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) 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	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. Knowing the triggers would ease the understanding of the content and motivations behind the data.
	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			
	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 orgnizational 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) L2 if information is available using standardised templates to make information easy to digest and interpret, and also standard vocabularies are available L3 if additionally SOPs specify KPIs to monitor	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			
	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			
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 inclusion/exclusion (I/E) criteria)
	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 L3 NA	
	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		
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)	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	https://www.cprd.com/data-quality	2	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	
	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		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	
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 dictionary) L3 if an automatised process for data linkage/mapping exists	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			
	If yes, which is the error rate when conducting the augmentation	Requested to DEAP and unable to provide		N/A		
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 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	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			
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 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, jusified why) 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	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			
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 L2 if details of established data processess by the provider are available L3 if SLA compliance is assessed and reported automatically	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		
	Possibility to collect additional data if needed	Requested to DEAP and unable to provide				
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 range 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

Item		Sub-item	Description	Origin of information	Maturity level grade	Maturity level criteria and definitions	Rationale
0	Data base identification	Country	SPAIN		*N/A*	*N/A*	*N/A*
		Data Access Provider	The data owner of the Valencia Health System Integrated Database (VID) is the Valencia regional government. Reseach teams at Fisabio, such as the HSRP Unit, can be considered as providers, as they are granted access to VID data on a project-basis, after 1. ethics committee approval of a research protocol and 2. data commission approval of the data extraction.	https://catalogues.ema.europa.eu/node/1077/administrative-details		*N/A*	
		Organisation type	EU Institution/Body/Agency Not-for-profit			*N/A*	
I	Rationale and scope for the RWD source creation	Primary purpose for which data are collected	Primary purpose of data collection is registry of daily clinical practice in the public healthcare system of the Valencia region. Data can also be used for research under the conditions explained above. Data in VID require expert data analyst review and manipulation, as well as data linkage procedures, before being ready for research purposes.	https://catalogues.ema.europa.eu/institution/3331380: HSRP Unit	*2*	*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.* *Knowing the triggers would ease the understanding of the content and motivations behind the data.*
		Criteria for the selection of the data being collected or integrated	Data are sourced of all general population covered by the universal public health care system in the Health Department of the Valencia Regional Government All data available in the VID (see IJE publication)	https://academic.oup.com/ije/article/49/3/740/5707448 https://catalogues.ema.europa.eu/node/1077/quantitative-descriptors		*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/.*	
		Data prompts	**Event triggering registration of a person in the data source, other:** Any contact with the health system triggers the registration of the person **Event triggering de-registration of a person in the data source:** Death, Emigration, Insurance coverage end **Event triggering creation of a record in the data source:** Most of them are created when a contact with the health system is produced. In other cases, such as pharmacy data, when the prescription is created or when the dispensing is produced. Each table of the data source has their own triggers	https://catalogues.ema.europa.eu/node/1077/data-flows-and-management		*L3 if the information is provided as Metadata (machine readable), including standrad formats, clear definitions and potentially some quality information*	
		Publications describing this RWD	https://academic.oup.com/ije/article/49/3/740/5707448				
