Case study	RWD source	Sample size estimation form the hypotetical trial protocol	Feasibility assessment *(yes/yes, with limitations/no)*	Rationale for the feasibility assessmernt	Limitations identified during the feasibility assessment and categorisation	Description of potential impact of the identified limitations on the study results
9 (Tolvaptan and Risk Associated to Hepatotoxicity in Autosomal Dominant Polycystic Kidney Disease)	**CPRD**	With an approximate estimated sample size of 552 (based on a 1:1 ratio between intervention and control groups, with 276 participants in each), and considering that CPRD includes data from approximately 4.4 million inhabitants (as of 2014), autosomal dominant polycystic kidney disease (ADPKD) affects 1 in 400 to 1,000 people, and tolvaptan at its 90mg and 60mg dosages is prescribed to approximately 2,273 patients with ADPKD in England, the target sample size is anticipated to be reached. [1,2]	Yes	Elements with high criticality are available and fairly reliable. Data recency of 3 months before extraction, reasonably enough for the research question. Based on published information, the sample size proposed in the hypothetical trial protocol seems achievable.	·Minor: From the DEAPs experience, they use diagnostic of alcohol abuse rather than use alcohol consumption in units. ·Minor: Kidney volume, if not directly available, could be assessed though an internal algorithm that assesses lab tests/diagnoses of renal function. ·Minor: Liver function test is available as requested by GP, but this will surely have testing bias (higher availability for those clinically worse or with symptoms). ·Minor: The DEAP stated will probably only have basline liver function enzyme levels available for the tolvaptan users as it is mandatory before starting, but not in the non-users. ·Minor: Treatment discontinuation not directly available but can be assessed using standard adherence calculation methods.	Given alcohol consumption detection is based on diagnostic codes, more severe cases are likely to be identified. So, alcohol use not associated with a particular disorder might be underrepresented and bias towards the most severe cases. As kidney volume is not readily available and needs to be derived as proxy, resulting values might be innacurate.

REFERENCES

[1] https://www.ncbi.nlm.nih.gov/books/NBK532934/

[2] Chong J, Harris T, Ong ACM. Regional variation in tolvaptan prescribing across England: national data and retrospective evaluation from an expert centre. Clin Kidney J. 2022 Aug 26;16(1):61-68. doi: 10.1093/ckj/sfac190. PMID: 36726434; PMCID: PMC9871855.

Scientific research question		Tolvaptan and Risk Associated to Hepatotoxicity in Autosomal Dominant Polycystic Kidney Disease						
Design elements	**Operationalization of definitions**	**Data elements for valid capture of variables**	**Criticality of the quality of the element**	**Extensiveness assessment (if applicable)**	**Reliability assessment (if applicable)**	**Coherence assessment (if applicable)**	**Timeliness assessment (if applicable)**	**Origin of information**
Study population	**Inclusion criteria**							
	18 years and older	Date of birth	*High*	100% have date of birth	Only year is available, this may impact precision.		As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors
	Autosomal Dominant Polycystic Kidney Disease	Diagnostic code Read Code (CPRD Gold) SNOMED (CPRD Aurum) Local EMIS® codesICD-10 for HES	*High*	Diagnostic codes available for 100% of patients, monthly for CPRD Gold and quaterly for CPRD Aurum			As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors
	Total kidney volume of 750 ml or more	Imaging procedure code plus imaging result of ultrasound	*High*	Available in HES (not in UU license). Especifically, they only have access to HES APC, but not HES DID				https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
	Creatinine clearance of 60 ml/min or more	Diagnostic Code Laboratory test	*High*	Diagnostic codes available for 100% of patients				
	Exclusion criteria							
	AST or ALT levels > 1.5x ULN at screening	Diagnostic code Laboratory test	*High*	Diagnostic codes available for 100% of patients. Laboratory values will not be available for non-users.				https://pmc.ncbi.nlm.nih.gov/articles/PMC9516205/#T1
	Total bilirubin or alkaline phosphatase > ULN	Diagnostic code Laboratory test	*High*	Diagnostic codes available for 100% of patients. Laboratory values will not be available for non-users.				
	Current or previous treatment with tolvaptan	Medication codes	*High*	100% of individuals have available information		As ATC codes are not available, a mapping to ATC will potentially be needed to extract study drugs information.The DEAP informed they already have the mapping ready, so this should not be a problem.		
