CPRD: Real-world Data Source for Epidemiologic Research

Shahinaz Gadalla, MD, PhD

Clinical Genetics Branch

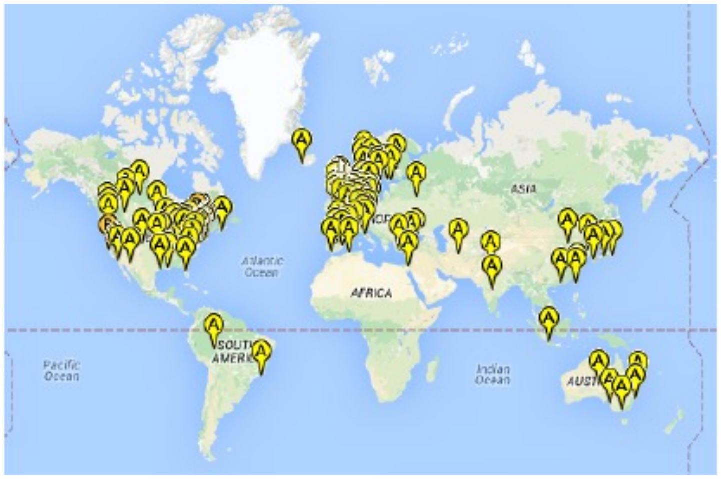
NIH NATIONAL CANCER INSTITUTE

Clinical Practice Research Datalink

CPRD

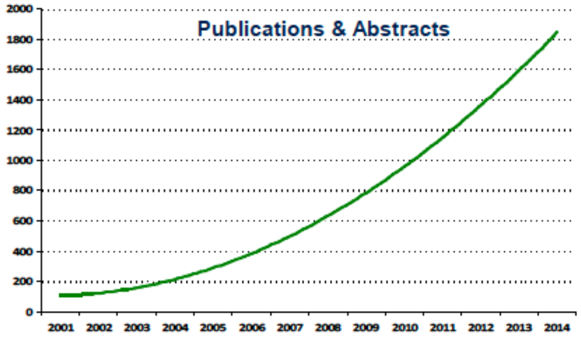
- UK governmental, not-for-profit research service organization
- Collects anonymized patient data from a network of GP practices across the UK
- The database includes 60 million patients (18 million currently registered)
- CPRD patient populations has similar age, sex distribution to the UK
- Death rates in CPRD population is similar to the national rates

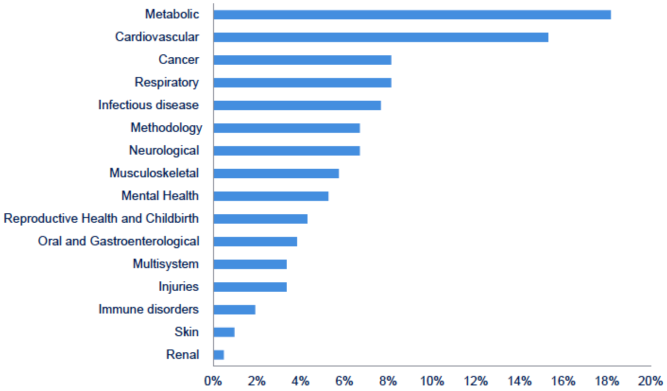
CPRD



Disease epidemiology
Drug safety
Drug use
Descriptive epidemiology
Care delivery
Infrastructure for prospective data collection & clinical trials

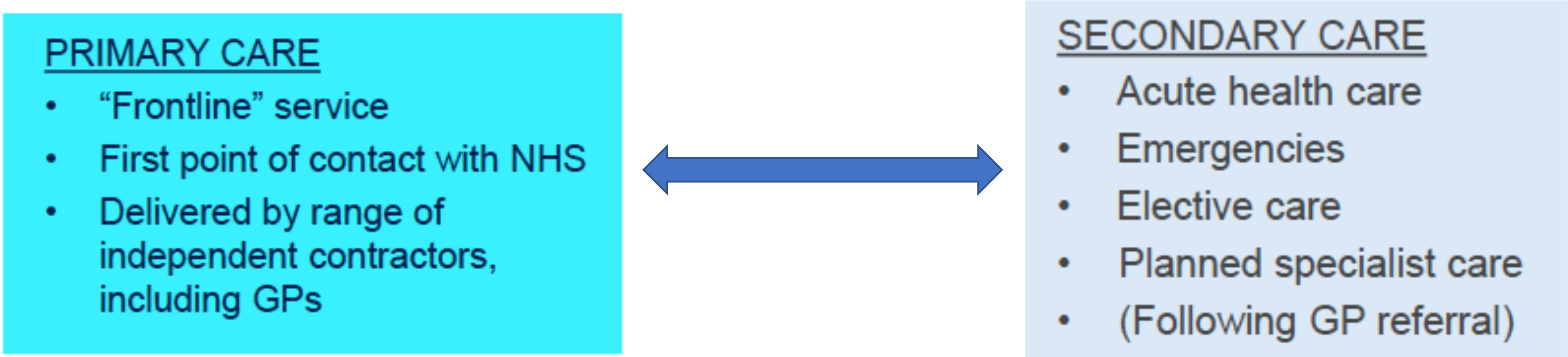
>3500 publications using CPRD
In-house expertise (>100 papers authored by CPRD staff)





