Instructions to navigate this file

Context

10 case studies were selected (see D1)

13 RWD sources have been selected for feasibility assessment, representing 6 countries

Among the RWD sources identified as potential contributors to the case studies outlined in Deliverable 1, pre-screening questions were asked (further detailed in the supplementary file accompanying the current deliverable). If these requirements were met, they underwent a three-step process to determine their feasibility.

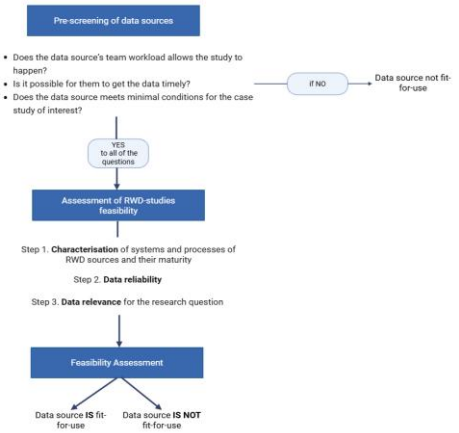
Feasibility assessment

Feasibility assessment is constituted by 3 steps, applying a standardized template consistently for each of the steps:

Step 1: First, we characterised the systems and processes of the RWD sources and their maturity, following the checklist proposed by the EMA.

Step 2: Second, we checked data reliability by using a set of metrics for data quality for RWD.

Step 3: T



Tabs

The orange tab contains the instructions for the contextualisation of this file.

Green tabs show Step 1 for each of the included RWD sources.

Blue tabs show Step 2 for each of the included RWD sources.

Purple tabs show Step 3, which should be performed per database and case study

For the tables of Step 1, 2 and 3, the column headers coloured in green, blue and purple respectively, are the ones that have been filled by the research team. Headers

Other helpful information and definitions

In column "I" in this tab you may find the **acronym** definition list.

For Step 1, sub-items listed as not found have remained like this if the Data Experts and Access Providers (DEAP) could not provide further information.

For Step 1, the maturity was assessed for the information that was available. In general terms, Levels of maturity were defined as follows:

Level 1: Documented – Basic information is provided through simple documentation or supporting links, with more extensive details available via standard operating procedures (SOPs), as well as key performance indicators (KPIs)

Level 2: Formalised – Information follows established or emerging standards. SOPs and KPIs align with recognised standards.

Level 3: Automated – Data quality is ensured by design, with information generated automatically by systems rather than entered manually. For example, in the case of data lineage, provenance information is generated by an ETL engine or derived from some executable electronic specification.

For Step 3, design elements were identified from the hypotetical target trial protocols of each case-study.

For Step 3, the criticality of the quality of each design element was classified as high or low. The criteria applied were as follows:

High - If the design element is essential to the design or integrity of the study. The design elements in this category were:

Inclusion/Exclusion, Primary Exposure, Primary Outcome, Intercurrent events of a While on, Principal stratum and Composite strategy, Secondary outcomes, Time of follow-up in the study, minimum time in the data source (if related to inclusion/exclusion criteria)

Low - If the design element is a comorbidity, general demographic aspect, descriptive or does not affect the viability of the study. The design elements in this category were:

Comorbidities, Descriptive variables, Intercurrent events of a Treatment Policy or Hypothetical strategy, Confounders, minimum time in the data source (if related to comorbidities)

For the FINAL FEASIBILITY ASSESSMENT, the sample size estimation from the hypotetical trial protocol was considered. Feasibility was categorised as "yes", "yes with limitations" and "no" (column D). Additionally, in column F, the limitations found during the feasibility assessment and their categorisation are reported. The criteria applied were:

Major: If the limitation significantly affects primary outcomes or key variables

If it affects statistical power or sample size

List of acronyms

Acronym	Meaning
ADHD	Attention Deficit Hyperactivity Disorder
ADPKD	Autosomal Dominant Polycystic Kidney Disease
AED	Emergency Admissions Data
AEMPS	Spanish Agency for Medicines
AP	Administrative and Patient Data
APC	Admitted Patient Care
ATC	Anatomical Therapeutic Chemical Classification System
A&E	Accident and Emergency
BIFAP	Spanish Database for Pharmacoepidemiological Research in the Public Sector
BIMCV	Valencian Community Biobank of Medical Images
BMI	Body mass index
CAPA	Corrective and Preventive Action
CCAA	Autonomous Communities
CDA-R2	Clinical Document Architecture Release 2
CDM	Common Data Model
CIS	Cancer Information System
CMBD/MBDS	Minimum Basic Data Set
CPD	Data Processing Centre
CPRD	Clinical Practice Research Datalink
CRC	Corporate Resource Catalog
DARWIN	Data Analysis and Real-World Interrogation Network
DQ	Data Quality
DHD	Dutch Hospital Data Foundation
DHDA	The Danish Health Data Authority
DHSC	Department of Health and Social Care
DEAP	Data Expert and Access Provider
DOI	Digital Object Identifier
DWH	Data Warehouse
EHR	Electronic Health Records
EMR	Electronic Medical Records
FEDRA	Spanish Farmacovigilance-Data on Adverse Reactions
FP	Family Pediatrician
FSE	Fascicolo Sanitario Elettronico
GDPR	General Data Protection Regulation
GIE	Comprehensive Study Management
GP	General practitioner
ICNARC	Intensive Care National Audit & Research Centre
ICD	International Classification of Diseases
ICD-CM	International Classification of Diseases-clinical modification
ICD-DA	International Classification of Diseases-Danish modification
ICPC	International Classification of Primary Care
ICU	Intensive Care Unit
IKNL	Netherlands Comprehensive Cancer Organisation
IMP	Index for Multiple Deprivation
IJE	International Journal of Epidemiology
IT	Information Technology
HCE	Healthcare Electronic Archive
HES	Hospital Episode Statistics
HL7	Health Level Seven
KPI	Key Performance Indicator
LBZ	Netherlands National Basic Register of Hospital Care
LOINC	Logical Observation Identifiers Names and Codes
LOPDDD	Organic Law on Personal Data Protection and Guarantee of Digital Rights
MACE	Major Adverse Cardiovascular Event

	If data availability or quality limits the study's feasibility to address the research question
	If the limitation may compromise the internal validity of the study
	If the limitation affects the sample size estimation
	If key subgroups (e.g., treatment-naive patients) are underrepresented
Potentially major:	If limitations are major, but have a solution or reasonable approach
Minor:	If the impact on the conclusions is small and can be mitigated
	If the issue is easily addressable or quantifiable
	If the impact on study timelines is limited and doesn't interfere with the key analyses
	If the limitation affects non-critical aspects of the study design, like data extraction procedures or non-essential controls
	If the limitation can be addressed using secondary analyses
	If the study sample is mostly representative but some minor groups are underrepresented

MHDA	Drugs Hospitalaries Dispensated in Ambulatory
MPID	Medicinal Product Identification
NCDM	Nordic Common Data Model
NCDR	National Cancer Data Repository
NCMP	Nordic Classification of Medical Procedures
NCR	National Cancer Registration (integraal kankercentrum Nederland)
NCRP	Nordic Classification of Radiological Procedures
NCRAS	National Cancer Registration and Analysis Service
NCSP	Nordic Classification of Surgical Procedures
NIHR	National Institute for Health and Care Research
NHS	National Health Service
NL	Netherlands
NOMESCO	Nordic Medico-Statistical Committee
NPU	Nomenclature, Properties, and Units
OMOP	The Observational Medical Outcomes Partnership
ONS	Office for national Statistics
OP	Outpatient
OTC	Over the Counter
PALGA	Netherlands National Pathology Database
PCP	Primary Care Physician
PDR	Patient Data Repository
RAE	Spanish Specialized Care Activity Registry
RDG	Research Data Governance
RKKP	Danish Clinical Quality Registries
RWD	Real-World Data
SFPT	Secure File Transfer Protocol
SIDIAP	Catalonia Information System for the Development of Research in Primary Care
SIERCV	Valencian Community Rare Disease Epidemiological Information System
SIP	Population Information System
SIV	Vaccination Information System
SLA	Service Level Agreement
SNOMED	Systematized Nomenclature of Medicine
SOP	Standard Operating Procedure
STIZON	Netherlands Foundation for Information Provision for Care and Research
THL	Finnish Institute of Health and Welfare
TRE	Trusted Research Enviroment
VID	The Valencia Health System Integrated Database

Step 1. DK Reg (Denmark)

Item		Sub-item	Description	Origin of information	Maturity level grade	Maturity level criteria and definitions	Rationale
0	Data base identification	Country	Denmark	https://english.sundhedsdatastyrelsen.dk/	*N/A*	*N/A*	*N/A*
		Data Access Provider	Danish Health Data Authority and Statistics Denmark	https://english.sundhedsdatastyrelsen.dk/			
		Organisation type	Regulatory Authority Works to ensure better health for the Danish citizens through the use of data and by creating digital coherence in the healthcare sector.	https://catalogues.ema.europa.eu/institution/3331256			
I	Rationale and scope for the RWD source creation	Primary purpose for which data are collected	Denmark has a large network of population-based medical databases containing routinely collected data, covering many aspects of life and health. Covers all patients from birth to death, across all hospitals and medical clinics in the country. During decades, register data covering the total Danish population from cradle to grave have been collected. Most of this information has been collected for administrative purposes. However, Danish legislation allows for researchers to utilise data for research of general relevance and importance. Denmark has a tax-based universal	https://ncrr.au.dk/danish-registers	*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 is collected from all hospitals and medical clinics in the country.	https://doi.org/10.2147/CLEP.S179083 https://healthcaredenmark.dk/national-strongholds/digitalisation/collection-and-sharing-of-health-data/		*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/.*	
		What triggers a record in the database	Event triggering registration of a person in the data source: Birth, Immigration Event triggering de-registration of a person in the data source: Death, Immigration Event triggering creation of a record in the data source: Danish registries is a set of tables with different events triggering a record in each table depending on the purpose of the registry	https://catalogues.ema.europa.eu/node/991/data-flows-and-management		*L3 if the information is provided as Metadata (machine readable), including standard formats, clear definitions and potentially some quality information*	
		Publications describing this RWD	https://doi.org/10.2147/CLEP.S179083				
II	Data collection or recording process	Description of data provider (geographical and oragnizational setting, nature of the data - reported by patients, HCP, etc)	The Danish Health Authorities provide systematic health data on volume of activity, economics, and quality of care for patients, health care professionals, researchers, and administrative staff. Maintains a wide range of medical databases. Sets national standards for digitization and data security, promotes coherent IT architecture within the health care system, and ensures availability of relevant and valid health data to benefit patient treatment and research. The EHR system is decentralised, and there are two different EHR systems used across the country.	https://doi.org/10.2147/CLEP.S179083	*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 Danish Health Data Authority is the body responsible for conceptualising and implementing health data governance. • A National Board for Health Data allows shared decision-making and strategy setting. • The data is stored at three key agents: the Danish Health Data Authority, the Danish Clinical Quality Registries (RKKP), and Statistics Denmark. • The biobanks and National Genome Center store biological material and genomic information. • The Regions are responsible for storing electronic health record (EHR) data in regional data warehouses. • Coverage of EHRs is complete, and healthcare providers are legally obliged to report to the regional data warehouses. Data is not exchanged between the two EHR systems directly, however, healthcare professionals are able to view their patients' EHR via the E-Journal, including data from other regions. Further detailed information on data storage, data management processes, architecture, key actors and overall governance can be found in the links.	https://tehdas.eu/app/uploads/2023/03/denmark-country-visit-factsheets-10-2022.pdf https://www.ehalsomyndigheten.se/globalassets/ehm/3_om-oss/sa-jobbar-vi-med-e-halsa/denmark_the-danish-framework-for-healthdata-6-1.pdf https://english.sundhedsdatastyrelsen.dk/digital-health-solutions/it-architecture https://english.sundhedsdatastyrelsen.dk/cyber-security www.ehalsomyndigheten.se/globalassets/ehm/3_om-oss/sa-jobbar-vi-med-e-halsa/denmark_the-danish-framework-for-healthdata-6-1.pdf		*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	Registrators: The GP, The hospital, The Pharmacy, The municipality Responsability: The Danish Health Data Authority Users: The health authorities, The health care system, The research, the public	https://www.ehalsomyndigheten.se/globalassets/ehm/3_om-oss/sa-jobbar-vi-med-e-halsa/denmark_the-danish-framework-for-healthdata-6-1.pdf			
		Key data elements captured (are they always recorded, are they optional, is there a planned coverage over time, ...)	Dispensing register, Patient register (inpatient and outpatient contacts, psychiatry included, diagnoses and procedures), Cancer register, Cause of death register, Birth register, Laboratory database, Vaccinations (dose and manufacturer available for covid-19 vaccines). All linked by unique Person Number, including other administrative data: death, migration, socioeconomic information (income, education)	Submission_ROC19_Annex V_Response template https://doi.org/10.2147/CLEP.S179083		*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)* *L2 if a structure checklist and dataset version control are available* *L3 is only aspirational. N/A*	*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				
		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)	Specific softwares have not been found. More details available in the link. Data checks might be done manually but researchers may program cleaning and quality checks in open source softwares (usually SASS and currently translated to R).	https://english.sundhedsdatastyrelsen.dk/Media/638657844560257530/Reference%20Architecture%20Sharing%20documents%20images.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	The Danish Health Data Authority (DHDA), shall approve standards, including data standards, classifications and interface standards, for IT applications in the health sector upon consultation with the national board of health IT.	https://english.sundhedsdatastyrelsen.dk/digital-health-solutions/it-architecture		*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 networks that are used to communicate between subsystems, whether these are cabled or wireless, often themselves provide some degree of protection against unauthorised access to data. However, the network itself cannot be expected to provide the required protection. Furthermore, since data can be sensitive, it may therefore be necessary to ensure additional protection.	https://english.sundhedsdatastyrelsen.dk/digital-health-solutions/it-architecture		*L3 N/A*	
V	Data management and governance	Data management principles being followed (e.g., GCP, ISO, FAIR, etc)	Data storage, key actors and overall governance are synthetized in the following link: www.ehalsomyndigheten.se/globalassets/ehm/3_om-oss/sa-jobbar-vi-med-e-halsa/denmark_the-danish-framework-for-healthdata-6-1.pdf The Danish Processing of Personal Data Act (persondataloven) The Danish Health Services Act (sundhedsloven) The Danish Medicines Act and the Danish Medical Devices Act (lægemiddelloven and lov om medicinsk udstyr) The Danish Social Services Act (serviceloven) Employment law regulations on control measures Registries undergo standard quality procedures at the data custodian. A metadata catalogue is being developed by the initiative 'Research Health Data Gateway' (En Indgang til Sundhedsdata). The metadata model being used in the is based on DCAT and ISO/IEC11179 and DCAT-AP DK OPEN DL. BEK nr 1695 af 14/12/2023 Executive Order no. 160 of 12 February 2013 on Standards for IT application in the Health Sector	https://catalogues.ema.europa.eu/node/991/administrative-details https://english.sundhedsdatastyrelsen.dk/Media/638657844593738618/Object%20Locating%20and%20Identification%201.0.4.3_en%20Public.pdf https://tehdas.eu/app/uploads/2023/03/denmark-country-visit-factsheets-10-2022.pdf https://english.sundhedsdatastyrelsen.dk/Media/638658829674899485/Executive%20Order%20on%20Standards%20for%20IT%20application.pdf	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)	Data storage, key actors and overall governance are synthetized in the following link: www.ehalsomyndigheten.se/globalassets/ehm/3_om-oss/sa-jobbar-vi-med-e-halsa/denmark_the-danish-framework-for-healthdata-6-1.pdf Quality control mechanisms in place include: reporting guidelines, training at point of collection, mandatory fields at data input point, validation of data at point of reception and feedback loops. The workflow, interoperability, architecture, error management, locating, and other data-related procedures are further detailed in the links. The networks that are used to communicate between subsystems, whether these are cabled or wireless, often themselves provide some degree of protection against unauthorised access to data. However, the network itself cannot be expected to provide the required protection. Furthermore, since data can be sensitive, it may therefore be necessary to ensure additional protection.	https://tehdas.eu/app/uploads/2023/03/denmark-country-visit-factsheets-10-2022.pdf https://english.sundhedsdatastyrelsen.dk/Media/638657844593738618/Object%20Locating%20and%20Identification%201.0.4.3_en%20Public.pdf		*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)	The aim of the Danish strategy for cyber and information security in the healthcare sector is to build the healthcare sector's capability and capacity to predict, prevent, detect and manage cyber and information security incidents by means a number of initiatives. The networks that are used to communicate between subsystems, whether these are cabled or wireless, often themselves provide some degree of protection against unauthorised access to data. However, the network itself cannot be expected to provide the required protection. Furthermore, since data can be sensitive, it may therefore be necessary to ensure additional protection. If a violation of data security is detected, the research institute is temporarily banned from accessing data for a certain amount of time. As stated under auditing documents, it should be possible to register which object locating systems produce data and, then, which applications consume this data. Periodic reviews (i.e. automatic reviews) of these log events can assist in detecting security issues, faulty allocation of rights, etc. Which systems produce data can be registered at low cost as a part of ensuring the persistence of location events. Registration of who uses the individual event data will lead to considerably larger overheads. The type of queries performed by the applications can be registered instead. In particular, in Statistics Denmark there is a secure server where researchers work. Download of files can be requested byt a confidentiality check needs to be performed before. Individual data must never exit the secure server. In Danish Health Data Authority, you need to login an intranet to work with the data. Download of files can be requested byt a confidentiality check needs to be performed before (to ensure there is no microdata). Only for exceptional cases microdata can be downloaded. In both cases, data is linked (if necessary for the research project) and anonymised. Counts <5 are masked as a rule. Data protection steps are illustrated here: www.ehalsomyndigheten.se/globalassets/ehm/3_om-oss/sa-jobbar-vi-med-e-halsa/denmark_the-danish-framework-for-healthdata-6-1.pdf	https://english.sundhedsdatastyrelsen.dk/cyber-security https://tehdas.eu/app/uploads/2023/03/denmark-country-visit-factsheets-10-2022.pdf https://english.sundhedsdatastyrelsen.dk/health-data-and-registers/research-services/the-secure-research-platform https://english.sundhedsdatastyrelsen.dk/health-data-and-registers/research-services https://www.dst.dk/Site/Dst/SingleFiles/GetArchiveFile.aspx?fi=83605109272&fo=0&ext=formid#:~:text=Statistics%20Denmark%20aims%20to%20maintain,IT%20and%20general%20financial%20resources. https://english.sundhedsdatastyrelsen.dk/health-data-and-registers/research-services/the-secure-research-platform			
		Auditing and DQ improvement procedures in place	Medical record review is a common reference standard used in validation studies to confirm the presence or absence of a disease. Other types of reference standards include patient self-reports, physicians' reports, autopsy reports, and alternative data sources with presumably higher data quality (such as clinical quality data, laboratory data, and pathology data). In general, data in health and clinical quality databases have high validity and completeness, because they are collected prospectively with the aim of quality control and clinical care. Consultants within each medical field register the data in clinical quality databases, further increasing the accuracy of the data. Although also registered prospectively, the validity of clinical data in the administrative databases may vary considerably among databases and within each database. Government-initiated systematic validation of personal demographic data, hospital admission data, and overall diagnoses within different clinical specialties. Investigator-driven systematic validation of individual diagnoses, examinations, procedures, and surgery codes within a specific clinical specialty Investigator-driven ad hoc validation of study-specific variables, the most common type of validation study As stated under auditing documents, it should be possible to register which object locating systems produce data and, then, which applications consume this data. Periodic reviews (i.e. automatic reviews) of these log events can assist in detecting security issues, faulty allocation of rights, etc. Which systems produce data can be registered at low cost as a part of ensuring the persistence of location events. Registration of who uses the individual event data will lead to considerably larger overheads. The type of queries performed by the applications can be registered instead.	https://doi.org/10.2147/CLEP.S179083 https://bmjopenquality.bmj.com/content/14/1/e003019 https://english.sundhedsdatastyrelsen.dk/Media/638657844593738618/Object%20Locating%20and%20Identification%201.0.4.3_en%20Public.pdf		*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	Health data in Denmark is updated in a timely manner e.g., it takes about 24 hours for EHR data to be sent to the regional data warehouse. Other register data have an update frequency between 1 day and 6 months.	https://tehdas.eu/app/uploads/2023/03/denmark-country-visit-factsheets-10-2022.pdf https://catalogues.ema.europa.eu/node/1145/administrative-details	*2*	*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 approximation 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	Data is harmonised when enters from the danish health authorities to Statistics Denmark. Values or missings are NOT imputed. Internal quality checks are conducted, but they are not acknowledged. Data is validated (unknown, not disclosed) and transformed in a time frame of 3 month approximately before being accessible for research use. In this process, variable names might be changed, especially for old data. This might be the reason for some discrepancy between Statistics Denmark and Danish Health Data Authorities. In case an error is found in the data, researchers are informed, and version of these data is traced. Information is spread to all the users of the system (so those having a login). In the linked document, administration, locating, filtering, integrity and error management of the data can be found. Also, interoperability information is present.	https://english.sundhedsdatastyrelsen.dk/Media/638657844593738618/Object%20Locating%20and%20Identification%201.0.4.3_en%20Public.pdf https://www.dst.dk/en/TilSalg/data-til-forskning https://www.dst.dk/en/TilSalg/data-til-forskning/danmarks-datavindue		*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	As stated under auditing documents, it should be possible to register which object locating systems produce data and, then, which applications consume this data. Periodic reviews (i.e. automatic reviews) of these log events can assist in detecting security issues, faulty allocation of rights, etc. Which systems produce data can be registered at low cost as a part of ensuring the persistence of location events. Registration of who uses the individual event data will lead to considerably larger overheads. The type of queries performed by the applications can be registered instead. Erroneous registrations, such as inaccurate positioning, can have enormous consequences in the applications. Therefore, it is important that these systems are aware of erroneous registrations. Some of the errors can be corrected in Layer 3 and some cannot be corrected until Layer 4, in which more information is available, e.g. information on how hospital beds can move. Erroneous registrations can be due to human error or inaccurate sensors. Such error events can entail a number of consequences in the affected applications; consequences that can only be corrected in the applications in question. Thus, it is important that a functionality for error correction exists. For example, the position of a hospital bed can lead to the bed being registered as "being cleaned". If this is due to an erroneous registration in the positioning system, there must be a way to correct the status of the bed.	https://english.sundhedsdatastyrelsen.dk/digital-health-solutions/it-architecture https://english.sundhedsdatastyrelsen.dk/cyber-security		*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)	Requested to DEAP and unable to provide		*N/A*		
VII	Data augmentation steps (e.g., imputation or linkage)	Is any augmentation happening in this datasource?	Data can be lined to the country data set by ID. Also, to the Danish Clinical Quality Registries (RKKP), specific disease cohorts (cancer, depression, ADHD, surgery, cardiac arrest, among others). Additionally, whole families can be linked (mother-father-children).	https://www.esundhed.dk/Dokumentation https://doi.org/10.2147/CLEP.S179083 Sundhedsvæsenets Kvalitetsinstitut - Kvalitetsdatabaser https://www.sundk.dk/about/ https://www.sundk.dk/kliniske-kvalitetsdatabaser/	*1*	*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	Deterministic			*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*	
		If yes, which algorithms and assumptions applied	Requested to DEAP and unable to provide		*N/A*		
		If yes, which is the error rate when conducting the augmentation	N/A		*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)	OTC drugs not available. Inpatient dispensing data set just implemented, unknown validity and with limited accessibility for now. No access to GP visits (contacts can be seen but not the reasons, prescribign, diagnoses, etc). No immigrations record, but from rpevious experience this represents minor losses. Erroneous registrations, such as inaccurate positioning, can have enormous consequences in the applications. Therefore, it is important that these systems are aware of erroneous registrations. Some of the errors can be corrected in Layer 3 and some cannot be corrected until Layer 4, in which more information is available, e.g. information on how hospital beds can move. Erroneous registrations can be due to human error or inaccurate sensors. Such error events can entail a number of consequences in the affected applications; consequences that can only be corrected in the applications in question. Thus, it is important that a functionality for error correction exists. For example, the position of a hospital bed can lead to the bed being registered as "being cleaned". If this is due to an erroneous registration in the positioning system, there must be a way to correct the status of the bed. In the following link, relevance, accuracy and reliability are reported for data from Denmark on: People, Labour and income, Economy, Social conditions, Education and research, Business, Transport, Culture and leisure, Environment and energy, and Hospitalization	https://english.sundhedsdatastyrelsen.dk/Media/638657844593738618/Object%20Locating%20and%20Identification%201.0.4.3_en%20Public.pdf https://www.dst.dk/en/Statistik/dokumentation/documentationofstatistics	*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 EWD source	Government-initiated systematic validation https://bmjopen.bmj.com/content/6/11/e012832.abstract https://www.tandfonline.com/doi/full/10.2147/CLEP.S332776 https://onlinelibrary.wiley.com/doi/10.1111/bcpt.12610			*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, NCDM (Nordic Common Data Model)	https://cataloques.ema.europa.eu/node/991/data-flows-and-management Submission_ROC19_Annex V_Response template https://www.ntnu.no/ojs/index.php/norepid/article/view/4053	*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 Cancer: ICD-O3 Dispensation of prescribed medicines: ATC Procedures: NOMESCO, NSCP, standard vocabulary in all Nordic countries Biomarker vocabulary: NPU for lab data Diagnosis / medical event vocabulary: ICD-10, ~~SNOMED for pathology data,~~ Danish version Medicinal product vocabulary: ATC	https://www.dovepress.com/article/download/47088		*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*	
X	The RWD source declared Service Level Agreements (SLA)	Guaranteed frequency of updates and incident response time (e.g., corrections in case of errors)	Requested to DEAP and unable to provide		*N/A*	*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	Videos of training webinars All new healthcare providers and staff members receive training to ensure data input is done correctly in the EHR system. • All regions have set up Regional Support Centres offering training and support to researchers for data access and analysis. • The Danish National Biobank offers a yearly course for PhD students on how to secure accessibility permissions and use the biobank samples efficiently. • The European Network Training Centre provides training on regulatory work. • Some institutes are establishing curricula on statistics and data analysis. Some training and capacity needs were identified: o Competencies and training skills to work with citizen-generated data (e.g., from wearables) o More training specific for healthcare staff on statistics and data analysis.	https://www.vejledningsfunktionen.dk/en/videos-of-webinars/ https://tehdas.eu/app/uploads/2023/03/denmark-country-visit-factsheets-10-2022.pdf	*2*	*L2 if details of established data processess by the provider are available*	
		Possibility to collect additional data if needed	It is possible to collect additional data. For example, when data is extracted from the database but more personal information is warranted, person identification number cvould be provided to link the extracted information, get the individuals' consent, and link especific surveyed information to the actual data.			*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 is only possible if the researcher is affiliated to an approved Danish research institute. • The regulatory framework for accessing and sharing data depends on the purpose for which data is requested. • The most important legal acts are: The Act on Research Ethics Review of Health Research Projects, The Health Act, and The Danish Data Protection Act. • GDPR is perceived to be interpreted differently between lawyers at national, regional and hospitallevel, sometimes causing challenges. • The need for ethical approval depends on the type of research project. For certain complex projects (e.g., extensive genome examinations without consent or stem cell research) the National Ethics Committee provides approval. This is also the case for certain data-based projects, where there is a risk of secondary findings. For other research projects it is the regional ethical committees that provide approval. University of Copenhaguen have a data processing agreement with the Danish Health Data Authority. The head of department is/are the data controllers. Each individual researching by using these data needs to take an exam on GDPR and confidentiality principles of Statistics Denmark. It needs to be renovated yearly. This certification procedure must happen regardless of the background training of the researcher. Guest researchers can also participate in such research with permission of the data controller (but the GDPR certification needs to be taken too). Within a department, Principal Investigators can be designated as referees of specific cohorts of data. Applicance for data can be done through this website: https://sundhedsdatastyrelsen.dk/data-og-registre/forskerservice/omlaegning-registre	https://tehdas.eu/app/uploads/2023/03/denmark-country-visit-factsheets-10-2022.pdf https://sundhedsdatastyrelsen.dk/data-og-registre/forskerservice/omlaegning-registre https://www.dst.dk/en/TilSalg/data-til-forskning/brugeradgang (user access information to the Denmark's Data Portal)	*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 N/A*	*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?	In case an error is found in the data, researchers are informed, and version of these data is traced. Information is spread to all the users of the system (so those having a login). A general email, phone and address are available. Also, in this website (https://sundhedsdatastyrelsen.dk/data-og-registre/forskerservice/omlaegning-registre), they offer requests for updated data and data conversion through email or phone.	https://sundhedsdatastyrelsen.dk/ https://sundhedsdatastyrelsen.dk/data-og-registre/forskerservice/kontakt	*3*	*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*