Treatment/exposure	Tolvaptan with symptomatic control	Medication codes	*High*	100% of individuals have available information. Symptomatic care may include antihypertensives, pain medication, treatment for kidneystones. As so, OTC codes might not be captured (pain-killers being often OTC).		As ATC codes are not available, a mapping to ATC will potentially be needed to extract study drugs information. The DEAP informed they already have the mapping ready, so this should not be a problem.		
Comparator group (if applicable)	Untreated with symptomatic control alone	Medication codes	*High*	100% of individuals have available information. Symptomatic care may include antihypertensives, pain medication, treatment for kidneystones. As so, OTC codes might not be captured (pain-killers being often OTC).	There might be some bias where we will over estimate those having more severe symptoms, or those only treated with OTC drugs.			
Key endpoint(s)	Elevations in liver enzyme levels (ALT and AST) of more than 3 times the upper limit of normal Total bilirubin > 2x ULN Alkaline phosphatase > 2x ULN	Diagnostic code Laboratory test Date of laboratory test	*High*	Diagnostic codes available for 100% of patients	Other liver injuries (drug induced liver injury/acute liver injury) will probably be prevented by stopping treatment early based on liver			https://pmc.ncbi.nlm.nih.gov/articles/PMC9516205/#T1
	Diagnosis of Drug-induced liver injury or Acute liver injury	Diagnostic code	*High*	Diagnostic codes available for 100% of patients				
Confounders	**Patient characteristics**							
	Age	Date of birth	*Low*	100% of individuals have date of birth	Only year is available, this may impact precision.		As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors
	Sex	Sex	*Low*	>99%. Sex categories in CPRD include unknown and indeterminate sex, but are never included in data extractions (<1% of records without sex information are excluded); they are extremely rare.				
	Baseline liver function (enzyme levels)	Laboratory test Date laboratory test Alcohol	*Low*	The DEAP stated will probably only have this available for the tolvaptan users as it is mandatory before starting, but not in the non-users	Liver function test is available as requested by GP, but this will surely have testing bias (higher availability for those clinically worse or with symptoms).			

	Alcohol use/addiction/use disorders	Diagnostic code Laboratory test	*Low*	Diagnostic codes available for 100% of patients From 2014 data, around 50% (men 18-67y); 80% (men and women 65+), Around 60% (women 18-64%) From the DEAPs experience, they use diagnostic of alcohol abuse rather than use alcohol consumption in units	Liver function test is available as requested by GP, but this will surely have testing bias (higher availability for those clinically worse or with symptoms).			https://doi.org/10.1093/ije/dyv098 https://www.tandfonline.com/doi/full/10.2147/CLEP.S477778#d1e326
	Smoking status	Smoking	*Low*	Around 70% (men 18-67y); 90% (men and women 65+), Around 80% (women 18-64%) *Data from 2014*				https://doi.org/10.1093/ije/dyv098
	Comorbidities							
	Any liver disease	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects in attending emergency room				
	Prior liver function-related incidents	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects in attending emergency room				
	Hypertension	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects in attending emergency room				
	Concomitant medication use with hepatotoxic potential							
	Statins	Medication code Date of prescription/dispensing	*Low*	100% of individuals have available information				
	Paracetamol	Medication code Date of prescription/dispensing	*Low*	100% of individuals have available information OTC codes might not be captured				
	Some antifungals (ketoconazole)	Medication code Date of prescription/dispensing	*Low*	100% of individuals have available information				
	Some antiepileptics (sodium valproate)	Medication code Date of prescription/dispensing	*Low*	100% of individuals have available information				
	Methotrexate	Medication code Date of prescription/dispensing	*Low*	100% of individuals have available information				
	Some antibiotics (vancomycin, piperacillin/tazobactam, amox/clav, ceftriaxone)	Medication code Date of prescription/dispensing	*Low*	100% of individuals have available information				
	Herbal medication	Herbal medication use	*Low*	Not available				
	Concomitant medication that interacts with tolvaptan							
	CYP3A4 inhibitors	Medication code Date of prescription/dispensing	*Low*	100% available codes				
	CYP3A4 inductors	Medication code Date of prescription/dispensing	*Low*	100% available codes				
	ADPKD progression							
	Total kidney volume	Procedure code Result of ultrasound	*Low*	Not directly available but could be assessed though an internal algorithm that assesses lab tests/diagnoses of renal function				https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf Provided by DEAP
	Baseline eGFR	Laboratory test Date laboratory test	*Low*					https://pmc.ncbi.nlm.nih.gov/articles/PMC5410977/
	(family) history of rapid disease progression	Family medical history	*Low*	Not available				