CPRD bibliography disease and conditions (2016)

Primary Care Service in the UK



PRIMARY CARE

- “Frontline” service
- First point of contact with NHS
- Delivered by range of independent contractors, including GPs

SECONDARY CARE

- Acute health care
- Emergencies
- Elective care
- Planned specialist care
- (Following GP referral)

General Practitioners (GPs) in the UK

- Main point of contact (93% all healthcare consultations)
- Patients registered with 1 GP
- “Gatekeepers” - control access
- Lifetime medical record
- Wide range of care
  - – Treatment of acute and chronic illness
  - – Preventive care
  - – Health education

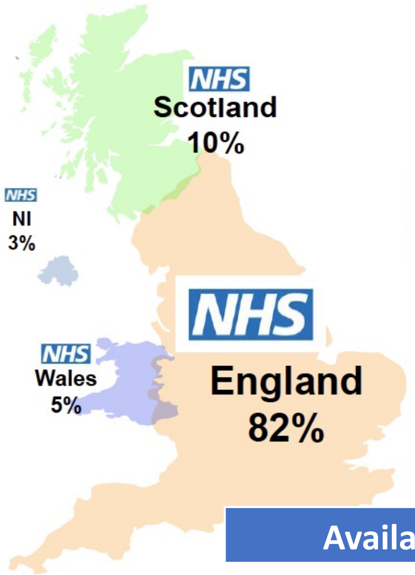
CPRD Primary Care Databases

CPRD Gold (7% of the UK population)

Vision® software

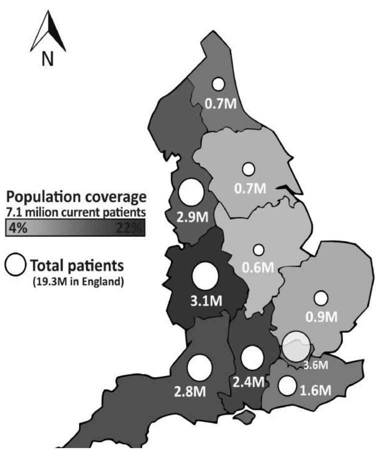
CPRD Aurum (13% of the UK population)

(EMIS Web® software)



Follow-up for current patients Median (IQR)	
12Yrs (4.5-24)	9Yrs (3.4-20.6)
25% of the patients have active records for 20+ years	
Read code mapped to SNOMED codes (April 2018)	SNOMED code
Available data before 1987	1987-Present

100%



Information: demographics, lifestyle factors, diagnosis and/or symptoms, immunization records, laboratory test results, prescription records

CPRD Linkages

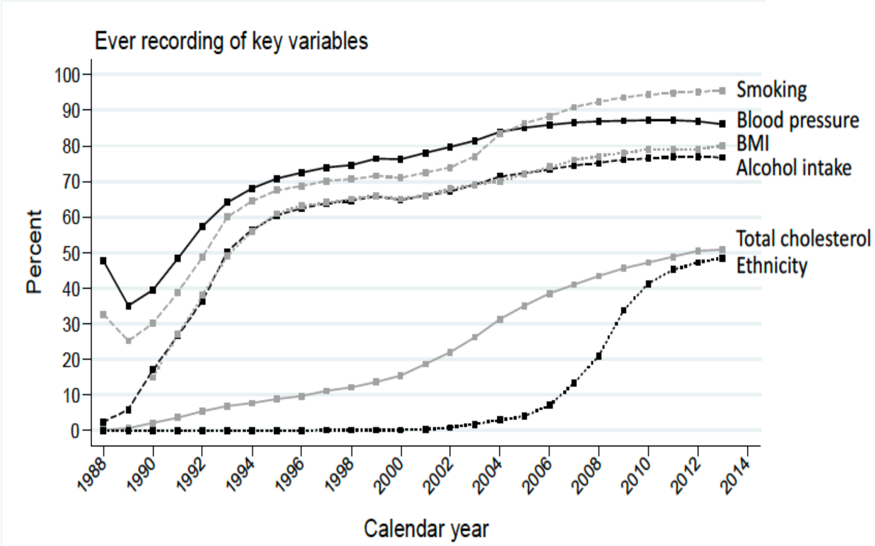
Area Level	NHS (National Health Service)	Cancer Registry (England)	Special links
Several indices of socio-economic status	Hospital Episode Statistics (HES; inpatient, outpatient)	Cancer registration	Mother-baby link
	Accident and Emergency data	Systemic anti-cancer therapy	Pregnancy register
	Death registration	National radiotherapy dataset	Covid-19 data
	Diagnostic imaging (type and body part tested, no image or results)		Ethnicity Record (‘Asian’, ‘black’, ‘mixed’, ‘white’, ‘other’, ‘unknown’)