Step 1. FI Reg (FL) KANTA PDR

Item		Sub-item	Description	Origin of information	Maturity level grade	Maturity level criteria and definitions	Rationale
0	Data base identification	Country	Finland		*N/A*	*N/A*	*N/A*
		Data Access Provider	Finnish Social and Health Data Permit Authority Findata, with register maintainers (for national registers Finnish Institute of Health and Welfare (THL), Social Insurance Institution of Finland (Kela), Statistics Finland, DVV)	https://findata.fi/en/data/#what-data-are-available-via-findata//https://pubmed.ncbi.nlm.nih.gov/22899561/ // https://findata.fi/en/data/#what-data-are-available-via-findata		*N/A*	
		Organisation type	Government agency			*N/A*	
I	Rationale and scope for the RWD source creation	Primary purpose for which data are collected	Originally, these data have been collected for administrative purposes (since 1950s), but their use for research, regulatory, health policy and organisational planning is enabled by the act on secondary use of health and social data. Linkage across different registers is possible with personal identification numbers which are pseudonymised when the data are provided for research use	https://tehdas.eu/tehdas1/packages/package-4-outreach-engagement-and-sustainability/tehdas-country-visits/	*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	All residents			*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/.*	
		What triggers a record in the database	Birth, use of healthcare service, dispensing event, death	https://journal.fi/finjehew/article/view/146124/94799// https://pubmed.ncbi.nlm.nih.gov/34321928/		*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://pubmed.ncbi.nlm.nih.gov/34321928/				
II	Data collection or recording process	Description of data provider (geographical and oragnizational setting, nature of the data - reported by patients, HCP, etc)	The data on these national registers is recorded by health care providers. Dispensing events are recorded in pharmacies, healthcare (primary and specialised) by the healthcare providers. Death certificates are written by physicians.	https://pubmed.ncbi.nlm.nih.gov/34321928/ //https://journal.fi/finjehew/article/view/146124/94799/	*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 Finnish Population Information System is a computerised national register that contains basic information about Finnish citizens and foreign citizens residing in Finland on a permanent or temporary basis. The Kanta Services are a set of digital services that store citizens' social welfare and health care data. The use of this data makes it easier to manage affairs relating to your health and wellbeing. In addition, the data supports social welfare and health care providers in their decision making. The Kanta Services are a nationwide solution that cover all of Finland. Kanta is used by both public and private providers of health care and social welfare services, and by pharmacies. The statistics on causes of death are based on data derived from the death certificates that are complemented with data on deaths from the Population Information System of the Population Register. The statistics on causes of death include all deaths in Finland or abroad of persons permanently resident in Finland at the time of their death. Investigating the cause of death and the related procedures including the production of statistics and archiving of death certificates is based on the Act (1973/459) and Decree (1973/948) on the investigation of the cause of death.	https://dvv.fi/en/population-information-system https://stat.fi/meta/til/ksyyt_en.html https://thl.fi/en/research-and-development/thl-biobank/for-researchers/application-process/access-to-national-register-data https://thl.fi/en/statistics-and-data/data-and-services/register-descriptions https://www.kanta.fi/en/research-and-knowledge-management 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	The Digital and Population Data Services Agency promotes the digitalisation of society, maintains population data, secures the availability of data, and provides services for the life events of its customers. For the Care register for Healthcare, every year, data suppliers (healthcare and social care units) are provided with a manual on the data content and on how to submit care notifications. More information is available in links in this and above sections.	https://dvv.fi/en/population-information-system https://stat.fi/meta/til/ksyyt_en.html https://thl.fi/en/research-and-development/thl-biobank/for-researchers/application-process/access-to-national-register-data https://thl.fi/en/statistics-and-data/data-and-services/register-descriptions https://www.kanta.fi/en/research-and-knowledge-management Provided by DEAP			
		Key data elements captured (are they always recorded, are they optional, is there a planned coverage over time, ...)	Prescriptions (written and dispensed), inpatient and outpatient contatcs, including diagnoses, procedures and specialties, Cancer register, Cause of death register, Birth register, Laboratory database, Vaccinations, death dates and causes of death, migration, socioeconomic information (income, education) linkable via personal identification number. Different data is available since: - Hospitaisation data, cancer registry, medical births register, special reimbursements, since 1972, - Dispensed / reimbursed prescriptions since 1995, - All prescriptions since 2014 (with full coverage since 2017), - Secialized healthcare outpatient admissions since 1998, - Primary care since 2011, - Lab measurements since 2014.	https://pubmed.ncbi.nlm.nih.gov/34321928/ // https://journal.fi/finjehew/article/view/146124/94799/		*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)* *L2 if a structure checklist and dataset version control are available* *L3 is only aspirational. N/A*	*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		*N/A*		
		If yes: does it follow any specific framework? Is there an assessment checklist? Are datasets versions traceable?	N/A		*N/A*		
IV	The data management infrastructure	List of systems used to manage the RWD (either for data collection, recording, processing, etc)	All the patient data recorded to the Kanta PDR are stored as XML docu-ments using Health Level Seven (HL7) Clinical Docu-ment Architecture Release 2 (CDA R2) format	https://academic.oup.com/jamia/article-abstract/13/1/30/781314?redirectedFrom=fulltext	*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)	The end users naturally employ backups and sum checks when data extractions are disposed to them. However, further information of this process at other stages is not available for the DEAP. As much as the data is originating from billing systems (healthcare contacts), there are internal audits in place but there is no consultable documentation on this.	Provided by DEAP	*1*	*L3 N/A*	
V	Data management and governance	Data management principles being followed (e.g., GCP, ISO, FAIR, etc)	As per GDPR the data subjects have the legal right to request the correction of their information. Researchers / other data users can notify the permit authority on assumed errors in the data	Provided by DEAP	*1*	*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)	All the patient data recorded to the Kanta PDR are stored as XML docu-ments using Health Level Seven (HL7) Clinical Docu-ment Architecture Release 2 (CDA R2) format	https://academic.oup.com/jamia/article-abstract/13/1/30/781314?redirectedFrom=fulltext	*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)	All data are analysed in audted remote use environments.	https://findata.fi/en/kapseli/regulation-on-secure-operating-environments/ Provided by DEAP	*2*		

		Auditing and DQ imporvement procedures in place	Medical record review is a common reference standard used in validation studies to confirm the presence or absence of a disease. Other types of reference standards include patient self-reports, physicians' reports, autopsy reports, and alternative data sources with presumably higher data quality (such as clinical quality data, laboratory data, and pathology data). In general, data in health and clinical quality databases have high validity and completeness, because they are collected prospectively with the aim of quality control and clinical care. Consultants within each medical field register the data in clinical quality databases, further increasing the accuracy of the data. Although also registered prospectively, the validity of clinical data in the administrative databases may vary considerably among databases and within each database. Government-initiated systematic validation of personal demographic data, hospital admission data, and overall diagnoses within different clinical specialties. Investigator-driven systematic validation of individual diagnoses, examinations, procedures, and surgery codes within a specific clinical specialty Investigator-driven ad hoc validation of study-specific variables, the most common type of validation study As stated under auditing documents, it should be possible to register which object locating systems produce data and, then, which applications consume this data. Periodic reviews (i.e. automatic reviews) of these log events can assist in detecting security issues, faulty allocation of rights, etc. Which systems produce data can be registered at low cost as a part of ensuring the persistence of location events. Registration of who uses the individual event data will lead to considerably larger overheads. The type of queries performed by the applications can be registered instead.	Provided by DEAP	*2*	*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 causes of death annually	https://catalogues.ema.europa.eu/node/1094/data-flows-and-management Provided by DEAP	*2*	*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	Values or missings are NOT imputed. Internal quality checks are conducted, but they are not acknowledged. Data validation procedures vary accorss registers. Variable names might be changed, especially for old data.	Provided by DEAP	*1*	*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*	
		Information about loss of precision during data manipulation steps	Requested to DEAP and unable to provide		*N/A*	*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)	Requested to DEAP and unable to provide		*N/A*		
VII	Data augmentation steps (e.g., imputation or linkage)	Is any augmentation happening in this datasource?	Family relationship linkage with different registers, linkage to FinnGen biobank data possible	Provided by DEAP	*1*	*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	Pseudonymised personal identification number (deterministic)	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.,*	
		If yes, which algorithms and assumptions applied	Direct for personal information, family relation linkage for paternity	Provided by DEAP			
		If yes, which is the error rate when conducting the augmentation	Not available	Provided by DEAP	*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)	Individuals declared legally dead are not included in the number of deceased in the statistics on causes of death. The statistics will lack the cause of death for some 100 to 400 individuals every year, because they cannot be provided with a death certificate. In 2023, the statistics lacked a death certificate in 284 deaths, corresponding to 0.5 per cent of all deaths. The data pertaining to causes of death are produced annually and completed by the end of the following year. More data quality characteristics are described in the link.	https://stat.fi/en/statistics/documentation/ksyyt#Accuracy,%20reliability%20and%20timeliness https://stat.fi/en/statistical-data	*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 EWD source	In general, there is ongoing work on data quality and structuring of data across Finland. Stakeholders • Registry data at THL are validated and checked through basic automated checks, manual checks, comparison to previous years and feedback loops. More automated checks are being developed.	https://tehdas.eu/tehdas1/packages/package-4-outreach-engagement-and-sustainability/tehdas-country-visits/		*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	Provided by DEAP	*2*	*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)	Diagnoses: ICD-10, ICD-9, ICPC Drugs: ATC Procedures: NOMESCO	https://academic.oup.com/jamia/article-abstract/13/1/30/781314?redirectedFrom=fulltext https://journal.fi/finjehew/article/view/146124 https://zenodo.org/records/13384860			
X	The RWD source declared Service Level Agreements (SLA)	Guaranteed frequency of updates and incident response time (e.g., corrections in case of errors)	Requested to DEAP and unable to provide		*N/A*	*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				
		Possibility to collect additional data if needed	It is possible to collect additional data. For example, when data is extracted from the database but more personal information is warranted, person identification number cvould be provided to link the extracted information, get the individuals' consent, and link especific surveyed information to the actual data.	Provided by DEAP	*2*		
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	Permissions can be applied via the Nationalpermit authority Findata. Individual data can be analysed only within aufdited remote use envionment. Suggested link to cell E40-41: https://findata.fi/en/	https://findata.fi/en/data/#what-data-are-available-via-findata https://thl.fi/en/statistics-and-data/data-and-services/research-use-and-data-permits	*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 N/A*	*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?	Requested to DEAP and unable to provide		*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*

Step 1. PEDIANET

Item		Sub-item	Description	Origin of information	Maturity level grade	Maturity level criteria and definitions	Rationale
0	Data base identification	Country	Italy		*N/A*	*N/A*	*N/A*
		Data Access Provider	SoSeTe (Società Servizi Telematici)	https://catalogues.ema.europa.eu/institution/3331353		*N/A*	
		Organisation type	Primary Care Medical Records			*N/A*	
I	Rationale and scope for the RWD source creation	Primary purpose for which data are collected	Pedianet is a national population database that contains anonymous patient-level data of paediatric population who received healthcare from family paediatricians (FPs) in Italy who were part of the PEDIANET network. The network links FPs distributed throughout several Italian regions designated by the Italian NHS, including Friuli-Venezia Giulia, Liguria, Lombardia, Piemonte, Veneto, Lazio, Marche, Toscana, Abruzzo, Campania, Sardegna, and Sicilia. Primary Care and Pediatric specialist Records and vaccines from public health. The database is maintained and owned by the Società Servizi Telematici Srl. The maintenance of the database is funded through different research projects. Studies carried out to date have been financed by public bodies (European Commission, Istituto Superiore di Sanità, AIFA, Consiglio Nazionale delle Ricerche, Regione Veneto, Aziende Socio Sanitarie, Istituto Zooprofilattico delle Venezie, etc.), or private groups such as pharmaceutical companies or international research groups.	https://link.springer.com/chapter/10.1007/978-3-030-51455-6_13	*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	Informed consent is required from the parents. Only participating pediatricians. The data is recorded during the medical examination carried out by the pediatrician. Pedianet is a national population database that contains anonymous patient-level data of paediatric population who received healthcare from family paediatricians (FPs) in Italy who were part of the PEDIANET network. In Italy, there is a tax-funded public healthcare system with universal access, and patients do not incur direct costs related to primary care visits. Informed consent is required from children's parents to enter the data in the database.			*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/.*	
		What triggers a record in the database	Triggering registration of a person in the data source: practice registration Triggering de-registration: death, los to follow-up or practice deregistration Triggering a record in the data source: visit to participating pediatrician			*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://link.springer.com/chapter/10.1007/978-3-030-51455-6_13				
II	Data collection or recording process	Description of data provider (geographical and oragnizational setting, nature of the data - reported by patients, HCP, etc)	Region of Veneto, Italy. The role of Pedianet not just as a database (especially for pharmacovigilance studies) but as an organised structure in which different competencies converge is essential. Pedianet is an independent network of family paediatricians established in 1998 to collect information from outpatient family paediatricians in Italy for clinical and epidemiological research (e.g., pharmacovigilance studies, studies on prescribing patterns, and studies of the efficiency of health services). The Pedianet database's beginning dates back to January 2000. The Pedianet system has the advantage of collecting data at a population level as a by-product of routine activities, therefore generating a far larger quantity of data than ad hoc studies.	https://link.springer.com/chapter/10.1007/978-3-030-51455-6_13	*1*	*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	Requested to DEAP and unable to provide		*N/A*	*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	Requested to DEAP and unable to provide		*N/A*		

		Key data elements captured (are they always recorded, are they optional, is there a planned coverage over time, ...)	Data collection started in January 2000 with a 3-year test period and is still ongoing. The data include demographic information as well as information on inpatient diagnoses (48,000,001), drug prescriptions (31,500,001), anthropometric measures (16,400,001), specialist medical examinations (12,000,001), and physical examinations or lab tests (16,000,001). Demographic data include year of birth, age, sex, region of residence, nationality, and information about the parents (e.g., nationality, smoking habits, educational level of the mother and the socioeconomic level, were recorded). In addition, the type of breastfeeding at 1, 3, 6, 9, and 12 months after birth, parity, Apgar score at 1, 5, and 10 min after birth, gestational age, birth weight, birth height, jaundice and family illnesses are recorded. Additional information on the health status of the mother may also be available but is not routinely documented. Date and cause of death of a patient are also recorded in the Pedianet database. Information on outpatient diagnoses and symptoms includes primary and ancillary diagnoses, the date of diagnosis, and diagnostic certainty. Diagnoses are coded using the International Classification of Diseases, version 9 (ICD-9) with at least four digits. Outpatient prescriptions, treatment, including immunizations, and diagnostic procedures (laboratory tests and physical examinations) are also recorded. Outpatient prescription information, both for reimbursed and non-reimbursed drugs, includes the date of prescription and the date of dispensation, the indication, the ATC code, the Italian MinSan code, the number of prescribed packages, and the dose prescribed. Information about physical examinations and laboratory tests is generally documented, including the measured value, the date, and if necessary, the reason for performing the examination or test.	https://link.springer.com/chapter/10.1007/978-3-030-51455-6_13	*1*	*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)* *L2 if a structure checklist and dataset version control are available* *L3 is only aspirational. N/A*	*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)	The Pedianet database has been registered according to the Italian law. The network links FPs distributed throughout several Italian regions designated by the Italian NHS, including Friuli-Venezia Giulia, Liguria, Lombardia, Piemonte, Veneto, Lazio, Marche, Toscana, Abruzzo, Campania, Sardegna, and Sicilia, and who use the same software (Junior Bit®) (Padova, Italy) in their professional practice.	https://link.springer.com/chapter/10.1007/978-3-030-51455-6_13	*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	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)	Requested to DEAP and unable to provide		*N/A*	*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)*	*Data management and governance impact reliability, as well as all quality dimensions for metadata.*
		Data management processes in place (DQ controls, KPIs, SOPs, etc)	Pedianet routinely checks their data by: 1. Checking the correspondence of data with information on the child's medical record 2. Contacting the treating doctor to ask for patient information and comparing this information with that in the Pedianet database 3.Checking for outliers 4.Ensuring that the medication prescribed to a patient is consistent with the diagnosis 5.Ensuring that the results of medical tests performed are consistent with the diagnoses. The data collected from the child's parents/tutors by paediatricians enters the dedicated cloud already encrypted and anonymised. Pedianet researchers do not know the process to anonymise the data and cannot know the owner of the data in any way.	https://link.springer.com/chapter/10.1007/978-3-030-51455-6_13	*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)	The data collected from the child's parents/tutors by paediatricians enters the dedicated cloud already encrypted and anonymised. Pedianet researchers do not know the process to anonymise the data and cannot know the owner of the data in any way.				
		Auditing and DQ improvement procedures in place	In regards of DQ improvements, in the near future, data will be directly transferred to the Pedianet database from the electronic health records (the so called Fascicolo Sanitario Elettronico, FSE), which collects all hospitalization, medical tests, and examinations done. From then, the validation of these data is no longer necessary, or will at least be faster.	https://link.springer.com/chapter/10.1007/978-3-030-51455-6_13		*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	https://catalogues.ema.europa.eu/node/1128/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	Requested to DEAP and unable to provide		*N/A*	*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	Requested to DEAP and unable to provide			*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)	Requested to DEAP and unable to provide				
VII	Data augmentation steps (e.g., imputation or linkage)	Is any augmentation happening in this datasource?	It is possible to identify family members (brothers and sisters) within the database. Veneto paediatricians can integrate their outpatient routine clinical care data with data from the regional vaccination registry, hospital admissions, emergency department visits, and the COVID-19 swab registry. They can access the patient's Electronic Health Records and integrate these data into their pediatric primary care database. Only children in Friuli Venezia Giulia can be linked to the immunization registry.	https://link.springer.com/chapter/10.1007/978-3-030-51455-6_13 https://catalogues.ema.europa.eu/node/1128/data-flows-and-management#darwin-data-source-linkage	*1*	*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	Sibiling	https://catalogues.ema.europa.eu/node/1128/quantitative-descriptors		*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	Requested to DEAP and unable to provide		*N/A*		
		If yes, which is the error rate when conducting the augmentation	Requested to DEAP and unable to provide			*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 major strength of the Pedianet database is its size. The database counts 265,000 Italian children aged from 0 to 16 years old. The main limitations of the database include the lack of information of the family environment, an underreporting of hospitalization and immunization data, and the lack of many OTC prescriptions. This makes Pedianet not fully appropriate for studies on non-specific outcomes. However, these studies can be performed through prospective observational cohort studies which have been successfully carried out using the Pedianet network.	https://link.springer.com/chapter/10.1007/978-3-030-51455-6_13	*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	Asthma, acute otitis media and liver injury diagnoses have been previously validated. Pedianet also informs internal procedures to validate their data: https://link.springer.com/chapter/10.1007/978-3-030-51455-6_13	https://www.sciencedirect.com/science/article/pii/S0264410X20301535?via%3Dihub https://pubmed.ncbi.nlm.nih.gov/12674038/ https://pubmed.ncbi.nlm.nih.gov/23190626/ https://pubmed.ncbi.nlm.nih.gov/28025733/		*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 (ETL frequency: every 4 months) OMOP (In progress, see link)	https://catalogues.ema.europa.eu/node/1128/data-flows-and-management#darwin-data-source-linkage	*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)	ICD-9, ATC, AIC, Free-text	https://catalogues.ema.europa.eu/node/1128/data-flows-and-management#darwin-data-source-linkage		*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*	
X	The RWD source declared Service Level Agreements (SLA)	Guaranteed frequency of updates and incident response time (e.g., corrections in case of errors)	Requested to DEAP and unable to provide		*N/A*	*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			*L2 if details of established data processess by the provider are available*	
		Possibility to collect additional data if needed	Patients can be re-contacted through the participating paediatrician to gather additional information.	https://link.springer.com/chapter/10.1007/978-3-030-51455-6_13	*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	Data are included in the database only after written informed consent is obtained from the parents of the child. Data are collected anonymously on a central server in Padua, where it is validated and prepared for research. Access to the database is allowed only for Pedianet researchers in the context of research projects that have been approved by both the Steering Committee and the Ethics Review Board (if required). It is not permitted to give third parties access to the data. Patient level data cannot be shared, however aggregated data may be shared with research partners, e.g., for pooled analysis. The coordination of the projects and data analysis is carried out by a scientific committee that includes internationally well-known paediatricians, epidemiologists, and researchers.	https://link.springer.com/chapter/10.1007/978-3-030-51455-6_13	*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 N/A*	*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?	General contact available: https://pedianet.it/	https://pedianet.it/	*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*

Step 1. CPRD

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/.*	*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		*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/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 inclusion/excl usion (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*	
		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)*	*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*	*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*	*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*	*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	Requested to DEAP and unable to provide			*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)	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*	*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		*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*	
		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*	*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* *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* *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*

Step 1. NCR

Item		Sub-item	Description	Origin of information	Maturity level grade	Maturity level criteria and definitions	Rationale
0	Data base identification	Country	the Netherlands		*N/A*	*N/A*	*N/A*
		Data Access Provider	Netherlands Comprehensive Cancer Organisation (IKNL)			*N/A*	
		Organisation type	Quality institute for oncological and palliative research and practice. National, regional, or municipal public founding	https://iknl.nl/en/about-iknl		*N/A*	
I	Rationale and scope for the RWD source creation	Primary purpose for which data are collected	The main goal of the Netherlands Comprehensive Cancer Organisation (IKNL) is to reduce the impact of cancer, from the personal to the societal level. With the Netherlands Cancer Registry (NCR) as its core activity, IKNL enables health care professionals, researchers, policy makers and others to reflect on cancer and on palliative care. Together with care professionals, researchers, patients, and policy makers we translate data into valuable insights to improve oncological and palliative care.	https://iknl.nl/en/about-iknl	*1*	*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	The NCR compiles clinical data of all individuals newly diagnosed with cancer in the Netherlands. Hospital inpatient care, hospital outpatient care.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP		*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/.*	
		What triggers a record in the database	Having performed a biopsy through PALGA (the national pathology database) Having a cancer diagnosis in LBZ **Event triggering registration of a person in the data source**: having performed a biopsy through PALGA (the national pathology database) and having a cancer diagnosis in LBZ **Event triggering de-registration of a person in the data source**: Persons are de-registered when they request this (they work with an opt-out system so everyone is included, and everyone can request to be taken out of the database), when they emigrate or die. **Event triggering creation of a record in the data source:** A group of data managers daily screen for new information of the patients registered in the data source. Persons (or actually tumors) are triggered as described in "event triggering registration of a person in the data source". We then register the relevant data for this person (tumor) after a set amount of time (typically 6-12 months after diagnosis). When the person develops another primary tumor, they go through the same process for that new tumor. Registration of patients is typically done about a year after the incidence date (the exact lag depends on the type of cancer). The vital status of patients is checked once per year.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP		*L3 if the information is provided as Metadata (machine readable), including standrad formats, clear definitions and potentially some quality information*	
		Publications describing this RWD	Not found PubMed, Google free search				
II	Data collection or recording process	Description of data provider (geographical and oragnizational setting, nature of the data - reported by patients, HCP, etc)	IKNL is a knowledge institute that is mostly government funded (by the Ministry of Health, Welfare and Sport)	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP	*1*	*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 is collected by well-trained data managers using coding manuals. The data entry application performs checks on the data that is entered, automatic checks are done on the database, as well as manual checks of random samples. A group of data managers is responsible for data quality and researchers in the organization can flag potential quality issues.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). 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 managers receive regular training. See also 'SOPs recording'. SOPs are kept up to date by discussing the items in the IKNL tumor boards, they are implemented in daily practice by regular training of the data managers (there is a training program during onboarding and frequent data managers meetings thereafter, data managers are also informed about changes in the SOPs by email. The most recent version of the SOPs are available online and are used during registration. Quality checks also include cross checks by direct colleagues.	Provided by DEAP			
		Key data elements captured (are they always recorded, are they optional, is there a planned coverage over time, ...)	Disease information, rare diseases, prescriptions of medicines, indication for use, procedures, clinical measurements, patinet-reported outcomes, unique identifier persons, diagnostic code, medicinal product information, quality of life measurements, sociodemographic information (age, gender). These are items grouped by: patient, tumor and treatment. I removed the PROMs/QoL as they are only captured on project basis. Coverage over time: there is a planned delay of 9 months delay (as data managers only have to access the EHR once per patient to capture the primary treatment plan), but in practice this is 1 to 2 years. Note: NCR only covers the primary plan, there is no information on follow-up (like PFS or secondary treatment -> I see this is addressed in row 31). The day of death is known by linkage to CBS. TNM recorded	Provided by DEAP (Always recorded: overview on 240918-itemset-long.pdf Unfortunately only in Dutch.)		*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	NA		*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/exclusion (I/E) criteria)*
		Is there a DQ assessment for data sources onboarded?	NA			*L2 if a structure checklist and dataset version control are available* *L3 is only aspirational. NA*	
		If yes: does it follow any specific framework? Is there an assessment checklist? Are datasets versions traceable?	NA				
IV	The data management infrastructure	List of systems used to manage the RWD (either for data collection, recording, processing, etc)	Data is collected manually from the EMR systems of the hospitals by data managers and entered into a database. IKNL uses its own application for this (RANK). The changes in the database are loaded into a datawarehouse (DWH) every night. The DWH keept a history of the registration, performs transformations on the data, etc.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP	*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	RANK is developed and maintained by the in-house Software Development department. They perform testing and quality control as well. The same applies to the DWH. This is based on a commercial application (from Microsoft).	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)	A history of the data is kept in a DWH. Backups are made as well each day.	Provided by DEAP		*L3 NA*	
V	Data management and governance	Data management principles being followed (e.g., GCP, ISO, FAIR, etc)	There are ongoing activities to make the NCR more FAIR. For example through introduction/use of (more) international standards (such as ATC, ICHI, SNOMED-CT).	Provided by DEAP	*1*	*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)	Data is collected by well-trained data managers using coding manuals. The data entry application performs checks on the data that is entered, automatic checks are done on the database, as well as manual checks of random samples. A group of data managers is responsible for data quality and researchers in the organization can flag potential quality issues.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP		*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)	IKNL has an IT department that is responsible for cyber security. There are also Information Security Officers that monitor this.	Provided by DEAP			

		Auditing and DQ imporvement procedures in place	IKNL is NEN-7510 certified. Quarterly internal audits are performed, as well as regular external audits. There is also a working group responsible for DQ. They perform checks on the data. Researchers can also signal potential DQ issues.	Provided by DEAP		*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	The data in the NCR (i.e. the data in the DWH) is updated daily (by overnight loading of the changes in the RANK database). Disease diagnosis: daily through PALGA (the national pathology database); remaining patients (those that do not receive a biopsy) are found by a yearly coupling with LBZ (https://www.dhd.nl/producten-diensten/registratie-data/ontdek-de-mogelijkheden-van-de-lbz). Other data: 6 months-1 year (depending on tumor type) after identification of a new primary tumor, a datamanager collects additional data around diagnosis and treatment from the EHRs of the hospitals where the patient was diagnosed and treated. The data in the OMOP-CDM is updated a few times per year.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP	*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	Data is registered manually by highly trained data managers so data that is entered into the NCR is already of high qualitym so cleaning is not required. Data transformations performed in the DWH are mainly creation of new variables (derived from existing NCR variable) and handling of changes of variables (or definition of variables) over time. A team consisting of data base administrators and experts on the NCR data do this in close collaboration with the clinical experts (tumor teams).	Provided by DEAP		*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 general, there are no manipulation steps that cause loss of precision. There may be loss of precision (loss of information) in the creation of specific variables. However, the original variables are also part of the NCR.	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)	A history of the DWH is maintained. Data manipulation that is performed by scripts has an accompanying justification.	Provided by DEAP			
VII	Data augmentation steps (e.g., imputation or linkage)	Is any augmentation happening in this datasource?	There are no data augmentation steps.	Provided by DEAP	*1*	*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	N/A			*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	N/A				
		If yes, which is the error rate when conducting the augmentation	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)	Cause of death not included. Comorbidity and cardiac events. Inclusion of only first line treatments. No data set has all information about a patient, you always need to make choices about what to collect. There is no cause of death and only first line treatment is registered. There is no registration of side effects and data about comorbidities is very incomplete. There are other minor quality issues in variables (sometimes variables are only registered in certain regions, or registration was nog mandatory in certain time periods so the data is incomplete). The department that handles data request knows these issues and they are communicated with the person that requests the data if this affects their research. There is no single overview of these DQ issues. NO information on recurrences	Provided by DEAP	*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	Possibility of data validation	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP		*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, ETL completed. IKNL uses its own data model for the NCR. Data deliveries to researchers are usually done as a csv file with accompanying data dictionary. Part of the NCR is also available in the OMOP-CDM. The data in the OMOP-CDM is updated a few times per year.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP	*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)	Prescription: ATC level 5, own vocabulary; Indication: ICD-O-3; Procedures vocabulary: own vocabulary; Diagnosis/medical event vocabulary: ICD-O-3. Starting: TNM	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP		*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*	
X	The RWD source declared Service Level Agreements (SLA)	Guaranteed frequency of updates and incident response time (e.g., corrections in case of errors)	Updates of the data are daily and occur in-house. Any issues with this can (and will) be handled immediately (during work hours).	Provided by DEAP	*2*	*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	Data deliveries are accompanied by a data dictionary. Data request are handled by a specialist at IKNL. They are available for additional questions about the data. There is also a general e-mail address for these questions.	Provided by DEAP		*L2 if details of established data processess by the provider are available*	
		Possibility to collect additional data if needed	Additional data can be collected by the data managers if there is additional funding available. There is a fee involved with this.	Provided by DEAP		*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	Acces to data trough an application form https://iknl.nl/en/ncr/apply-for-data	https://iknl.nl/en/ncr-data	*1*	*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 group of data managers is responsible for data quality and researchers in the organization can flag potential quality issues.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP/ DEAP	*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*