Intercurrent events	Treatment discontinuation	Medication code Date of drug discontinuation	*Low*	Not directly available but can be assessed using standard adherence calculation methods				
	Tolvaptan initiation in control group	Medication code Date of prescription/dispensing	*High*	100% of individuals have available information				https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
	All cause mortality	Diagnostic code Date of death	*High*	Diagnostic codes available for 100% of patients 8% of the whole population (irrespective of vital status) has a date of death recorded; 100% of death people have a date of death By 2013, 98.8% of deaths were in agreement with the Office of National Statistics, within ±30 days				https://onlinelibrary.wiley.com/doi/full/10.1002/pds.4747 https://www.sciencedirect.com/science/article/abs/pii/S1386505619306252

	Initiation of any medication with known hepatoxicity effects	Medication code Date of prescription/dispensing	*Low*					
Follow-up time needed per patient in the study	36 months	1 week screening + 36 months of treatment (including follow-up)	*High*				As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	
Minimum time in the data source for lookback assessment	1 year	1 year of lookback	*High*				As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	
Others	Permitted concomitant therapies: Antihypertensive medication, pain medication, treatment for kidney stone Prohibited concomitant therapies: diuretics	Medication code Date of prescription/dispensing	*Low*	100% of individuals have medicatin information available				

	Estimated sample size: Approx 522 people with ADPKD.			Considering that CPRD includes data from approximately 4.4 million inhabitants (as of 2014), autosomal dominant polycystic kidney disease (ADPKD) affects 1 in 400 to 1,000 people, and tolvaptan at its 90mg and 60mg dosages is prescribed to approximately 2,273 patients with ADPKD in England, the target sample size is anticipated to be reached.				

Dimension	Sub-dimension	Metrics	Description	Origin of information
Timeliness	Currency	How often is the database updated (i.e., frequency of updates)	GOLD: monthly; Aurum: quaterly	https://academic.oup.com/ije/article/44/3/827/632531
		The time gap between the latest available data and date when data is delivered to user. (i.e., how up-to-date data are when it reach the user)	1 month plus lag of delivery for CPRD GOLD, and 3 months plus lag of delivery for CPRD Aurum	Provided by DEAP
		The time elapsed from when a user requests the data to when they actually receive it	Requested to DEAP and unable to provide	
		Median time (years) between first and last available records for unique individuals	5.89 years	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors
Extensiveness	Coverage	Percentage of a target population present in a database	CPRD-GOLD 2,894,922 current acceptable patients (i.e. registered at currently contributing practices that use Vision software, excluding transferred out, deceased patients and those flagged by CPRD as not acceptable for clinical research for data quality issues) equal to **4.32%** based on the UK population estimates of 67,026,300 from the Office of National Statistics (July 2024). CPRD-AURUM 16,585,135 Current acceptable patients (i.e. registered at currently contributing practices2, excluding transferred out and deceased patients) equal to **24.27%** percentage UK population coverage (67,026,300) (september 2024).	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors https://www.cprd.com/doi/cprd-gold-november-2024-dataset https://www.cprd.com/doi/cprd-aurum-september-2024-dataset https://jech.bmj.com/content/76/10/880
	Completness	% of subjects in the data with a recorded birth date	Percentage not provided (only year of birth available)	
		% of subjects in the data, irrespective of vital status, that have a recorded date of death	A date of death is recorded for 100% of individuals who are known to have died	https://zenodo.org/records/13384860
		% of subjects in the data with a record of sex	100%	https://zenodo.org/records/13384860
		% of subjects in the data who had an event with a code for the event	100% (86% of the emergency room setting)	https://zenodo.org/records/13384860 https://www.cprd.com/cprd-linked-
		% of subjects in the data who had a prescription/dispensing with a recorded code for the medicine	100%	https://zenodo.org/records/13384860
		% of subjects in the data who got vaccinated with a recorded code for the vaccine	A register of vaccination with a code for the vaccine is recorded for 100% of individuals who are known to have been vaccinated	https://zenodo.org/records/13384860
		Others: BMI	BMI completeness increased over calendar timefrom 37% in 1990–1994 to 77% in 2005–2011, was higher among female and increased with age	https://bmjopen.bmj.com/content/3/9/e003389