II	Data collection or recording process	Description of data provider (geographical and oragnizational setting, nature of the data - reported by patients, HCP, etc)	The Valencia Health System Integrated Database (VID) is a set of multiple, public, population-wide electronic databases for the Valencia Region, the fourth most populated Spanish region, with ~5 million inhabitants. VID provides exhaustive longitudinal information including sociodemographic and administrative data (sex, age, nationality, etc.), clinical (diagnoses, procedures, diagnostic tests, imaging, etc.), pharmaceutical (prescription, dispensation) and healthcare utilization data from hospital care, emergency departments, specialized care (including mental and obstetrics care), primary care and other public health services.	https://academic.oup.com/ije/article/49/3/740/5707448 https://catalogues.ema.europa.eu/node/1077/quantitative-descriptors	*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	Data owner (Valencia government) and data users (Fisabio) comply with the regulations on the protection of personal data and the guarantee of digital rights, specifically Organic Law 3/2018 of December 5, on the Protection of Personal Data and the Guarantee of Digital Rights, and Regulation (EU) 2016/679 of the European Parliament and of the Council of April 27, 2016 (General Data Protection Regulation – GDPR). Likewise, both organisations are obliged to implement the necessary technical and organizational measures to ensure the security and integrity of personal data and to prevent its alteration, loss, or unauthorized processing or access.	https://www.san.gva.es/es/web/investigacio/solicitud-datos-sia-gaia https://fisabio.san.gva.es/es/registro-de-actividades-de-tratamiento-de-datos/ Provided by DEAP		*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	Data extraction: data is gathered from various databases within the VID system and extracted based on a previously approved research protocol by an Institutional Review Board (IRB). Participants: IT technicians from the Valencia Region Health Department. Data cleansing and linkage: extracted data is reviewed (consistency and quality checks) before database linkage. Participants: Data analysts from the Health Services and Policy Research (HSRP) Unit at Fisabio. Security measures: throughout the process, data is pseudonymized and protected with access controls and obfuscation procedures to ensure privacy and regulatory compliance. Participants: IT technicians from the Valencia Region Health Department. Dissemination: research findings are published in peer-reviewed journals, included in the HMA EMA Catalogue, and disseminated via various social media platforms. Participants: Researchers from the HSRP Unit at Fisabio.	Provided by DEAP	*N/A*		
		Key data elements captured (are they always recorded, are they optional, is there a planned coverage over time, ...)	Disease information (information about COVID-19 test in RedMIVA), Rare diseases, Pregnancy and/or neonates, Hospital admission/or discharge, ICU admission, Cause of death, Prescriptions of medicines, Dispensing of medicines, Contraception measures covered by public health system, Indication for use, Medical devices, Administration of vaccines, Procedures, Clinical measurements, Healthcare provider, Patient-generated data, Units of healthcare utilisation, Unique identifier for persons, Diagnostic codes, Medicinal product information (Active ingredient(s), ATC and product codes, INN, Presentation, Dosage regime, Formulation, Package size, Strength, Prescribed duration), Lifestyle factors (alcohol use, tobacco use), Sociodemographic information (age, country of origin, deprivation index, gender, health area, socioeconomic status) Included databanks: - SIP (basic information of VHS coverage and sociodemographic data) - ABUCASIS (ambulatory medical record): includes GAIA (prescription and dispensation) and SIA (diagnoses, history, lab results, lifestyle habits) - ORION (hospital medical record): includes MBDS (diagnoses and procedures, discharge and admission data) and AED (triage data, diagnoses, tests and procedures in public emergency rooms) - CRC (corporate info: physician information and geographical and functional organization of health services) - RedMIVA (results of microbiological analysis) - SIV (vaccination information: type, manufacturer, batch number, number of doses, location and administration date, adverse reactions related to vaccines, rejected vaccinations and, if applicable, risk groups) - CIS (cancer information: incidence, prevalence, tumour site and tumour type) - SIER-CV (epidemiological information on rare diseases: incidence, prevalence, patient characteristics, geographical distribution...): includes the Congenital Anomalies Registry (prevalence of congenital anomalies in the region and the exposure to teratogen agents) - BIMCV (a digital biobank of medical images) All ~~data sources~~ databanks that composes VID (01_SIP, 02_PCV, 03_CEX, 04_MBDS, 05_AED, 06_DIAGNOSES, 07_GAIA, 08_SIV, 09_MDR, 10_PMR, 11_EOS, 12_TESTS, 13_CONG and 14_REDMIVA) have a linking ID number that identifies uniquely each person.	https://catalogues.ema.europa.eu/node/1077/data-flows-and-management#darwin-data-source-linkage https://academic.oup.com/ije/article/49/3/740/5707448 https://catalogues.ema.europa.eu/node/1077/quantitative-descriptors	2	*L3 if additionally SOPs specify KPIs to monitor*	
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)*	*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 inclusion/excl usion (I/E) criteria)*
		Is there a DQ assessment for data sources onboarded?	N/A			*L2 if a structure checklist and dataset version control are available*	
		If yes: does it follow any specific framework? Is there an assessment checklist? Are datasets versions traceable?	N/A			*L3 is only aspirational. N/A*	