Step 1. PHARMO

Item		Sub-item	Description	Origin of information	Maturity level grade	Maturity level criteria and definitions	Rationale
0	Data base identification	Country	The Netherlands	https://catalogues.ema.europa.eu/node/9	*N/A*	*N/A*	*N/A*
		Data Access Provider	The PHARMO Institute for Drug Outcomes Research (PHARMO Institute)	https://catalogues.ema.europa.eu/node/997/administrative-details		*N/A*	
		Organisation type	Laboratory/Research/Testing facility	https://catalogues.ema.europa.eu/node/997/administrative-details		*N/A*	
1	Rationale and scope for the RWD source creation	Primary purpose for which data are collected	The PHARMO Data Network is a population-based network of healthcare databases and combines data from different healthcare settings in the Netherlands. These different settings, including general practitioner, in- and out-patient pharmacy, clinical laboratory, hospitals, cancer registry, pathology registry and perinatal registry, are linked on a patient level through validated algorithms.	https://catalogues.ema.europa.eu/node/997/administrative-details	*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	All patients registered at the contributing healthcare providers are included, unless the patient requested to opt out.	https://www.dovepress.com/existing-data-sources-for-clinical-epidemiology-the-pharmo-database-ne-peer-reviewed-fulltext-article-CLEP		*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*	
		What triggers a record in the database	TRIGGERING REGISTRATION Birth Disease diagnosis Insurance coverage start Start of treatment or practice registration TRIGGERING A RECORD Multiple prompts depending on healthcare setting (e.g. hospital discharge, specialist visit, medicinal product dispensing etc.) TRIGGERING DEREGISTRATION Death Emigration Loss to follow-up Practice deregistration	https://catalogues.ema.europa.eu/node/997/administrative-details			
		Publications describing this RWD	Completeness and Representativeness of the PHARMO General Practitioner (GP) Data: A Comparison with National Statistics: https://www.dovepress.com/completeness-and-representativeness-of-the-pharmo-general-practitioner-peer-reviewed-fulltext-article-CLEP Existing Data Sources for Clinical Epidemiology: The PHARMO Database Network: https://www.dovepress.com/existing-data-sources-for-clinical-epidemiology-the-pharmo-database-ne-peer-reviewed-fulltext-article-CLEP Cohort profile: the PHARMO Perinatal Research Network (PPRN) in the Netherlands: a population-based mother–child linked cohort : https://bmjopen.bmj.com/content/10/9/e037837 A population-based linked cohort of cancer and primary care data: A new source to study the management of cancer in primary care: https://onlinelibrary.wiley.com/doi/10.1111/ecc.13529 First Year of Life Medication Use and Hospital Admission Rates: Premature Compared with Term Infants: https://www.jpeds.com/article/S0022-3476(12)01461-8/abstract	https://catalogues.ema.europa.eu/node/997/administrative-details https://pharmo.nl/resource-library/			
II	Data collection or recording process	Description of data provider (geographical and oragnizational setting, nature of the data - reported by patients, HCP, etc)	The PHARMO Institute, part of Lumanity, as a scientific research organisation dedicated to the study of epidemiology, drug utilisation, drug safety, health outcomes, and utilisation of healthcare resources. The PHARMO Institute maintains a large and high quality Database Network and works closely with (inter)national medical universities and European healthcare database partners. Through its studies with longitudinal and real-life patient data, the PHARMO Institute contributes to risk management, outcomes research and provides solutions for decision makers in market access, health economics and health outcomes domains.	https://catalogues.ema.europa.eu/node/997/administrative-details	*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	PHARMO conducts studies in accordance with the ENCePP Guide on Methodological Standards in Pharmacoepidemiology and the ENCePP Code of Conduct. Lumanity is ISO 9001:2015 certified. Standard operating procedures, work instructions and checklists are used to guide the conduct of a study. These procedures and documents include internal quality audits, rules for secure and confidential data storage, methods to maintain and archive project documents, rules and procedures for execution and quality control of R or SAS programming, standards for writing protocols and reports, and requirements for senior scientific review of key study documents. Our SOPs include: 0500 Quality policy 0510 Setup quality manual - v2.1 0520 QMS training - v4 0530 QMS checks - v2.3 0531 Compliance assessment - v2.3 0532 Effectiveness assessment - v2.2 0540 QMS adjustment - v2.2 0550 Procedures P 0511 Information Security Incidents 1000 Research P 1100 Research and data analysis I 1101 Request fits PHARMO - v5 I 1101.1 Classifying an opportunity - v4 I 1101.2 SAS feasibility assessment - v5 C 1101.3 Checklist Request fits PHARMO - v3.3 I 1102 Proposal agreed - v5 C 1102.1 Checklist Proposal and Investment - v3.5 I 1103 Contract agreed - v5.1 I 1103.1 Subcontract agreed - v1.4 I 1103.2 Assignment of a project team - v3.4 I 1103.3 Master Service Agreement (MSA) agreed - v1.1 C 1103.4 Checklist (sub)contract/MSA - v5.1 I 1104 Project management - v7 C 1104.1 Checklist Project management prepared - v3.4 I 1105 Protocol/Statistical Analysis Plan agreed - v5.3 C 1105.1 Checklist Protocol/Statistical Analysis Plan - v3.3 I 1106 Programming finalised - v7.1 I 1106.1 Programming execution - v6.1 I 1106.2 Programming execution in multi-country studies - v1.2 I 1106.3 Methodology knowledge base - v3.2 C 1106.4 Checklist Programming - v4.1 I 1107 Report agreed - v6 C 1107.1 Checklist Report - v3.3 I 1108 Dissemination study results - v6 C 1108.1 Checklist Abstract - v3.3 C 1108.2 Checklist Poster presentation - v3.3 C 1108.3 Checklist PowerPoint presentation - v3.3 C 1108.4 Checklist Outline manuscript - v3.3 C 1108.5 Checklist Manuscript - v3.3 I 1109 Research project finalised - v5.1				
	How SOPs are implemented and monitored	PHARMO has an established CAPA management process to identify, correct and mitigate deviations, this is described in the CAPA management SOP (Document Number L-SOP-QA-007, v5.0). Deviations are flagged to Central Compliance team, who investigate and log them, recurring deviations are also monitored for effectiveness and any repeating route causes are escalated to management for further investigation and preventative actions. Re-training is instigated on repeat deviations as necessary. SOP available on request.				

L2 if information is available using standardised templates to make information easy to digest and interpret, and also standard vocabularies are available

		Key data elements captured (are they always recorded, are they optional, is there a planned coverage over time, ...)	General Practitioner (GP) Database (in-house)Data from electronic patient records registered by GPs of all patients enrolled at the GPDiagnoses Symptoms Laboratory test results Referrals to specialists Healthcare product/drug prescriptions Out-patient Pharmacy Database (in-house)Data on all GP or specialist prescribed healthcare products dispensed by the community pharmaciesType of product Date Strength Dosage regimen Quantity Route of administration Prescriber specialty Costs In-patient Pharmacy Database (in-house)Data on all drug dispensing from the hospital pharmacy, given during hospitalizationType of product Start and end date of use Strength Dosage regimen Route of administration Prescriber specialty Clinical Laboratory Database (in-house)Results of tests performed on clinical specimens, requested by GPs or specialistsDate and time of testing Test result Unit of measurement Type of clinical specimen Hospital Database (external)Data on all hospitalizations for more than 24 hours or for which a bed is required, out-patient visits and high budget impact medication. Data are obtained from the Dutch Hospital Data Foundation.bDiagnoses In- and out-patient procedures High budget impact medication Admission, discharge and visit dates Perinatal Registry (external)Data on pregnancies, birth and neonatal outcomes. Data are obtained from Perined.cInformation on mothers, eg: ▪Maternal age ▪Obstetric history ▪Parity Information on pregnancies, eg: ▪Mode of conception ▪Mode of delivery Information on children, eg: ▪Birth weight ▪Gestational age ▪Apgar score Pathology Registry (external)Excerpts of histological, cytological and autopsy examinations. Data are obtained from PALGA.dSummary of pathology report PALGA diagnosis, structured along five classification axes: 1 Topography 2 Morphology 3 Function 4 Procedure 5 Diseases Cancer Registry (external)Data on all newly diagnosed cancer cases. Data are obtained from the Dutch Comprehensive Cancer Organization.eCancer diagnosis Tumor staging Tumor site Morphology Comorbidity at diagnosis Initial treatment	https://catalogues.ema.europa.eu/node/997/administrative-details https://pmc.ncbi.nlm.nih.gov/articles/PMC7196787/		*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)* *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/excl usion (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)	The collection, processing, linkage and anonymisation of the data is performed by STIZON. STIZON is an independent, ISO/IEC 27001 certified foundation, which acts as a Trusted Third Party (TTP) between the data sources and the PHARMO Institute.	https://catalogues.ema.europa.eu/node/997/administrative-details	*2*	*L1 if information is available as free text and/or online link(s)* *L2 if the hardware or software implementation complies with recognised quality standards that can be reported*	*Essential for reliability regarding data alterations resulting from system accidents, software errors or malicious intervention.*
		Software testing and software quality control in place	Our policy is managed by Ilionx in line with ISO 27001 certification https://www.ilionx.com/en/').	https://www.ilionx.com/en			

		Measures to prevent accidental physical data alterations (e.g.: backups, redundant systems, checksums)	We expect that Ilionx policy addresses the following: • Identification: Regularly review and monitor sources for patches and updates, including vendor notifications, security bulletins, and industry alerts. • Assessment: Evaluate the relevance and criticality of identified patches, considering factors such as severity, impact on operations, and compatibility. • Testing: Implement a testing phase in a controlled environment to ensure patches do not disrupt normal operations or introduce new vulnerabilities. • Approval: Obtain necessary approvals before deploying patches to production systems. • Deployment: Schedule and deploy patches based on criticality • Verification: Confirm successful deployment and functionality of the patched systems. • Documentation: Maintain detailed records of all patching activities, including identification, assessment, testing, deployment, and verification. Emergency Patching In the event of a critical vulnerability that poses an immediate threat, the following steps should be taken: • Immediate Assessment: Evaluate the threat level and impact. • Rapid Testing: Expedite testing in a controlled environment. • Emergency Deployment: Apply the patch to production systems immediately if testing is successful. • Post-Implementation Review: Conduct a review to ensure the patch has been applied correctly and document the process. Regular audits are conducted to ensure compliance with this policy.			*L3 NA*	
V	Data management and governance	Data management principles being followed (e.g., GCP, ISO, FAIR, etc)	ISO 9001, ISO27001 certifications	https://catalogues.ema.europa.eu/node/997/administrative-details	2	*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)	Once the linked data are made available to PHARMO, the quality of the data is assessed by data acceptance tests. The contents of these yearly tests differ per database but include quality indicators such as level of missing data, values within a reasonable range and appropriate coding used.	https://www.dovepress.com/existing-data-sources-for-clinical-epidemiology-the-pharmo-database-ne-peer-reviewed-fulltext-article-CLEP			
		Measures to prevent data alterations by unauthorised parties (cybersecurity)	Data is decrypted on use with authenticated user accounts. PHARMO has clearly defined data backup, archival and data retrieval standard operating procedures as well as ISO27001 and ISO9001 certifications ensuring clear and ongoing adherence to regulatory standards and client contracts / agreements. Data Confidentiality and Privacy are rigorously trained within PHARMO through our dedicated Compliance training system Metacompliance, this includes Information Security Policy which describes how to report incidents Specifically, PHARMO provide GDPR awareness training for staff and staff data breach obligations are stated in Global Data Classification & Data Privacy Policy v2.1. Any security incident that is confirmed to be a data breach is managed by Cyber Resilience, IT and the DPO. If a data breach is confirmed then the data controller is informed within the timeframe agreed in the applicable MSA. If the breach meets the reporting threshold to regulatory authorities, this needs to be done within 72 hours of discovery. The DPO will liaise with the data controller to establish if the breach is likely to result in a high risk of adversely affecting individuals and for a plan to inform data subject, should it be required. Once the DPO and Data controller has confirmed that no further damage or attack is occurring and that all communication and mediations have been completed they will define the incident as resolved. All data breaches would be logged in a central Security Incident Log.				
		Auditing and DQ imporvement procedures in place	Annual PEN and Business Continuity and Disaster recovery testing is conducted. In addition, several parts of our business carry out regular testing through and Illionix support. Testing of 1 VM + 1 Azure SQL DB (randomly selected) is performed every month. For PHARMO, we use a third party (nSEC/Resilience) to perform penetration, infrastructure and application testing. The processes and expertise of nSEC/Resilience can be found here: https://www.nsec-resilience.com/.	https://www.nsec-resilience.com/			
VI	Data manipulation steps	Frequency of data updates	The frequency of data collection varies per healthcare provider but is at least on an annual basis. Linkage of the different data sources is yearly; the lag time is about 1 year.	https://www.dovepress.com/existing-data-sources-for-clinical-epidemiology-the-pharmo-database-ne-peer-reviewed-	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*	*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	Data transformation, mapping, and cleaning is performed on a per-project basis depending on the databases required for the study and the desired common data model. To optimize data quality and completeness (which depends on the degree of structure and the extent to which variables are collected as part of routine clinical care), PHARMO can develop algorithms to derive or impute missing values. We can also examine alternative sources of input to address missingness (e.g., enhancement via free text note review, image reassessment, use of natural language processing or artificial intelligence/machine learning).				

		Information about loss of precision during data manipulation steps	A rigorous quality control process is in place to ensure the validity of the data following any manipulation. To ensure the validity of the data, PHARMO compares the distribution of patient characteristics across sources. In case of incongruence of results, we perform a quality check to ensure that the right data was extracted and analyzed (i.e., equivalent approach for cohort selection). The following mitigations can also be employed: -Cohort identification (limitations of ICD-10 codes, drug launched in multiple indications or used off-label) --> addressed through validation of diagnosis through review of other data elements (e.g., diagnostic journey/combination of tests received, clinical lab values, free text GP notes) -Outcomes definition (adaption of clinical trial-based case definitions to the context of observational studies) --> addressed through establishing clear distinctions to differentiate cases from non-cases; defining event-identifying code/ algorithm based on routinely collected healthcare data, in consultation with local clinician (i.e., may vary country-by-country) -Signal validation (ascertainment of the outcome/exposure (e.g., is a Major Adverse Cardiovascular Event [MACE] event due to drug use or just indigestion?))--> addressed through employment of multiple methods to evaluate outcomes; convening clinician panel to review patient profile and interpret whether an observation is a 'true' case; if uncertainty remains, re-identify & consent the patient to speak directly with the treating physician; calibrating against previously reported rates or rates observed in other countries			*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)	Data manipulation is carefully logged to be able to re-trace the methodology in case of any errors.				
VII	Data augmentation steps (e.g., imputation or linkage)	Is any augmentation happening in this datasource?	Databases are linked with external registries such as the Cancer Registry, Pathology Registry and Perinatal Registry. Record Linkage. General Practitioner Database, In-Patient Pharmacy Database, Clinical Laboratory Database, Hospital Database, Cancer Registry, Pathology Registry, Perinatal Registry, and others upon request. Dispensing records from out-patient pharmacies (ie community pharmacies) are linked to hospital admission records. Drug utilization could thus be linked to clinical outcomes.	EMA Catalogue and https://www.dovepress.com/existing-data-sources-for-clinical-epidemiology-the-pharmo-database-ne-peer-reviewed-fulltext-article-CLEP	*2*	*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	The different data sources are linked on a patient level through probabilistic linkage based on validated algorithms. Linkage to the perinatal registry is further explained here: https://www.valueinhealthjournal.com/article/S1098-3015(16)31489-9/fulltext	https://www.dovepress.com/existing-data-sources-for-clinical-epidemiology-the-pharmo-database-ne-peer-reviewed-fulltext-article-CLEP https://www.valueinhealthjournal.com/ar		*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	Linkage includes three major steps: 1. Identifying overlapping personal identifying variables 2. Linkage: Blocking (pairing of patients with similar gender and date of birth); Determining Proability; (calculation of the probability that both records below to the same patients; variables includes first initial, first letter last name, soundex code of last name, zip code); Matching (selection of record pair with the highest cumulative weight value above threshold) 3. Face validity				
		If yes, which is the error rate when conducting the augmentation	A quality control mechanism is in place to resolve errors or remove data for records which cannot be linked			*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 study specifically assessing representativeness and completeness of PHARMO GP data was conducted (see link). The PHARMO GP data are representative of the Dutch population with regard to the demographic characteristics and diagnoses in primary care. Medication data in the PHARMO GP data are more complete than national statistics, and differences are related to reimbursement. The results of this cross-sectional study showed that the PHARMO GP data are representative of the Dutch population with regard to the demographic characteristics and diagnoses in primary care. Medication data in the PHARMO GP data are more complete than national statistics. The sex distribution in the GP population was representative of the Dutch population; half of the people were male (49.7% vs 49.6% [std.diff: 0.00]). For 2006, the std.diffs for the different pharmacological subgroups ranged from −0.12 to 0.24. Only the subgroup "viral vaccines" differed between the PHARMO GP data and the Statistics Netherlands (CBS) data. Its use was more complete in the PHARMO GP data than in the Statistics Netherlands (CBS) data (3.3% vs 0.2% [absolute std.diff: 0.24]). For 2012, the std.diffs ranged from −0.08 to 0.30. Only the subgroups "viral vaccines" and "hormonal contraceptives for systemic use" differed between the PHARMO GP data and the Netherlands. In 2018, the overall use of hormonal contraceptives for systemic use was 2.1% in the Dutch population and 7.7% based on the PHARMO GP data. Less than 0.1% of the Dutch population was vaccinated with a viral vaccine in 2018 according to information from the National Health Care Institute of the Netherlands. Based on PHARMO GP data, this was 3.3%.	Overbeek JA, Swart KMA, Houben E, Penning-van Beest FJA, Herings RMC. Completeness and Representativeness of the PHARMO General Practitioner (GP) Data: A Comparison with National Statistics. Clin Epidemiol. 2023 Jan 5;15:1-11. doi: 10.2147/CLEP.S389598. PMID: 36636730; PMCID: PMC9830053.	*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 EWD source	Validation of the linkage against name and address information for a sample of the patients resulted in a sensitivity and specificity of 0.98.: van Herk-sukel MP, van de Poll-franse LV, Lemmens VE, et al. New opportunities for drug outcomes research in cancer patients: the linkage of the Eindhoven Cancer Registry and the PHARMO record linkage system. Eur J Cancer. 2010;46(2):395–404. doi: 10.1016/j.ejca.2009.09.010			*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, ...)	Data is in its native format but can be harmonized to a common data model on request, including ConcePTION, OMOP, or bespoke		*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)	General Practitioner (GP) Database (in-house)□Diagnoses and symptoms: ICPC □rug prescriptions: WHO ATC Classification System Out-patient Pharmacy Database (in-house)□ispensings: □ational product classification □HO ATC Classification System In-patient Pharmacy Database (in-house)□ispensings: National product classification WHO ATC Classification System Clinical Laboratory Database (in-house)□CIA Coding System □DINC (partly) Hospital Database (external)□iagnoses: WHO ICD □rocedures: DHD registration system for procedures □edication: Dutch classification system Pathology Registry (external)□iagnosis codes are a combination of diagnostic terms (localization, acquisition technique, abnormality) and related to the SNOMED coding system. Cancer Registry (external)□umor staging: TNM-classification □umor site and morphology: WHO ICD-O	https://pmc.ncbi.nlm.nih.gov/articles/PMC7196787/		*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*	
X	The RWD source declared Service Level Agreements (SLA)	Guaranteed frequency of updates and incident response time (e.g., corrections in case of errors)	Data are linked on annual basis. Quality control is provided on a per project basis	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	An epidemiologist from the PHARMO Institute can be made available to support data	Provided by DEAP		*L2 if details of established data processess by the provider are available*	
		Possibility to collect additional data if needed	Additional data can be collected via INSZO and linked via STIZON on an as needed basis	https://inszo.nl/		*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 the PHARMO Database Network is, by governance regulations of the data collection, restricted to researchers of the PHARMO Institute and academic affiliates. Each data request is checked against privacy and company policies, and requires approval of the privacy and governance board. The terms and conditions and a data application form are available on the PHARMO website (www.pharmo.com). This endeavor is in line with the policy and mission of the PHARMO Institute to contribute to a better understanding of the use, safety, effectiveness and cost of pharmaceuticals as used in real-life. The application form can be found on the PHARMO website (www.pharmo.com) and should be submitted together with a study protocol. Applications are checked against the policies that apply for use of data from the PHARMO Database Network and as agreed upon with the contributing healthcare providers. Funding for academic research is not provided by the PHARMO Institute and should be obtained by the researcher taking into account the data access fees for use of the data. Upon approval of the data application by the CC, researchers are provided access to the data at the PHARMO Institute offices.	https://www.dovepress.com/existing-data-sources-for-clinical-epidemiology-the-pharmo-database-ne-peer-reviewed-fulltext-article-CLEP	*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?	General online formulaire, postal address, phone and email contact. Complaints are managed by Recording complaint, Follow-up with client and internally including CAPA. Defining the true cause of the problem is important. Root cause analysis is key for investigating the problem and ensuring that appropriate corrective and preventive measures are put in place. The process includes: • Issue - Defining existing or potential problem or non-conformance • Background and Root cause Analysis -Determining the true cause of the problem • Developing action plan to -□orrect the problem -□revent (re-)occurrence • Implementing plan and dating the actions • Reviewing the action and evaluating effectiveness	https://pharmo.nl/contact-us/	*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*

Step 1. BIFAP

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	AEMPS (Spanish Agency for Medicines)			*N/A*	
		Organisation type	EU Institution/Body/Agency Not-for-profit Regulatory Authority			*N/A*	
1	Rationale and scope for the RWD source creation	Primary purpose for which data are collected	To serve as source of information for independent studies on drug safety and support of medicines regulation activities. BIFAP is a non-profit research project funded by the Spanish Agency for Medicines and Medical Devices (AEMPS). Information is collected by PCPs. Information on hospital discharge diagnoses is linked to patients included in BIFAP for a subset of periods and regions participating in the database. All information on prescriptions of medicines by the PCP is incorporated and linked by the PCP to a health problem (episode of care), and information on the dispensation of medicines at pharmacies is extracted from the e-prescription system that is widely implemented in Spain. The project started in 2001 and nine participant autonomous regions send their data to BIFAP every year. BIFAP includes anonymized clinical and prescription/dispensing data. From several regions, hospitalization data can be linked.		*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 recorded by family doctors and primary care paediatricians in Electronic Medical Records (HCE-AP) provided by the Autonomous Communities (CCAA) that voluntarily participate through collaboration agreements.	https://catalogues.ema.europa.eu/node/955/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/.* *L3 if the information is provided as Metadata (machine readable), including standrad formats, clear definitions and potentially some quality information*	
		What triggers a record in the database	**Event triggering registration of a person in the data source, other:** Upon registration with a primary care physician within the Spanish NHS (≈98,9% of the Spanish population) in the 9 out of the 17 Spanish regions that contribute data. **Event triggering de-registration of a person in the data source:** Death, Emigration **Event triggering creation of a record in the data source:** In every encounter with the general practitioner/paediatrician. Hospital admission and pharmacy dispensation will also trigger the creation of a record.	https://catalogues.ema.europa.eu/node/955/data-flows-and-management			
		Publications describing this RWD	https://pubmed.ncbi.nlm.nih.gov/32337840/ https://www.bifap.org/scientific-publications?lang=en				
II	Data collection or recording process	Description of data provider (geographical and oragnizational setting, nature of the data - reported by patients, HCP, etc)	Agencia Española de Medicamentos y Productos Sanitarios (Spanish Agency for Medicines and Medical Devices, AEMPS). It is the regulatory agency of the Spanish administration that oversees the quality, safety, efficacy and correct information of medicines and medical devices in Spain, as well as cosmetics, from their research to their use, in the interests of the protection and promotion of human health, animal health and the environment. The AEMPS also is the coordinating centre for the Spontaneous Reporting Scheme in Spain and also administer the database supporting this program (FEDRA).	https://catalogues.ema.europa.eu/institution/3331474	*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	Yes , the document linked describes the flows and the different kinds of processing to which the data is subjected, the management of access to the data, the use of the data by virtue of the user type of the database, and the characteristics of the exploitation of the data.	https://www.bifap.org/data-governance?lang=en		*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 Collection: Data is gathered from various sources, including the Autonomous Communities (CCAA) and AEMPS. Participants: CCAA, AEMPS, Information Systems Technicians. Data Cleansing and Structuring: The collected data is cleaned (errors and inconsistencies are removed) and structured to make it usable and organized. Participants: BIFAP Technicians at AEMPS. Standardized Reporting: A subset of the data is extracted and formatted into standardized reports for easier access and dissemination. Participants: AEMPS Technicians, Technicians from participating Autonomous Communities, Collaborating Physicians. Research Studies: The structured data is extracted for research purposes and used by researchers for analysis, with results being shared and disseminated.Participants: Researchers, BIFAP Technicians at AEMPS. Security Measures: Throughout the process, data is pseudonymized and protected with access controls and obfuscation procedures to ensure privacy and compliance. Participants: AEMPS Technicians. Dissemination: Research results are published in peer-reviewed journals, on the HMA EMA Catalogue and shared via the BIFAP website. Participants: Researchers, BIFAP Technicians at AEMPS.	https://www.bifap.org/data-governance?lang=en			

		Key data elements captured (are they always recorded, are they optional, is there a planned coverage over time, ...)	**(HCE-AP) data** **PATIENT DATA:** patient pseudonym unique identifier, gender, date of birth, date of entry into the database, date of deletion from database, cause of deletion (includes patient's death), status in database **VISITS:** date **CONSTRAINTS:** diagnosis with dates, code, and descriptor **HISTORY:** diagnosis with dates, code, and descriptor **GENERAL PATIENT DATA:** includes data such as tobacco, alcohol, blood pressure,body mass index, etc. The date of collection and type are recorded. **EPISODES OR DIAGNOSIS:** descriptor with date, code. **COMMENTS ASSOCIATED WITH THE EPISODE:** date, observations **INTER-VISITS:** dates, medical specialty, motivation, results **VACCINES:** date, code, antigens. **ANALYTICS:** request and result dates, type, determination, value, units, ranges **PRESCRIPTIONS OF PRIMARY HEALTHCARE:** dates, type, drug code, number of containers, dosage. **HCE-AP** **HOSPITAL DISCHARGE DIAGNOSES:** admission reasons recorded and coded by the RAE-CMBD system: includes dates of admission and discharge, type of discharge, primary and secondary diagnoses at hospital discharge. **DATA CONCERNING ELECTRONIC DISPENSATIONS OF MEDICINAL PRODUCTS PRESCRIBED IN PRIMARY HEALTHCARE SETTINGS:** identification of the dispensed drug code, dates, type, number of containers, dosage.	https://www.bifap.org/data-governance?lang=en		*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	BIFAP draws on the data provided by the autonomous communities that voluntarily participate through collaboration agreements. These agreements detail the commitments of the Autonomous Community and the Spanish Agency of Medicines and Medical Devices, as well as other operational aspects related to data processing, monitoring commissions, etc. Each autonomous community gathers data in each own database before sending them to BIFAP. Private partners are not included.	https://bifap.aemps.es/autonomous-communities?lang=en	*2*	*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?	Each Autonomous Community has the right to find the basic structure and collection of data. When this data is extracted and pseudonymised , it is sent to BIFAP. There this data is cleaned and standardized	https://pubmed.ncbi.nlm.nih.gov/32337840/ https://bifap.aemps.es/autonomous-communities?lang=en		*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?	No, it follows an internal structure (BIFAP Common Data Model)	https://pubmed.ncbi.nlm.nih.gov/32337840/		*L3 is only aspirational. NA*	
IV	The data management infrastructure	List of systems used to manage the RWD (either for data collection, recording, processing, etc)	"Comprehensive Study Management (GIE) software, different operating modules are made available to the researcher"	https://www.bifap.org/data-governance?lang=en	*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	https://www.bifap.org/data-governance?lang=en	*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)	The transfer of data from the Autonomous Communities to the BIFAP Data Processing Centre (CPD) in the AEMPS is carried out via an SFTP server enabled by the AEMPS for this purpose where each IT manager in the Autonomous Community has a protected folder to store the data. To maintain absolute confidentiality and privacy on the pseudonymised data files to which they have access, which they may not copy or use for any purpose other than the study, nor disclose or assign to anyone outside the research team, even for conservation purposes. This commitment shall be maintained upon completion of the study. As an additional measure, data storage is performed on encrypted hard disks in order to avoid any opening of the files in the event of disk theft	https://www.bifap.org/data-governance?lang=en	*2*	*L3 N/A*	
V	Data management and governance	Data management principles being followed (e.g., GCP, ISO, FAIR, etc)	Data protection regulations in force in Spain: General Data Protection Regulation, GDPR. Royal Decree LOPDDD	https://www.bifap.org/data-governance?lang=en	*3*	*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)	Minimum quality for research: - Invalid or not coded gender - Invalid date of birth (before 1800 or after follow-up start date) - Age over 115 years at end of follow-up date - Tracking start date on or after tracking end date - No patient clinical records - HCE-AP with "inactive" record without administrative deletion record. - Existence of data in the HCE-AP with a date prior to the start of follow-up or after the end of follow-up date. - Administrative deletion due to transfer to another primary care quota Identifyind data: ensure effective anonymisation Data on prescribed or dispensed medicinal products: purged and addpated for research use Data related to clinical and diagnostic events: structuring of the diagnostic coding is carried out Free text information is used in BIFAP for better event characterisation, event validation, or event identification that is not properly encoded. EMRs received during the annual extraction procedure are reviewed and the following checks and actions take place: Dates of start of follow-up and end of follow-up are defined for each patient. Only the information dated within this period is used for research. These dates are defined based on quality of clinical and administrative data available in the EMR. If needed, the first registered date of death (from administrative or clinical data) is defined as the end of follow-up. KPI: numeouse publications >75% of electronic prescription dispensation records in 2018 44% of patients with linkage to hospital discharged data in 2018 8.1 million patients in 2018			*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)	Physical and logical data security measures to prevent re-identification and access by unauthorized third parties are summarised below: a) The BIFAP database is subject to all physical access controls applied by the Ministry of Health under the National Security Scheme b) As an additional measure, data storage is performed on encrypted hard disks in order to avoid any opening of the files in the event of disk theft. c) Access to the computers where the database is accessed requires a password. Only 4 people who are currently part of the BIFAP's IT Unit can access as password users. d) Access to equipment is only possible at AEMPS headquarters. Access to the AEMPS headquarters is recorded at check-in and check-out. e) The computer security measures applied at the software level are subject to the security criteria applied by the Ministry of Health in all its information systems. f) All activity performed by users accessing BIFAP data is monitored by generating Programming Reports: with each output file the application keeps a report on the activity carried out by the user, in order to track and make it possible to trace the queries made in BIFAP with the tools, as well as the delivery of Structured Data Files for statistical analysis to external researchers. "The transfer of data from the Autonomous Communities to the BIFAP Data Processing Centre (CPD) in the AEMPS is carried out via an SFTP server enabled by the AEMPS for this purpose where each IT manager in the Autonomous Community has a protected folder to store the data."	https://www.bifap.org/data-governance?lang=en			
		Auditing and DQ imporvement procedures in place	An DPIA has been carried out using the ASSI-GDPR Tool (Information Systems Security Audit for compliance with the General Data Protection Regulation).	https://www.bifap.org/data-governance?lang=en (Annexes)		*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	Every 1 year	Provided by DEAP	*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	Requested to DEAP and unable to provide		*N/A*	*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	Requested to DEAP and unable to provide			*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)	Requested to DEAP and unable to provide				

