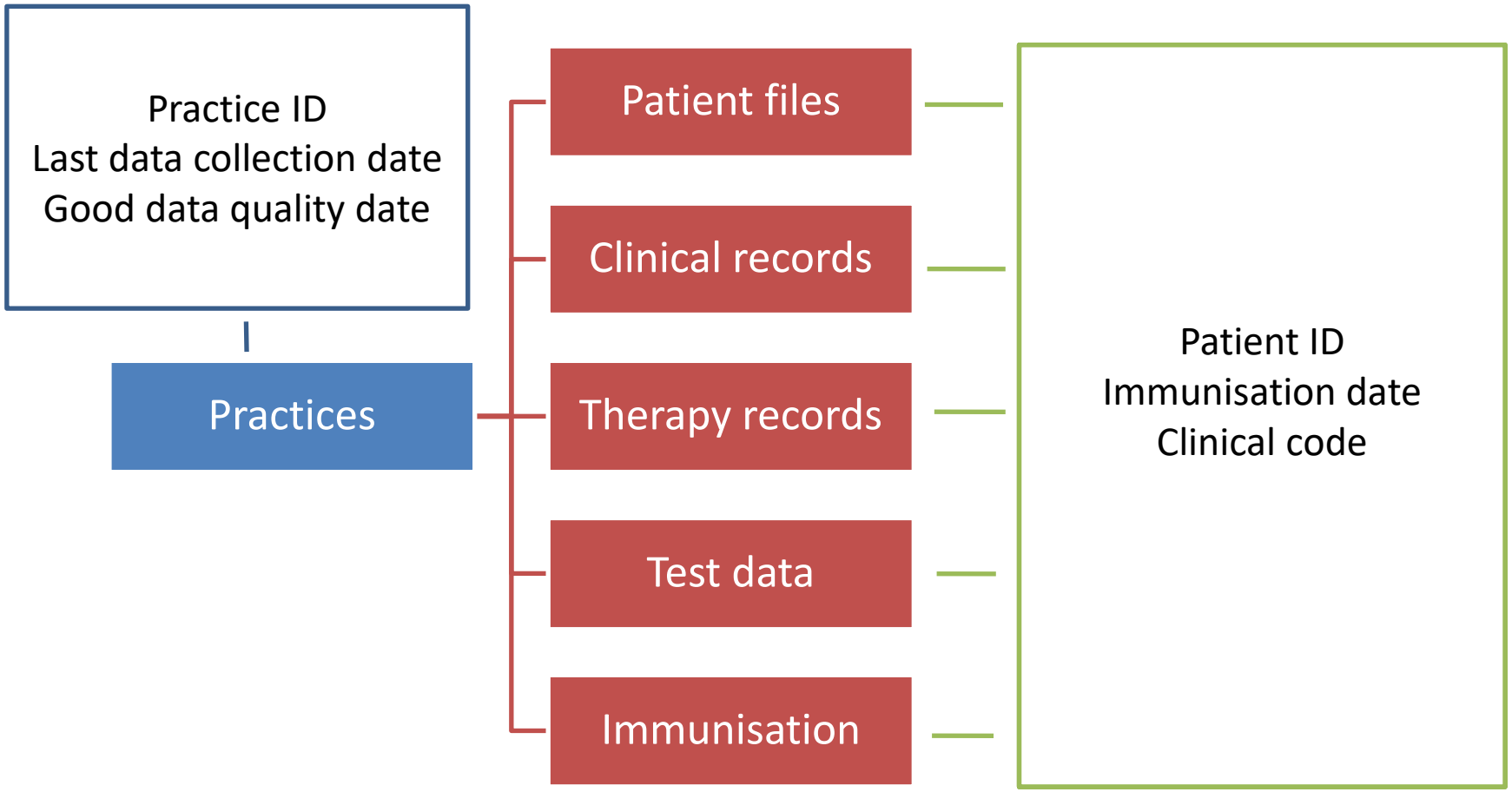
Table 11 Top keywords: medical conditions