IV	The data management infrastructure	List of systems used to manage the RWD (either for data collection, recording, processing, etc)	The IJE publication provides full detail of the different systems used to gather RWD in VID.	https://academic.oup.com/ije/article/49/3/740/5707448	*1*	*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	The HSRP Unit in Fisabio as DAP has measures in place to ensure full traceability and data protection for data handled by the research team. Measures adopted by the data owner (Valencia regional government) are regulated by european and national regulations about data privacy and data protection.	Provided by DEAP		*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)	The HSRP Unit in Fisabio as data provider has measures in place to ensure full traceability and data protection for data handled by the research team. Measures adopted by the data owner (Valencia regional government) are regulated by european and national regulations about data privacy and data protection.	Provided by DEAP		*L3 N/A*	
V	Data management and governance	Data management principles being followed (e.g., GCP, ISO, FAIR, etc)	Data owner is subjet to european and national regulations on data management.	Provided by DEAP	*2*	*L1 if information is available as free text and/or online link(s)*	*Data management and governance impact reliability, as well as all quality dimensions for metadata.*
		Data management processes in place (DQ controls, KPIs, SOPs, etc)	Only in some cases, such as the MBDS and the AED records, are data subject to a consolidation and quality check process before data are available for research	https://academic.oup.com/ije/article/49/3/740/5707448		*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*	
		Measures to prevent data alterations by unauthorised parties (cybersecurity)	Data owner implements all safety regulations imposed by legal mandate to ensure data safety. With regard to data managed by HSRP Unit, these are stored in a secure server permanently, and access is tightly restricted only to data analysts and senior researers within the team that work in the project.	https://www.san.gva.es/ca/web/sanidad/client-vpn			
		Auditing and DQ imporvement procedures in place	In February 2025, a public body (Oficina Autonómica de Auditoria e Inspección Sanitaria de la Comunidad Valenciana) was created to audit Valencian Health System. No repots have been published yet.	https://www.gva.es/es/inicio/atencion_ciudadano/buscadores/departamentos/detalle_departamentos?id_dept=28011		*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*	
VI	Data manipulation steps	Frequency of data updates	Monthly data updating	https://catalogues.ema.europa.eu/node/1077/data-flows-and-management	*1*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*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	MBDS and AED records are reviewed by data coding specialists to improve coding accuracy and data quality. In the context of a research project, in the case of HRSP Unit all steps with regard to data cleaning etc are tracked and stored safely in the internal server.	https://academic.oup.com/ije/article/49/3/740/5707448		*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*	
		Information about loss of precision during data manipulation steps	In the context of data for a specific research project, HSRP Unit has measures in place to ensure full traceability and data protection.	Provided by DEAP		*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*	
		Lineage information (e.g., justifcation of data manipulation, track of changes and versions)	In the context of data for a specific research project, HSRP Unit has measures in place to ensure full traceability and data protection.	Provided by DEAP			
VII	Data augmentation steps (e.g., imputation or linkage)	Is any augmentation happening in this datasource?	All the data tables above stated belong to the core VID database. Father-child linkage is perfomed.	https://catalogues.ema.europa.eu/node/1077/data-flows-and-management#darwin-data-source-linkage Provided by DEAP	*N/A*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*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	Probabilistic	Provided by DEAP		*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 dictionary)*	
		If yes, which algorithms and assumptions applied	Father-child or family unit linkage is based on residency address.	Provided by DEAP			
		If yes, which is the error rate when conducting the augmentation	Unknown	Provided by DEAP		*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)	Incompleteness of early data from AED records or coding reliability of diagnostic information in the EMR Different datasets cover different periods (ABUCASIS from 2009, ORION from 2008, AED reliable since 2017, RedMIVA from 2008, SIV reliable since 2005, CIS from 2004, SIER-CV reliable since 2012) Lacking data on in-hospital pharmaceutical prescription (pending to be integrated as part of the ORION information system)	https://academic.oup.com/ije/article/49/3/740/5707448	*2*	*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 RWD source	There are many publications using VID data.			*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, ...)	ConcePTION, OMOP	https://catalogues.ema.europa.eu/sites/default/files/cdm-etl-spec/0_3_VID_Catalogue_RTL_specifications_0.pdf	*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)	Cause of death: ICD-10-CM (ICD10-ES or Spanish clinical modification) Indication: ICD-10-CM / ICD-9-CM Procedures: ICD-10-CM / ICD-9-CM Diagnosis / medical event vocabulary: ICD-10-CM / ICD-9-CM Prescriptions of medications: ATC Dispensing of medicines: ATC Medicinal product: ATC / Other Oncology: ICD-O3	https://catalogues.ema.europa.eu/sites/default/files/cdm-etl-spec/0_3_VID_Catalogue_RTL_specifications_0.pdf		*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, justified why)* *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)	Defined on a reserch contract-basis, corrections in the case of errors is a common procedure.	*Provided by DEAP*	*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	N/A	*N/A*	*N/A*	*L2 if details of established data processess by the provider are available*	