VII	Data augmentation steps (e.g., imputation or linkage)	Is any augmentation happening in this datasource?	Linked data sources into BIFAP common data model: • EMRs from Primary Care • Diagnosis Tests of Covid-19 (SARS-CoV2 positive test results) during the COVID pandemic • Hospital Diagnosis at in patients discharge in a subset of the participating regions and calendar periods • Medicines Dispensed at Community Pharmacies for the total BIFAP population • Vaccines Covid-19 administered, National Registry, for the total BIFAP population • All vaccines administered, National Registry, for the total BIFAP population • Causes of Death recorded in the national registry • Hospital Pharmacies dispensing Data in a subset of the participating regions	https://catalogues.ema.europa.eu/node/955/data-flows-and-management	2	*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	Linkagae strategy: deterministic			*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*	
		If yes, which algorithms and assumptions applied	None	Provided by DEAP			
		If yes, which is the error rate when conducting the augmentation	Unknown	Provided by DEAP			
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)	Not aware of any issue	Provided by DEAP	2	*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 RWD source	https://www.bifap.org/scientific-publications?lang=en Validation of digestive cancer: https://www.mdpi.com/2077-0383/13/2/361 COVID-19 validation: https://pmc.ncbi.nlm.nih.gov/articles/PMC11102056/				
IX	The RWD source representation	Description of data model or models used (OMOP, FHIR, ...)	OMOP, CONCEPTION, BIFAP DATA MODEL	https://catalogues.ema.europa.eu/node/955/data-elements-collected	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)	Cause of death: ICD-10-CM / ICD-9-CM Indication: SNOMED CT Procedures: ICD-10-CM / ICD-9-CM / SNOMED CT Diagnosis / medical event: ICD-10-CM / ICD-9-CM / SNOMED CT / Not coded (Free text) Prescriptions of medicines: ATC / SNOMED Dispensing vocabulary: ATC / SNOMED Medicinal product vocabulary: SNOMED	https://catalogues.ema.europa.eu/node/955/data-elements-collected ICPC: https://pubmed.ncbi.nlm.nih.gov/32337840/			
X	The RWD source declared Service Level Agreements (SLA)	Guaranteed frequency of updates and incident response time (e.g., corrections in case of errors)	Requested to DEAP and unable to provide		N/A	*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	For researchers registered on the BIFAP website, there are training courses online (Introduction to BIFAP data; Medication data in BIFAP; Diagnosis data in BIFAP; Creation of variables in BIFAP) before using BIFAP data, as well as aggregated data and a list of available standard variables to use in research studies. Governance and auditing documents are public, and an automatic online process is in place to submit studies for evaluation by the Scientific Commitee. Help desk contact available.	Provided by DEAP	2	*L2 if details of established data processess by the provider are available*	
		Possibility to collect additional data if needed	Possible to request reextractions, but not found if collection of new data. Additional data from linked sources are under evaluation (Hospital Pharmacies dispensing Data; Specialist Prescriptions; Link mother-child; etc).	Provided by DEAP		*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	AEMPS Administrator All other functions of the AEMPS Autonomous Communities: Members of the Advisory Committee and staff authorised by each member of the BIFAP Advisory Committee in the collaborating Autonomous Communities. Collaborating physicians Registered Researchers: Healthcare professionals or health research professionals	https://www.bifap.org/learn-more?lang=en	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 N/A*	*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	A general email and phone number are available. Review of clinical histories for case confirmation must be conducted at the AEMPS offices. Consequently, feedback from researchers/data consumers who validate recorded outcomes is frequent facilitated through personal communication and collaboration.	https://www.bifap.org/ Provided by DEAP	2	*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*

Step 1. SIDIAP

Item		Sub-item	Description	Origin of information	Maturity level grade	Maturity level criteria and definitions	Rationale
0	Data base identification	Country	Spain	https://www.sidiap.org/index.php/ca/	*N/A*	*N/A*	*N/A*
		Data Access Provider	Fundació Institut Universitari per a la Recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol)	https://www.sidiap.org/index.php/ca/		*N/A*	
		Organisation type	Educational Institution Laboratory/Research/Testing facility Not-for-profit	https://www.sidiap.org/index.php/ca/		*N/A*	
I	Rationale and scope for the RWD source creation	Primary purpose for which data are collected	Generate new knowledge and practical evidence to promote,advance and manage Primary Care research in Catalonia and other areas	https://cataloques.ema.europa.eu/institution/50154	*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 from 328 primary care centres (database of population-wide primary care electronic health records) managed by the Catalan Health Institute in Catalonia, Spain	https://academic.oup.com/ije/article/51/6/e324/6567646		*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/.*	
		What triggers a record in the database	**Event triggering registration of a person in the data source:** Birth, Immigration, Practice registration **Event triggering de-registration of a person in the data source:** Death, Emigration, Practice deregistration **Event triggering creation of a record in the data source:** Any data registered by a healthcare professional can be available	https://cataloques.ema.europa.eu/node/1019/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/51/6/e324/6567646 https://researchonline.lshtm.ac.uk/id/eprint/856930/1/Construction%20and%20validation%20of%20a%20scoring%20system%20for%20the%20selection%20of%20high-quality%20data%20in%20a%20Spanish%20population%20primary%20care%20database%20%28SIDIAP%29.pdf				
II	Data collection or recording process	Description of data provider (geographical and oragnizational setting, nature of the data - reported by patients, HCP, etc)	Primary Health Care University Research Institute Jordi Gol (IDIAP Jordi Gol). The IDIAP aims to generate new knowledge and practical evidence to promote,advance and manage Primary Care research in Catalonia and other areas,by means of training,dissemination of results and translation of research findings into clinical practice.	https://www.sidiap.org/index.php/en/qui-som-en/el-sidiap-en	*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 are captured from ECAP, software of EHR in Primary Care. We are not aware if there is a specific SOP for that. As for the creation of the database, we do not have any document published anywhere on how it is done. When it comes to downloading the data, we always do the same processes, in the same order.	Provided by DEAP	*1*	*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 downloads of the data from the original tables, the corrections (if necessary), the unifications of units are all done by the team and quality controls are carried out to check that the download has properly done.	Provided by DEAP			
		Key data elements captured (are they always recorded, are they optional, is there a planned coverage over time, ...)	Key data: rare diseases, pregnancy and/or neonates, hospital admission and/or discharge, ICU admission, prescriptions of medicines, dispensing of reimbursed medicines, contraception, administration of vaccines, procedures, clinical measurements, healthcare provider, units of healthcare utilisation, unique identifier for persons, diagnostic codes, medicinal product information (active ingredient(s), dose, package size, strength), lifestyle factors (alcohol use, frequency of exercise, tobacco use), sociodemographic information (age, country of origin, deprivation index, gender, living in rural area, pharmaceutical copayment)	https://cataloques.ema.europa.eu/node/1019/data-elements-collected https://www.sidiap.org/index.php/es/dades-3/farmacs	*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*		*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)	Raw data is manually collected by > 30.000 professionals from Institut Català de la Salut through patients' electronic health records (EHRs). Then, SIDIAP systematically collects data and structures them in data domains, each containing the person's pseudo-anonymized identifier (allowing linkage between them). Each person's pseudonymized ID is unique for each project	https://doi.org/10.1093/ije/dyac068	*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	We have quality controls throughout the data extraction, transformation and loading (ETL) process, which we execute sequentially during each of the database creation phases. The software used for this purpose, implemented in-house, is kept up to date on an ongoing basis.	Provided by DEAP	*1*	*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)	Data are stored on servers in an access-controlled data centre. A small group of users can access the servers. There is a minimum of three copies of the data, one in a backup cabin, the second on a cloud server and the third on password-protected hard disks.	Provided by DEAP		*L3 N/A*	
V	Data management and governance	Data management principles being followed (e.g., GCP, ISO, FAIR, etc)	Data protection regulations in force in Spain: General Data Protection Regulation, GDPR. LOPDDD (Organic Law 3/2018)	https://academic.oup.com/ije/article/51/6/e324/6567646?login=true	*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)	The SIDIAP database contains pseudonymised data emerged from the primary care electronic health records (I) from approximately 300 primary care practices around Catalonia. All these practices use the same I software, and all primary care health professionals receive similar training on the correct use of the software for optimal coding regarding clinical management of their patients. For each study, the local research team and SIDIAP data managers develop a data specification and extraction protocol based on the approved protocol. Specific data quality checks are performed on a study-per-study basis. Patients are regarded eligible to be included in a study if they are registered and can be followed in the database. Study data are processed using SQL and Python by the data management team and analysed by the research team.	Provided by DEAP		*L2 if standard best practices are being used and a direct impact on DQ is reported. There are SOPs and data management processes that adhere to the standards. The representation of metadata follows FAIR standards*	
		Measures to prevent data alterations by unauthorised parties (cybersecurity)	Encriptation process to request data	https://www.sidiap.org/index.php/en/serveis-en/recursos-investigador-en			
		Auditing and DQ imporvement procedures in place	Internal and external validation processes are carried out to determine the data quality of the SIDIAP information at each data update. These include stratifying the data by geographical regions and year in order to identify differences in data collection that need to be harmonized (e.g. recording of a specific information under different codes)	https://doi.org/10.1093/ije/dvac068		*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	From 2025, once a year (starting each January with data up to previous December)	Provided by DEAP	*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	Once data are extracted from primary sources, a process of verification, homologation and data management is performed in order to build a standardized data repository. After that, multiple processes (selection, depuration, data quality control, creation of variables, missing data management) are conducted before introducing transformed data into SIDIAP	https://www.sciencedirect.com/science/article/abs/pii/S0025775312001339?via%3Dihub		*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	The quality checks mentioned above analyse possible losses of accuracy after the data transformation process by comparing the downloaded source data with the transformed data, as well as the transformed data from different updates or versions of the database.	Provided by DEAP	*1*	*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)	All code used for database generation is versioned with track changes and stored in a Git-based code hosting and collaboration tool.	Provided by DEAP			
VII	Data augmentation steps (e.g., imputation or linkage)	Is any augmentation happening in this datasource?	Linkage with data augmentation with other data sources: - CMBD-URG (Hospital Emergency Room) - CMBD-AH (Hospital Discharges) - MHDA (Drugs Hospitalaries Dispensated in Ambulatory) - Pharmacies dispensations, EHR and Labaratories datasets	https://www.sidiap.org/index.php/ca/dades/informacio-disponible	*1*	*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	Records linked at a patient level between datasets. Possible linkage: linkage is performed on a project by project basis	https://cataloques.ema.europa.eu/node/1019/data-flows-and-management		*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	No algorithms applied within the database, but some research teams use different algorithms in their studies	Provided by DEAP https://doi.org/10.1136/bmjopen-2022-071335 https://medinform.jmir.org/2022/11/e37976 https://doi.org/10.1016/j.bone.2022.116469	*2*		
		If yes, which is the error rate when conducting the augmentation	Not available	Provided by DEAP	*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)	Data missingness (recent measurement of a variable of interest not being available), under-reporting of certain variables (mental disorders), lacking granularity for certain research questions due to the primary care nature of database	https://academic.oup.com/ije/article/51/6/e324/6567646#387259150	*2*	*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	The quality of a wide number of data captured in SIDIAP (e.g. cancer, Alzheimer's disease, dementia, cardiovascular risk factors and musculoskeletal disorders) has been demonstrated (see references 13-16, 20, 21, 23 & 24 of Data Resource Profile paper)	https://academic.oup.com/ije/article/51/6/e324/6567646#387259150			
IX	The RWD source representation	Description of data model or models used (OMOP, FHIR, ...)	TrineTX, ConcepTION, OMOP	https://cataloques.ema.europa.eu/node/1019/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)	Procedures: ICD-10-CM / ICD-9-CM Diagnosis / medical event: ICD-10-CM Prescriptions of medicines: ATC Dispensing vocabulary: ATC Medicinal product vocabulary: ATC level 7 / RxNorm	https://cataloques.ema.europa.eu/node/1019/data-flows-and-management			
X	The RWD source declared Service Level Agreements (SLA)	Guaranteed frequency of updates and incident response time (e.g., corrections in case of errors)	Data are downloaded on an annual basis, generating one instance of the database for each year (end date 31/12). Possible errors detected are analysed and after assessing their severity, we decide whether it is necessary to include the corrections in the same instance or whether it is possible to wait for the next update of the data.	Provided by DEAP	*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	It depends. If it is a 'classic' SIDIAP project, the specification document and the quality control are delivered. In the case of CDM, nothing is handed over. In principle, as the data are analysed in IDIAP, no external support is given, but if necessary, it is given to the IDIAP research groups.	Provided by DEAP			
		Possibility to collect additional data if needed	If the additional data are data that are collected in the system, the possibility of adding them to the database could be assessed / considered. If they are data from a researcher or source outside the organisation, they cannot be added to the generic database of SIDIAP, they can only be linked to the data of the project itself (as long as it is done with SIDIAP 'Classic').	Provided by DEAP			
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	Researchers from public institutions can request data access if they comply with certain requirements, requiring approval both from SIDIAP's scientific committee and ethical committee	https://academic.oup.com/ije/article/51/6/e324/6567646#387259150	*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 N/A*	*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?	An e-mail address and a telephone number are available.	https://www.sidiap.org/index.php/en/solicituds-en	*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*

Step 1. VID

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 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://cataloques.ema.europa.eu/node/1077/data-flows-and-management#darwin-data-source-linkage https://academic.oup.com/ije/article/49/3/740/5707448 https://cataloques.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)* *L2 if a structure checklist and dataset version control are available* *L3 is only aspirational. N/A*	*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				
		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)	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)* *L2 if the hardware or software implementation complies with recognised quality standards that can be reported* *L3 N/A*	*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			
		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			
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)* *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)	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			
		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.qva.es/es/inicio/atencion_ciudadano/buscadores/departamentos/detalle_departamentos?id_dept=28011			
VI	Data manipulation steps	Frequency of data updates	Monthly data updating	https://cataloques.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* *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*	*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			

		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://cataloques.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)* *L3 if an automatised process for data linkage/mapping exists*	
		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			
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://cataloques.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* *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)	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://cataloques.ema.europa.eu/sites/default/files/cdm-etl-spec/0_3_VID_Catalogue_RTL_specifications_0.pdf			
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*

Step 2. Dk Reg (Denmark)

Dimension	Sub-dimension	Metrics	Description	Origin of information
Timeliness	Currency	How often is the database updated (i.e., frequency of updates)	Continously updated prescription, hospital and population data –complete with short latency Validated death and cancer data: yearly updates with 1–2 year lag	Statistics Denmark and the National Health Data Authority, NHDA (https://www.dst.dk/en/TilSalg/data-til-forskning/generelt-om-data/dokumentation-af-data) https://catalogues.ema.europa.eu/node/1145/data-flows-and-management
		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 year and 3 months, to 2 years and 6 months	Provided by DEAP
		The time elapsed from when a user requests the data to when they actually receive it	3-7 months. If the cohort is already extracted and used within a specific approved purpose for other projects, no lag in delivery.	Provided by DEAP
		Median time (years) between first and last available records for unique individuals	37.2 years	https://catalogues.ema.europa.eu/node/1145/quantitative-descriptors
Extensiveness	Coverage	Percentage of a target population present in a database	All residents in DK (100%). Active population size in the data source is 5.8M.	https://catalogues.ema.europa.eu/node/1145/quantitative-descriptors
	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; rarely, there might be some months of lag.	Provided by DEAP
		% of subjects in the data with a record of sex	Requested to DEAP and unable to provide	
		% of subjects in the data who had an event with a code for the event	Requested to DEAP and unable to provide (for hosp data at least 1 diagnosis should be present)	Provided by DEAP
		% of subjects in the data who had a prescription/dispensing with a recorded code for the medicine	Requested to DEAP and unable to provide (only incomplete due to human error, good completness)	Provided by DEAP
		% of subjects in the data who got vaccinated with a recorded code for the vaccine	Requested to DEAP and unable to provide (self-reported vaccinations by patients, or reported by physicians, sometimes stored in dispensing datasets, ...)	Provided by DEAP
Reliability	Accuracy	The population distribution in the data source aligns with that of the country	As the source includes population at a national level, this assessment is not applicable.	https://catalogues.ema.europa.eu/node/991/quantitative-descriptors
		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)	Requested to DEAP and unable to provide	
		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)	Requested to DEAP and unable to provide	
	Precision	Exposures codes precision level, including medicines and vaccines (e.g., active principle, therapeutic group, ...)	Active principle (ATC level 5 codes)	https://academic.oup.com/ije/article/46/3/798/2447869
		Precision of date of birth (e.g., day, month, year)	Day, month, year	https://www.esundhed.dk/Dokumentation
		Precision of date of death (e.g., day, month, year)	Day, month, year	https://www.esundhed.dk/Dokumentation
		Precision of date of the event/diagnosis (e.g., day, month, year)	PPVs health events in the Danish National Registry range from 15% to 100%	Provided by DEAP
		Precision of date of the exposure (e.g., day, month, year)	Requested to DEAP and unable to provide	
	Traceability	Provenance of event records	Hospital inpatient care, hospital outpatient care, emergency room, primary care For procedures: procedures during hospitalisation	https://catalogues.ema.europa.eu/node/1145/administrative-details
		Provenance of medicines/vaccines records	Dispensing in community pharmacy	https://catalogues.ema.europa.eu/node/1145/administrative-details
Coherence	Format coherence	For dates, formatting constraint being followed	Exact format not provided. Date of birth/death: character, length 8	https://www.esundhed.dk/Dokumentation?rid=11&tid=53&vid=1315 https://www.esundhed.dk/Dokumentation?rid=17
		For sex, formatting constraint being followed	M (male), K (female)	https://www.esundhed.dk/Dokumentation?rid=5&tid=7&vid=63
	Relational coherence	% of records with the Person ID in the PERSONS table	Requested to DEAP and unable to provide	
	Semantic coherence - to determine	For EVENTS definitions, codelists/data dictionaries being employed according to external standards	ICD10DA, SNOMED For procedures: NCSP, NCMP, NCRP, PROCDA	https://catalogues.ema.europa.eu/node/1145/administrative-details
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	ATC, RxNorm	https://catalogues.ema.europa.eu/node/1145/administrative-details
	Uniqueness	Number of records flagged as potential duplicates	Requested to DEAP and unable to provide	

Step 2. FI Reg (FL) KAN

Dimension	Sub-dimension	Metrics	Description	Origin of information
Timeliness	Currency	How often is the database updated (i.e., frequency of updates)	Varies between registers (monthly-annual)	Provided by DEAP by consulting register maintainers
		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)	From a few months to 2 years, depending on the needed.	Provided by DEAP
		The time elapsed from when a user requests the data to when they actually receive it	4-12 months	Provided by DEAP
		Median time (years) between first and last available records for unique individuals	Specific number of years is unknown but data subjects are followed from birth/immigration to death/emigration.	Provided by DEAP
Extensiveness	Coverage	Percentage of a target population present in a database	>99%	Although these are nationwide data, it is possible to deny the use of personal information for reasons outlined in the act on secondary use of health and social data. Until now, the number of these persons has been small, less than 1% of population.
	Completeness	% of subjects in the data with a recorded birth date	100%	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
		% 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	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
		% of subjects in the data with a record of sex	100%	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
		% of subjects in the data who had an event with a code for the event	99.97%	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
		% of subjects in the data who had a prescription/dispensing with a recorded code for the medicine	99.99%	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
		% 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 98.67% of individuals who are known to have been vaccinated	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
Reliability	Accuracy	The population distribution in the data source aligns with that of the country	As the source includes population at a national level, this assessment is not applicable.	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
		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)	<0.5%	Provided by DEAP. Checks of diagnosis data in our previous studies accross different therapeutic areas
		Records of healthcare events (diagnoses, prescriptions, admissions, etc) with logical inconsistencies (e.g., and admission occurs after death)	<1%	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
		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)	End of treatment episodes derived based on dispensing date and dispensed amount	
	Precision	Exposures codes precision level, including medicines and vaccines (e.g., active principle, therapeutic group, ...)	Active principle (ATC level 5 codes)	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
		Precision of date of birth (e.g., day, month, year)	Month, year (Day can be obtained if justified by the research question	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
		Precision of date of death (e.g., day, month, year)	Day, month, year	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
		Precision of date of the event/diagnosis (e.g., day, month, year)	Day, month, year	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
		Precision of date of the exposure (e.g., day, month, year)	Day, month, year (dispensing date). End date must be calculated based on data)	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
	Traceability	Provenance of event records	Primary care, specialised health care (inpatient and outpatient); intensive care unit (limited)	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
		Provenance of medicines/vaccines records	Both prescribed and dispensed included, vaccinations also from primary care records	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
Coherence	Format coherence	For dates, formatting constraint being followed	Character, length 8: ddmmyyyy	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
		For sex, formatting constraint being followed	0 Unknown 1 Man ☐ 2 Woman 3 Not known / cannot be defined☐ 9 undefined	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
	Relational coherence	% of records with the Person ID in the PERSONS table	100%	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
	Semantic coherence	For EVENTS definitions, codelists/data dictionaries being employed according to external standards	ICD-10 (https://koodistopalvelu.kanta.fi/codeserver/pages/classification-view-page.xhtml?classificationKey=23&versionKey=58), ICPC (https://koodistopalvelu.kanta.fi/codeserver/pages/classification-view-page.xhtml?classificationKey=210&versionKey=282)	https://koodistopalvelu.kanta.fi/codeserver/pages/classification-view-page.xhtml?classificationKey=210&versionKey=282
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	ATC, nordic product number (vnr)	
	Uniqueness	Number of records flagged as potential duplicates	0%	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017

Step 2. PEDIANET

Dimension	Sub-dimension	Metrics	Description	Origin of information
Timeliness	Currency	How often is the database updated (i.e., frequency of updates)	Twice a year (every 6 months)	https://catalogues.ema.europa.eu/node/1128/quantitative-descriptors
		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)	At least 6 months plus the lag of delivery	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	14 years	https://catalogues.ema.europa.eu/node/1128/quantitative-descriptors
Extensiveness	Coverage	Percentage of a target population present in a database	3% of the Italian Paeditraic Population (<15 years)	https://catalogues.ema.europa.eu/node/1128/quantitative-descriptors
	Completeness	% of subjects in the data with a recorded birth date	Month and year of birth are recorded for 100% of subjects; day is defaulted to the 15th for anonymization	
		% of subjects in the data, irrespective of vital status, that have a recorded date of death	0%	https://zenodo.org/records/13384860 Provided by DEAP
		% 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	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
Reliability	Accuracy	The population distribution in the data source aligns with that of the country	It is as expected, and in previous studies it has been reported to be representative in gender and age distributions with their national statistics, even if it enrolls only a sample of the population (pediatrics). To note, there is a higher proportion of toddlers and children (28 days to 12 y) active in respect to the population size, compared to other agebands. Active population size (302204): Neonate: 920 (0.3%) Infants and toddlers (28 days – 23 months): 24572 (8.1%) Children (2 to < 12 years): 182785 (60.5%) Adolescents (12 to < 18 years): 93927 (31.1%)	https://catalogues.ema.europa.eu/node/1128/quantitative-descriptors https://www.sciencedirect.com/science/article/pii/S0264410X20301535 https://www.ema.europa.eu/en/documents/report/observational-data-real-world-data-subgroup-report_en.pdf
		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)	Date values before birth: 0-4.8%	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)	Requested to DEAP and unable to provide	
	Precision	Exposures codes precision level, including medicines and vaccines (e.g., active principle, therapeutic group, ...)	Active principle (ATC level 5 codes)	Provided by DEAP
		Precision of date of birth (e.g., day, month, year)	Day, month, year	Provided by DEAP
		Precision of date of death (e.g., day, month, year)	Day, month, year	Provided by DEAP
		Precision of date of the event/diagnosis (e.g., day, month, year)	Day, month, year	Provided by DEAP
		Precision of date of the exposure (e.g., day, month, year)	Day, month, year	Provided by DEAP
	Traceability	Provenance of event records	primary care diagnosis, emergency, hospital diagnosis, exemption, primary care event, diagnosis_event_hospitalisation_automatically_referred_to_PC, vaccination_centre_event, specialist diagnosis	https://zenodo.org/records/13384860
		Provenance of medicines/vaccines records	prescription in primary care, administration at a paediatrician, administration by public health authority	https://zenodo.org/records/13384860
Coherence	Format coherence	For dates, formatting constraint being followed	Character, length 6: YYmmdd	Provided by DEAP
		For sex, formatting constraint being followed	M (male), F (female)	Provided by DEAP
	Relational	% of records with the Person ID in the PERSONS table	100%	https://zenodo.org/records/13384860
	Semantic coherence - to determine	For EVENTS definitions, codelists/data dictionaries being employed according to external standards	ICD9CMP (is an ad hoc coding system taking into account exemption codes too), Free text	Provided by DEAP
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	ATC, MPID	https://zenodo.org/records/13384860
	Uniqueness	Number of records flagged as potential duplicates	Requested to DEAP and unable to provide	

Step 2. CPRD

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 https://www.sciencedirect.com/science/article/pii/S2666776224001534#appsec1
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	

_x000D

Step 2. NCR

Dimension	Sub-dimension	Metrics	Description	Origin of information
Timeliness	Currency	How often is the database updated (i.e., frequency of updates)	NCR updates are daily. However, data is registered 6-12 months after diagnosis so there is a lag there. Vital status is indeed checked once per year.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP/ DAP
		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 to 2 years, as data managers only have access to EHR once per patient to capture the primary treatment plan	Provided by DEAP
		The time elapsed from when a user requests the data to when they actually receive it	~2 months	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP/ DAP
		Median time (years) between first and last available records for unique individuals	0.7 years	https://catalogues.ema.europa.eu/node/952/quantitative-descriptors
Extensiveness	Coverage	Percentage of a target population present in a database	>95% coverage of the total population in The Netherlands. >= 18 y Population size: 3,677,269	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP
	Completness	% of subjects in the data with a recorded birth date	100%	Provided by DEAP
		% 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	Provided by DEAP
		% of subjects in the data with a record of sex	100%	Provided by DEAP
		% of subjects in the data who had an event with a code for the event	100%	Provided by DEAP
		% of subjects in the data who had a prescription/dispensing with a recorded code for the medicine	99.27% of registered chemotherapies have an ATC code. 0.73% of registered chemotherapies are either coded as "intensive chemotherapy" (the majority, mainly for hematology) or "trial medication" (for all cancer patients in the past 5 years)	Provided by DEAP
		% of subjects in the data who got vaccinated with a recorded code for the vaccine	None	Provided by DEAP
Reliability	Accuracy	The population distribution in the data source aligns with that of the country	As NCR is a disease registry, it reflects only the population affected by colorectal cancer, rather than the general population of the country. Yearly participants to the registry: -2019: 1,574,506 -2020: 1,338,052 -2021: 1,632,493	https://iknl.nl/getmedia/046b1a45-b673-4117-9320-9fd8e1823bd6/Monitor-darmkanker-2021-UK-definitieve-versie.pdf https://www.rivm.nl/sites/default/files/2021-10/Monitor_bevolkingsonderzoek_darmkanker_2020_eng.pdf
		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)	Requested to DEAP and unable to provide	
		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)	Requested to DEAP and unable to provide	
	Precision	Exposures codes precision level, including medicines and vaccines (e.g., active principle, therapeutic group, ...)	Active principle (ATC level 5 codes)	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP
		Precision of date of birth (e.g., day, month, year)	Age at diagnosis, the NCR contains date of birth (day, month, year), but this is generally not shared in a data request, instead age at diagnosis is shared, for example	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP/ DAP
		Precision of date of death (e.g., day, month, year)	Day, month, year	Provided by DEAP
		Precision of date of the event/diagnosis (e.g., day, month, year)	Day, month, year; but usually not shared in a data request, instead interval since diagnosis (or a different interval) is shared	Provided by DEAP
		Precision of date of the exposure (e.g., day, month, year)	Day, month, year	Provided by DEAP
	Traceability	Provenance of event records	EMR. Death, so that is from CBS (it is retrieved from the EHR as well if known, but will be checked during linkage with CBS)	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP
		Provenance of medicines/vaccines records	EMR	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP
Coherence	Format coherence	For dates, formatting constraint being followed	Requested to DEAP and unable to provide	
		For sex, formatting constraint being followed	Requested to DEAP and unable to provide	
	Relational coherence	% of records with the Person ID in the PERSONS table	100%	Provided by DEAP
	Semantic coherence - to determine whether the database uses a standardised dictionary	For EVENTS definitions, codelists/data dictionaries being employed according to external standards	Indication: ICD-O; Procedures vocabulary: own vocabulary; Diagnosis/medical event vocabulary: ICD-O Stage: TNM	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP/ DAP
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	Prescription: ATC level 5, own vocabulary;	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP
	Uniqueness	Number of records flagged as potential duplicates	NO but If one patients gets two different tumors, they would get two entries in the NCR and therefore there are more records in the source data. These two tumors will be connected to the same single patient.	DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP

Step 2. PHARMO

Dimension	Sub-dimension	Metrics	Description	Origin of information
Timeliness	Currency	How often is the database updated (i.e., frequency of updates)	Annual	Provided by DEAP https://www.dovepress.com/article/download/53399
		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 to 13 months	Provided by DEAP
		The time elapsed from when a user requests the data to when they actually receive it	At least 1 month (ethics approval)	Provided by DEAP
		Median time (years) between first and last available records for unique individuals	12 years	https://catalogues.ema.europa.eu/node/997/quantitative-descriptors
Extensiveness	Coverage	Percentage of a target population present in a database	40% of Dutch population. Active population size is of approximately 7M.	Provided by DEAP
	Completness	% of subjects in the data with a recorded birth date	100% available through Perinatal Registry	Provided by DEAP
		% 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	Provided by DEAP
		% of subjects in the data with a record of sex	~99%	Provided by DEAP
		% of subjects in the data who had an event with a code for the event	80-90% coverage of hospital events (admissions/discharges from 1998, specialist visits from 2014)	Provided by DEAP
		% of subjects in the data who had a prescription/dispensing with a recorded code for the medicine	25% coverage of out-patient pharmacy from; 10% coverage of in-patient pharmacy from 1985; 80-90% coverage of high cost medicines after 2017	Provided by DEAP
		% of subjects in the data who got vaccinated with a recorded code for the vaccine	From the total of individuals who are known to have been vaccinated: 25% coverage of out-patient pharmacy; 10% coverage of in-patient pharmacy from 1985; 80-90% coverage of high cost medicines after 2017	Provided by DEAP https://pmc.ncbi.nlm.nih.gov/articles/PMC9830053/
Reliability	Accuracy	The population distribution in the data source aligns with that of the country	Population sex and age distribution are aligned with country demographics, especially for hospital data. Estimated percentage of the population covered by the data source in the catchment area: 23% on average upon linkage, but it depends on which data sources covered in the PHARMO Data Network are needed to be linked. When solely considering hospital data for instance, 80% of hospitals are covered. The GP data we have access to is representative of the Netherlands and covers ~20-25% of the Dutch population. The most current population size by age in the Netherlands can be found on this website https://www.cbs.nl/nl-nl/visualisaties/dashboard-bevolking/bevolkingspiramide .	https://pmc.ncbi.nlm.nih.gov/articles/PMC9830053/#f0003 https://www.dovepress.com/article/download/53399 https://catalogues.ema.europa.eu/node/997/quantitative-descriptors
		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)	PHARMO GP data are representative of the Dutch population with regard to diagnoses in primary care. Medication data in the PHARMO GP data are more complete than national statistics.	https://pmc.ncbi.nlm.nih.gov/articles/PMC9830053/
		Records of healthcare events (diagnoses, prescriptions, admissions, etc) with logical inconsistencies (e.g., and admission occurs after death)	Logical inconsistencies are expected to be 0% as they are checked during quality control processes as described in Step 1	Provided by DEAP
		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)	Indication of treatment is partially available in high-cost medicines database	Provided by DEAP https://www.dovepress.com/article/download/53399
	Precision	Exposures codes precision level, including medicines and vaccines (e.g., active principle, therapeutic group, ...)	Active principle (ATC level 5 codes)	https://www.dovepress.com/article/download/53399
		Precision of date of birth (e.g., day, month, year)	Year	Provided by DEAP
		Precision of date of death (e.g., day, month, year)	Day, month, year	Provided by DEAP
		Precision of date of the event/diagnosis (e.g., day, month, year)	Day, month, year	Provided by DEAP
		Precision of date of the exposure (e.g., day, month, year)	Day, month, year	Provided by DEAP
	Traceability	Provenance of event records	Hospital discharge records, Primary care medical records, Hospital inpatient care, Primary care – specialist level (e.g. paediatricians), Secondary care – specialist level (ambulatory), Hospital outpatient care	https://catalogues.ema.europa.eu/node/997/administrative-details
		Provenance of medicines/vaccines records	Pharmacy dispensing records, In-patient pharmacy data, Drug prescription records	https://catalogues.ema.europa.eu/node/997/administrative-details
Coherence	Format coherence	For dates, formatting constraint being followed	Requested to DEAP and unable to provide	Provided by DEAP
		For sex, formatting constraint being followed	M (male), F (female), O (other/unspecified)	Provided by DEAP
	Relational coherence	% of records with the Person ID in the PERSONS table	Relational coherence varies by data extraction, depending on the data banks required. This is checked on a project-by-project basis.	Provided by DEAP
	Semantic coherence	For EVENTS definitions, codelists/data dictionaries being employed according to external standards	ICD-10, ICD-9, ICPC, WCIA, LOINC, DHD registration system for procedures, SNOMED	Provided by DEAP
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	ATC, national product classsification	Provided by DEAP https://catalogues.ema.europa.eu/node/997/administrative-details
	Uniqueness	Number of records flagged as potential duplicates	Removed during probablistic linkage process	Provided by DEAP

Step 2. BIFAP

Dimension	Sub-dimension	Metrics	Description	Origin of information
Timeliness	Currency	How often is the database updated (i.e., frequency of updates)	Once a year	https://www.bifap.org/data-governance?lang=en
		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)	At least 6 months plus the lag of delivery	Provided by DEAP
		The time elapsed from when a user requests the data to when they actually receive it	5 to 6 months (including 1-2 months for Institutional Rewvie Board's approval)	Provided by DEAP
		Median time (years) between first and last available records for unique individuals	10 years	https://catalogues.ema.europa.eu/node/955/quantitative-descriptors
Extensiveness	Coverage	Percentage of a target population present in a database	98.9% of the Spanish population is registered in the Spanish NHS 36% of the total Spanish population in the NHS (46.6 Million) 95% of the population of the nine included regions Up to 2018, 57.6% active patients of the participating regions (14 million)	https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.5006 AEMPS Internal statistics
	Completness	% of subjects in the data with a recorded birth date	100%	https://es.slideshare.net/slideshow/protecto-bifap/13799690
		% of subjects in the data, irrespective of vital status, that have a recorded date of death	Among the patients with a recorded administrative death in BIFAP, 33.6% had a death date that matched the National Death Registry, and 84.8% were recorded within the same 30-day period.	https://zenodo.org/records/13384860 Validation against Death National Registry among people aged 35-80 years, partially provided in Abstract #1375. Volume33, IssueS2. Supplement: Abstracts of ISPEs 2024, 40th international conference, 24–28 August 2024, Germany. November 2024. e5891;
		% 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	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 ADVANCE database characterisation and fit for purpose assessment for multi-country studies on the coverage, benefits and risks of pertussis vaccinations. Sturkenboom M et al. Vaccine. 2020 Dec 22;38 Suppl 2:B8-B21. doi: 10.1016/j.vaccine.2020.01.100. Epub 2020 Feb 12. PMID: 32061385. Age-specific vaccination coverage estimates for influenza, human papillomavirus and measles containing vaccines from seven population-based healthcare databases from four EU countries - The ADVANCE project. Braeye T et al. Vaccine. 2020 Apr 3;38(16):3243-3254. doi: 10.1016/j.vaccine.2020.02.082. Epub 2020 Mar 12. PMID: 32171573
Reliability	Accuracy	The population distribution in the data source aligns with that of the country	Up to the end of 2018, BIFAP includes data from 7566 PCPs (6419 general practitioners and 1147 pediatricians). This yields a total of 12 million (2.3 million pediatric) patients for studies. Out of them, 8 million contains up-to-date information the year 2018 ("active" patients), representing 57.6% of the participating regions (14 million) and 17% of the total Spanish population (46.6 million). Now, all these counts have increased. Population age distribution are aligned with a developed country demographics reported by the National Statistics Institute (INE), although the yough adults are slightly less represented in BIFAP than country population due to the less frequent health seeking behaviour. Active population:□ Paediatric Population (< 18 years):□664591 (11.2%) Neonate: 81731 (0.3%) Infants and toddlers (28 days – 23 months):□16259 (0.5%) Children (2 to < 12 years): 1410787 (5.9%) Adolescents (12 to < 18 years):□955814 (4.5%) Adults (18 to < 46 years):□930964 (25.0%) Adults (46 to < 65 years):□089694 (21.4%) Elderly (≥ 65 years):□698925 (15.6%) Adults (65 to < 75 years):□784108 (7.5%) Adults (75 to < 85 years):□215680 (5.1%) Adults (85 years and over): 699137 (2.9%)	https://catalogues.ema.europa.eu/node/955/quantitative-descriptors Maciá-Martínez MA, Gil M, Huerta C, Martín-Merino E, Álvarez A, Bryant V, Montero D; BIFAP Team. Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria (BIFAP): A data resource for pharmacoepidemiology in Spain. Pharmacoepidemiol Drug Saf. 2020 Oct;29(10):1236-1245. doi: 10.1002/pds.5006. Epub 2020 Apr 26. PMID: 32337840. https://www.ine.es/
		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)	Internal quality checks include criteria to identify and eliminate implausible or clearly erroneous data.	Provided by DEAP
		Records of healthcare events (diagnoses, prescriptions, admissions, etc) with logical inconsistencies (e.g., and admission occurs after death)	Date values outside observation periods: 0-8.7% Date values before birth: 0% Date values after death: 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)	Pregnancy	https://pubmed.ncbi.nlm.nih.gov/31749191/
	Precision	Exposures codes precision level, including medicines and vaccines (e.g., active principle, therapeutic group, ...)	Active principle (ATC level 5 codes)	https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2023.1207976/full https://pmc.ncbi.nlm.nih.gov/articles/PMC5993167/
		Precision of date of birth (e.g., day, month, year)	Day, month and year (in house), but for researching pourpose only the year of birth or death are obtained and the month-year of birth is considered for research on babies/children. Date of birth can serve as the source of indirect re-identification.	https://zenodo.org/records/13384860
		Precision of date of death (e.g., day, month, year)	day, month and year. For researching pourpose only the year of death are obtained. Date of dearth can serve as the source of indirect re-identification.	https://www.bifap.org/data-governance?lang=en
		Precision of date of the event/diagnosis (e.g., day, month, year)	Day, month, year	Provided by DEAP
		Precision of date of the exposure (e.g., day, month, year)	Day, month, year	Provided by DEAP
	Traceability	Provenance of event records	Primary care medical records/Hospital discharge records	https://catalogues.ema.europa.eu/node/955/administrative-details

		Provenance of medicines/vaccines records	Medicines record from Pharmacy dispensing records and Vaccination records from vaccines administrations in every public healthcare setting (for most of participating regions)	https://catalogues.ema.europa.eu/node/955/administrative-details
Coherence	Format coherence	For dates, formatting constraint being followed	Character, length 8: yyyymmdd	Provided by DEAP
		For sex, formatting constraint being followed	M (male), F (female)	Provided by DEAP
	Relational coherence	% of records with the Person ID in the PERSONS table	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	ICD-10-CM ICD-9-CM Not coded (Free text) SNOMED CT	https://catalogues.ema.europa.eu/node/955/data-elements-collected
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	ATC SNOMED National drug code	https://catalogues.ema.europa.eu/node/955/data-elements-collected
	Uniqueness	Number of records flagged as potential duplicates	Requested to DEAP and unable to provide	

Step 2. SIDIAP

Dimension	Sub-dimension	Metrics	Description	Origin of information
Timeliness	Currency	How often is the database updated (i.e., frequency of updates)	6 months	https://catalogues.ema.europa.eu/node/1019/data-flows-and-management
		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)	At least 6 months plus the lag of delivery	Provided by DEAP
		The time elapsed from when a user requests the data to when they actually receive it	Depends on workload and projects timelines, 2-3 months approximately	Provided by DEAP
		Median time (years) between first and last available records for unique individuals	15 years	https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors
Extensiveness	Coverage	Percentage of a target population present in a database	5.8 million patients covered by the Catalan Institute of Health (approximately 78% of the Catalan population)	Provided by DEAP
	Completness	% of subjects in the data with a recorded birth date	100%	https://zenodo.org/records/13384860
		% 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, it is always available in Primary Care EHR	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 individuals known to have been vaccinated, 100% had the vaccine type recorded and 63% had an ATC code 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). SIDIAP is highly representative of the population of Catalonia in terms of geographical, age and sex distributions. Active population size: Paediatric Population (< 18 years): 1011295 (12.6%) Term newborn infants (0 – 27 days): 1238 (0.02%) Infants and toddlers (28 days – 23 months): 77469 (1.0%) Children (2 to < 12 years): 544330 (6.8%) Adolescents (12 to < 18 years): 388258 (4.8%) Adults (18 to < 46 years): 2105451 (26.2%) Adults (46 to < 65 years): 1636162 (20.3%) Elderly (≥ 65 years): 1144459 (14.2%) Adults (65 to < 75 years): 571570 (7.1%) Adults (75 to < 85 years): 377566 (4.7%) Adults (85 years and over): 195323 (2.4%)	https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.ine.es/ https://academic.oup.com/ije/article/51/6/e324/6567646 García-Gil Mdel M, Hermosilla E, Prieto-Alhambra D, Fina F, Rosell M, Ramos R, Rodriguez J, Williams T, Van Staa T, Bolíbar B. Construction and validation of a scoring system for the selection of high-quality data in a Spanish population primary care database (SIDIAP). Inform Prim Care. 2011;19(3):135-45. doi: 10.14236/jhi.v19i3.806. PMID: 22688222.
		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% Data values after death: 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)	Mother-child link (probabilistic)	https://www.sciencedirect.com/science/article/pii/S1532046424001655?via%3Dihub
	Precision	Exposures codes precision level, including medicines and vaccines (e.g., active principle, therapeutic group, ...)	Active principle (ATC level 5 codes)	https://www.sciencedirect.com/science/article/pii/S1532046424001655?via%3Dihub
		Precision of date of birth (e.g., day, month, year)	Day, month and year Only month and year can be provided for research purposes to avoid re-identification	Provided by DEAP
		Precision of date of death (e.g., day, month, year)	Day, month and year	https://zenodo.org/records/13384860
		Precision of date of the event/diagnosis (e.g., day, month, year)	Day, month and year	Provided by DEAP
		Precision of date of the exposure (e.g., day, month, year)	Month and year for dispensing/reimbursement. Day/month/year for prescription	https://zenodo.org/records/13384860
	Traceability	Provenance of event records	Primary care, Emergency, Hospital, Specialist and ICU	https://www.sidiap.org/index.php/es/dades-3/farmacs
		Provenance of medicines/vaccines records	Dispensing, reimbursed (administered for vaccines)	
Coherence	Format coherence	For dates, formatting constraint being followed	Requested to DEAP and unable to provide	https://zenodo.org/records/13384860
		For sex, formatting constraint being followed	Requested to DEAP and unable to provide	https://catalogues.ema.europa.eu/node/1019/data-elements-collected
	Relational coherence	% of records with the Person ID in the PERSONS table	100%	https://zenodo.org/records/13384860
	Semantic coherence - to determine	For EVENTS definitions, codelists/data dictionaries being employed according to external standards	ICD-10-CM, ICD-9-CM	https://zenodo.org/records/13384860
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	ATC code (100%), MPID (100%)	https://zenodo.org/records/13384860
	Uniqueness	Number of records flagged as potential duplicates	Requested to DEAP and unable to provide	https://catalogues.ema.europa.eu/node/1019/data-elements-collected

Step 2. VID

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%) Adults (85 years and over): 147000 (3.1%)	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 determine	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	

Step 3. CPRD-C1

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://cataloques.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. 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.	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://cataloques.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

Step 3. VID-C1

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*	Identificable				
	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.				

Step 3. NCR-C2

Scientific research question :	**What is the hazard ratio (HR) of time to death in patients with primary stage 4 non-small-cell lung cancer with a PD-L1 expression of <1% receiving dual immunotherapy (nivolumab + ipilimumab)+ chemotherapy versus pembrolizumab + chemotherapy, regardless of treatment discontinuation or treatment switch?**							
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**							
	age >18	Date of birth	*High*	100% of individuals have available date of birth	As the format is not known, precision can not be evaluated. Low impact in the study?		Since 1989 OMOP-CDM since 1992. Daily updates	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP
	Histologically confirmed squamous or non-squamous cell lung cancer	Pathology database and diagnosis code ICD-O-3	*High*	100% of individuals have available information	Once year all cancer dx are reviewed to identify cancer patients that did not have a biopsy and pathology finding.		Since 1989 OMOP-CDM since 1992. Daily updates	
	Primary stage IIIb or 4 NSCLC	TNM	*High*	Recorded	6% missing. TNM is reliable as Data is collected by well-trained data managers using coding manuals. The data entry application performs checks on the data that is entered, automatic checks are done on the database, as well as manual checks of random samples. A group of data managers is responsible for data quality and researchers in the organization can flag potential quality issues.	No formal consistent assessment has been made against Dx codes or imaging	Since 1989 OMOP-CDM since 1992. Daily updates	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP. https://pmc.ncbi.nlm.nih.gov/articles/PMC11484730/
	PDL1 expression <1%	Lab data	*High*	Recorded	No missing data			https://pmc.ncbi.nlm.nih.gov/articles/PMC11484730/
	ECOG performance score 0 or 1	ECOG score	*High*	Recorded	15% missing			https://pmc.ncbi.nlm.nih.gov/articles/PMC11484730/
	No actionable genetic mutation	genetic mutation	*Low*	Actionable genetic aberrations are recorded when relevant, provided testing was performed.				https://pubmed.ncbi.nlm.nih.gov/35461050/
	Exclusion criteria							
	Previous systemic anti-cancer treatment-autoimmune disease or severe infectious disease (e.g. HIV)		*High*	Only first line treatments.	Previous-anticancer treatment can be detected from previous patient records in the NCR			
	A previous other malignancy		*High*		Can be detected from previous patient records in the NCR			
Treatment/exposure	Nivolumab (360 mg IV/3 wk) + ipilimumab (1 mg/kg IV/ 6 wk) + 2x histology-based, platinum doublet chemotherapy (IV/ 3wk) Chemotherapy: Carboplatin (area under the concentration–time curve [AUC] 6]) plus paclitaxel (200 mg/m²) or nab-paclitaxel (100 mg/m²) on days 1, 8, and 15 (SCLC) Carboplatin (AUC 5 or 6) plus pemetrexed (500 mg/m²) or cisplatin (75 mg/m²) plus pemetrexed (500 mg/m²) (NSCLC).	Hospital prescription	*High*	Prescription first line treatment, dose not registered				DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP
Comparator group (if applicable)	Pembrolizumab (200 mg IV/3wk) for max 35 cycles + 4x chemotherapy (IV/3wk). Chemotherapy: The choice of chemotherapy is on the discretion of the treating physician	Hospital prescription	*High*	Prescription first line treatment, dose not registered				
Key endpoint(s)	Time to Death (from any cause)	Date of death	*High*	A date of death is recorded for 100% of individuals who are known to have died	Vital status checked once per year. As the date of death is registered it will be possible to calculate.		Vital status checked once per year	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP
Confounders	Performance status (ECOG)	ECOG score	*Low*		Missing in 15% of cases			
	Histology (squamous, non-squamous)	Pathology	*Low*		Pathology database (PALGA). Some cancer patients did not have a biopsy and pathology			
	Age >18	Date of birth	*Low*	100% of individuals have available information	As the format is not known, precision can not be evaluated. Low impact in the study? Population age groups recorded.		Since 1989 OMOP-CDM since 1992. Daily updates	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP

	Sex	Sex	*Low*	100% of individuals have available information	No information on gender classification			Provided by DEAP
	Site of metastasis	Site of metastasis	*Low*		Metastases site at diagnosis not at follow-up			
	Stage IIIB of IV	TNM	*Low*	TNM is missing for the 6% of subjects	TNM is reliable as Data is collected by well-trained data managers using coding manuals. The data entry application performs checks on the data that is entered, automatic checks are done on the database, as well as manual checks of random samples. A group of data managers is responsible for data quality and researchers in the organization can flag potential quality issues.	No formal consistent assessment has been made against Dx codes or imaging	Since 1989 OMOP-CDM since 1992. Daily updates	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP. https://pmc.ncbi.nlm.nih.gov/articles/PMC11484730/
Intercurrent events	Treatment discontinuation	Stop date	*Low*	100% of individuals have available information				DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP
	Anti cancer treatment swicth	only first line treatment	*Low*	100% of individuals have available information	As only first line treatment is recorded it won't be posible to diferentiate descontinuation than switch			
Follow-up time needed per patient in the study	min 3 years-max 5 years	Follow-up until date of death or date of emigration	*High*				The median length of follow-up per patient is approximately 9 months	
Minimum time in the data source for lookback assessment	Unspecified	Unspecified	*Low*				The median length of follow-up per patient is approximately 9 months	

	Estimated sample size: 244 (1:1 ratio of saline and mRNA Covid-19 vaccine, thus 122 per group)			NCR includes 5,000 patients with stage IV NSCLC and 1,000 with stage III NSCLC. Since 2021, 100 patients have been treated with nivolumab + ipilimumab, compared to 3,000 patients receiving pembrolizumab. The sample size for pembrolizumab is adequate, while the size for nivolumab + ipilimumab could be limited.			https://pubmed.ncbi.nlm.nih.gov/37833206/	38% PDL1<1% and 74% stage III/IV and 37% (PDL1<1% and stage III/IV)

Step 3. BIFAP-C3

Scientific research question		**Dapagliflozin and Major Adverse Cardiovascular Events in Type 2 Diabetes**						
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**							
	Adult >40 years of age at the time of eligibility screening	Date of birth (month/years)	*High*	100% have date of birth available		100% of patients have DOB captured in the same month and year format (MM/YYYY)		Maciá-Martínez MA, Gil M, Huerta C, Martín-Merino E, Álvarez A, Bryant V, Montero D; BIFAP Team. Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria (BIFAP): A data resource for pharmacoepidemiology in Spain. Pharmacoepidemiol Drug Saf. 2020 Oct;29(10):1236-1245. doi: 10.1002/pds.5006. Epub 2020 Apr 26. PMID: 32337840.
	Diagnosis of type diabetes 2	Diagnostic code	*High*	Diagnostic codes available for 100% of patients. In existing publications, 515,701 subjects with type 2 diabetes were identified.		For all diagnosis: in BIFAP SNOMED are utilized (for primary care recording of the majority of the participating regions), ICD-9 and ICD-10 (for diagnosis at hospital discharge).		Rodríguez-Miguel A, Fernández-Fernández B, Ortiz A, Gil M, Rodríguez-Martín S, Ruiz-Hurtado G, Fernández-Antón E, Ruilope LM, de Abajo FJ. Glucose-Lowering Drugs and Primary Prevention of Chronic Kidney Disease in Type 2 Diabetes Patients: A Real-World Primary Care Study. Pharmaceuticals (Basel). 2024 Sep 29;17(10):1299. doi: 10.3390/ph17101299. PMID: 39458940; PMCID: PMC11510410.
	Established ACVD as a history or diagnosis of ischemic heart disease, ischemic cerebrovascular disease or peripheral arterial disease during the 1 year eligibility	Diagnostic code	*High*	Diagnostic codes available for 100% of patients.			Data is updated once a year so we should be able to have this information for the year prior randomisation. However, the delay to data delivery should be accounted for. The median lenght of follow-up per patient is 10 years, which accounts for the time needed to patients to accomplish the elegibility criteria.	**Papers including estimation of the precision of ischemic stroke:** • de Abajo FJ, Rodríguez-Martín S, Rodríguez-Miguel A, Gil MJ. Risk of ischemic stroke associated with calcium supplements with or without vitamin D: a nested case-control study. J Am Heart Assoc. 2017;6(5): pii:e005795. https://doi.org/10.1161/JAHA.117.005795. • García-Poza P, de Abajo FJ, Gil MJ, Chacón A, Bryant V, García- Rodríguez LA. Risk of ischemic stroke associated with non-steroidal anti-inflammatory drugs and paracetamol: a population-based casecontrol study. J Thromb Haemost. 2015;13(5):708-718. https://doi. Org/10.1111/jth.12855. • Barreira-Hernández D, Rodríguez-Martín S, Gil M, Mazzucchelli R, Izquierdo-Esteban L, García-Lledó A, Pérez-Gómez A, Rodríguez-Miguel A, de Abajo FJ. Risk of Ischemic Stroke Associated with Calcium Supplements and Interaction with Oral Bisphosphonates: A Nested Case-Control Study. Journal of Clinical Medicine. 2023; 12(16):5294. https://doi.org/10.3390/jcm12165294 • Rodríguez-Martín S, Barreira-Hernández D, Gil M, García-Lledó A, Izquierdo-Esteban L, De Abajo FJ. Influenza Vaccination and Risk of Ischemic Stroke: A Population-Based Case-Control Study [published online ahead of print, 2022 Sep 7]. Neurology. 2022;10.1212 • Alqdwah-Fattouh R, Rodríguez-Martín S, Barreira-Hernández D, et al. Selective Serotonin Reuptake Inhibitors and Risk of Noncardioembolic Ischemic Stroke: A Nested Case-Control Study. Stroke. 2022;53(5):1560-1569. doi:10.1161/STROKEAHA.121.036661 **Papers including estimation of the precision of venous thromboembolism** Martín-Merino E, Petersen I, Hawley S, et al. Risk of venous thromboembolism among users of different anti-osteoporosis drugs: a population-based cohort analysis including over 200,000 participants from Spain and the UK. Osteoporos Int. 2018;29(2):467-478. https:// doi.org/10.1007/s00198-017-4308-5.
	High ACVD risk defined as no established ACVD, age ≥ 55 years in men and ≥ 60 in women and one orf more of the following: - History or diagnosis of dyslipidemia - Current lipid lowering therapy - History or diagnosis of hypertension - Current anti-hypertensive medication use prescribed for blood pressure lowering - Current tobacco use or within 1 year prior to randomisation	Sex Date of birth Diagnostic code (ICD-10 or equivalent) Smoking habits Prescription/dispensing data Indication linked to drug use	*High*	100% have date of birth, sex, diagnostic codes and drug codes available. Smoking available 55% of records. In existing publications, 57.1% of a total of 515,701 T2D patients had dyslipidemia and 75.9% hypertension. Also, ~33,765 ex-smokers and 130,462 current smokers were identified. A study reported 56.3% missingness of smoking reporting in diabetic patients.	Drug use is not linked to a specific indication. Similarly, smoking status may also be biased, as the criterion is current use or use within one year prior to randomization; therefore, patients who smoked before this period would be classified as non-smokers.		The median lenght of follow-up per patient is 10 years, which accounts for the time needed to patients to accomplish the elegibility criteria.	Martín-Merino E, Calderón-Larrañaga A, Hawley S, Poblador-Plou B, Llorente-García A, Petersen I, Prieto-Alhambra D. The impact of different strategies to handle missing data on both precision and bias in a drug safety study: a multidatabase multinational population-based cohort study. Clin Epidemiol. 2018 Jun 5;10:643-654. doi: 10.2147/CLEP.S154914. PMID: 29892204; PMCID: PMC5993167. Rodríguez-Miguel A, Fernández-Fernández B, Ortiz A, Gil M, Rodríguez-Martín S, Ruiz-Hurtado G, Fernández-Antón E, Ruilope LM, de Abajo FJ. Glucose-Lowering Drugs and Primary Prevention of Chronic Kidney Disease in Type 2 Diabetes Patients: A Real-World Primary Care Study. Pharmaceuticals (Basel). 2024 Sep 29;17(10):1299. doi: 10.3390/ph17101299. PMID: 39458940; PMCID: PMC11510410. Quesada JA, Orozco-Beltran D. Analysis of missing data in electronic health records of people with diabetes in primary care in Spain: A population-based cohort study. Int J Med Inform. 2025 Feb;194:105722. doi: 10.1016/j.ijmedinf.2024.105722. Epub 2024 Nov 23. PMID: 39586146.
	Exclusion criteria							
	Treatment with SGLT2i or DPP-4i in the last year prior randomisation	Prescription/dispensing data	*High*	ATC codes 100% available. In existing publications, these drug groups have already been studied.			Data is updated once per year so we should be able to have this information for the year prior randomisation. However, the delay to data delivery should be accounted for. The median lenght of follow-up per patient is 10 years, which accounts for the time needed to patients to accomplish the elegibility criteria.	Rodríguez-Miguel A, Fernández-Fernández B, Ortiz A, Gil M, Rodríguez-Martín S, Ruiz-Hurtado G, Fernández-Antón E, Ruilope LM, de Abajo FJ. Glucose-Lowering Drugs and Primary Prevention of Chronic Kidney Disease in Type 2 Diabetes Patients: A Real-World Primary Care Study. Pharmaceuticals (Basel). 2024 Sep 29;17(10):1299. doi: 10.3390/ph17101299. PMID: 39458940; PMCID: PMC11510410. Study report of SAFEGUARD proyect aimed at assessing safety of antidiabetic drugs: Community Research and Development Information Service (CORDIS), European Commission. Final Report Summary—SAFEGUARD (Safety Evaluation of Adverse Reactions in Diabetes). Final Publishable Summary Report. Erasmus Universitair Medisch Centrum Rotterdam, Netherlands. 2015. https://cordis.europa.eu/ docs/results/282/282521/final1-safeguard_final-publishable-summaryreport. pdf Accessed December 12, 2019.
	Treatment with pioglitazone or rosiglitazone treatment in the last year prior randomisation	Prescription/dispensing data	*High*	ATC codes 100% available.			Data is updated twice a year so we should be able to have this information for the year prior randomisation. However, the delay to data delivery should be accounted for. The median lenght of follow-up per patient is 10 years, which accounts for the time needed to patients to accomplish the elegibility criteria.	