Reliability	Accuracy	The population distribution in the data source aligns with that of the country	Population distribution as expected based on the statistics of the general population of England. Previous literature acknowledges some potential overrepresentation of minority ethnic groups. There is a study ongoing in regards to CPRD representativeness (see link). Active population size by ageband: -Paediatric Population (< 18 years): 519902 (13.1%) -Children (2 to < 12 years): 287819 (8.3%) -Adolescents (12 to < 18 years): 200949 (5.1%) -Adults (18 to < 46 years): 1061418 (26.7%) -Adults (46 to < 65 years): 725924 (18.3%) -Elderly (≥ 65 years): 587470 (14.8%) -Adults (65 to < 75 years): 303212 (7.6%) -Adults (75 to < 85 years): 205960 (5.2%) -Adults (85 years and over): 78298 (2.0%)	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors https://jech.bmj.com/content/76/10/880 https://pophealthmetrics.biomedcentral.com/articles/10.1186/s12963-023-00302-0 https://www.cprd.com/approved-studies/representativeness-clinical-practice-research-datalink-cprd-primary-care-databases
		Records of diagnostics, exposures or medical observations that do not agree with common expectations and knowledge or feasible ranges (e.g., pregnancy records in males, a human with 4 arms, systolic pressure higher than 250mmHg, etc)	A data cleaning procedure is performed to avoid inconsistencies and other unfeasible data (see link) Rate of adherence among metformin new users is lower than rates determined in previous UK studies Nearly all patients who had elevated HbA1c labs or hypoglycemic treatments also had a type 2 diabetes diagnosis code Completeness for hyper-cholesterolemia and anemia diagnoses is modest even when the presence of treatments and lab results indicated the conditions were likely present (51%-59% and 58%-70%, respectively)	https://www.cprd.com/sites/default/files/2023-02/CPRD%20Aurum%20Glossary%20Terms%20v2.pdf https://www.sciencedirect.com/science/article/pii/S2214623720300351?via%3Dihub#s0055 https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135
		Records of healthcare events (diagnoses, prescriptions, admissions, etc) with logical inconsistencies (e.g., and admission occurs after death)	Data values after death: 0% (from DEAP experience, some event dates may occur after censoring) Date values before birth: 0.02%	https://zenodo.org/records/13384860 https://www.cprd.com/sites/default/files/2023-02/CPRD%20Aurum%20Glossary%20Terms%20v2.pdf
		Variables that are based in imputation, derivation or inference (e.g., end of treatment date is derived from treatment start date and treatment cycle length)	Mother-baby id, pregnancy, ethnicity	https://onlinelibrary.wiley.com/doi/10.1002/pds.5135 https://www.cprd.com/cprd-algorithm-derived-data
	Precision	Exposures codes precision level, including medicines and vaccines (e.g., active principle, therapeutic group, ...)	Active principle (ATC level 5 codes)	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
		Precision of date of birth (e.g., day, month, year)	Year (Month/year only for children)	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
		Precision of date of death (e.g., day, month, year)	Day, month, year	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
		Precision of date of the event/diagnosis (e.g., day, month, year)	Day, month, year	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
		Precision of date of the exposure (e.g., day, month, year)	Day, month, year	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
	Traceability	Provenance of event records	Primary care medical records, Emergency room, Intensive care unit, Hospitalisation (ER/ICU, HOSP only through linked data. UU only has access to HES admited patient care)	https://catalogues.ema.europa.eu/node/1026/administrative-details
		Provenance of medicines/vaccines records	Primary care medical records (Prescription medicines, No dispensing medicines)	https://catalogues.ema.europa.eu/node/1026/administrative-details
Coherence	Format coherence	For dates, formatting constraint being followed	Date of birth: MM/YY Other dates: DD/MM/YYYY (Death, events/diagnosis/exposure) Character, length 5 or 10	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
		For sex, formatting constraint being followed	Mapping: Lookup SEX Type: INTEGER, Format:1, 1♂(male) 2♀(female) 3□(indeterminate) 4□(unknown)	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
	Relational coherence	% of records with the Person ID in the PERSONS table	98.2-100%	https://zenodo.org/records/13384860
	Semantic coherence - to determine whether the database uses a standardised dictionary	For EVENTS definitions, codelists/data dictionaries being employed according to external standards	Read Code (CPRD Gold): these are used for diagnoses; from April 2018, Read codes are prospectively mapped to SNOMED CT codes SNOMED (CPRD Aurum) Local EMIS® codesICD-10 for HES Medcodeid (unique code for the medical term selected by the GP)	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf https://www.cprd.com/sites/default/files/2024-08/CPRD%20Aurum%20Data%20Specification%20v3.5.pdf
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	Prodcodeid (unique code for the treatment selected by the GP), SNOMED for some immunisations No ATC codes available in the raw data but ATC for active substances link is available at the Utrecht University	https://zenodo.org/records/13384860

	Uniqueness	Number of records flagged as potential duplicates	Requested to DEAP and unable to provide	