		Possibility to collect additional data if needed	VID data is linkable to a set of databases, as defined in the IJE reference.	*Provided by DEAP*	*1*	*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	Ethics approval by an accredited ethical research committee is required to access the data for research purposes. After that, the regional data commission has to aprove the data extraction. Access to data for researchers has no financial cost but is covered by research ethics and authorization processes.	https://academic.oup.com/ije/article/49/3/740/5707448	*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 range 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?	No.		*N/A*	*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*

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://iech.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://iech.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, 1M (male) 2F (female) 3I (indeterminate) 4U (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	

Dimension	Sub-dimension	Metrics	Description	Origin of information
Timeliness	Currency	How often is the database updated (i.e., frequency of updates)	Data banks are updated daily according to clinical practice. MBDS and AED are updated every 6 months	https://zenodo.org/records/13384860
		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 day to 6 months (depending on the data bank) plus the lag of delivery	Provided by DEAP
		The time elapsed from when a user requests the data to when they actually receive it	Between 8 and 14 months	Provided by DEAP
		Median time (years) between first and last available records for unique individuals	12 years	https://catalogues.ema.europa.eu/node/1077/quantitative-descriptors
Extensiveness	Coverage	Percentage of a target population present in a database	Aproximately 98% of the 5 million inhabitants of the region of Valencia, with an annual birth cohort of 48000 newborns, representing 10.7% of the Spanish population and around 1% of the European population	https://academic.oup.com/ije/article/49/3/740/5707448 https://zenodo.org/records/13384860
	Completness	% of subjects in the data with a recorded birth date	100%	
		% 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%	https://zenodo.org/records/13384860
		% of subjects in the data who had a prescription/dispensing with a recorded code for the medicine	ATC code (100%), MPID (100%)	https://zenodo.org/records/13384860
		% of subjects in the data who got vaccinated with a recorded code for the vaccine	From the total of individuals known to have been vaccinated, 100% had the vaccine batch number recorded and 100% had the vaccine type available	https://zenodo.org/records/13384860
Reliability	Accuracy	The population distribution in the data source aligns with that of the country	Population age distribution are aligned with a developed country demographics reported by the National Statistics Institute (INE). To bear in mind, information about people with no contact with the healthcare system or attending the private health sector is not represented. Active population size: Paediatric Population (< 18 years): 461000 (9.6%) Preterm newborn infants (0 – 27 days): 2700 (0.1%) Term newborn infants (0 – 27 days): 33000 (0.7%) Infants and toddlers (28 days – 23 months): 99000 (2.1%) Children (2 to < 12 years): 521000 (10.9%) Adolescents (12 to < 18 years): 263000 (5.5%) Adults (18 to < 46 years): 679000 (14.2%) Adults (46 to < 65 years): 748000 (15.6%) Elderly (≥ 65 years): 996000 (20.8%) Adults (65 to < 75 years): 517000 (10.8%) Adults (75 to < 85 years): 332000 (6.9%)	https://zenodo.org/records/13384860 https://www.ine.es/ García-Sempere A, Orrico-Sánchez A, Muñoz-Quiles C, Hurtado I, Peiró S, Sanfélix-Gimeno G, Diez-Domingo J. Data Resource Profile: The Valencia Health System Integrated Database (VID). Int J Epidemiol. 2020 Jun 1;49(3):740-741e. doi: 10.1093/ije/dyz266. PMID: 31977043; PMCID: PMC7394961.
		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)	Requested to DEAP and unable to provide	
		Records of healthcare events (diagnoses, prescriptions, admissions, etc) with logical inconsistencies (e.g., and admission occurs after death)	Data values before birth: 0-0.1% Data values after death: 0-0%	https://zenodo.org/records/13384860
		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)	In VID no imputation, derivation or inference is performed unless required for a specific project.	https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2023.1207976/full Provided by DEAP
	Precision	Exposures codes precision level, including medicines and vaccines (e.g., active principle, therapeutic group, ...)	Active principle (ATC level 5) and national product codes are available.	Provided by DEAP
		Precision of date of birth (e.g., day, month, year)	Day, month and year	Provided by DEAP
		Precision of date of death (e.g., day, month, year)	Day, month and year	Provided by DEAP
		Precision of date of the event/diagnosis (e.g., day, month, year)	Day, month and year	https://zenodo.org/records/13384860
		Precision of date of the exposure (e.g., day, month, year)	Day, month and year	Provided by DEAP
	Traceability	Provenance of event records	Primary care, Emergency, Hospital, Specialist and ICU	https://zenodo.org/records/13384860
		Provenance of medicines/vaccines records	Prescription, dispensation, vaccine information system	https://catalogues.ema.europa.eu/node/1077/administrative-details
Coherence	Format coherence	For dates, formatting constraint being followed	yyyy/mm/dd Uncertain variable format and length.	Provided by DEAP