	Acute cardiovascular event in the last year prior randomisation	Diagnostic code (ICD-10 or equivalent) Emergency room and/or hospitalisation diagnoses	*High*	100% records have a diagnostic code. This is likely to be a diagnosis requiring admission to the hospital. In BIFAP, hospital information can be linked for a part of their population. Hospital information in BIFAP includes dates of admission and discharge, type of discharge, primary and secondary diagnoses at hospital discharge. So, an acute cardiovascular event will only be picked if it	Hospital information in BIFAP includes dates of admission and discharge, type of discharge, primary and secondary diagnoses at hospital discharge. So, an acute cardiovascular event will only be picked if it constituted one of the main reasons for admission. It is deemed to be highly reliable; however, some in-hospital events might be missed.		Data is updated twice a year so we should be able to have this information for the year prior randomisation. However, the delay to data delivery should be accounted for. The median lenght of follow-up per patient is 10 years, which accounts for the time needed to patients to accomplish the elegibility criteria.	
	Diagnosis Type 1 diabetes any time before randomisation	Diagnostic code (ICD-10 or equivalent)	*High*	Diagnostic codes available for 100% of patients	In published studies using BIFAP, patients with T1D were detected by using insulin in monotherapy as aproxy.		Data is updated twice a year so we should be able to have this information for the year prior randomisation. However, the delay to data delivery should be accounted for. The median lenght of follow-up per patient is 10 years, which accounts for the time needed to patients to accomplish the elegibility criteria.	Rodríguez-Miguel A, Fernández-Fernández B, Ortiz A, Gil M, Rodríguez-Martín S, Ruiz-Hurtado G, Fernández-Antón E, Ruilope LM, de Abajo FJ. Glucose-Lowering Drugs and Primary Prevention of Chronic Kidney Disease in Type 2 Diabetes Patients: A Real-World Primary Care Study. Pharmaceuticals (Basel). 2024 Sep 29;17(10):1299. doi: 10.3390/ph17101299. PMID: 39458940; PMCID: PMC11510410.
Treatment/exposure	Dapagliflozin	Medication codes	*High*	ATC codes 100% available.				
Comparator group (if applicable)	DPP4i	Medication codes	*High*	ATC codes 100% available.				
Key endpoint(s)	Time to first MACE (non-fatal MI, stroke, all-cause death)	Date of death Date of diagnosis Diagnostic code	*High*	Among the patients with a recorded administrative death in BIFAP, 33.6% had a death date that matched the National Death Registry, and 84.8% were recorded within the same 30-day period. A non-random pattern of missingness (MNAR) was observed due to incomplete or inaccurate recording of the cause of death, with a tendency to preferentially register cardiovascular-related deaths. Consequently, adjustments using statistical methods for MNAR should be considered in the TTE protocol.				Papers on nonfatal acute myocardial infarction precisión/validacion estimation • de Abajo FJ, Gil MJ, García Poza P, et al. Risk of nonfatal acute myocardial infarction associated with non-steroidal antiinflammatory drugs, non-narcotic analgesics and other drugs used in osteoarthritis: a nested case-control study. Pharmacoepidemiol Drug Saf. 2014;23 (11):1128-1138. https://doi.org/10.1002/pds.3617. • de Abajo FJ, Gil MJ, Rodríguez A, et al. Allopurinol use and risk of non-fatal acute myocardial infarction. Heart. 2015;101(9):679-685. https://doi.org/10.1136/heartjnl-2014-306670. • Rodríguez-Martín S, de Abajo FJ, Gil M, et al. Risk of acute myocardial infarction among new users of allopurinol according to serum urate level: a nested case-control study. J Clin Med. 2019;8(12):pii:E2150. https://doi.org/10.3390/jcm8122150. • Mazzucchelli R, Rodríguez-Martín S, García-Vadillo A, et al. Risk of acute myocardial infarction among new users of chondroitin sulfate: A nested case-control study. PLoS One. 2021;16(7):e0253932. Published 2021 Jul 12. doi:10.1371/journal.pone.0253932 • de Abajo FJ, Rodríguez-Martín S, Barreira D, et al. Influenza vaccine and risk of acute myocardial infarction in a population-based case-control study. Heart. 2022;108(13):1039-1045. Published 2022 Jun 10. doi:10.1136/heartjnl-2021-319754 • Mazzucchelli R, Rodríguez-Martín S, García-Vadillo A, et al. Risk of acute myocardial infarction among new users of bisphosphonates: a nested case-control study. Osteoporos Int. 2020;31(12):2403-2412. doi:10.1007/s00198-020-05538-2 Papers on ischemic stroke precisión/validacion estimation • de Abajo FJ, Rodríguez-Martín S, Rodríguez-Miguel A, Gil MJ. Risk of ischemic stroke associated with calcium supplements with or without vitamin D: a nested case-control study. J Am Heart Assoc. 2017;6(5): pii:e005795. https://doi.org/10.1161/JAHA.117.005795. • García-Poza P, de Abajo FJ, Gil MJ, Chacón A, Bryant V, García- Rodríguez LA. Risk of ischemic stroke associated with non-steroidal anti-inflammatory drugs and paracetamol: a population-based casecontrol study. J Thromb Haemost. 2015;13(5):708-718. https://doi. Org/10.1111/jth.12855. • Barreira-Hernández D, Rodríguez-Martín S, Gil M, Mazzucchelli R, Izquierdo-Esteban L, García-Lledó A, Pérez-Gómez A, Rodríguez-Miguel A, de Abajo FJ. Risk of Ischemic Stroke Associated with Calcium Supplements and Interaction with Oral Bisphosphonates: A Nested Case-Control Study. Journal of Clinical Medicine. 2023; 12(16):5294. https://doi.org/10.3390/jcm12165294 • Rodríguez-Martín S, Barreira-Hernández D, Gil M, García-Lledó A, Izquierdo-Esteban L, De Abajo FJ. Influenza Vaccination and Risk of Ischemic Stroke: A Population-Based Case-Control Study [published online ahead of print, 2022 Sep 7]. Neurology. 2022;10.1212 • Alqdwah-Fattouh R, Rodríguez-Martín S, Barreira-Hernández D, et al. Selective Serotonin Reuptake Inhibitors and Risk of Noncardioembolic Ischemic Stroke: A Nested Case-Control Study. Stroke. 2022;53(5):1560-1569. doi:10.1161/STROKEAHA.121.036661 Precision of administrative death records evaluated in the following published poster: Validation against Death National Registry among people aged 35-80 years, partially provided in Abstract #1375. Volume33, IssueS2. Supplement: Abstracts of ISPEs 2024, 40th international conference, 24–28 August 2024, Germany. November 2024. e5891;
Confounders	Age at index	Date of birth (month/years)	*Low*	100% have date of birth available				Rodríguez-Miguel A, Fernández-Fernández B, Ortiz A, Gil M, Rodríguez-Martín S, Ruiz-Hurtado G, Fernández-Antón E, Ruilope LM, de Abajo FJ. Glucose-Lowering Drugs and Primary Prevention of Chronic Kidney Disease in Type 2 Diabetes Patients: A Real-World Primary Care Study. Pharmaceuticals (Basel). 2024 Sep 29;17(10):1299. doi: 10.3390/ph17101299. PMID: 39458940; PMCID: PMC11510410. Mar Martín-Pérez et al. Multiple vertebral fractures after antiosteoporotic medications discontinuation: A comparative study to evaluate the potential rebound effect of denosumab PMID: 39521365 DOI: 10.1016/j.bone.2024.117325 Quesada JA, Orozco-Beltran D. Analysis of missing data in electronic health records of people with diabetes in primary care in Spain: A population-based cohort study. Int J Med Inform. 2025 Feb;194:105722. doi: 10.1016/j.ijmedinf.2024.105722. Epub 2024 Nov 23. PMID: 39586146. Martín-Merino E et al. Cessation rate of anti-osteoporosis treatments and risk factors in Spanish primary care settings: a population-based cohort analysis Arch Osteoporos (2017) 12:1-12. DOI 10.1007/s11657-017-0331-6

Gender: female or male	Sex	*Low*	100% have date of birth, sex, diagnostic codes and drug codes available.				
Frailty	Diagnostic code (ICD-10 or equivalent)	*Low*		In a published study in specific regions of Spain, 26.2% of 855 were found to be frail by using Fried criteria. Other studies report ~10%. Institucionalized patients went up to 68.8%. This can be taken as reference to benchmark to.			Rivas-Ruiz F, Machón M, Contreras-Fernández E, Vrotsou K, Padilla-Ruiz M, Díez Ruiz AI, de Mesa Berenguer Y, Vergara I; Group GIFEA. Prevalence of frailty among community-dwelling elderly persons in Spain and factors associated with it. Eur J Gen Pract. 2019 Oct;25(4):190-196. doi: 10.1080/13814788.2019.1635113. Epub 2019 Oct 22. PMID: 31637940; PMCID: PMC6853242. Jürschik P, Nunin C, Botigué T, Escobar MA, Lavedán A, Viladrosa M. Prevalence of frailty and factors associated with frailty in the elderly population of Lleida, Spain: the FRALLE survey. Arch Gerontol Geriatr. 2012 Nov-Dec;55(3):625-31. doi: 10.1016/j.archger.2012.07.002. Epub 2012 Jul 31. PMID: 22857807. González-Vaca J, de la Rica-Escuín M, Silva-Iglesias M, Arjonilla-García MD, Varela-Pérez R, Oliver-Carbonell JL, Abizanda P. Frailty in INstitutionalized older adults from ALbacete. The FINAL Study: rationale, design, methodology, prevalence and attributes. Maturitas. 2014 Jan;77(1):78-84. doi: 10.1016/j.maturitas.2013.10.005. Epub 2013 Oct 16. PMID: 24189222.
Obesity: defined as a separate diagnosis and/or BMI greater than or equal to 30.	BMI or weight and height Diagnostic code (ICD-10 or equivalent)	*Low*	In a published study on 515,701 T2D patients 312,383 were fount to be obese by using BMI > or = 30kg/m^2. Another study reported 35.4% missingness of BMI in diabetic patients.				Rodríguez-Miguel A, Fernández-Fernández B, Ortiz A, Gil M, Rodríguez-Martín S, Ruiz-Hurtado G, Fernández-Antón E, Ruilope LM, de Abajo FJ. Glucose-Lowering Drugs and Primary Prevention of Chronic Kidney Disease in Type 2 Diabetes Patients: A Real-World Primary Care Study. Pharmaceuticals (Basel). 2024 Sep 29;17(10):1299. doi: 10.3390/ph17101299. PMID: 39458940; PMCID: PMC11510410. Quesada JA, Orozco-Beltran D. Analysis of missing data in electronic health records of people with diabetes in primary care in Spain: A population-based cohort study. Int J Med Inform. 2025 Feb;194:105722. doi: 10.1016/j.ijmedinf.2024.105722. Epub 2024 Nov 23. PMID: 39586146.
Heart transplant	Diagnostic code (ICD-10 or equivalent) Procedure code	*Low*					
Microvascular complications: mono-/polyneuropathy, eye complications, Diabetic foot/Peripheral angiopathy, nephropathy, Diabetes with several-/unspecified complications	Diagnostic code (ICD-10 or equivalent)	*Low*					
Severe hypoglycemia	Diagnostic code (ICD-10 or equivalent) Laboratory values	*Low*	In previous studies on patients with diabetes in BIFAP, The proportion of individuals with at least one missing value was 76.0%. Regarding diabetes control measures, 10.8% of records had missing glycated hemoglobin values, and 21.4% had missing basal blood glucose values.				Quesada JA, Orozco-Beltran D. Analysis of missing data in electronic health records of people with diabetes in primary care in Spain: A population-based cohort study. Int J Med Inform. 2025 Feb;194:105722. doi: 10.1016/j.ijmedinf.2024.105722. Epub 2024 Nov 23. PMID: 39586146.
Keto-/lactate acidosis	Diagnostic code (ICD-10 or equivalent) Laboratory values	*Low*					
Lower limb amputations	Diagnostic code (ICD-10 or equivalent) or procedure code	*Low*					
Chronic obstructive pulmonary disease (COPD)	Diagnostic code (ICD-10 or equivalent) Medication code and date (as proxies)	*Low*	100% medication codes available.	COPD frenquency and comparison between diabetic and non-diabetic patients has beed assessed and reported in a paper referred in column 'I'			Arias Fernández, L., Pardo Seco, J., Cebey-López, M. et al. Differences between diabetic and non-diabetic patients with community-acquired pneumonia in primary care in Spain. BMC Infect Dis 19, 973 (2019). https://doi.org/10.1186/s12879-019-4534-x
Cancer	Diagnostic code (ICD-10 or equivalent) Medication code and date (as proxies)	*Low*	A previous study reported 0 missings on cancer recording in diabetic patients. 100% medication codes available.	Validation studies have been performed on validation of colorectal cancer diagnosis in BIFAP, so these are deemed to eb reliable (PPV>92% and NPV 100%). Also, an algorithm to detect digestive cancer has been developed and validated.			Quesada JA, Orozco-Beltran D. Analysis of missing data in electronic health records of people with diabetes in primary care in Spain: A population-based cohort study. Int J Med Inform. 2025 Feb;194:105722. doi: 10.1016/j.ijmedinf.2024.105722. Epub 2024 Nov 23. PMID: 39586146. Gil M, Rodríguez-Miguel A, Montoya-Catalá H, González-González R, Álvarez-Gutiérrez A, Rodríguez-Martín S, García-Rodríguez LA, de Abajo FJ. Validation study of colorectal cancer diagnosis in the Spanish primary care database, BIFAP. Pharmacoepidemiol Drug Saf. 2019 Feb;28(2):209-216. doi: 10.1002/pds.4686. Epub 2018 Dec 12. PMID: 30548462. Fernández-Antón E, Rodríguez-Miguel A, Gil M, Castellano-López A, de Abajo FJ. Development and Validation of Case-Finding Algorithms for Digestive Cancer in the Spanish Healthcare Database BIFAP. J Clin Med. 2024 Jan 9;13(2):361. doi: 10.3390/jcm13020361. PMID: 38256495; PMCID: PMC10816118.
Major organ specific bleeding	Diagnostic code (ICD-10 or equivalent) Procedure code	*Low*					
Bariatric surgery	Diagnostic code (ICD-10 or equivalent) or procedure code	*Low*					

Chronic kidney disease (CKD) stages 1-4	Diagnostic code (ICD-10 or equivalent) Laboratory values	*Low*	Cases whose first abnormal record was albuminuria/proteinuria or a diagnosis of CKD and without a record of eGFR could not be classified. Neither albuminuria nor proteinuria were exhaustively recorded, which may lead to a certain under-recording of CKD in its early stages. Imaging or hisological findings are not recorded.	In previous studies using BIFAP, they classified CKD using the KDIGO criteria (except imaging or histological findings). Also, they defined as "chronic kidney insufficiency" cases of CKD at stages from G3a to G5. The following information was concluded from a previous study using BIFAP, which may help to assess the plausability of our results: "Regarding the epidemiology of CKD in primary care and among patients with type 2 diabetes, we found the following: (1) over the study period, the incidence rate of CKD was stable overall; (2) in patients older than 70 years, the incidence rate was higher in females than in males; and (3) the factors more strongly associated with incident CKD were the antecedents of gout, hyperuricemia, hyperkalemia, hypertension, heart failure, hyperparathyroidism, and prior isolated abnormal values of eGFR or proteinuria/albuminuria." In T2D patients, levels of serum creatinine or eGFR were recorded in the database for each subject every year, on average.	Around 10% of cases were detected by albuminuria/proteinuria or a recorded diagnosis of CKD but with no data on eGFR or creatinine at the index date. Thus, these two variables seem not to be concordantly recorded.	Levels of serum creatinine or eGFR were recorded in the database for each subject every year, on average	Rodríguez-Miguel A, Fernández-Fernández B, Ortiz A, Gil M, Rodríguez-Martín S, Ruiz-Hurtado G, Fernández-Antón E, Ruilope LM, de Abajo FJ. Glucose-Lowering Drugs and Primary Prevention of Chronic Kidney Disease in Type 2 Diabetes Patients: A Real-World Primary Care Study. Pharmaceuticals (Basel). 2024 Sep 29;17(10):1299. doi: 10.3390/ph17101299. PMID: 39458940; PMCID: PMC11510410.
End stage kidney disease (CKD stage 5)	Diagnostic code (ICD-10 or equivalent) Laboratory values Procedure codes (i.e., dialysis)	*Low*	Cases whose first abnormal record was albuminuria/proteinuria or a diagnosis of CKD and without a record of eGFR could not be classified. Neither albuminuria nor proteinuria were exhaustively recorded, which may lead to a certain under-recording of CKD in its early stages. Imaging or hisological findings are not recorded.	In previous studies using BIFAP, they classified CKD using the KDIGO criteria (except imaging or histological findings). Also, they defined as "chronic kidney insufficiency" cases of CKD at stages from G3a to G5. The following information was concluded from a previous study using BIFAP, which may help to assess the plausability of our results: "Regarding the epidemiology of CKD in primary care and among patients with type 2 diabetes, we found the following: (1) over the study period, the incidence rate of CKD was stable overall; (2) in patients older than 70 years, the incidence rate was higher in females than in males; and (3) the factors more strongly associated with incident CKD were the antecedents of gout, hyperuricemia, hyperkalemia, hypertension, heart failure, hyperparathyroidism, and prior isolated abnormal values of eGFR or proteinuria/albuminuria." In T2D patients, levels of serum creatinine or eGFR were recorded in the database for each subject every year, on average.	Around 10% of cases were detected by albuminuria/proteinuria or a recorded diagnosis of CKD but with no data on eGFR or creatinine at the index date. Thus, these two variables seem not to be concordantly recorded.	Levels of serum creatinine or eGFR were recorded in the database for each subject every year, on average	Rodríguez-Miguel A, Fernández-Fernández B, Ortiz A, Gil M, Rodríguez-Martín S, Ruiz-Hurtado G, Fernández-Antón E, Ruilope LM, de Abajo FJ. Glucose-Lowering Drugs and Primary Prevention of Chronic Kidney Disease in Type 2 Diabetes Patients: A Real-World Primary Care Study. Pharmaceuticals (Basel). 2024 Sep 29;17(10):1299. doi: 10.3390/ph17101299. PMID: 39458940; PMCID: PMC11510410.
All separate GLD (glucose-lowering drugs): biguanides (metformin), sulfonylurea, sulfonamides, alfa glucoside inhibitors, thiazolidinediones, other blood glucose lowering drugs, insulin	Medication codes	*Low*	100% have date of birth, sex, diagnostic codes and drug codes available.				
Drugs to prevent CVD: I) Antihypertensives (Angiotensin-converting-enzyme inhibitors, Angiotensin receptor blockers, Beta-blockers, Low-/high ceiling diuretics, Aldosterone antagonists, Thiazide diuretics); II) Ca channel blockers; III) Digitoxin/digoxin, IV) Antiarrhythmics (flekanide, amiodarone); V) Statins; VI) Anticoagulants (Warfarin); and VII) Antiplatelet agents (Low dose acetylsalicylic acid, Receptor P2Y12 antagonists, Other antiplatelets)	Medication codes	*Low*	100% have date of birth, sex, diagnostic codes and drug codes available.				
Corticosteroids	Medication codes	*Low*					
Weigh loss drug	Medication codes	*Low*					

Intercurrent events	Treatment discontinuation	Date of drug discontinuation	*High*		In BIFAP, the information regarding the date of dispensation, the number of dispensed packages and the number of doses per package is recorded. However, duration of each dispensed package might be only calculated if the doctor wrote prescription instructions (i.e. posology). If it is not available, there is a multistep algorithm to derive the duration of the package dispensed, based on the date of dispensation at the pharmacy, the distance between subsequent prescriptions, the mode or median of prescription duration in the data set, or imputing 30 days. Consequently, treatment discontinuation can be determined whenever algorithms are based on these data points.			
	Treatment switch	Date of drug discontinuation Date of drug start	*High*		In BIFAP, the information regarding the date of dispensation, the number of dispensed packages and the number of doses per package is recorded. However, duration of each dispensed package might be only calculated if the doctor wrote prescription instructions (i.e. posology). If it is not available, there is a multistep algorithm to derive the duration of the package dispensed, based on the date of dispensation at the pharmacy, the distance between subsequent prescriptions, the mode or median of prescription duration in the data set, or imputing 30 days. Consequently, treatment discontinuation can be determined whenever algorithms are based on these data points.			
	Addition of another antihyperglycemic therapy	Medication codes	*High*	100% have date of birth, sex, diagnostic codes and drug codes available.				
	Non-CV death	Diagnostic code Date of death	*High*		Depending on the instance utilized, BIFAP may or may not have the cause of death data linked (i.e., the National Registry of Mortality). In instances where such data is linked, the cause of death will be recorded using diagnostic codes (ICD-10 or equivalent).Mortality data is updated with a one-year delay relative to the present time. For research purposes only the year of death is available. This can impact precision and the time sequence of outcomes.			
Follow-up time needed per patient in the study	5 years	6 years (including recruitment and follow-up)	*High*	The median time between first and last available records for any individual is 10 years. For active individuals, 12 years. Thus, presumably 6 years of follow-up time will be available.			The median time between first and last available records for any individual is 10 years. For active individuals, 12 years. Seems the needed timeliness will be met.	https://catalogues.ema.europa.eu/node/955/quantitative-descriptors
Minimum time in the data source for lookback assessment	1 year	1 year of lookback	*High*				Data is updated once a year so we should be able to have this information for the year prior randomisation. However, the delay to data delivery should be accounted for. The median lenght of follow-up per patient is 10 years.	https://catalogues.ema.europa.eu/node/955/quantitative-descriptors

Others	Rescue medications: acute use of insulin or sulfonylureas	Medication codes Treatment duration	*Low*	In BIFAP, the information regarding the date of dispensation, the number of dispensed packages and the number of doses per package is recorded. However, duration of each dispensed package might be only calculated if the doctor wrote prescription instructions (i.e. posology). If it is not available, there is a multistep algorithm to derive it, to calculate the duration of the package dispensed, based on the date of dispensation at the pharmacy, the distance between subsequent prescriptions, the mode or median of prescription duration in the data set, or imputing 30 days.	Depending on what is available, the accuracy of the duration measurement can be impacted.			

	Estimated sample size: Approx. 13,341 participants			Considering that BIFAP includes data from approximately 14 million inhabitants (up to 2018), the target sample size is anticipated to be reached				

Step 3. CPRD-C3

Scientific research question		**Dapagliflozin and Major Adverse Cardiovascular Events in Type 2 Diabetes**						
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**							
	Adult >40 years of age at the time of eligibility screening	Date of birth (years)	*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	https://cataloques.ema.europa.eu/node/1026/quantitative-descriptors
	Diagnosis of type diabetes 2	Diagnostic code	*High*	Diagnostic codes available for 100% of patients	A diagnosis code of type 2 diabetes is likely to be correct where present (correctness 99%). From DEAP experience, diabetes mellitus without type specification occurs frequently as well; usually insulin in monotherapy is used to assess T1D and NIADS for, T2D.	Nearly all patients who had elevated HbA1c labs or hypoglycemic treatments also had a type 2 diabetes diagnosis code (concordance >90%). In CPRD seems lab values and prescriptions are less likely to be missing than diagnoses.		CPRD Aurum database: Assessment of data quality andcompleteness of three important comorbidities, including diabetes: https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135 *"Among 37 502 patients in CPRD Aurum, correctness of type 2 diabetes,hyperlipidemia, and anemia diagnoses was high (99%, 93%, and 97%, respectively).Completeness was only high for type 2 diabetes (94%-98%);"*
	Established ACVD as a history or diagnosis of ischemic heart disease, ischemic cerebrovascular disease or peripheral arterial disease during the 1 year eligibility	Diagnostic code	*High*	Diagnostic codes available for 100% of patients			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	https://cataloques.ema.europa.eu/node/1026/quantitative-descriptors
	High ACVD risk defined as no established ACVD, age ≥ 55 years in men and ≥ 60 in women and one orf more of the following: - History or diagnosis of dyslipidemia - Current lipid lowering therapy - History or diagnosis of hypertension - Current anti-hypertensive medication use prescribed for blood pressure lowering - Current tobacco use or within 1 year prior to randomisation	Sex Date of birth Diagnostic code (ICD-10 or equivalent) Smoking habits Prescription/dispensing data Indication linked to drug use	*High*	Smoking present for 89.7% of records. Only prescription medicines (100%) Not especific linked indication to drug use; but codes Dx are used as a aproximation			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	https://cataloques.ema.europa.eu/node/1026/quantitative-descriptors
	Exclusion criteria							
	Treatment with SGLT2i or DPP-4i in the last year prior randomisation	Prescription/dispensing data	High	Only prescription medicines (100%)		As ATC codes are not available, a mapping to ATC will potentially be needed to extract study drugs information.	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	https://cataloques.ema.europa.eu/node/1026/quantitative-descriptors
	Treatment with pioglitazone or rosiglitazone treatment in the last year prior randomisation	Prescription/dispensing data	*High*	Only prescription medicines (100%)			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	https://cataloques.ema.europa.eu/node/1026/quantitative-descriptors
	Acute cardiovascular event in the last year prior randomisation	Diagnostic code (ICD-10 or equivalent) Emergency room and/or hospitalisation diagnoses	*High*	Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects attending emergency room			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	Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC) https://academic.oup.com/ije/article/46/4/1093/3072145?login=true
	Diagnosis Type 1 diabetes any time before randomisation	Diagnostic code (ICD-10 or equivalent)	*High*	Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects attending emergency room	From DEAP experience, diabetes mellitus without type specification occurs frequently as well; usually insulin in monotherapy is used to assess T1D and NIADS for, T2D.		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	https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135

Treatment/exposure	Dapagliflozin	Medication codes	*High*	100% of individuals have available information		Nearly all patients who had elevated HbA1c labs or hypoglycemic treatments also had a type 2 diabetes diagnosis code. 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.		https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135
Comparator group (if applicable)	DPP4i	Medication codes	*High*	100% of individuals have available information		As ATC codes are not available, a mappint 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.		
Key endpoint(s)	Time to first MACE (non-fatal MI, stroke, all-cause death)	Date of death Date of diagnosis	*High*	8% of the whole population (irrespective of vital status) has a date of death recorded; 100% of death people have a date of death Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects in attending emergency room				
Confounders	Age at index	Date of birth (month/years)	*Low*	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	https://cataloques.ema.europa.eu/node/1026/quantitative-descriptors
	Gender: female or male	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.				
	Frailty	Diagnostic code (ICD-10 or equivalent)	*Low*	In previous studies the Charlson comorbidity index has been used, as well as dementia. In CPRD also eFI index is available (see link), among others.				https://www.sciencedirect.com/science/article/pii/S2214623720300351?via%3Dihub#s0055 https://www.cprd.com/approved-studies/exploring-role-electronic-frailty-index-efi-using-routine-primary-care-electronic
	Obesity: defined as a separate diagnosis and/or BMI greater than or equal to 30.	BMI or weight and height Diagnostic code (ICD-10 or equivalent)	*Low*	In previous studies it seems obesity has been defined by BMI and also as diagnose. Both strategies might be considered to capture obesity.		A significant proportion of cases of hyperlipidemia will be missed if the investigator relies solely on diagnosis codes to select patients.		https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135 https://www.sciencedirect.com/science/article/pii/S2214623720300351?via%3Dihub#s0055
	Heart transplant	Diagnostic code (ICD-10 or equivalent) Procedure code	*Low*	Diagnostic codes available for 100% of patients				
	Microvascular complications: mono-/polyneuropathy, eye complications, Diabetic foot/Peripheral angiopathy, nephropathy, Diabetes with several-/unspecified complications	Diagnostic code (ICD-10 or equivalent)	*Low*	Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects attending emergency room				
	Severe hypoglycemia	Diagnostic code (ICD-10 or equivalent) Laboratory values	*Low*	Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects attending emergency room	A significant proportion of lab data lacking a normal range were missing units.	A significant proportion of lab data lacking a normal range had values inconsistent with units provided.		https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135
	Keto-/lactate acidosis	Diagnostic code (ICD-10 or equivalent) Laboratory values	*Low*	DX codes 100% Laboratory values	A significant proportion of lab data lacking a normal range were missing units.	A significant proportion of lab data lacking a normal range had values inconsistent with units provided.		https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135
	Lower limb amputations	Diagnostic code (ICD-10 or equivalent) or procedure code	*Low*	Diagnostic codes available for 100% of patients				
	Chronic obstructive pulmonary disease (COPD)	Diagnostic code (ICD-10 or equivalent) Medication code and date (as proxies)	*Low*	Diagnostic codes available for 100% of patients Drug codes are available for 100% of patients Diagnostic codes are available for 86% subjects in attending emergency room				

	Cancer	Diagnostic code (ICD-10 or equivalent) Medication code and date (as proxies)	*Low*	Diagnostic codes available for 100% of patients Drug codes are available for 100% of patients Diagnostic codes are available for 86% subjects in attending emergency room				
	Major organ specific bleeding	Diagnostic code (ICD-10 or equivalent) Procedure code	*Low*	Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects in attending emergency room				
	Bariatric surgery	Diagnostic code (ICD-10 or equivalent) or procedure code	*Low*	Diagnostic codes available for 100% of patients				
	Chronic kidney disease (CKD) stages 1-4	Diagnostic code (ICD-10 or equivalent) Laboratory values	*Low*	Diagnostic codes available for 100% of patients	A significant proportion of lab data lacking a normal range were missing units.	A significant proportion of lab data lacking a normal range had values inconsistent with units provided.		https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135
	End stage kidney disease (CKD stage 5)	Diagnostic code (ICD-10 or equivalent) Laboratory values Procedure codes (i.e., dialysis)	*Low*	Diagnostic codes available for 100% of patients	A significant proportion of lab data lacking a normal range were missing units.	A significant proportion of lab data lacking a normal range had values inconsistent with units provided.		
	All separate GLD (glucose-lowering drugs): biguanides (metformin), sulfonylurea, sulfonamides, alfa glucoside inhibitors, thiazolidinediones, other blood glucose lowering drugs, insulin	Medication codes	*Low*	Drug codes are available for 100% of patients		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.		
	Drugs to prevent CVD: I) Antihypertensives (Angiotensin-converting-enzyme inhibitors, Angiotensin receptor blockers, Beta-blockers, Low-/high ceiling diuretics, Aldosterone antagonists, Thiazide diuretics); II) Ca channel blockers; III) Digitoxin/digoxin, IV) Antiarrhythmics (flekanide, amiodarone); V) Statins; VI) Anticoagulants (Warfarin); and VII) Antiplatelet agents (Low dose acetylsalicylic acid, Receptor P2Y12 antagonists, Other antiplatelets)	Medication codes	*Low*	Drug codes are available for 100% of patients		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.		
	Corticosteroids	Medication codes	*Low*	Drug codes are available for 100% of patients		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.		
	Weigh loss drug	Medication codes	*Low*	Drug codes are available for 100% of patients		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.		
Intercurrent events	Treatment discontinuation	Date of drug discontinuation	*High*	In CPRD information on treatment duration is available. This may help defining drug exposure-related variables.	Previous studies have analysed metformin discontinuation and adherence. As CPRD has prescription data, it is unknown whether the patient took the prescription.			https://www.sciencedirect.com/science/article/pii/S2214623720300351?via%3Dihub#s0095
	Treatment switch	Date of drug discontinuation Date of drug start	*High*	In CPRD information on treatment duration is available. This may help defining drug exposure-related variables.				
	Addition of another antihyperglycemic therapy	Medication codes	*High*	Drug codes are available for 100% of patients				
	Non-CV death	Diagnostic code (ICD-10 or equivalent) Date of death	*High*	Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects in attending emergency room				

Follow-up time needed per patient in the study	5 years	6 years (including recruitment and 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.	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors
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.	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors

	Estimated sample size: Approx. 13,341 participants			Considering that CPRD includes data from approximately 4.4 million inhabitants (as of 2014), the target sample size is anticipated to be reached.				