Item		Sub-item	Description	Origin of information	Maturity level grade	Maturity level criteria and definitions	Rationale
0	Data base identification	Country	United Kingdom (UK)	https://www.cprd.com/	N/A	*N/A*	*N/A*
		Data Access Provider	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 (DHSC). The DHSC is the legal 'controller' of the data which they hold.	https://www.cprd.com/		*N/A*	
		Organisation type	Government-funded, and not-for-profit cost recovery organisation.	https://www.cprd.com/introduction-cprd		*N/A*	
I	Rationale and scope for the RWD source creation	Primary purpose for which data are collected	Supporting retrospective and prospective public health studies and interventional research.	https://www.cprd.com/introduction-cprd	*3*	*L1 if information is available as free text and/or online link(s)*	*Relevant for all DQ dimensions (reliability, extensiveness, coherence and timeliness) as it provides a general understandin g of the strengths and limitations of an RWD source.*
		Criteria for the selection of the data being collected or integrated	The CPRD collates routinely collected anonymised electronic health record data from general practices who have agreed at a practice level to provide data on a monthly basis. Centers can join under request by means a form available online to request joining the network. Specific criteria are not specified/not found. All patients registered with the participating practices are included in the dataset, unless they have individually requested to opt out of data sharing, by asking their GP to amend their registration details on the system to disable the extraction of their data	https://www.cprd.com/join-growing-network-practices-contributing-cprd https://doi.org/10.1093/ije/dyv098		*L2 if information is available using standardised templates to make information easy to digest and interpret (the EMA recommends to check this tool as reference: REQueST Tool and its vision paper [Internet]. EUnetHTA. 2019. Available from: 721 https://www.eunethta.eu/request-tool-and-its-vision-paper/.* *L3 if the information is provided as Metadata (machine readable), including standrad formats, clear definitions and potentially some quality information*	*Knowing the triggers would ease the understanding of the content and motivations behind the data.*
		What triggers a record in the database	**Event triggering registration of a person in the data source:** Practice registration **Event triggering de-registration of a person in the data source:** Death, Practice deregistration **Event triggering creation of a record in the data source:** Patient has contact with a GP practice	https://catalogues.ema.europa.eu/node/1026/data-flows-and-management			
		Publications describing this RWD	https://academic.oup.com/ije/article/44/3/827/632531 https://doi.org/10.1093/ije/dyv098 https://doi.org/10.1093/ije/dyz034				
II	Data collection or recording process	Description of data provider (geographical and oragnizational setting, nature of the data - reported by patients, HCP, etc)	They are the regulator of medicines, medical devices and blood components for transfusion in the UK. Tha nature of the date is provided by GPs	https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency/about	*2*	*L1 if information is available as free text and/or online link(s)*	*Essential to understand extensiveness and to assess reliability (that can be affected by errors or biases in the collection process). Also, essential to evaluate SOP for data collection or recording practices that may impact coherence (e.g., where "curation at source" is involved and provide hard constraints for timeliness).*
		Standard Operating Procedures (SOPs) recording	The SOPs for data collection, qualy control and research use are detaill in the links	https://www.cprd.com/safeguarding-patient-data https://www.cprd.com/data-access		*L2 if information is available using standardised templates to make information easy to digest and interpret, and also standard vocabularies are available*	
		How SOPs are implemented and monitored	The responsible party of each of the following procedures are: - GPs are responsible for Data collection - NHS is responsible for De-identification and linkage - CPRD is responsible for Quality and anomysation for research - The DHSC is the legal 'controller' of the data which they hold. We have not found further details on monitoring proecdures.	https://www.cprd.com/safeguarding-patient-data			
		Key data elements captured (are they always recorded, are they optional, is there a planned coverage over time, ...)	The CPRD primary care database includes data on demographics, symptoms, tests and laboratory results, diagnoses, therapies (immunisations, prescriptions and prescription duration), health-related behaviours and lifestyle variables (such as smoking, alcohol consumption, and height and weight), referrals to secondary care and hospital admissions. For over half of patients, linkage with datasets from secondary care, disease-specific cohorts and mortality records enhance the range of data available for research. Diagnoses, symptoms and signs are also available from intensive care unit, hospitalisation and emergency room. For further details please visit the link on "CPRD GOLD Data Specification" and "CPRD Aurum Data Specification".	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf https://www.cprd.com/sites/default/files/2024-08/CPRD%20Aurum%20Data%20Specification%20v3.5.pdf https://pmc.ncbi.nlm.nih.gov/articles/PMC4521131/ https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135 https://www.cprd.com/cprd-linked-data#HES%20Accident%20and%20Emergency%20data		*L3 if additionally SOPs specify KPIs to mo*	