		For sex, formatting constraint being followed	STRING 1 character, M (male), F (female)	Provided by DEAP
	Relational coherence	% of records with the Person ID in the PERSONS table	100%. This is controlled at extraction. Data must have all Person IDs in their persons table to be used for a study.	https://zenodo.org/records/13384860
	Semantic coherence - to	For EVENTS definitions, codelists/data dictionaries being employed according to external standards	ICD -10-CM (The ICD-10-CM used is the ICD10-ES (Spanish clinical modification), ICD-9-CM,	Provided by DEAP
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	ATC code (100%), MPID (0%) MIPD (100%)	Provided by DEAP
	Uniqueness	Number of records flagged as potential duplicates	Requested to DEAP and unable to provide	

Scientific research question		Effectiveness of RNA vaccine against COVID-19 in healthy individuals or stable chronic conditions						
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**							
	16 years of age or older.	Date of birth (years)	*High*	100% have date of birth	As people equal or over 16 years, only year is available, this may slightly impact precision.		As data in aurum is updated quartely, is to bear in mind that information extracted will be at least 4 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 https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf and DAP
	Preexisting stable medical conditions may be included, provided their condition does not require hospitalization within six weeks prior to enrolment.	Diagnostic code (ICD or equivalent) Date of diagnosis Date of the hospitalisation admission Date of the hospitalisation discharge	*High*	Diagnostic codes available for 100% of patients (no DEAP experience yet with relevant codelist for this)	In the case of previously hospitalized COVID-19 cases, data from hospitalization may be unreliable from April 1st 2021 to January 31st 2022		The median lenght of follow-up per patient is approximately 6 years and 13 years for active individuals, which accounts for the time needed to patients to accomplish the elegibility criteria.	Provided by DEAP
	Exclusion criteria							
	Received any medication intended to prevent COVID-19.	Medication code Date of prescription/dispensing Indication	*High*	Medication codes are available for 100% of patients	Active principle is the level of detail of medication (apparently enough to decipher any medication intended to prevent COVID-19 infection)	If medicine codes are defined as ATC codes, a mapping to ATC will potentially be needed to extract study drugs information in CPRD. The DEAP informed a mapping is readily available.		
	Immunocompromised individuals with known or suspected immunodeficiency, as determined by medical history, laboratory tests, or physical examination.	Medication code Date of prescription/dispensing Medication dosages Diagnostic code Date of diagnosis Laboratory test results (e.g., CD4 levels)	*High*	Diagnostic codes available for 100% of patients Prescription medicines for 100% of patients (dispensing not available, unknown missingness for dosage) Lab test results will be missing often, but can be part of the case definition				https://pmc.ncbi.nlm.nih.gov/articles/PMC10806788/pdf/bmjopen-2023-073866.pdf DAP
	Participants who have previously received any COVID-19 vaccine.	Vaccination code Administration date	*High*	Vaccine codes available for 100% of patients	Data from patients receiving Novavax, Janssen and Valneva may be unreliable, as these vaccines have not entered yet or have entered UK very lately			https://pmc.ncbi.nlm.nih.gov/articles/PMC10806788/pdf/bmjopen-2023-073866.pdf
Treatment/exposure	Vaccine BNT162b2 (2 doses separated by 21 days)	Vaccination code Vaccine manufacturer Type of vaccine Administration dates Vaccine dose	*High*	Vaccine codes available for 100% of patients				
Comparator group (if applicable)	Saline Placebo (Normal saline (0.9% sodium chloride solution for injection) in 2 doses separated by 21 days	Administration dates Concentration of saline Route of administration Medication code	*High*	Medication codes are available for 100% of patients. However, the missingness of route of administration and concentration is unknown	Saline solution as a placebo has the ATC code V07AB. Howeve	If medicine codes are defined as ATC codes, a mapping to ATC will potentially be needed to extract study drugs information in CPRD. The DEAP informed a mapping is readily available.		
Key endpoint(s)	First ocurrence of laboratory-confirmed COVID-19 infection	Medication code Date of prescription/dispensing Indication	*High*	Vaccine codes available for 100% of patients PCR, Antibody and antigen tests for COVID are available.			As data in aurum is updated quartely, is to bear in mind that information extracted will be at least 4 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	https://pmc.ncbi.nlm.nih.gov/articles/PMC10806788/pdf/bmjopen-2023-073866.pdf https://zenodo.org/records/13384860
Confounders	Age	Date of birth (years)	*Low*	100% have date of birth	In people equal or over 16 years, only year is available, this may slightly impact precision.		As data in aurum is updated quartely, is to bear in mind that information extracted will be at least 4 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	
	Reason of end of administrative follow-up	Event triggering de-registration of a person in the data source	*Low*	Diagnostic codes available for 100% of individuals				
	Sex/Gender	Sex	*Low*	100% of individuals have sex information available. Sex categories in CPRD include unknown and indeterminate sex, but are never included in data extractions; they are extremely rare.				