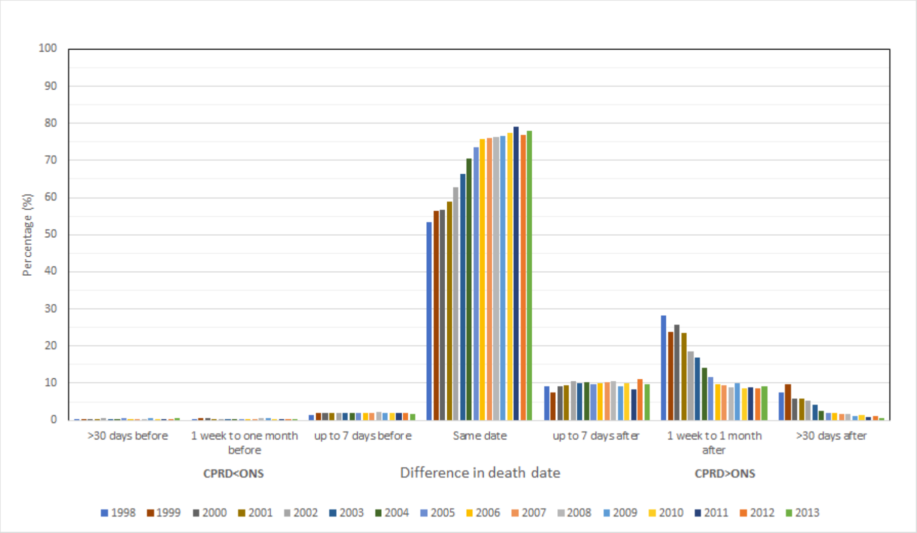
Period of Coverage & data coding for Different Datasets

Database	Period of coverage	Coding
Primary care	1987-2024	Read/SNOMD
HES Admitted Patient Care (inpatient)	1997-2023	ICD-10
Death Register	1998-2024	ICD-9 & ICD-10
Cancer Registry	1990-2018	ICD-O-3 & ICD-10