Step 3. Dk Reg-C4

Scientific research question		**Risk of gastrointestinal bleeding associated with use of rivaroxaban**						
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**							
	Patients over 75 years old	Date of birth	*High*	100% of individuals have available information. Except in emergencies, Denmark's approximately 3,600 GPs (20% of the physician workforce) are the first point of contact for patients				https://www.dovepress.com/the-danish-health-care-system-and-epidemiological-research-from-health-peer-reviewed-fulltext-article-CLEP
	Presence of NVAF	Diagnostic code	*High*	100% of subjects in the data who had a diagnosis have diagnostic code	Reliability of demographic data, hospital admission data, and overall diagnoses is deemed to be high as standard validation procedures are in place.	Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.		Provided by DEAP https://catalogues.ema.europa.eu/node/1019/administrative-details https://www.dovepress.com/the-danish-health-care-system-and-epidemiological-research-from-health-peer-reviewed-fulltext-article-CLEP
	Exclusion criteria							
	History of using VKA or any DOAC in the year prior to randomisation.	Medication code Date of prescription/dispensing	*High*	100% of subjects in the data who had a prescription/dispensing have medicine code				Provided by DEAP
	Any lesion or condition, if considered to be a significant risk for major bleeding. This may include current or recent gastrointestinal ulceration, presence of malignant neoplasms at high risk of bleeding, recent brain or spinal injury, recent brain, spinal or ophthalmic surgery, recent intracranial haemorrhage, known or suspected oesophageal varices, arteriovenous malformations, vascular aneurysms or major intraspinal or intracerebral vascular abnormalities.	Diagnostic code	*High*	100% of subjects in the data who had a diagnosis have diagnostic code		Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.		Provided by DEAP https://catalogues.ema.europa.eu/node/1019/administrative-details
	Concomitant treatment of acute coronary syndrome with antiplatelet therapy in patients with a prior stroke or a transient ischaemic attack (TIA)	Diagnostic code Medication code Date of prescription/dispensing	*High*	100% of subjects in the data who had a prescription/dispensing have medicine code	Date of last medication intake for each type of medication was calculated as the date of last acquisition plus amount (package size multiplied by number of packages) divided by the prescribed daily dose. If the prescribed daily dose was not recorded, then the defined daily dose (as defined by WHO) was used as a proxy for a streamlined and standardised assumption.			Provided by DEAP
	Concomitant treatment of coronary artery disease / peripheral artery disease with ASA in patients with previous haemorrhagic or lacunar stroke, or any stroke within a month	Diagnostic code Medication code Date of prescription/dispensing	*High*	100% of subjects in the data who had a prescription/dispensing have medicine code	Date of last medication intake for each type of medication was calculated as the date of last acquisition plus amount (package size multiplied by number of packages) divided by the prescribed daily dose. If the prescribed daily dose was not recorded, then the defined daily dose (as defined by WHO) was used as a proxy for a streamlined and standardised assumption.			Provided by DEAP
	Hepatic disease associated with coagulopathy and clinically relevant bleeding risk including cirrhotic patients with Child Pugh B and C	Diagnostic code	*High*			Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.		Provided by DEAP https://catalogues.ema.europa.eu/node/1019/administrative-details
Treatment/exposure	Rivaroxaban	Medication code	*High*	100% of subjects in the data who had a prescription/dispensing have medicine code				Provided by DEAP
Comparator group (if applicable)	Apixaban	Medication code	*High*	100% of subjects in the data who had a prescription/dispensing have medicine code				Provided by DEAP

Key endpoint(s)	Time to first major GI bleeding	Diagnostic code Date of diagnostic	*High*	100% of subjects in the data who had a diagnosis have diagnostic code		Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.	Median time (years) between first and last available records for unique individuals: >30 years	Provided by DEAP: Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk - https://www.sciencedirect.com/science/article/abs/pii/S0895435614000845?via%3Dihub https://catalogues.ema.europa.eu/node/991/quantitative-descriptors
Confounders	Thrombocytopenia	Diagnostic code Laboratory values	*Low*	Moderate availability. Lab values NPU codes. The laboratory data set includes the date and type of test, its result and the biological material tested.		Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.		Provided by DEAP https://catalogues.ema.europa.eu/node/1019/administrative-details
	Hypertension	Diagnostic code Medication code as proxy	*Low*	100% of subjects in the data who had a diagnosis or who had a prescription/dispensing have code	High PPV,-sensitivity PPV=93.5 (89.2–96.2); Se= 84.2 (78.9–88.4)	Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.		Provided by DEAP https://catalogues.ema.europa.eu/node/1019/administrative-details https://www.dovepress.com/the-danish-health-care-system-and-epidemiological-research-from-health-peer-reviewed-fulltext-article-CLEP https://www.dovepress.com/article/download/98182
	History of stroke/TIA	Diagnostic code	*Low*	100% of subjects in the data who had a diagnosis have diagnostic code		Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.		https://pubmed.ncbi.nlm.nih.gov/17478969/ https://karger.com/ned/article/28/3/150/210588/Validity-of-Stroke-Diagnoses-in-a-National
	History of major bleeding event	Diagnostic code	*Low*	100% of subjects in the data who had a diagnosis have diagnostic code		Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.		Provided by DEAP https://catalogues.ema.europa.eu/node/1019/administrative-details
	Presence of malignancy	Diagnostic code	*Low*	Cancer registry is available.		Cancer codes are ICDO3 Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.		Provided by DEAP https://catalogues.ema.europa.eu/node/1019/administrative-details
	Hepatic impairment	Diagnostic code Laboratory values	*Low*	A laboratory dataset is available, with presumably high completeness. The laboratory data set includes the date and type of test, its result and the biological material tested.		Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.		Provided by DEAP https://catalogues.ema.europa.eu/node/1019/administrative-details
	History of pulmonary embolism (PE) or deep venous thrombosis (DVT)	Diagnostic code	*Low*	100% of subjects in the data who had a diagnosis have diagnostic code		Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.		Provided by DEAP https://catalogues.ema.europa.eu/node/1019/administrative-details"
	History of peptic ulcer diseases	Diagnostic code	*Low*	100% of subjects in the data who had a diagnosis have diagnostic code	High PPV	Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.		Positive predictive value of peptic ulcer diagnosis codes in the Danish National Patient Registry https://pubmed.ncbi.nlm.nih.gov/28503076/

	Concomitant use of medicines that modify haemostasis or increase the gastrointestinal bleeding risk such as nonsteroidal anti-inflammatory drugs, corticosteroids, selective serotonin reuptake inhibitors, antiplatelet drugs	Medication code Date of prescription/dispensing	*Low*	100% of subjects in the data who had a prescription/dispensing have medicine code	Date of last medication intake for each type of medication was calculated as the date of last acquisition plus amount (package size multiplied by number of packages) divided by the prescribed daily dose. If the prescribed daily dose was not recorded, then the defined daily dose (as defined by WHO) was used as a proxy for a streamlined and standardised assumption.			Provided by DEAP
Intercurrent events	Treatment discontinuation	Date of drug discontinuation Medication code Date of prescription/dispensing	*High*	Not readily available, needs to be drawn from the date of dispensing/prescription, or from its duration	Date of last medication intake for each type of medication was calculated as the date of last acquisition plus amount (package size multiplied by number of packages) divided by the prescribed daily dose. If the prescribed daily dose was not recorded, then the defined daily dose (as defined by WHO) was used as a proxy for a streamlined and standardised assumption.			https://www.thelancet.com/action/showPdf?pii=S2666-7568%2821%2900170-7
	Treatment switch to another DOAC	Medication code Date of drug discontinuation Date of drug start	*High*	100% of subjects in the data who had a prescription/dispensing have medicine code	Date of last medication intake for each type of medication was calculated as the date of last acquisition plus amount (package size multiplied by number of packages) divided by the prescribed daily dose. If the prescribed daily dose was not recorded, then the defined daily dose (as defined by WHO) was used as a proxy for a streamlined and standardised assumption.			
	Switch to vitamin K antagonist	Medication code Date of drug discontinuation Date of drug start	*High*	100% of subjects in the data who had a prescription/dispensing have medicine code	Date of last medication intake for each type of medication was calculated as the date of last acquisition plus amount (package size multiplied by number of packages) divided by the prescribed daily dose. If the prescribed daily dose was not recorded, then the defined daily dose (as defined by WHO) was used as a proxy for a streamlined and standardised assumption.			
	Non-bleeding death	Diagnostic code Cause of death Date of death	*High*	A date of death is recorded for 100% of individuals who are known to have died Cause of death registry is available.		Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.		Provided by DEAP https://catalogues.ema.europa.eu/node/1019/administrative-details
Follow-up time needed per patient in the study	2 years	2 years of follow-up	*High*				Median time (years) between first and last available records for unique individuals: >30 years	https://catalogues.ema.europa.eu/node/991/quantitative-descriptors
Minimum time in the data source for lookback assessment	1 year	1 year	*High*				Median time (years) between first and last available records for unique individuals: >30 years	https://catalogues.ema.europa.eu/node/991/quantitative-descriptors

	Estimated sample size: Approx. 45,493 participants			Considering that the Danish population includes approximately 5.9 million inhabitants (as of 2023), the target sample size is anticipated to be reached.				

Step 3. SIDIAP-C4

Scientific research question		**Risk of gastrointestinal bleeding associated with use of rivaroxaban**						
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**							
	Patients over 75 years old	Date of birth	*High*	100% available				
	Presence of NVAF	Diagnostic code	*High*	100% of subjects in the data who had a diagnosis have diagnostic code Llifestyle factors such as smoking or alcohol consumption, are also recorded, with unknown completeness. Previous studies reported approximately 23% missingness for alcohol, and 28% for BMI.		Diagnoses and drugs follow UMLS ontologies.		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/
	Exclusion criteria							
	History of using VKA or any DOAC in the year prior to randomisation.	Medication code Date of prescription/dispensing	*High*	100% of subjects in the data who had a prescription/dispensing have medicine code		Diagnoses and drugs follow UMLS ontologies.	Median time (years) between first and last available records for unique individuals: 15.00 Median time (years) between first and last available records for unique active individuals (alive and currently registered): 16.00	https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/
	Any lesion or condition, if considered to be a significant risk for major bleeding. This may include current or recent gastrointestinal ulceration, presence of malignant neoplasms at high risk of bleeding, recent brain or spinal injury, recent brain, spinal or ophthalmic surgery, recent intracranial haemorrhage, known or suspected oesophageal varices, arteriovenous malformations, vascular aneurysms or major intraspinal or intracerebral vascular abnormalities.	Diagnostic code	*High*	100% of subjects in the data who had a diagnosis have diagnostic code Llifestyle factors such as smoking or alcohol consumption, are also recorded, with unknown completeness. Previous studies reported approximately 23% missingness for alcohol, and 28% for BMI.	It is impossible to assess the timing when a person stopped smoking, and also smoking intensity is not recorded. The most important limitation is the under-registration of GI haemorrhages in CMBD database, as it captures diagnoses at hospital discharge, but in our setting most GI haemorrhages are attended and treated in short-stay hospital wards of the Emergency Departments which do not routinely register those diagnoses in the CMBD database	Diagnoses and drugs follow UMLS ontologies. Smoking is classified into never, ex- or current smoking Alcohol is classified into no/mild-moderate-high/at risk drinker		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/ https://www.reumatologiaclinica.org/en-the-association-between-smoking-development-articulo-S1699258X20302035 https://pmc.ncbi.nlm.nih.gov/articles/PMC10540223/#s5
	Concomitant treatment of acute coronary syndrome with antiplatelet therapy in patients with a prior stroke or a transient ischaemic attack (TIA)	Diagnostic code Medication code Date of prescription/dispensing	*High*	100% of subjects in the data who had a prescription/dispensing have medicine code		Diagnoses and drugs follow UMLS ontologies.		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/
	Concomitant treatment of coronary artery disease / peripheral artery disease with ASA in patients with previous haemorrhagic or lacunar stroke, or any stroke within a month	Diagnostic code Medication code Date of prescription/dispensing	*High*	100% of subjects in the data who had a prescription/dispensing have medicine code		Diagnoses and drugs follow UMLS ontologies.		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/
	Hepatic disease associated with coagulopathy and clinically relevant bleeding risk including cirrhotic patients with Child Pugh B and C	Diagnostic code	*High*	100% of subjects in the data who had a diagnosis have diagnostic code		Diagnoses and drugs follow UMLS ontologies.		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/
Treatment/exposure	Rivaroxaban	Medication code	*High*	100% of subjects in the data who had a prescription/dispensing have medicine code		Diagnoses and drugs follow UMLS ontologies.		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/
Comparator group (if applicable)	Apixaban	Medication code	*High*	100% of subjects in the data who had a prescription/dispensing have medicine code		Diagnoses and drugs follow UMLS ontologies.		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/
Key endpoint(s)	Time to first major GI bleeding	Diagnostic code Date of diagnostic	*High*	100% of subjects in the data who had a diagnosis have diagnostic code	CMBD-AH captures diagnoses at hospital discharge, and CMBD-URG may capture most GI haemorrhages. Both are available.	Diagnoses and drugs follow UMLS ontologies.	Median time (years) between first and last available records for unique individuals: 15.00 Median time (years) between first and last available records for unique active individuals (alive and currently registered): 16.00	https://pubmed.ncbi.nlm.nih.gov/37781690/ https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://pmc.ncbi.nlm.nih.gov/articles/PMC10540223/#s5
Confounders	Thrombocytopenia	Diagnostic code Laboratory values	*Low*	100% of subjects in the data who had a diagnosis have diagnostic code Some tests' performance might be recorded but not their results		Diagnoses and drugs follow UMLS ontologies.		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/

	Hypertension	Diagnostic code Medication code as proxy	*Low*	100% of subjects in the data who had a diagnosis have diagnostic code	Clinical measurements are recorded, with unknown reliability	Diagnoses and drugs follow UMLS ontologies.		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/
	History of stroke/TIA	Diagnostic code	*Low*	100% of subjects in the data who had a diagnosis have diagnostic code		Diagnoses and drugs follow UMLS ontologies.		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/
	History of major bleeding event	Diagnostic code	*Low*	100% of subjects in the data who had a diagnosis have diagnostic code		Diagnoses and drugs follow UMLS ontologies.		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/
	Presence of malignancy	Diagnostic code	*Low*	100% of subjects in the data who had a diagnosis have diagnostic code		Diagnoses and drugs follow UMLS ontologies.		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/
	Hepatic impairment	Diagnostic code Laboratory values	*Low*	100% of subjects in the data who had a diagnosis have diagnostic code Some tests' performance might be recorded but not their results				
	History of pulmonary embolism (PE) or deep venous thrombosis (DVT)	Diagnostic code	*Low*	100% of subjects in the data who had a diagnostic have diagnostic code		Diagnoses and drugs follow UMLS ontologies.		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/
	History of peptic ulcer diseases	Diagnostic code	*Low*	100% of subjects in the data who had a diagnosis have diagnostic code		Diagnoses and drugs follow UMLS ontologies.		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/
	Concomitant use of medicines that modify haemostasis or increase the gastrointestinal bleeding risk such as nonsteroidal anti-inflammatory drugs, corticosteroids, selective serotonin reuptake inhibitors, antiplatelet drugs	Medication code Date of prescription/dispensing	*Low*	100% of subjects in the data who had a prescription/dispensing have medicine code		Diagnoses and drugs follow UMLS ontologies.		https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors https://www.sidiap.org/index.php/ca/
Intercurrent events	Treatment discontinuation	Date of drug discontinuation Medication code Date of prescription/dispensing	*High*	Not readily available, needs to be drawn from the date of dispensing/prescription, or from its duration				
	Treatment switch to another DOAC	Medication code Date of drug discontinuation Date of drug start	*High*	Not readily available, needs to be drawn from the date of dispensing/prescription, or from its duration				
	Switch to vitamin K antagonist	Medication code Date of drug discontinuation Date of drug start	*High*	Not readily available, needs to be drawn from the date of dispensing/prescription, or from its duration				
	Non-bleeding death	Diagnostic code Cause of death Date of death	*High*	Diagnostic codes and date of death are 100% available. Cause of death not available. A date of death is recorded for 100% of individuals who are known to have died	Cause of death might need to be inferred since it is not recorded in the data source. This may have limited accuracy. However, previous studies using SIDIAP have assessed death due to specific causes.			https://pubmed.ncbi.nlm.nih.gov/37781690/ https://scientiasalut.gencat.cat/handle/11351/6224
Follow-up time needed per patient in the study	2 years	2 years of follow-up	*High*				Median time (years) between first and last available records for unique individuals: 15.00 Median time (years) between first and last available records for unique active individuals (alive and currently registered): 16.00	https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors
Minimum time in the data source for lookback assessment	1 year	1 year	*High*				Median time (years) between first and last available records for unique individuals: 15.00 Median time (years) between first and last available records for unique active individuals (alive and currently registered): 16.00	https://catalogues.ema.europa.eu/node/1019/quantitative-descriptors

	Estimated sample size: Approx. 45,493 participants			Considering that SIDIAP includes data from approximately 5.8 million inhabitants, with 11,962 patients with non-valvular atrial fibrillation (NVAF) claimed a prescription of anticoagulation between 2011 and 2014 identified in previous literature, the target sample size is anticipated to be reached.				

Step 3. PEDIANET-C8

Scientific research question		**Evaluate the effectiveness of nirsevimab against RSV-respirat**						
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**							
	Infants ≤12 months of age at the beginning of the RSV season	Date of birth	*High*	Available for 100% of patients	High	According to previous studies, RSV seasonality is considered between October 1st and March 31st	As data is updated twice yearly, is to bear in mind that information extracted will be at least 6 months old. Median time between first and last records for unique active individuals is ~14 years. Since deadline of the project is in spring 2026, whole data of 2025-2026 could not be included. In case of extension of project, data of 2025-2026 season will be available	Provided by DEAP
	Born at ≥29 weeks gestational age	Gestational age at birth	*High*	Unknown exact missingness, but is expected to be minimal, as there is a registered exemption from the payment of medication for this subgroup	High		As data is updated twice yearly, is to bear in mind that information extracted will be at least 6 months old. Median time between first and last records for unique active individuals is ~14 years	
	Exclusion criteria							
	Ongoing RSV infection	Diagnostic codes Laboratory test results	*High*	Event codes available for 100% of patients				
	Participation in another RSV prophylaxis trial	Electronic data regarding informed consent document	*High*	Not available				
	No informed consent from the parents	Electronic data regarding informed consent document	*High*	Informed consent is already required from children's parents to enter the data in the database and to have Pedianet data linked to other databases, making this criterion potentially redundant				https://pmc.ncbi.nlm.nih.gov/articles/PMC10196108/pdf/fped-11-1143735.pdf
	The mother has received RSV vaccine during the pregnancy	Medication code Mother-baby ID Date of start of pregnancy Date of end of pregnancy	*High*	The search of medication codes aims to exclude presence of nirsevimab ATC code. For those patients born since 2024, there is a field regarding maternal vaccination for family pediatricians to complete (however, maternal RSV inmunization is rare in Italy since it is not reimbursed apart from some specific local health units)	Sibling linkage is explicitely described, but no information is available about mother-offspring linkage			Provided by DEAP
	The infant is eligible to receive palivizumab	Gestational age at birth Diagnostic codes	*High*	Unknown exact missingness of gestational age at birth, but is expected to be minimal, as there is a registered exemption from the payment of medication for this subgroup Event codes (in this case, of predisponent conditions that make an infant eligible to receive palivizumab) available for 100% of patients	High		As data is updated twice yearly, is to bear in mind that information extracted will be at least 6 months old. Median time between first and last records for unique active individuals is ~14 years	
Treatment/exposure	Nirsevimab	Medication code	*High*	Medication codes available for 100% of patients				
Comparator group (if applicable)	Standard care (no injection)	Medication code	*High*	The search of medication codes aims to exclude presence of nirsevimab ATC code				
Key endpoint(s)	Hospitalization for RSV-associated LRTI (hospital admission with an RSV-positive test result, during the RSV season)	Diagnostic codes Laboratory test results Hospitalization admission Date of hospitalization admission	*High*	Event codes available for 100% of patients	According to previous studies with same database analysing pediatric LRTI, the total number of bronchiolitis tested for pathogens in patients which are only visited on Emergency Departments (and not requiring hospitalization) is unknown, which may understimate the incidence of RSV-associated LRTI. Nevertheless, all children being hospitalized with LRTI are tested for a panel of viruses including RSV, so the overall missingness of RSV-associated LRTI is expected to be minimal.		As data is updated twice yearly, is to bear in mind that information extracted will be at least 6 months old. Median time between first and last records for unique active individuals is ~14 years	https://pmc.ncbi.nlm.nih.gov/articles/PMC10196108/pdf/fped-11-1143735.pdf (published analysis on prior versions of PEDIANET) https://zenodo.org/records/13384860 Information provided by DAP
	Hospitalization for LRTI	Diagnostic codes Hospitalization admission	*High*	Event codes available for 100% of patients				https://zenodo.org/records/13384860
	Medically-attended LRTI	Diagnostic codes	*High*	Event codes available for 100% of patients				https://zenodo.org/records/13384860
Confounders	Age at the start of the RSV season	Date of birth	*Low*	Available for 100% of patients				Provided by DEAP
	Gestational age	Gestational age	*Low*	Unknown exactly missingness				Provided by DEAP
	Congenital cardiac diseases	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients				
	Bronchopulmonary dysplasia	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients				
	Neuromuscular disorders	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients				
	Lung malformations	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients				
	Congenital or acquired immunodeficiency	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients				
	Receipt of immunosuppressive therapy, such as anti-cancer chemotherapy or radiation therapy, within the preceding 6 months; or long-term systemic corticosteroid therapy (prednisone or equivalent for more than 2 consecutive weeks within the past 3 months)	Medication code Date of prescription/dispensing Date of discontinuation or duration of treatment	*Low*	Medication codes available for 100% of patients Unknown missingness for exemption from medical treatment, although it is expected to be mnimal	Exemption from medical treatment will be used to exclude children with cancer and other diseases that require immunosupressive therapy Duration will be determined using a proxy based on diagnosis needing corticosteroides for more than 2 weeks and drugs prescriptions			Provided by DEAP
	Down syndrome	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients				
	Previous RSV infection	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients				
	Receipt of nirsevimab [only for the control group]	Medication code	*Low*	Medication codes available for 100% of patients				

	Death	Date of death	*High*	Diagnostic codes available for 100% of patients Life status available, although 0% of patients have date of death recorded				
Follow-up time needed per patient in the study	6 months	6 months (including recruitment and follow-up)	*Low*				As data is updated twice yearly, is to bear in mind that information extracted will be at least 6 months old. Median time between first and last records for unique active individuals is ~14 years, so this time-window seems achievable.	
Minimum time in the data source for lookback assessment	At least 9 months of feedback based on the exclusion criteria "the mother has received RSV vaccine during the pregnancy"	9 months	*High*				As data is updated twice yearly, is to bear in mind that information extracted will be at least 6 months old. Median time between first and last records for unique active individuals is ~14 years, so this time-window seems achievable.	
	Estimated sample size: 7,408 participants			Considering that PEDIANET includes data on 24,572 toddlers aged between 28 days and 23 months, and ~ 30K birth every year, the target sample sizes seems feasible.				Provided by DEAP

Step 3. CPRD-C5

Scientific research question		**Comparison of single-device vilanterol/fluticasone furoate with other inhaled corticosteroid-long-acting beta agonist single-device combinations in the risk of pneumonia in adolescents with asthma**						
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**							
	12-17 at treatment initiation (adolescents)	Date of birth (years)	*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
	Asthma diagnosis	Diagnostic code Date of diagnostic	*High*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				https://bmjopen.bmj.com/content/7/8/e017474.abstract
	Step-up from ICS alone to ICS+LABA	Medication code Date of prescription/dispensing Duration of treatment	*High*	100% medication code 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication	As ATC codes are not available, a mappint 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.		
	Exclusion criteria							
	Previous pneumonia diagnosis within the previous year	Diagnostic code Date of diagnostic	*High*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room			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.	
	LABA treatment within the previous year	Medication code Date of prescription/dispensing	*High*	100%	Only prescription date is available	As ATC codes are not available, a mappint 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.	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.	
	Long-term hospitalisation within the previous year	Date of admission Date of discharge Duration of hospitalisation	*High*	Not available for the core data source; available with data augmentation (linkage) only Known issues in CPRD that could affect extensiveness include: Provisional HES data are monthly publications of HES data (these data may be incomplete or contain errors for which no adjustments have yet been made by HES), It is also probable that clinical data are not complete, which may affect the last two months of any given period.	Known issues in CPRD that could affect reliability include: Unfinished episodes, at the end of the fiscal year ("month 13"), the annual data is refreshed and known data quality issues are corrected, prior to locking the annual published data, HRG files provided to CPRD with high level of null values for variable hes_yr compared to Set22	Hospitalisations refer to the total period of inpatient hospital stay from admission to discharge. When a hospitalisation spans the end of the HES year, it is artificially modelled as two hospitalisations. Known issues in CPRD that concern coherence include: • Invalid/missing dates depicted as 15/10/1582 or 01/01/1600 • Episodes where admission date precedes the epistart date • Explicit duplicate records which vary only by unique episode identifier (epikey) • Maternity records may have inconsistencies which need to be considered when using the data • Counts produced from provisional data are likely to be lower than those generated for the same period in the final dataset. There may also be errors due to coding inconsistencies that have not yet been investigated and corrected.	Unfinished episodes might be found at the end of the fiscal year ("month 13") 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.	

Treatment/exposure	Vilanterol-flucatisone furoate combination	Medication code Date of prescription/dispensing	*High*	100% medication code 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication	As ATC codes are not available, a mappint 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)	Other ICS-LABA combinations	Medication code Date of prescription/dispensing	*High*	100% medication code 100% date of prescription (only prescription is available)		As ATC codes are not available, a mappint 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.		
Key endpoint(s)	Time to first ocurrence of pneumonia	Diagnostic code Date of diagnostic	*High*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				
Confounders	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.				
	age	Date of birth (years)	*Low*	100% have date of birth	Only year is available, this may impact precision.			
	season	Date of diagnostic Date of prescription/dispensing	*Low*	100% Date of diagnostic 100% Date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
	calendar year	Date of diagnostic Date of prescription/dispensing	*Low*	100% Date of diagnostic 100% Date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
	age at 1st asthma diagnosis	Date of birth Date of diagnostic Diagnostic code	*Low*	100% have date of birth	Only year is available, this may impact precision.			
	diabetes	Diagnostic codes Date of diagnostic	*Low*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				
	rheumatological diseases	Diagnostic codes Date of diagnostic	*Low*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				
	malignancies	Diagnostic codes Date of diagnostic	*Low*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				
	cardiovascular diseases	Diagnostic codes Date of diagnostic	*Low*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				
	n of hospital days (any cause)	Date of admission Date of discharge Duration of hospitalisation	*Low*					
	n of outpatient visits	Date of visit Setting of visit	*Low*					
	Down/intellectual disabilities/ congenital malformations	Diagnostic codes Date of diagnostic	*Low*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				
	psychotropics (or specific groups)	Medication code Date of prescription/dispensing	*Low*	100% medication code 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			

	asthma exacerbations	Medication code Date of prescription/dispensing Diagnostic code Date of diagnosis	*Low*	100% medication code 100% date of prescription Admission or emergency room diagnostic code only available in HES (not disposable for the current study, would need data augmentation, i.e., linkage).	Prescription of a medication does not guarantee the subject collected and took the medication Exacerbations leading to hospital admission or emergency room might be needed for a more reliable capturing of exacerbation.			
	oral corticosteroids	Medication code Date of prescription/dispensing	*Low*	100% medication code 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
	biologicals	Medication code Date of prescription/dispensing	*Low*	100% medication code 100% date of prescription If these medications are dispensed/prescribed inhospital, availability is unknown	Prescription of a medication does not guarantee the subject collected and took the medication			
	SABA	Medication code Date of prescription/dispensing	*Low*	100% medication code 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
	leukotriene receptor antagonists	Medication code Date of prescription/dispensing	*Low*	100% medication code 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
	antibiotic use	Medication code Date of prescription/dispensing	*Low*	100% medication code 100% date of prescription (only prescription is available) If used inhospital, information might not available	Prescription of a medication does not guarantee the subject collected and took the medication			
	pneumococcal vaccination	Medication code Date of prescription/dispensing	*Low*	Vaccine codes available for 100% of patients 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
	influenza vaccination	Medication code Date of prescription/dispensing	*Low*	Vaccine codes available for 100% of patients 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
	death	Date of death	*Low*	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 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://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.4747
Intercurrent events	Treatment discontinuation	Medication code Date of prescription/dispensing Duration of treatment	*High*	Prescription information available 100% medication codes and date of prescription	In CPRD information on treatment duration of the prescription is available. This may help defining drug exposure-related variables. Prescription of a medication does not guarantee the subject collected and took the medication			
	Switch to another ICS+LABA (i.e., budesonide/formoterol, or salmeterol/fluticasone propionate)	Medication code Date of prescription/dispensing Duration of treatment	*High*	Prescription information available 100% medication codes and date of prescription	In CPRD information on treatment duration of the prescription is available. This may help defining drug exposure-related variables.			
	Oral corticosteroids use	Medication code Date of prescription/dispensing	*Low*	Prescription information available 100% medication codes and date of prescription				
	Add on SABA, LAMA, leukotriene receptor antagonist, biologics	Medication code Date of prescription/dispensing Duration of treatment	*Low*	Prescription information available 100% medication codes and date of prescription	In CPRD information on treatment duration of the prescription is available. This may help defining drug exposure-related variables.Prescription of a medication does not guarantee the subject collected and took the medication			
	Rescue medications	Medication code Date of prescription/dispensing	*Low*	Prescription information available 100% medication codes and date of prescription				

	Non-pneumonia death	Date of death Diagnostic code Date of diagnosis	*High*	8% of the whole population (irrespective of vital status) has a date of death recorded; 100% of death people have a date of death Diagnostic codes available for 100% of patients 100% Date of diagnostic Where the exact date of death or the cause is importantin a CPRD study, it may be advisable to include the individually linked national ONS death registration data				https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.4747
Follow-up time needed per patient in the study	up to a maximum of 2 years after the randomisation day	Date of exit of the database	*High*				The median lenght of follow-up per patient is approximately 6 years and 13 years for active individuals	
Minimum time in the data source for lookback assessment	1 year	1 year	*High*				The median lenght of follow-up per patient is approximately 6 years and 13 years for active individuals	

	Estimated sample size: Approx. 26,750 participants (13,375 per group)			Considering that CPRD includes data from approximately 4.4 million inhabitants (as of 2014), the target cohort size is anticipated to be achievable.				