	Ethnicity	Ethnicity	*Low*	27.1% of all patients in the CPRD (1990–2012) have ethnicity recorded. This proportion rises to 78.3% for patients registered since April 2006. In CPRD-HES, 81.7% of currently registered patients in the UK had ethnicity recorded in primary care. For patients with multiple ethnicity records, mismatched ethnicity within individual primary and secondary care datasets was<10%.	Not available (according to DAP)			https://pmc.ncbi.nlm.nih.gov/articles/PMC4245896/pdf/fdt116.pdf https://pophealthmetrics.biomedcentral.com/articles/10.1186/s12963-023-00302-0
	Socioeconomic status	Socioeconomic status	*Low*	Socioeconomic status seems to be gathered via Index of Multiple Deprivation, having unknown missingness				https://pubmed.ncbi.nlm.nih.gov/29190680/
	Smoking	Smoking habits	*Low*	Smoking present for 89.7% of records.				
	BMI	BMI or weight and height	*Low*	BMI completeness increased over calendar time from 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 https://www.sciencedirect.com/science/article/pii/S2666776224001534#appsec1
	Comorbidities (e.g., diabetes, heart disease, immunosuppression)	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients				
	Prior SARS-CoV-2 infection history	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients				
	Prior vaccination history (e.g., influenza)	Vaccination codes	*Low*	Vaccine codes available for 100% of patients				
	Healthcare access/utilization	Visits to any healthcare resource Date of visits	*Low*	Diagnostic codes available for 100% of patients				
	Medicines	Medication code	*Low*	Prescription medicines for 100% of patients (dispensing not available, unknown missingness for dosage)				
	Occupation (i.e., healthcare worker)	Occupation	*Low*	Not available				
	Household size	Household size	*Low*	Unknown missingness, although two derived binary variables are described: living alone and cohabitation				https://pubmed.ncbi.nlm.nih.gov/29190680/
	Geographic location (e.g., urban vs rural, state or region)	Rural-Urban classification	*Low*	94% Aurum; 78% Gold				https://jech.bmj.com/content/jech/76/10/880.full.pdf
Intercurrent events	Missing or ineligible for second dose of target vaccine	Date of first target vaccine	*Low*	Vaccine codes available for 100% of patients.				
	A third (booster) dose of target vaccine up to 3 months after second dose	Number of doses of target vaccine Date of second target vaccine Date of third target vaccine Vaccination code	*Low*	Vaccine codes available for 100% of patients				
	Third (booster) dose of non-target covid vaccine	Number of doses of non-target COVID vaccine Vaccination code	*Low*	Vaccine codes available for 100% of patients	Data from patients receiving Novavax, Janssen and Valneva may be unreliable, as these vaccines have not entered yet or have entered UK very lately			
	Receipt of any non-covid vaccine dose following treatment completion	Type of non-covid vaccine Administration date of a particular non-covid vaccine Number of doses of target vaccine Date of second target vaccine Date of third target vaccine	*Low*	Non-covid vaccine information with unknown missingness	Some non-covid vaccines information are not available at least for period 2017-20: measles-containing vaccine (1st dose, DTP3 (3rd dose), Hib3 (3rd dose), HepB3 (3rd dose), Pol3 (3rd dose), neumococcal conjugate (2nd dose), varicella (1st dose), BCG, HPV, rotavirus, meningococcal			https://zenodo.org/records/13384860
	Receipt of any other preventative COVID-19 treatment following treatment completion	Medication code Date of prescription/dispensing Indication	*Low*	Prescription medicines for 100% of patients (dispensing not available, unknown missingness for dosage) Indications of medicines are not available, but could be derived indirectly				
	Death	Death date	*High*			98.2% of deaths in the Office of National Statistics data are recorded in the CPRD GOLD primary care data, while agreement onthe exact date of death increased over time to 78.0% in 2013.	As data in aurum is updated quartely, is to bear in mind that information extracted will be at least 4 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.4747
Follow-up time needed per patient in the study	3 months	3 months (including recruitment and follow-up)	*Low*				As data in aurum is updated quartely, is to bear in mind that information extracted will be at least 4 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 https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
Minimum time in the data source for lookback assessment	6 weeks	6 weeks	*High*				The median lenght of follow-up per patient is approximately 6 years and 13 years for active individuals	

	Estimated sample size: Approx. 44,000 participants			As CPRD includes 4.4 million inhabitants (data from 2014), cohort size is expected to be reached.				https://doi.org/10.1093/ije/dyv098