Ref: https://cprd.com/linked-data

Data Availability & Accuracy of Selected Variables

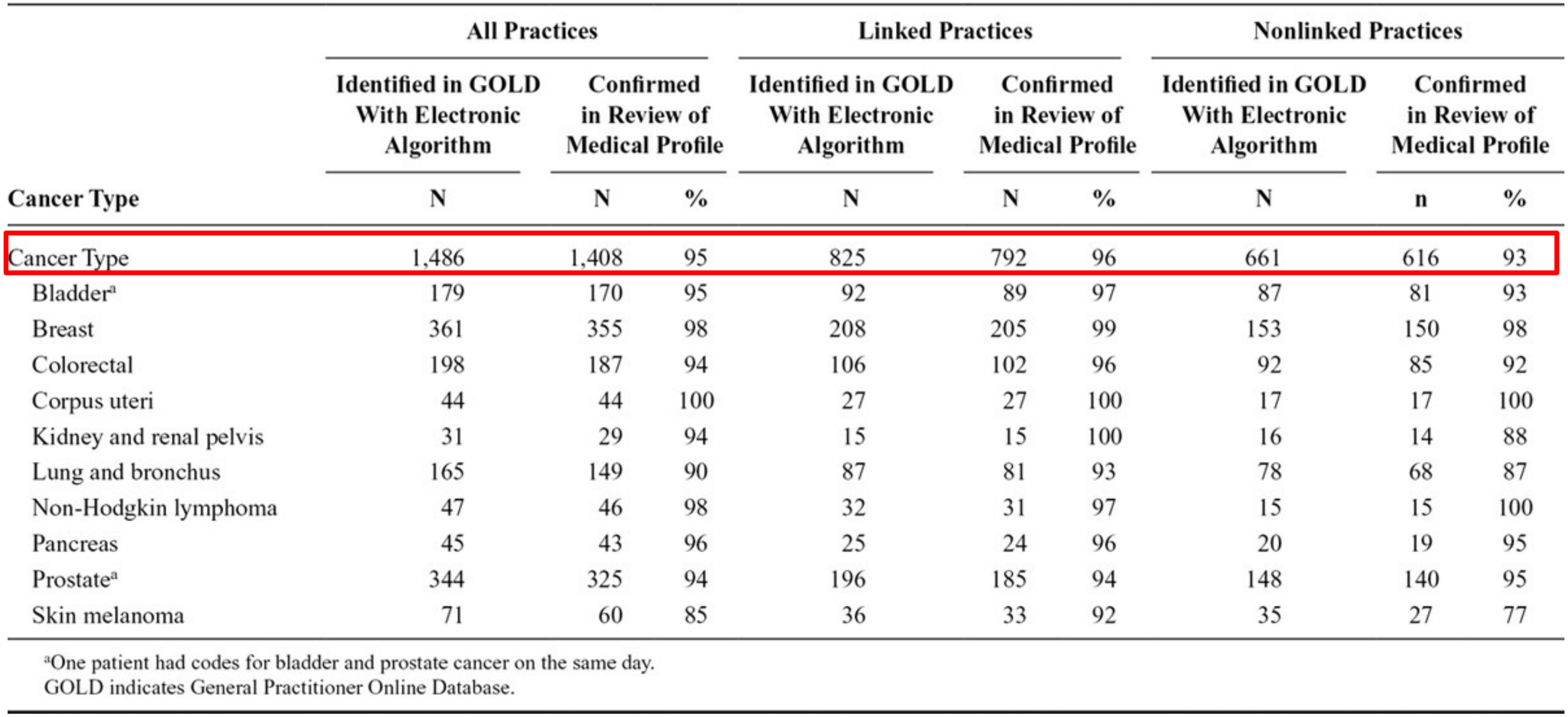




Herret et *al., Int J Epi. 2015*

Gallagher et al. PharmacoepidemiolDrug Saf. 2019

Validity of Cancer Diagnosis in CPRD



	All Practices			Linked Practices			Nonlinked Practices		
	Identified in GOLD With Electronic Algorithm	Confirmed in Review of Medical Profile		Identified in GOLD With Electronic Algorithm	Confirmed in Review of Medical Profile		Identified in GOLD With Electronic Algorithm	Confirmed in Review of Medical Profile	
Cancer Type	N	N	%	N	N	%	N	n	%
Cancer Type	1,486	1,408	95	825	792	96	661	616	93
Bladder[a]	179	170	95	92	89	97	87	81	93
Breast	361	355	98	208	205	99	153	150	98
Colorectal	198	187	94	106	102	96	92	85	92
Corpus uteri	44	44	100	27	27	100	17	17	100
Kidney and renal pelvis	31	29	94	15	15	100	16	14	88
Lung and bronchus	165	149	90	87	81	93	78	68	87
Non-Hodgkin lymphoma	47	46	98	32	31	97	15	15	100
Pancreas	45	43	96	25	24	96	20	19	95
Prostate[a]	344	325	94	196	185	94	148	140	95
Skin melanoma	71	60	85	36	33	92	35	27	77

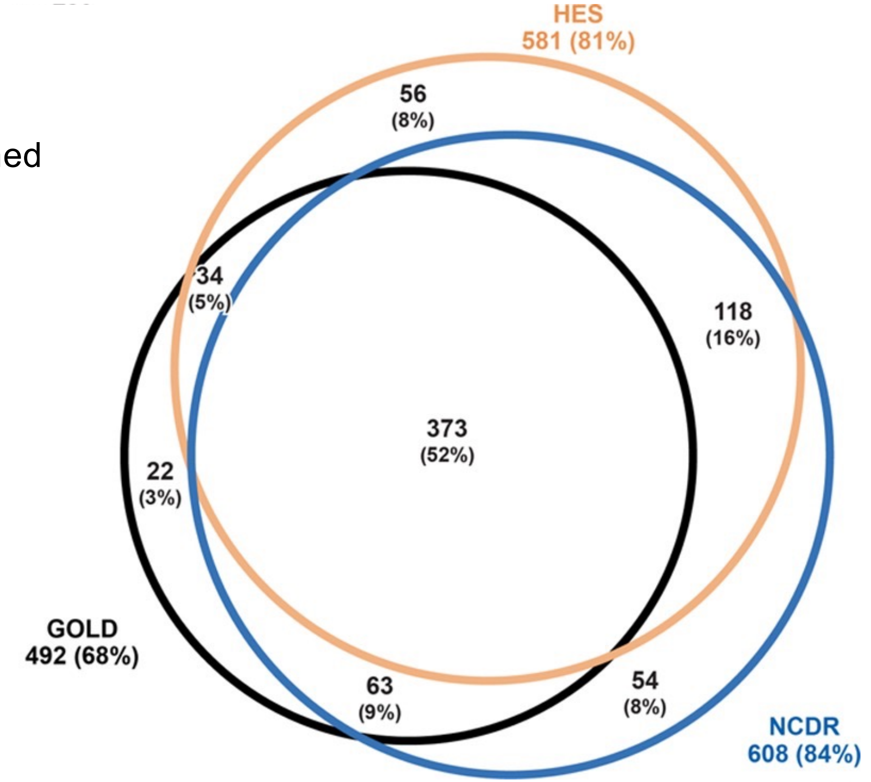
[a] One patient had codes for bladder and prostate cancer on the same day.
GOLD indicates General Practitioner Online Database.