Rank	Keyword	Occurrences	Rank	Keyword	Occurrences
1	Smoking	328	16	Cardiovascular diseases	96
2	Diabetes mellitus	223	17	Myocardial infarction	94
3	Hypertension	223	18	Chronic obstructive lung disease	90
4	Non-insulin dependent diabetes mellitus	179	19	Heart failure	86
5	Depression	167	20	Cerebrovascular accident	85
6	Stroke	165	21	Rheumatoid arthritis	83
7	Asthma	158	22	Epilepsy	81
8	Diabetes mellitus, type 2	155	23	Breast cancer	78
9	Cancer risk	150	24	Fracture	75
10	Cardiovascular risk	147	25	Psoriasis	75
11	Cardiovascular disease	133	26	Gastrointestinal haemorrhage	69
12	Obesity	129	27	Hip fracture	68
13	Heart infarction	126	28	Osteoporosis	68
14	Pregnancy	125	29	Colorectal cancer	65
15	Ischaemic heart disease	104	30	Fractures, bone	65

Table 12 Top keywords: medications/substances

Rank	Keyword	Occurrences	Rank	Keyword	Occurrences
1	Non-steroid anti-inflammatory agent	182	16	Proton pump inhibitor	75
2	Acetylsalicylic acid	154	17	Warfarin	71
3	Metformin	150	18	Antidiabetic agent	69
4	Corticosteroid	143	19	Anticonvulsive agent	68
5	Hydroxymethylglutaryl coenzyme a reductase inhibitor	138	20	Serotonin uptake inhibitor	65
6	Insulin	133	21	Calcium channel blocking agent	64
7	Antidepressant agent	124	22	Antibacterial agents	62
8	β adrenergic receptor blocking agent	108	23	Hydroxymethylglutaryl-coA reductase inhibitors	62
9	Hypoglycemic agents	91	24	Oral antidiabetic agent	61
10	Anti-inflammatory agents, non-steroidal	90	25	Paracetamol	56
11	Dipeptidyl carboxypeptidase inhibitor	88	26	Diuretic agent	53
12	Antihypertensive agent	82	27	Ibuprofen	52
13	Neuroleptic agent	80	28	Simvastatin	52
14	Antibiotic agent	77	29	Tricyclic antidepressant agent	52
15	Hemoglobin A1c	75	30	Diclofenac	50

What does the data look like?



Patient file

patid	gender	yob	mob	marital	famnum	frd	crd	regstat	reggap	internal	tod	toreason	deathdate	accept
112001	1	1967	.	9	4578	01/01/2000	01/01/2000	0	0	0	31/01/2020	1	31/01/2020	1
212001	1	1999	.	1	12658	06/12/1999	06/12/1999	0	0	0	.	0	.	1
312001	2	2002	.	1	1478	07/05/2005	07/05/2005	0	0	0	.	0	.	1
412002	1	1991	.	2	9512	08/01/2001	08/01/2001	0	0	0	11/10/2007	2	.	1
512002	2	2017	4	2	14785	09/12/2020	09/12/2020	0	0	0	.	0	.	1
612002	2	1958	.	11	6547	10/12/1999	10/12/1999	0	0	0	17/04/2019	16	.	1

patid	pracid	gender	yob	mob	emis_ddate	regstartdate	patienttyp	regenddate	acceptable	cprd_ddate	uts	lcd	region
122001	22001	1	1967	.	31/01/2020	01/01/2000	1	31/01/2020	1	31/01/2020	.	29/03/2022	1
222001	22001	1	1999	.	.	06/12/1999	1	.	1	.	.	29/03/2022	1
322001	22001	2	2002	.	.	07/05/2005	1	.	1	.	.	29/03/2022	1
422002	22002	1	1991	.	.	08/01/2001	1	11/10/2007	1	.	.	01/03/2022	6
522002	22002	2	2017	4	.	09/12/2020	1	.	1	.	.	01/03/2022	6
622002	22002	2	1958	.	.	10/12/1999	1	17/04/2019	1	.	.	01/03/2022	6

Accounting for GPs getting used to computers

3 stages of practice software use:

- As patient register
- To issue prescriptions
- As full digital medical record

UTS (‘up to standard’) date

Clinical file

patientid	diagnosis_~e	read_code	description
276	11feb2014	Z4A..00	Discussion
276	21apr2014	1J4..00	Suspected UTI
276	15sep2014	J521.11	Irritable bowel syndrome
276	15sep2014	9N3A.00	Telephone triage encounter
276	15sep2014	ZL5..00	Referral to doctor
276	18oct2014	66R..00	Repeat prescription monitoring
276	18oct2014	2I19.00	Discomfort
276	15dec2014	1J4..00	Suspected UTI
276	15dec2014	6A...00	Patient reviewed
276	21dec2014	16C5.00	C/O - low back pain
276	02feb2015	16C5.00	C/O - low back pain
276	07feb2015	9N36.00	Letter from specialist
276	04may2015	K594.00	Irregular menstrual cycle
276	24may2015	9N36.00	Letter from specialist
276	25oct2015	K190.00	Urinary tract infection, site not specified
276	16feb2016	9ND..00	Incoming mail processing
276	14feb2017	K5A2z00	Menopausal symptoms NOS
276	25jul2017	K5A2z00	Menopausal symptoms NOS
276	25jul2017	8HP2.00	Refer for microbiological test
276	25jul2017	J16y400	Dyspepsia
276	25jul2017	4131.00	Blood test requested
276	25jul2017	R004200	[D]Light-headedness
276	12aug2017	1B53.00	Dizziness present
276	13aug2017	8HT3.00	Referral to audiology clinic
276	07oct2017	9E45.00	Life assurance report requested
276	04nov2017	9E45.00	Life assurance report requested
276	05dec2017	9N36.00	Letter from specialist
276	05mar2018	66U..11	Hormone replacement therapy
277	29jun2004	14A2.00	H/O: hypertension
277	07nov2012	1D14.00	C/O: a rash
277	07nov2012	G2...00	Hypertensive disease
277	07nov2012	9N79.00	New patient consultation
277	07nov2012	68R..00	New patient screen
277	07nov2012	G2...00	Hypertensive disease
277	11feb2013	G2...00	Hypertensive disease
277	13apr2013	9OW..00	New patient screen admin.

Therapy file

patientid	rx_date	drugcode	drug_description	bnf
134	13may2018	89567996	olanzapine	04020100
134	13may2018	94409998	sodium valproate	04080100
135	13dec2005	97300998	influenzainactive surf antign	14040900
136	17jan2013	97131998	amoxicillin	05010103
136	04apr2013	96999998	codeine phosphate	03090100
136	11jan2014	97235998	co-proxamol	04070100
136	21dec2014	89344997	diclofenac sodium	10010100
136	03jun2015	93638998	zopiclone	04010100
136	10jun2015	94447998	fluoxetine	04030300
136	24jun2015	93638998	zopiclone	04010100
136	08jul2015	93638997	zopiclone	04010100
136	08jul2015	94447998	fluoxetine	04030300
136	28feb2016	97131997	amoxicillin	05010103
137	24sep2017	97158998	chloramphenicol	11030100
137	02dec2017	99747998	oral rehydration salts	09020102
137	02dec2017	94571997	chloramphenicol	12010100
137	26dec2017	94794998	amoxicillin	05010103
137	19mar2018	94571997	chloramphenicol	12010100
137	11jun2018	94794998	amoxicillin	05010103
138	04mar2013	91190996	alendronate sodium	06060200
138	10mar2013	92677998	tramadol	04070200
138	10mar2013	96181998	hypromellose+ dextran	11080100
138	10mar2013	97201998	spironolactone	02020300
138	11mar2013	93638998	zopiclone	04010100
138	16mar2013	95417996	prednisolone	06030200
138	25mar2013	93638998	zopiclone	04010100
138	26mar2013	97998998	ferrous sulphate	09010101
138	26mar2013	92677998	tramadol	04070200

What do I need to work with CPRD?

Statistical / epidemiological skills

- ‘big data’, millions of records + complex data structure

Clinical input

- GP or specialist input to put data in context / understand why thing (aren’t) recorded

Time & funding

- Detailed data application necessary to get access to data

CPRD data applications

Data access subject to protocol approval by Research Data Governance (RDG) Process

- Review by Expert Review Committees and/or Central Advisory Committee
- Guidance available on cprd.com/research-applications