Scientific research question		**Effectiveness of RNA vaccine against COVID-19 in healthy individuals or stable chronic conditions**						
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**							
	16 years of age or older.	Date of birth (years)	*High*	Age is available for 100% of individuals in VID				
	Preexisting stable medical conditions may be included, provided their condition does not require hospitalization within six weeks prior to enrolment.	Diagnostic code (ICD or equivalent) Date of diagnosis Date of the hospitalisation admission Date of the hospitalisation discharge	*High*	Diagnostic codes are available in 100% of individuals	Due to lag period for inpatient data, hospitalization within six weeks prior to enrolment may not be easily checked		Data obtained instantaneously if outpatient context, needs up to 6 months for update if inpatient context	Provided by DEAP
	Exclusion criteria							
	Received any medication intended to prevent COVID-19.	Medication code Date of prescription/dispensing Indication	*High*	Medication: ATC code (100%), except for inpatient medication data, not available in VID.	Active principle is the level of detail of medication (apparently enough to decipher any medication intended to prevent COVID-19 infection)			Provided by DEAP
	Immunocompromised individuals with known or suspected immunodeficiency, as determined by medical history, laboratory tests, or physical examination.	Medication code Date of prescription/dispensing Medication dosages Diagnostic code Date of diagnosis Laboratory test results (e.g., CD4 levels)	*High*	Diagnositc codes: 100% Medication: ATC code (100%), except for inpatient medication data, not available in VID.				
	Participants who have previously received any COVID-19 vaccine.	Vaccination code Administration date	*High*	Vaccines: Lot number (100%), Vaccine type (100%)				
Treatment/exposure	Vaccine BNT162b2	Vaccination code Vaccine manufacturer Type of vaccine Administration dates Vaccine dose	*High*	Vaccines: Lot number (100%), Vaccine type (100%)				
Comparator group (if applicable)	Saline Placebo (Normal saline (0.9% sodium chloride solution for injection)	Administration dates Concentration of saline Route of administration Medication code	*High*	Medication: ATC code (100%)	Saline solution as a placebo has the ATC code V07AB. However, it might be administrated due to various indications and could be scarcely recorded as a placebo. This element is not reliably captured in RWD.			
Key endpoint(s)	First ocurrence of laboratory-confirmed COVID-19 infection	COVID-19 test result COVID-19 test date COVID-19 test type Vaccination date Vaccine code	*High*	Vaccines: Lot number (100%), Vaccine type (100%)			Vaccine information is updated daily, buy COVID-19 infection status may need up to 6 months for update if inpatient context	Provided by DEAP
Confounders	Age	Date of birth (years)	*Low*	Unknown missingness				
	Reason of end of administrative follow-up	Event triggering de-registration of a person in the data source	*Low*	Diagnostic codes: 100%				
	Sex/Gender	Sex	*Low*	100% of individuals have available information				https://zenodo.org/records/13384860
	Ethnicity	Ethnicity	*Low*	Unknown missingness				
	Socioeconomic status	Socioeconomic status	*Low*	Socioeconomic status is indirectly recorded via 2 binary outcomes with unknown missingness: yearly income (< 18,000 vs > 18,000) and risk of social inclusion (yes vs no)				https://pmc.ncbi.nlm.nih.gov/articles/PMC6372152/pdf/pone.0211681.pdf
	Smoking	Smoking habits	*Low*	Unknown missingness				
	BMI	BMI or weight and height	*Low*	Unknown missingness				
	Comorbidities (e.g., diabetes, heart disease, immunosuppression)	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients				
	Prior SARS-CoV-2 infection history	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients				
	Prior vaccination history (e.g., influenza)	Vaccination codes	*Low*	Vaccine codes available for 100% of patients				
	Healthcare access/utilization	Visits to any healthcare resource Date of visits	*Low*	Diagnostic codes available for 100% of patients				
	Medicines	Medication code	*Low*	Medication: ATC code (100%)				
	Occupation (i.e., healthcare worker)	Occupation	*Low*	Employment status is recorded, but exact occupation is not available				
	Household size	Household size	*Low*	Not available				
	Geographic location (e.g., urban vs rural, state or region)	Rural-Urban classification	*Low*	Health area is recorded, but urban vs rural classification is not available				
Intercurrent events	Missing or ineligible for second dose of target vaccine	Date of first target vaccine	*Low*	Identifiable				
	A third (booster) dose of target vaccine up to 3 months after second dose	Number of doses of target vaccine Date of second target vaccine Date of third target vaccine Vaccination code	*Low*	Vaccines: Lot number (100%), Vaccine type (100%)				