Margulis et al., Epidemiology. 2018

Completeness of Cancer Diagnosis in CPRD Gold

Total 720 confirmed cancer cases

Gold+HES=92%

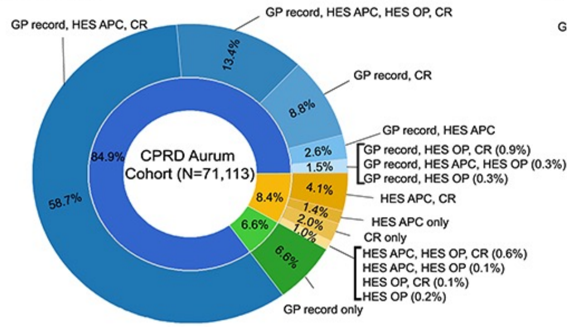


Margulis et al., Epidemiology. 2018

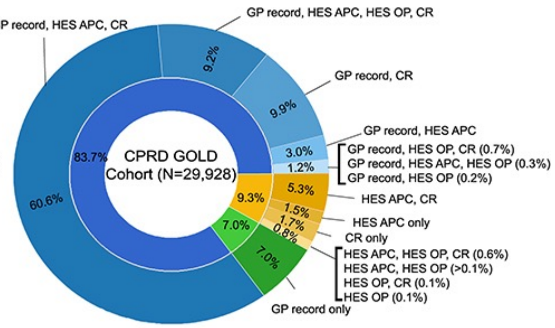
Comparison of Breast Cancer Diagnosis
Primary care and linkage databases