Features	BASIC*	STANDARD*	FULL
Price	£45,000	£82,500	£363,000

CPRD data add-ons / linkages

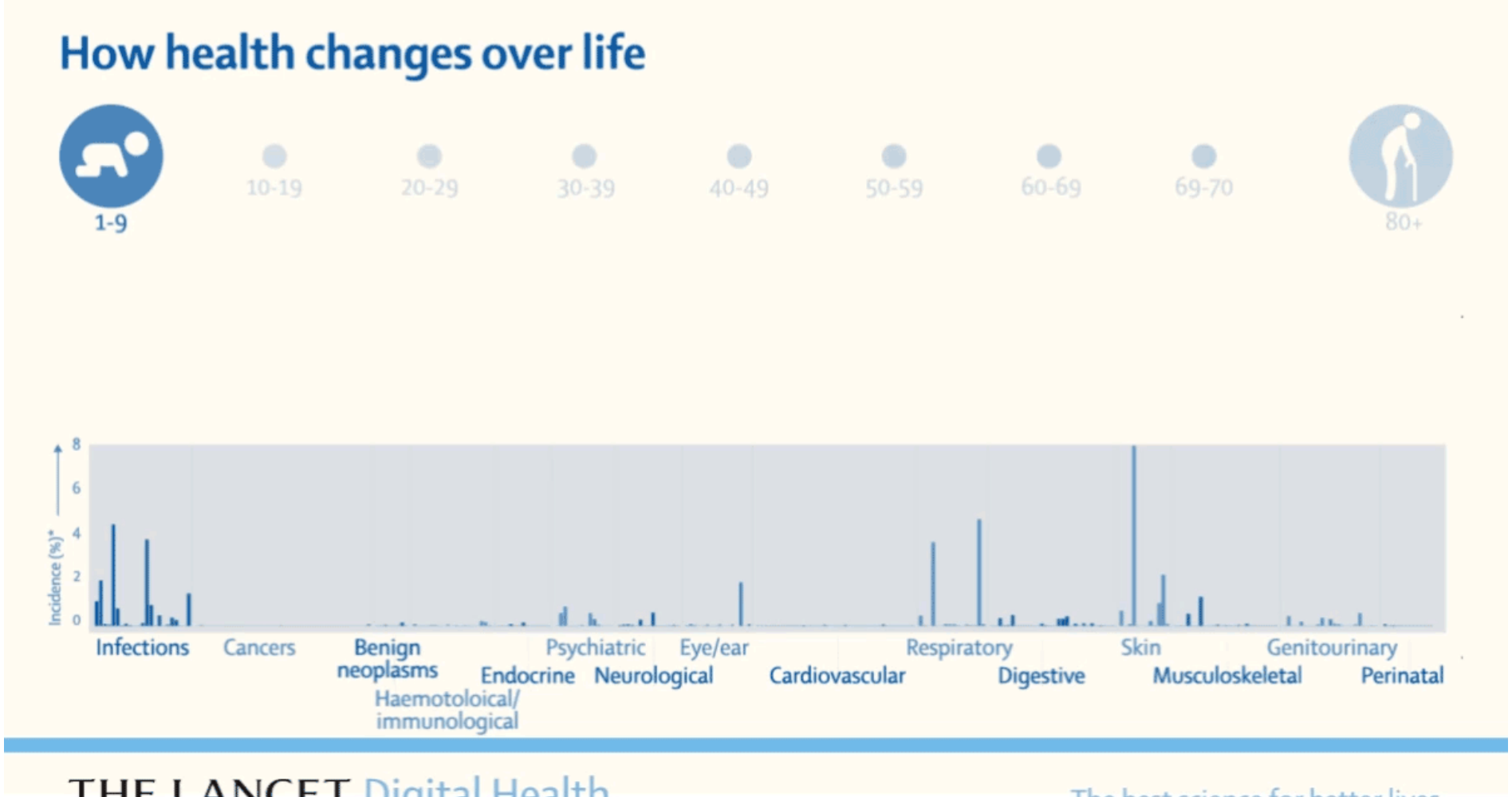
CPRD algorithm derived data:

- Mother-baby link
- Pregnancy Register
- Ethnicity Records

Linked data:

- COVID-19 data (lab tests, hospitalisations, intensive care)
- Small area level data (deprivation measures, rural-urban classification)
- Data from NHS England (hospital contacts, diagnostic imaging, death records)
- Cancer data from National Disease Registration Service (cancer registration, therapy, radiotherapy)

Chronological map of human disease



V. Kuan et al. The Lancet Digital Health 2019