	Third (booster) dose of non-target covid vaccine	Number of doses of non-target COVID vaccine Vaccination code	*Low*	Vaccines: Lot number (100%), Vaccine type (100%)				
	Receipt of any non-covid vaccine dose following treatment completion	Type of non-covid vaccine Administration date of a particular non-covid vaccine Number of doses of target vaccine Date of second target vaccine Date of third target vaccine	*Low*	Vaccines: Lot number (100%), Vaccine type (100%)	In previous data instances non-covid vaccines information were not extracted for the period 2017-20 (e.g. measles-containing vaccine, Hib3, HepB3, neumococcal conjugate, varicella, BCG, HPV, rotavirus, meningococcal			https://zenodo.org/records/13384860
	Receipt of any other preventative COVID-19 treatment following treatment completion	Medication code Date of prescription/dispensing Indication	*Low*	Medication: ATC code (100%) Indications of medicines are not available, but could be derived indirectly				
	Death	Death date	*High*	A date of death is recorded for 100% of individuals who are known to have died 100% diagnostic codes				
Follow-up time needed per patient in the study	3 months	3 months (including recruitment and follow-up)	*Low*	No problem in VID			Time to data access for research has been reported in previous Steps.	
Minimum time in the data source for lookback assessment	6 weeks	6 weeks	*High*				The median length of follow-up per patient is approximately 12 years	
	Estimated sample size: Approx. 44,000 participants			Considering that VID includes data from approximately 5 million inhabitants, the target sample size is anticipated to be reached.				

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
1 (mRNA vaccine against COVID-19)	**CPRD**	With an approximate estimated sample size of 44,000 (1:1 ratio of saline and mRNA Covid-19 vaccine), and considering that CPRD includes data from 4.4 million inhabitants (as of 2014), the target sample size is anticipated to be reached. Furthermore, experimental exposure is expected to occur frequently.	Yes, with limitations on the identification of a design element	Elements with high criticalty are available, **except placebo**. Data recency of 3 months before extraction, reasonably enough for the research question. The time elapsed from when a user requests the data to when they actually receive it is unknown. Sample size is achievable.	·Potentially major: In the case of previously hospitalized COVID-19 cases, data from hospitalization may be unreliable from April 1st 2021 to January 31st 2022 ·Potentially major: The use of placebo is not reliably captured in RWD. ·Minor: Data from patients receiving Novavax, Janssen and Valneva may be unreliable, as these vaccines have not entered yet or have entered UK later ·Minor: Dispensing is not available, only prescription ·Minor: Dose number is not available	As data from hospitalizations for previously hospitalized COVID-19 cases might be unreliable during the mentioned period, some individuals who were actually hospitalized might be underdetected (misclassification). As placebo is not used in real world practice hence is not reliably captured in RWE, using it may lead to misclassification of of exposure of the comparator group. Consider replacing by "non-treated" subjects. As Novavax, Janssen and Valneva vaccines have entered UK very lately, their uncompleteness may slightly impact extensiveness.
	VID	With an approximate estimated sample size of 44,000 (1:1 ratio of saline and mRNA Covid-19 vaccine), and considering that VID includes data from approximately 5 million inhabitants, the target sample size is anticipated to be reached. Furthermore, experimental exposure is expected to occur frequently.	Yes, with limitations on the identification of a design element	Elements with high criticalty are available, **except placebo**. Data recency of 6 months before extraction, reasonably enough for the research question. The time elapsed from when a user requests the data to when they actually receive it is ~1 year, which should be accounted for the study performance. Sample size is achievable.	·Potentially major: The use of placebo is not reliably captured in RWD. ·Minor: Inpatient medication not available ·Minor: Missingness of some confounders is unknown	As placebo is not used in real world practice hence is not reliably captured in RWE, using it may lead to misclassification of of exposure of the comparator group. Consider replacing by "non-treated" subjects. The unavailability of inpatient medication is not expected to have an impact to the current study since the exposures in this study (vaccines) are administered in outpatient care. Underestimation of some counfounders may exist as missingness is unknown.