~100,000 patients with breast cancer record in any CPRD source between 2004-2019

Aurum



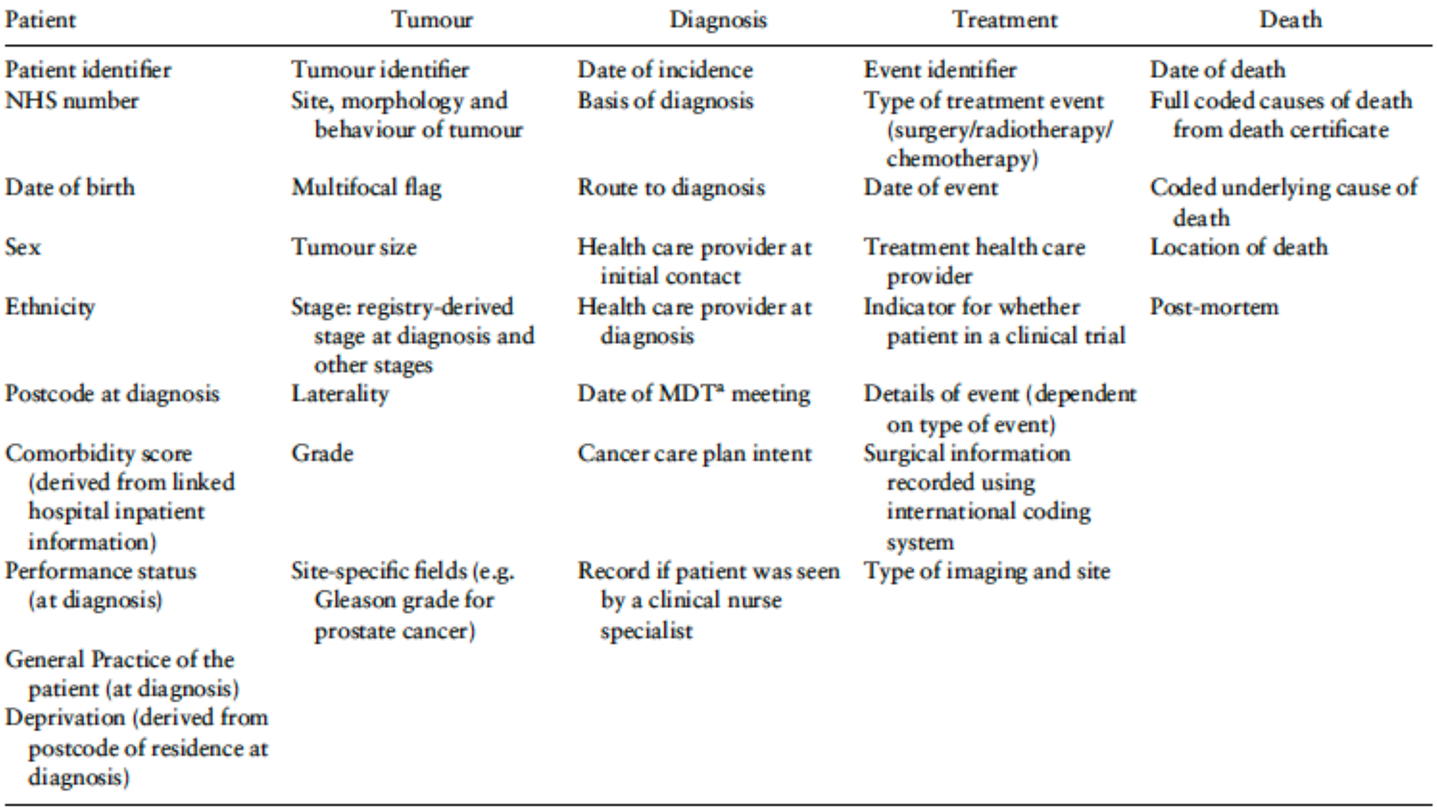
Gold



90% of the patients are found in primary care records (similar for Gold and Aurum)
The completeness of primary records were lowest (~70%) for youngest (<30 years) and oldest (80+)
>80 had record in both primary care and HES
~88% had records in both primary care and cancer registry
<20% of the cases had records in HES OP