III	The selection of RWD sources and their onboarding *(Applies to RWD sources that integrate or repurpose other RWD sources)*	Criteria to accept or exclude a datasource	N/A		*N/A*	*L1 if information about selection criteria or DQ performance is available as free text and/or online link(s)* *L2 if a structure checklist and dataset version control are available* *L3 is only aspirational. NA*	*When data are provided by a data aggregator, ensure that all the available evidence related to systems and processes potentially affecting DQ (extensiveness and reliability especially) can be followed. Provide information of impact on both reliability and evidence (as well as other dimensions if relative constraints are formulated in*
		Is there a DQ assessment for data sources onboarded?	N/A				
		If yes: does it follow any specific framework? Is there an assessment checklist? Are datasets versions traceable?	N/A				
IV	The data management infrastructure	List of systems used to manage the RWD (either for data collection, recording, processing, etc)	EMIS Web® electronic patient record system software for CPRD Aurum Vision® software for CPRD GOLD (From April 2018, Read codes are prospectively mapped to SNOMED CT codes by Vision)	https://www.cprd.com/primary-care-data-public-health-research https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135 https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf	*2*	*L1 if information is available as free text and/or online link(s)*	*Essential for reliability regarding data alterations resulting from system accidents, software errors or malicious intervention.*
		Software testing and software quality control in place	Requested to DEAP and unable to provide		*N/A*	*L2 if the hardware or software implementation complies with recognised quality standards that can be reported*	
		Measures to prevent accidental physical data alterations (e.g.: backups, redundant systems, checksums)	CPRD is oblied to complete an annual NHS Data Security and Protection Toolkit assessment to demonstrate that it meets the required standard for holding data securely. We are unsure of what this toolkit entails. information is broad and might be only available when you buy/contract the service.	https://www.cprd.com/safeguarding-patient-data https://www.dsptoolkit.nhs.uk/	*2*	*L3 NA*	

V	Data management and governance	Data management principles being followed (e.g., GCP, ISO, FAIR, etc)	Requested to DEAP and unable to provide		*N/A*	*L1 if information is available as free text and/or online link(s)* *L2 if standard best practices are being used and a direct impact on DQ is reported. There are SOPs and dat amanagement processes that adhere to the standards. The representation of metadata follows FAIR standards* *L3 if data management and governance is implemented in the data platforms 'Digital Quality Measures' (DQMs) so that reports of performance and deviations are automated. Submitted metadata are generated "by design". Basically, if everything in L2 is automatised and generated by default*	*Data management and governance impact reliability, as well as all quality dimensions for metadata.*
		Data management processes in place (DQ controls, KPIs, SOPs, etc)	**Check:** the volume of data downloaded against that supplied data volumes are in the expected range all data elements received are of the correct type, length and format **Our range of validation and quality checks include:** Collection-level validation ensures integrity by checking that data received from practices contain only expected data files and ensures that all data elements are of the correct type, length and format. Duplicate records are identified and removed. Transformation-level validation checks for referential integrity between records ensure that there are no orphan records included in the database (for example, that all event records link to a patient). Research-quality-level validation covers the actual content of the data. CPRD provides a patient-level data quality metric in the form of a binary 'acceptability' flag. This is based on recording and internal consistency of key variables including date of birth, practice registration date and transfer out date. In addition to checks undertaken by the CPRD teams before the data is released, researchers using the data are advised to undertake study-specific checks themselves.	https://www.cprd.com/data-quality	*2*		
		Measures to prevent data alterations by unauthorised parties (cybersecurity)	Single study dataset licence – where a study dataset defined by an approved research application will be prepared by CPRD, and access granted to researchers via the CPRD Trusted Research Environment (TRE). As UU, they have a multistudy license; so data is extracted by UU themselves. The TRE is not used by UU at this moment; we use our own secure TRE for research purposes	https://www.cprd.com/cprd-safe-our-trusted-research-environment			