Hagberg et al., Clin Epidemiol. 2023

Key Data in the Cancer Registry

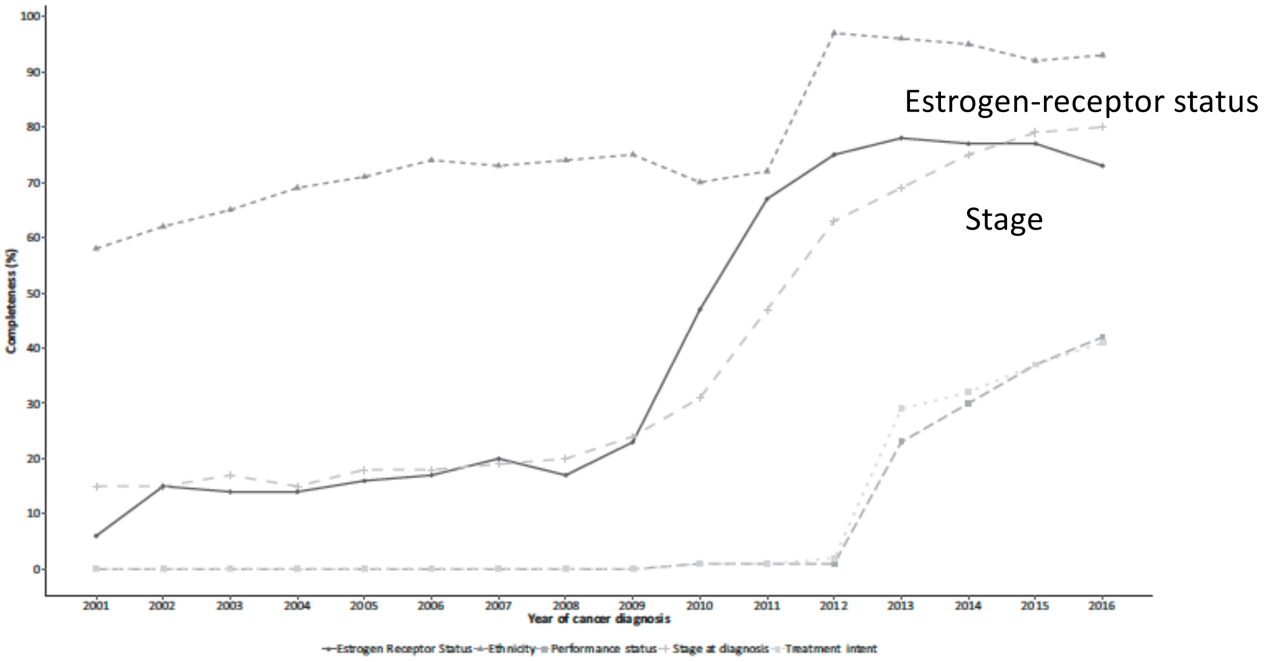


Patient	Tumour	Diagnosis	Treatment	Death
Patient identifier	Tumour identifier	Date of incidence	Event identifier	Date of death
NHS number	Site, morphology and behaviour of tumour	Basis of diagnosis	Type of treatment event (surgery/radiotherapy/chemotherapy)	Full coded causes of death from death certificate
Date of birth	Multifocal flag	Route to diagnosis	Date of event	Coded underlying cause of death
Sex	Tumour size	Health care provider at initial contact	Treatment health care provider	Location of death
Ethnicity	Stage: registry-derived stage at diagnosis and other stages	Health care provider at diagnosis	Indicator for whether patient in a clinical trial	Post-mortem
Postcode at diagnosis	Laterality	Date of MDT[a] meeting	Details of event (dependent on type of event)	
Comorbidity score (derived from linked hospital inpatient information)	Grade	Cancer care plan intent	Surgical information recorded using international coding system	
Performance status (at diagnosis)	Site-specific fields (e.g. Gleason grade for prostate cancer)	Record if patient was seen by a clinical nurse specialist	Type of imaging and site	
General Practice of the patient (at diagnosis)				
Deprivation (derived from postcode of residence at diagnosis)				

Full data dictionary is available here: https://www.cprd.com/sites/default/files/2022-02/CPRD%20Cancer%20Registration%20Dictionary%20set%2021%20v10.1.pdf

Hanson et al., Int J Epidemiol. 2020

Cancer Registry Data Completeness from 2001 to 2016



Estrogen-receptor status

Stage

Hanson et al., Int J Epidemiol. 2020

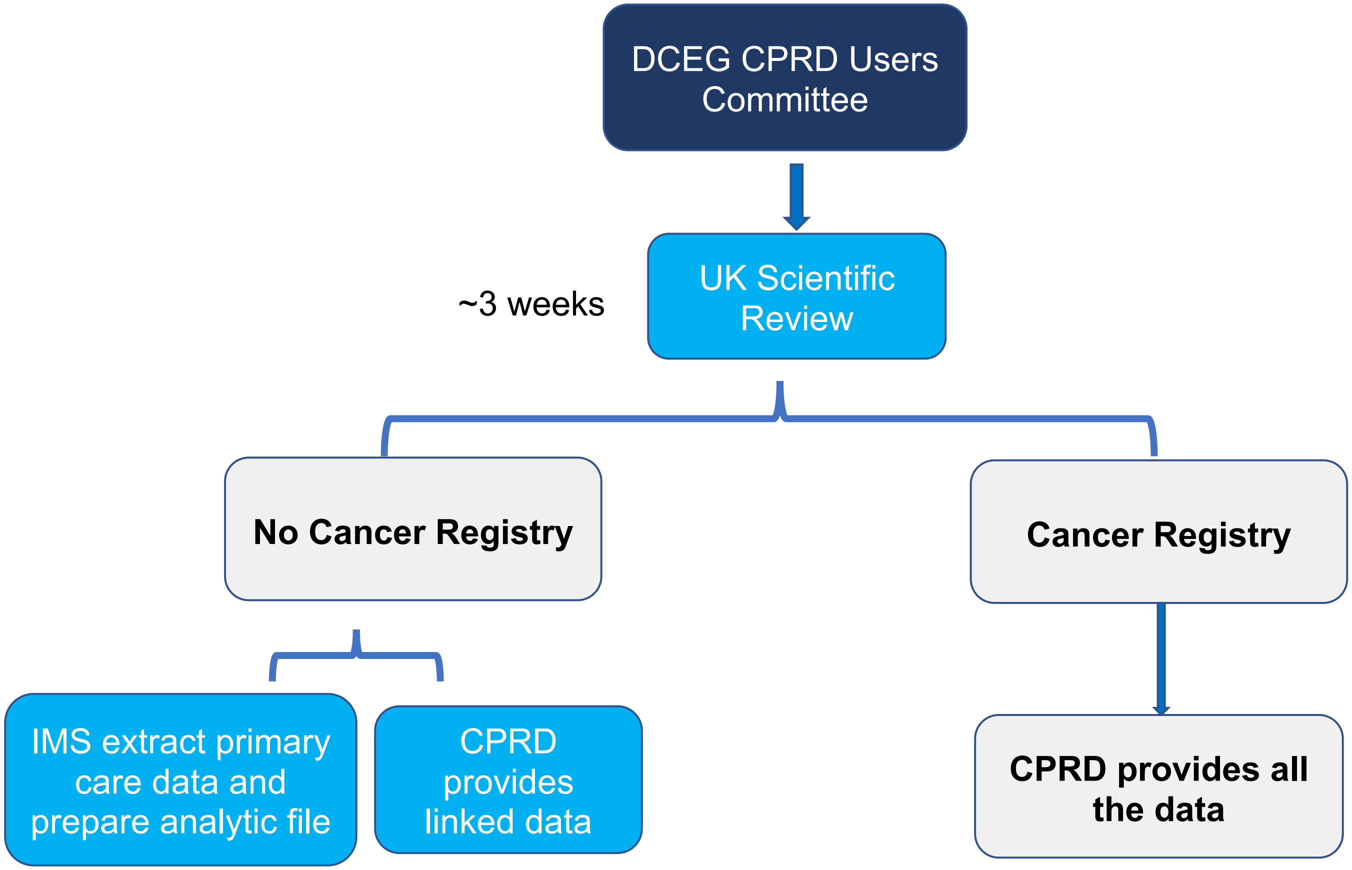
Features of CPRD Applicable to Epidemiology Research

- Large size
- Up-to-date
- Long follow-up
- All ages
- Full disease spectrum
- Cancer record is near complete
- Benign tumors available
- Population representation
- High quality data
- Available linkages

DCEG Use of CPRD (FY16-FY21)

- Annual License: unlimited number of studies
- Cost sharing model: interested branches with OD help
  - CGB
  - MEB
  - BB
  - IIB
  - ITEB
- CPRD Users committee: representative from branch users
- Data management: IMS (dedicated analysts)

Process to Access CPRD



DCEG CPRD Users Committee

UK Scientific Review

~3 weeks

No Cancer Registry

Cancer Registry

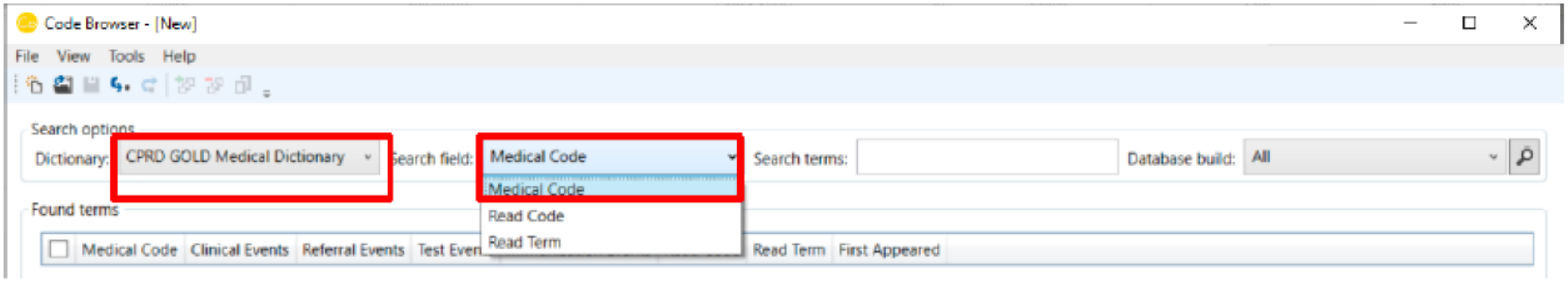
IMS extract primary care data and prepare analytic file

CPRD provides linked data

CPRD provides all the data

Challenges & Opportunities: Coding system

New, extensive, and inclusive (diagnosis, details, symptoms, signs, complications)



Dictionary: CPRD GOLD Medical Dictionary

Search field: Medical Code

Search terms:

Database build: All

DCEG actions:

- Developed a code repository
- Developed algorithms to define smoking, alcohol use, and obesity status

Studied Cancer Sites & Study Population

Cancers

- Burkitt lymphoma
- Biliary tract cancer
- Liver cancer
- Ovarian cancer
- Endometrial cancer
- Gastric cancer
- Prostate cancer
- Lung cancer

Special population

- Patients with myotonic dystrophy
- Transgender individuals

Methodological Studies

- Combining incident and prevalent cohorts in survival analysis
- Conversion of CPRD Aurum to OMOP Common Data Model (NLM)

Accomplishments

- 11 lead PIs and many collaborators
- 7 requests cancer registry linkage
- 17 manuscripts (published or in-press)
- Two Ph.D. dissertations
- IRA funding

DCEG CPRD Users

Lead PIs

IIB:

Sam Mbulaiteye

Jill Koshiol

Meredith Shield

Sarah Jackson

MEB:

Katherine McGlynn

Constnza Camargo

ITEB:

Gretchen Gierach

Tere Landi

BB:

Ruth Pfeiffer

Barry Graubard

Hormuzd Katki

CGB:

Nico Wentzensen

Shahinaz Gadalla

IMS

Emily Carver

David Ruggieri

Current & Former Fellows

Youjin Wang

Rotana Alsaggaf

Monica D’Ary

Kara Michels

Sarah Irvin

Emily Pearce

Ana Best

Minkyo Song

Jack Murphy

Michael Kebede

Rebecca Landy

CPRD Users Committee Members