		Auditing and DQ imporvement procedures in place	Sensitive mortality data Operational management issues Data destruction Acces control Information transfer Risk managment Operational transfer	https://digital.nhs.uk/services/data-access-request-service-dars/data-sharing-audits/2021/post-audit-review-cprd			
VI	Data manipulation steps	Frequency of data updates	GOLD: monthly; Aurum: Quarterly	https://catalogues.ema.europa.eu/node/976/data-flows-and-management	*2*	*L1 if free-text information, links or publications are available reporting all the mentioned features* *L2 if Tests performed follow some standard or shared set of tests, that can be re-used across RWD sources. Key performance indicators (KPIs) for data cleaning (e.g., data duplications, mislabelling, etc.) are provided. Data mapping tables and algorithms are described with a standard characterisation of their performance. Lists of standard test batteries used to detect loss of accuracy or precision are provided. All lineage information is provided as metadata associated to the dataset* *L3 if information about data onboarding is directly provided by the platform, e.g.: * Transaction logs are available including deviations and actions that required manual intervention Actual data transformation code is accessible and verifiable. Quality checks and KPIs reported are automatically generated by the data platform (e.g.: unit testing) Lineage information is automatically generated by the processing platform*	*Impacts reliability both in terms of accuracy (possible errors) and precision (i.e., the degree of approximatio n by which data represents reality). Essential to ensure traceability of information. Also impacts coherence and potentially timeliness.*
		Data transformations performed, data mapping steps, data cleaning	Requested to DEAP and unable to provide		*N/A*		
		Information about loss of precision during data manipulation steps	Requested to DEAP and unable to provide				
		Lineage information (e.g., justifcation of data manipulation, track of changes and versions)	Each dataset has a digital object identifier (DOI) to trace specific database versions	https://www.cprd.com/digital-object-identifiers-dois-datasets	*2*		
VII	Data augmentation steps (e.g., imputation or linkage)	Is any augmentation happening in this datasource?	Patient-level data from consenting practices are linked via a trusted third party—the Health and Social Care Information Centre—to a range of other data sources. Established linkages include Hospital Episode Statistics (HES), covering Admitted Patient Care (APC), Accident & Emergency (A&E), and Outpatient (OP) data; Office for National Statistics (ONS) mortality records, including causes of death; and multiple deprivation indices such as the Index of Multiple Deprivation (IMD), Townsend index, Carstairs index, and Rural-Urban classification. Linkages also extend to disease registries, including the National Cancer Intelligence Network and tumour-level records from the National Cancer Data Repository (NCDR) submitted to ONS by the England Cancer Registries, as well as the Myocardial Ischaemia National Audit Project. Additional linkages are planned (see CPRD website), and researchers can request bespoke linkage for individual studies.	https://catalogues.ema.europa.eu/node/1026/data-flows-and-management https://www.cprd.com/cprd-linked-data https://pmc.ncbi.nlm.nih.gov/articles/PMC4521131/ https://www.cprd.com/sites/default/files/2022-02/Linkage%20Source%20Documentation_set22_1.20.pdf	*2*	*L1 if free-text information, links or publications are available reporting all the mentioned features* *L2 if algorithms are published and their performance documented. Information on which values result from imputation is provided as part of the dataset (e.g., presented in metadata or a data*	*Data augmentation steps impact accuracy (reliability) and extensiveness. We consider here data transformations that produce new information subject to reliability issues: e.g.: imputation of missing values, or extraction of codes via natural language processing.*
		If yes, which are the methods applied	For linkage to the HES datasets, ONS Death, NCRAS, ICNARC and Mental Health data, the trusted third party use an eight-step process to match patients using some or all of the following: NHS number, date of birth, sex and postcode. It is explained in the attached link	https://www.cprd.com/sites/default/files/2022-02/Linkage%20Source%20Documentation_set22_1.20.pdf			

		If yes, which algorithms and assumptions applied	It is explained in the attached link	https://www.cprd.com/sites/default/files/2022-02/Linkage%20Source%20Documentation_set22_1.20.pdf		*dictionary)*	
		If yes, which is the error rate when conducting the augmentation	Requested to DEAP and unable to provide		*N/A*	*L3 if an automatised process for data linkage/mapping exists*	
VIII	Known quality issues and independent QA assessment of the RWD source	Known DQ issues (e.g., poor overall completeness in Q3 2020 due to COVID-19)	A significant proportion of lab data lacking a normal range were missing units or had values inconsistent with units provided. A significant proportion of cases of hyperlipidemia or anemia will be missed ifthe investigator relies solely on diagnosis codes to select patients. Researchers should consider using available treatments, supporting codes, and lab data to supplement diagnosis codes and enhance case capture when studying anemia, diabetes and hyperlidiemia using CPRD. In rpevious articles, CPRD assumed that, for anemia, diabetes or hyperlimidemia, lab and prescription data were less likely than GP entered diagnosis codes to be missing or miscoded, as prescriptions must be entered into the electronic recordto be issued and lab data with a normal range are likely to be elec-tronically transferred from the laboratory. As CPRD has prescription data, it is unknown whether the patient took the prescription.	https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135 https://www.sciencedirect.com/science/article/pii/S2214623720300351?via%3Dihub#s0055	*1*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*Explicit description of known DQ issues, as well as external validation performed (all dimensions affected)*
		Validation studies and publications resulting from this EWD source	Useful publications on the quality of CPRD data for research	https://www.cprd.com/data-quality		*L2 if standard procedures are set for external/internal validation of the data* *L3 if the mechanism provided includes notification of automatically detected DQ issues*	
IX	The RWD source representation	Description of data model or models used (OMOP, FHIR, ...)	OMOP and CONCEPTION	https://catalogues.ema.europa.eu/node/1026/data-flows-and-management	*3*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*Descriptive of the intended coherence DQ of a dataset and its metadata.*
		Data ontology (dictionaries and vocabularies) being used, and if in standard formats that allow mapping across different languages (e.g., UMLS)	Medcodeid (unique code for the medical term selected by the GP), Prodcodeid (unique code for the treatment selected by the GP), Read (for diagnoses; from April 2018, Read codes are prospectively mapped to SNOMED CT codes by Vision), Snomed (added to clinical, immunisation, referral and test tables) Read Code (CPRD Gold) SNOMED (CPRD Aurum) Local EMIS® codes and ICD-10 for HES	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf https://www.cprd.com/sites/default/files/2024-08/CPRD%20Aurum%20Data%20Specification%20v3.5.pdf https://pmc.ncbi.nlm.nih.gov/articles/PMC4521131/ https://www.cprd.com/cprd-linked-data#HES%20Accident%20and%20Emergency%20data		*L2 if the description refers to a model such as OMOP, I2B2, FHIR, others, or an extension of them. Data dictionaries are standard (and if non-standard.* *L3 if a standard CDM is used, the datasource has been mapped to one or more than one CDM, and if data dictionaries are provided using standard formats that facilitate the mappings across different vocabularies and across languages*	
X	The RWD source declared Service Level Agreements (SLA)	Guaranteed frequency of updates and incident response time (e.g., corrections in case of errors)	Monthly		*1*	*L1 if free-text information and links are available reporting all the mentioned features*	*Descriptive of guaranteed timeliness and possible variations of extensivenes s/reliability provided.*
		Processes and resources accompanying the data, such as documentation, training materials or help desk contact	Requested to DEAP and unable to provide		*N/A*	*L2 if details of established data processess by the provider are available*	
		Possibility to collect additional data if needed	Requested to DEAP and unable to provide			*L3 if SLA compliance is assessed and reported automatically*	
XI	The RWD source licensing and restrictions	Data use agreements that may limit data use or access (consent, limitations of use), accessibility policies, licensing constraints, standard policies of use, data retention	Access to CPRD data, including UK Primary Care Data, and linked data such as Hospital Episode Statistics, is subject to protocol approval via CPRD's Research Data Governance (RDG) Process.	https://www.cprd.com/data-access	*2*	*L1 if free-text information and links are available reporting all the mentioned features* *L2 if policies and licensing are standardised to a broad ranae of RWD* *L3 NA*	*Descriptive of aspects that can limit extensiveness and coherence in downstream data aggregations.*
XII	Feedback	Is there a data ecosystem in place so that quality assessment by data consumers can provide feedback to improve the data collection and production process, thus allowing a continuous monitoring and improvement of DQ?	A general email and address are available	https://www.cprd.com/contact	*1*	*L1 if a person of contact is provided for Q&A* *L2 if the contact provided allows tracking of issues and follow-up* *L3 if the mechanism provided includes notification of automatically detected DQ issues*	*Descriptive of feedback mechanisms in place to improve all aspects of DQ*