Step 3. FinReg-C5

Scientific research question		**Comparison of single-device vilanterol/fluticasone furoate with other inhaled corticosteroid-long-acting beta agonist single-device combinations in the risk of pneumonia in adolescents with asthma**						
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**							
	12-17 at treatment initiation (adolescents)	Date of birth	*High*	Available for 100% of subjects				
	Asthma diagnosis	Diagnostic codes Date of diagnostic	*High*		Records of healthcare diagnosis with logical inconsistencies are <1%, so data is deemed to be reliable.	Diagnoses (ICD-10) and drugs follow UMLS ontologies.		
	Step-up from ICS alone to ICS+LABA	Medication code Date of prescription/dispensing Duration of treatment	*High*	Dispensing information available.	End of treatment episodes is derived based on dispensing date and dispensed amount, so the accuracy will be impacted by these two variables	Diagnoses and drugs follow UMLS ontologies.		
	Exclusion criteria							
	previous pneumonia diagnosis within the previous year	Diagnostic codes Date of diagnostic	*High*	Hospital admission and discharge diagnosis are available. Primary care diagnosis data and antibiotc dispensings are also available.		Diagnoses and drugs follow UMLS ontologies.	Population registers have a long history in Finland, with population information having been registered since the 1530s.	
	LABA treatment within the previous year	Medication code Date of dispensing	*High*	Dispensing information available.		Diagnoses and drugs (ATC) follow UMLS ontologies.	Population registers have a long history in Finland, with population information having been registered since the 1530s.	
	long-term hospitalisation within the previous year	Date of admission Date of discharge Duration of hospitalisation	*High*	Hospital admission and discharge diagnosis are available.			Population registers have a long history in Finland, with population information having been registered since the 1530s.	
Treatment/exposure	Vilanterol-flucatisone furoate combination	Medication code Date of prescription/dispensing	*High*	Dispensing information available.	Records of healthcare diagnosis with logical inconsistencies are <1%, so data is deemed to be reliable.	Diagnoses and drugs follow UMLS ontologies.		
Comparator group (if applicable)	Other ICS-LABA combinations	Medication code Date of prescription/dispensing	*High*		Records of healthcare diagnosis with logical inconsistencies are <1%, so data is deemed to be reliable.	Diagnoses and drugs follow UMLS ontologies.		
Key endpoint(s)	Time to first ocurrence of pneumonia	Diagnostic codes Date of diagnosis	*High*	Hospital admission and discharge diagnosis are available.	Records of healthcare diagnosis with logical inconsistencies are <1%, so data is deemed to be reliable.	Diagnoses and drugs follow UMLS ontologies.		
Confounders	sex	Sex	*Low*	Available for 100% of subjects. However, there are the "undefined" and "unknown" categories.	Available for 100% of subjects. However, there are the "undefined" and "unknown" categories. (Although these categories exist, they are extremely rare in this age group)			
	age	Date of birth	*Low*	Available for 100% of subjects	Date of birth is provided with month and year			
	season	Date of diagnosis Date of prescription/dispensing	*Low*	All prescriptions and diagnoses have exact date season therefore available				Provided by DEAP
	calendar year	Date of diagnosis Date of prescription/dispensing	*Low*	Dates precision includes year. We anticipate calendar year will be available for all the records.				
	age at 1st asthma diagnosis	Date of birth Date of diagnosis Diagnostic codes	*Low*	Date of birth available for 100% of subjects		Diagnoses and drugs follow UMLS ontologies.	Children born in Finland are registered at birth (medical birth register)	Provided by DEAP
	diabetes	Diagnostic codes Date of diagnosis	*Low*			Diagnoses and drugs follow UMLS ontologies.		
	rheumatological diseases	Diagnostic codes Date of diagnosis	*Low*			Diagnoses and drugs follow UMLS ontologies.		
	malignancies	Diagnostic codes Date of diagnosis	*Low*			Diagnoses and drugs follow UMLS ontologies.		
	cardiovascular diseases	Diagnostic codes Date of diagnosis	*Low*			Diagnoses and drugs follow UMLS ontologies.		
	n of hospital days (any cause)	Date of admission Date of discharge Duration of hospitalisation	*Low*	Duration of hospitalisation is not readily available but can be calculated from the admission and discharge dates.				

	n of outpatient visits	Date of visit Setting of visit	*Low*	Use of healthcare services is a record trigger, so the number of visits could be assessed with no problem. Setting can be identified, and actual visits can be differentiate from dispensings.				Provided by DEAP
	Down/intellectual disabilities/ congenital malformations	Diagnostic codes Date of diagnosis	*Low*			Diagnoses and drugs follow UMLS ontologies.		
	Psychotropics (or specific groups)	Medication code Date of prescription/dispensing	*Low*			Diagnoses and drugs follow UMLS ontologies.		
	Asthma exacerbations	Medication code Date of prescription/dispensing Diagnostic code Date of diagnosis	*Low*	Exacerbations inhospital not available. For outpatients, if there is a codelist to identify them, they can be picked. Hospital admission and discharge diagnosis are available.		Diagnoses and drugs follow UMLS ontologies.		Provided by DEAP
	Oral corticosteroids	Medication code Date of prescription/dispensing	*Low*	Date and code available for 100% of patients with dispensings. No inhospital drug use available.	Active substance ATC available, so high precision is expected.	Diagnoses and drugs follow UMLS ontologies.		
	Biologicals	Medication code Date of prescription/dispensing	*Low*	Date and code available for 100% of patients with dispensings. No inhospital drug use available.	Active substance ATC available, so high precision is expected.	Diagnoses and drugs follow UMLS ontologies.		
	SABA	Medication code Date of prescription/dispensing	*Low*	Date and code available for 100% of patients with dispensings. No inhospital drug use available.	Active substance ATC available, so high precision is expected.	Diagnoses and drugs follow UMLS ontologies.		
	Leukotriene receptor antagonists	Medication code Date of prescription/dispensing	*Low*	Date and code available for 100% of patients with dispensings. No inhospital drug use available.	Active substance ATC available, so high precision is expected.	Diagnoses and drugs follow UMLS ontologies.		
	Antibiotic use	Medication code Date of prescription/dispensing	*Low*	Date and code available for 100% of patients with dispensings. No inhospital drug use available.	Active substance ATC available, so high precision is expected.	Diagnoses and drugs follow UMLS ontologies.		
	Pneumococcal vaccination	Medication code Date of prescription/dispensing	*Low*		Active substance ATC available, so high precision is expected.	Diagnoses and drugs follow UMLS ontologies.		
	Influenza vaccination	Medication code Date of prescription/dispensing	*Low*		Active substance ATC available, so high precision is expected.	Diagnoses and drugs follow UMLS ontologies.		
	Death	Date of death	*Low*	Death and cause of death are available. From previous publications, it seems a cause of death register exists.				Tolonen H, Salomaa V, Torppa J, Sivenius J, Immonen-Räihä P, Lehtonen A; FINSTROKE register. The validation of the Finnish Hospital Discharge Register and Causes of Death Register data on stroke diagnoses. Eur J Cardiovasc Prev Rehabil. 2007 Jun;14(3):380-5. doi: 10.1097/01.hjr.0000239466.26132.f2. PMID: 17568236.
Intercurrent events	Treatment discontinuation	Medication code Date of prescription/dispensing Duration of treatment	*High*	Date and code available for 100% of patients with dispensings. Duration of treatment is not directly available but can be estimated.	End of treatment episodes is derived based on dispensing date and dispensed amount, so the accuracy will be impacted by these two variables death/migration/discontinuation/switch, then it is available but there is no data on other reasons (inadequate symptom control etc)			
	Switch to another ICS+LABA (i.e., budesonide/formoterol, or salmeterol/fluticasone propionate)	Medication code Date of prescription/dispensing Duration of treatment	*High*	Date and code available for 100% of patients with dispensings. No inhospital drug use available.	End of treatment episodes is derived based on dispensing date and dispensed amount, so the accuracy will be impacted by these two variables			
	Oral corticosteroids use	Medication code Date of prescription/dispensing	*Low*	Date and code available for 100% of patients with dispensings. No inhospital drug use available.	Active substance ATC available, so high precision is expected.	Diagnoses and drugs follow UMLS ontologies.		
	Add on SABA, LAMA, leukotriene receptor antagonist, biologics	Medication code Date of prescription/dispensing Duration of treatment	*Low*	Date and code available for 100% of patients with dispensings. No inhospital drug use available.	Active substance ATC available, so high precision is expected.	Diagnoses and drugs follow UMLS ontologies.		
	Rescue medications	Medication code Date of prescription/dispensing	*Low*					
	Non-pneumonia death	Date of death Diagnostic code Date of diagnosis	*High*	Death and cause of death are available. From previous publications, it seems a cause of death register exists.				https://stat.fi/en/statistics/ksyyt , and description of the individual-level data : https://aineistokatalogi.fi/catalog/studies/778c33bf-aceb-423f-89d9-e5abb5a0585c

Follow-up time needed per patient in the study	Up to a maximum of 2 years after the randomisation day	Date of exit of the database	*High*				Population registers have a long history in Finland, with population information having been registered since the 1530s.	
Minimum time in the data source for lookback assessment	1 year	1 year	*High*					

	Estimated sample size: Approx. 26,750 participants (13,375 per group)			Considering that Finland has a population exceeding 5 million (380,000 individuals aged 12-17y, 10.5% with asthma) and that up to 18,293 users of vilanterol-fluticasone furoate were identified among 233,261 patients with chronic asthma or similar chronic obstructive pulmonary diseases as of 2021, the target sample size is achievable but might be difficult to be reached.				

Step 3. CPRD-C6

Scientific research question		**Comparison of sacubitril/valsartan with ACE inhibitors in the risk of angioedema and other safety events**						
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**							
	Age >18 years old	Date of birth (years)	*High*	100% have date of birth	Only year is available, this may impact precision.			
	Patients with Heart Failure	Diagnostic code Date of diagnosis	*High*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				
	Exclusion criteria							
	Documented history or diagnosis of angioedema (either ACEI/ARB-induced or hereditary/idiopathic angioedema) any time before the screening visit	Diagnostic code Date of diagnosis Date of prescription/dispensation Medication code	*High*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room			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.	
	Treatment with both ACEIs and ARBs in the month before or at the screening visit	Date of prescription/dispensation Medication code	*High*	100% medication code 100% date of prescription (only prescription is available)		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.	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.	
	Current acute decompensated heart failure, defined as exacerbation of chronic heart failure manifested by signs and symptoms that may require intravenous therapy at the screening visit	Diagnostic code Date of diagnosis Date of visit to emergency room Date of admission Date of prescription/dispensation Medication code	*High*	Diagnostic codes available for 100% of patients 100% Date of diagnostic 100% medication code 100% date of prescription (only prescription is available) Diagnostic codes are available for 86% subjects in attending emergency room		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.		
	Diagnosis of peripartum- or chemotherapy- induced cardiomyopathy within 1 year before the screening visit.	Diagnostic code Date of diagnosis Date of prescription/dispensation Medication code Date of delivery	*High*	Diagnostic codes available for 100% of patients 100% Date of diagnostic 100% medication code 100% date of prescription (only prescription is available)		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.	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.	
	Diagnosed with severe hepatic impairment, biliary cirrhosis or cholestasis (Child-Pugh C classification) any time before the screening visit	Diagnostic code Date of diagnosis Laboratory test results	*High*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				
	Patients at their second or third trimester of pregnancy at the screening visit	Date of conception	*High*					
	Potassium levels > 5.4 mmol/l at the screening visit	Laboratory value Laboratory result date	*High*					
	Systolic Blood Pressure (SBP) <100 mmHg at the screening visit	Clinical measurement value Clinical measurement date	*High*					
Treatment/exposure	Sacubitril/Valsartan	Date of prescription/dispensation Medication code	*High*	Prescription medicines (100%), prescription date associated with the event, as entered by the GP		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. The theoretical end date could be calculated based on the date of prescription and the duration of treatment.		chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf

Comparator group (if applicable)	ACEi or ARB	Date of prescription/dispensation Medication code	*High*	Prescription medicines (100%)		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)	Time to occurrence of angioedema	Diagnostic code Date of diagnosis	*High*	Diagnostic codes anad dates available for 100% of patients Diagnostic codes are available for 86% subjects attending emergency room			Seem achievable. Median time (years) between first and last available records for unique individuals: 5.89 years Median time (years) between first and last available records for unique active individuals (alive and currently registered): 13.35 years	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors Provided by DEAP
Confounders	Age	Date of birth (years)	*Low*	100% have date of birth	Only year is available, this may impact precision.			
	Sex/gender	Sex at birth	*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.				
	Race	Race	*Low*		Not available			chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
	Smoking status	Smoking	*Low*	Smoking present for 89,7% of records. Depends on study window and look-back period, but should indeed be ok for recent years				
	History of diabetes	Diagnostic code Date of diagnosis	*Low*	Diagnostic codes available for 100% of patients 100% Date of diagnostic				
	DDP-4 inhibitors	Medication code Date of prescription/dispensation	*Low*	Prescription medicines (100%)		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.		
	History of ACE associated cough	Diagnostic code Date of diagnosis Medication code Date of prescription/dispensation	*Low*	Prescription medicines (100%) and 100% of diagnostic codes.	If a specific diagnostic code exists, cough should not be an issue, whether it is ACE related or not. In previous studies, they identified ACEi-cough was defined as an event of cough when this occurred while on treatment with ACEi.			
	Heart or renal transplant	Procedure code	*Low*					
	Seasonal allergies	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients 100% Date of diagnostic				
	Tissue plasminogen activators	Laboratory value Date laboratory	*Low*	unlikely registered				Provided by DEAP
	Localized tissue trauma	Diagnostic code	*Low*	100% diagnostic codes. If a codelist is present they can assess				Provided by DEAP
	Lymphoproliferative or autoimmune diseases	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients 100% Date of diagnostic				
	Estrogen-containing oral contraceptives or estrogen replacement therapy	Medication code Date of prescription/dispensation	*Low*	Prescription medicines (100%)		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.		
	Infections	Diagnostic code	*Low*	Diagnostic codes available for 100% of patients 100% Date of diagnostic				
	Stress	Diagnostic code	*Low*	Surely registered , but would not trust the value of such info				Provided by DEAP

	NSAIDS, acetylsalicylic acid	Medication code Date of prescription/dispensation	*Low*	Prescription medicines (100%). OTC not available.		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.		
	Insecticide use, administration of an intravenous contrast agent, exacerbation of pre-existing lower extremity edema	Medication code Date of prescription/dispensation Diagnostic code Date of diagnosis	*Low*	Insecticide exposure not available. Prescription 100%. Exacerbation of edema of extremity might be captured if a specific diagnostic codelist is in place.				Provided by DEAP
Intercurrent events	Treatment discontinuation	Medication code Date of prescription/dispensation	*High*	Not readily available, needs to be drawn from the date of dispen	As CPRD has prescription data, it is unknown whether the patient took the prescription (previous studies have analysed medications (metformin) discontinuation and adherence).			https://www.sciencedirect.com/science/article/pii/S2214623720300351?via%3Dihub#s0095
	Treatment switch	Medication code Date of prescription/dispensation	*High*	Not readily available, needs to be drawn from the date of dispensing/prescription, or from its duration				
	Addition of any of the three HF medications (ACEi, ARB, SV) if not the treatment of the group	Medication code Date of prescription/dispensation	*High*					
	All-cause death	Date of death	*High*	By 2013, 98.8% of deaths were in agreement with the Office of National Statistics, within ±30 days. 8% of the whole population (irrespective of vital status) has a date of death recorded; 100% of persons who died have a recorded date of death				https://www.sciencedirect.com/science/article/abs/pii/S1386505619306252 https://onlinelibrary.wiley.com/doi/full/10.1002/pds.4747
Follow-up time needed per patient in the study	Up to a maximum of 5 years	5 years	*High*	Available. Median time (years) between first and last available records for unique individuals: 5.89 years Median time (years) between first and last available records for unique active individuals (alive and currently registered): 13.35 years			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 Provided by DEAP
Minimum time in the data source for lookback assessment	1 year	1 year	*High*	Available. Median time (years) between first and last available records for unique individuals: 5.89 years Median time (years) between first and last available records for unique active individuals (alive and currently registered): 13.35 years			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 Provided by DEAP

	Estimated sample size: Approx. 30,784 participants			Considering that CPRD includes data from approximately 4.4 million inhabitants (as of 2014), the target cohort size is anticipated to be achievable.				

Step 3. PHARMO-C6

Scientific research question		Comparison of sacubitril/valsartan with ACE inhibi						
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**							
	Age >18	Date of birth (years)	*High*	70-100% complete				https://pmc.ncbi.nlm.nih.gov/articles/PMC9830053/
	Patients with Heart Failure	Diagnostic code Date of diagnosis	*High*	80-90% coverage of hospital events (admissions/discharges from 1998, specialist visits from 2014)		UMLS diagnostic and medicine codes are available.		Provided by DEAP
	Exclusion criteria							
	Documented history or diagnosis of angioedema (either ACEI/ARB-induced or hereditary/idiopathic angioedema) any time before the screening visit	Diagnostic code Date of diagnosis Date of prescription/dispensation Medication code	*High*	80-90% coverage of hospital events (admissions/discharges from 1998, specialist visits from 2014)		UMLS diagnostic and medicine codes are available.	Average longitudinality of 12 years.	Provided by DEAP
	Treatment with both ACEIs and ARBs in the month before or at the screening visit	Date of prescription/dispensation Medication code	*High*			UMLS diagnostic and medicine codes are available.	Average longitudinality of 12 years.	
	Current acute decompensated heart failure, defined as exacerbation of chronic heart failure manifested by signs and symptoms that may require intravenous therapy at the screening visit	Diagnostic code Date of diagnosis Date of visit to emergency room Date of admission Date of prescription/dispensation Medication code	*High*	80-90% coverage of hospital events (admissions/discharges from 1998, specialist visits from 2014)		UMLS diagnostic and medicine codes are available.		Provided by DEAP
	Diagnosis of peripartum- or chemotherapy- induced cardiomyopathy within 1 year before the screening visit.	Diagnostic code Date of diagnosis Date of prescription/dispensation Medication code Date of delivery	*High*			UMLS diagnostic and medicine codes are available.	Average longitudinality of 12 years.	
	Diagnosed with severe hepatic impairment, biliary cirrhosis or cholestasis (Child-Pugh C classification) any time before the screening visit	Diagnostic code Date of diagnosis Laboratory test results	*High*	Tests and test results are available in PHARMO. Unknown missingness		UMLS diagnostic and medicine codes are available.	Average longitudinality of 12 years.	
	Patients at their second or third trimester of pregnancy at the screening visit	Date of conception	*High*					
	Potassium levels > 5.4 mmol/l at the screening visit	Laboratory value Laboratory result date	*High*					https://catalogues.ema.europa.eu/node/997/quantitative-descriptors
	Systolic Blood Pressure (SBP) <100 mmHg at the screening visit	Clinical measurement value Clinical measurement date	*High*					https://catalogues.ema.europa.eu/node/997/quantitative-descriptors
Treatment/exposure	Sacubitril/Valsartan	Date of prescription/dispensation Medication code	*High*	70-100% complete 12,598 patients initiating either sac/val or ACEIs with linked hospital data between December 2015 - June 2021		UMLS diagnostic and medicine codes are available.		
Comparator group (if applicable)	ACEi or ARB	Date of prescription/dispensation Medication code	*High*	70-100% complete		UMLS diagnostic and medicine codes are available.		
Key endpoint(s)	Time to occurrence of angioedema	Diagnostic code Date of diagnosis	*High*	80-90% coverage of hospital events (admissions/discharges from 1998, specialist visits from 2014)	Data on angioedema has been validated by clinician as part of previous study	UMLS diagnostic and medicine codes are available.	Seems achievable as median follow-up in the database is 12 years.	Provided by DEAP https://catalogues.ema.europa.eu/node/997/quantitative-descriptors
Confounders	Age	Date of birth (years)	*Low*	70-100% complete				https://pmc.ncbi.nlm.nih.gov/articles/PMC9830053/
	Sex	Sex at birth	*Low*	70-100% complete				https://pmc.ncbi.nlm.nih.gov/articles/PMC9830053/
	Race	Race	*Low*	<40% complete				
	Smoking status	Smoking	*Low*	40-69% complete				
	History of diabetes	Diagnostic code Date of diagnosis	*Low*	70-100% complete Around 90%		UMLS diagnostic and medicine codes are available.		https://www.valueinhealthjournal.com/article/S1098-3015(23)05991-0/fulltext

	DDP-4 inhibitors	Medication code Date of prescription/dispensation	*Low*	70-100% complete		UMLS diagnostic and medicine codes are available.		
	History of ACE associated cough	Diagnostic code Date of diagnosis Medication code Date of prescription/dispensation	*Low*	70-100% complete		UMLS diagnostic and medicine codes are available.	Average longitudinality of 12 years.	
	Heart or renal transplant	Procedure code	*Low*	70-100% complete		UMLS diagnostic and medicine codes are available.		
	Seasonal allergies	Diagnostic code	*Low*	70-100% complete	If a diagnostic code exists, this will be picked. If not, questionable reliability			
	Tissue plasminogen activators	Laboratory value Date laboratory	*Low*	40-69% complete 25% coverage of out-patient pharmacy from; 10% coverage of in-patient pharmacy from 1985; 80-90% coverage of high cost medicines after 2017				
	Localized tissue trauma	Diagnostic code	*Low*	70-100% complete	If a diagnostic code exists, this will be picked. If not, questionable reliability			
	Lymphoproliferative or autoimmune diseases	Diagnostic code	*Low*	70-100% complete		UMLS diagnostic and medicine codes are available.		
	Estrogen-containing oral contraceptives or estrogen replacement therapy	Medication code Date of prescription/dispensation	*Low*	70-100% complete		UMLS diagnostic and medicine codes are available.		
	Infections	Diagnostic code	*Low*	70-100% complete		UMLS diagnostic and medicine codes are available.		
	Stress	Diagnostic code	*Low*	40-69% complete Expected to be under-reported/under-recorded	Expected to be under-reported/under-recorded			Provided by DEAP
	NSAIDS, acetylsalicylic acid	Medication code Date of prescription/dispensation	*Low*	If prescribed; OTC not available. 25% coverage of out-patient pharmacy from; 10% coverage of in-patient pharmacy from 1985; 80-90% coverage of high cost medicines after 2017		UMLS diagnostic and medicine codes are available.		
	Insecticide use, administration of an intravenous contrast agent, exacerbation of pre-existing lower extremity edema	Medication code Date of prescription/dispensation Diagnostic code Date of diagnosis	*Low*	Insecticide exposure not available. Prescription 100%. Exacerbation of edema of extremity might be captured if a specific diagnostic codelist is in place.				Provided by DEAP
Intercurrent events	Treatment discontinuation	Medication code Date of prescription/dispensation	*High*	Not readily available, needs to be drawn from the date of dispensing/prescription, or from its duration				
	Treatment switch	Medication code Date of prescription/dispensation	*High*	Not readily available, needs to be drawn from the date of dispensing/prescription, or from its duration				
	Addition of any of the three HF medications (ACEi, ARB, SV) if not the treatment of the group	Medication code Date of prescription/dispensation	*High*	70-100% complete 25% coverage of out-patient pharmacy from; 10% coverage of in-patient pharmacy from 1985; 80-90% coverage of high cost medicines after 2017		UMLS diagnostic and medicine codes are available.		
	All-cause death	Date of death	*High*	A date of death is recorded for 100% of individuals who are known to have died through Mortality Register				https://catalogues.ema.europa.eu/node/997/data-elements-collected Provided by DEAP
Follow-up time needed per patient in the study	Up to a maximum of 5 years	5 years	*High*				Average longitudinality of 12 years. Data are available with an approximately 1-year lag depending on the databases required	https://catalogues.ema.europa.eu/node/997/quantitative-descriptors
Minimum time in the data source for lookback assessment	1 year	1 year	*High*				Average longitudinality of 12 years. Data are available with an approximately 1-year lag depending on the databases required	Provided by DEAP

	Estimated sample size: Approx. 30,784 participants			Considering that Pharmo includes data from 40% of the Dutch population, the target sample size is anticipated to be reached.				

Step 3. VID-C7

Scientific research question		**Safety of paternal exposure to valproate at conception and the risk of long-term neurodevelopment outcomes in the offspring.**						
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**							
	Males of 18 years of age older	Date of birth Sex	*High*	Date of birth if available in VID for virtually all indviduals For sex 100% of individuals have available information				
	Participants must have a female partner with which they intend to conceive.	Presence of family linkage Type of linkage Sex	*High*	Mother-child- Father have been linked by deterministic linkage. 67% of livebirths linked to the mother can be linked to the father. Diagnostic codes: 100% Sex 100% of individuals have available information	Deterministic linkage is used to obtain mother-child pairs, so deemed to be highly reliable. Father-child linkage was obtained using a deterministic approach, with direct and indirect linkage. Direct linkage was performed linking children with their male health cardholder or legal guardian, which had to comply with other criteria such as being 15y or older at the time of birth of the index child and their surnames should be compatible. Then we used household address combined with sex, age and surnames criteria, excluding households where more than one male were retrieved. This linkage is only possible for live births. So, only females experiencing a pregnancy ending in a live birth will be linked to the male partner.		This linkage is available from 2010 to 2024	https://www.eurolinkcat.eu/loadFile.aspx?filename=D2.4%20report%20submitted(2).pdf https://catalogues.ema.europa.eu/node/1077/quantitative-descriptors The paper on father-child linkage is under development at the moment.
	Diagnosis of generalized epilepsy in males	Diagnostic code Date of diagnosis	*High*	Diagnostic codes: 100% Previous studies describe 65 million people worldwide have epilepsy, about 400,000 patients with epilepsy in Spain, and an annual incidence of epilepsy among adults of 37.7 cases/100,000 inhabitants, and more than 50% being men. The most common types of seizures and epilepsy are generalized seizures and epilepsy of unknown etiology. Estimates of the incidence of generalized epilepsies in the United States are at 7.7 per 100,000 person-years.	Diagnostic codes are granular enough in VID to distinguish generalized epilepsy from other types of epilepsy (i.e., have diagnostic codes with one number or more after the dot, following the ICD format).	Diagnoses and drugs follow UMLS ontologies.		Quintana M, Sánchez-López J, Mazuela G, Santamarina E, Abraira L, Fonseca E, Seijo I, Álvarez-Sabin J, Toledo M. Incidence and mortality in adults with epilepsy in northern Spain. Acta Neurol Scand. 2021 Jan;143(1):27-33. doi: 10.1111/ane.13349. Epub 2020 Oct 13. PMID: 32969054. Villanueva V, Carreño M, Gil-Nagel A, Serrano-Castro PJ, Serratosa JM, Toledo M, Álvarez-Barón E, Gil A, Subías-Labazuy S. Identifying key unmet needs and value drivers in the treatment of focal-onset seizures (FOS) in patients with drug-resistant epilepsy (DRE) in Spain through Multi-Criteria Decision Analysis (MCDA). Epilepsy Behav. 2021 Sep;122:108222. doi: 10.1016/j.yebeh.2021.108222. Epub 2021 Aug 6. PMID: 34371462. https://www.ncbi.nlm.nih.gov/books/NBK546611/
	Exclusion criteria							
	Male not have any known contraindication for either valproate or levetiracetam use.	Sex Diagnostic code Date of diagnosis	*High*	Diagnostic codes: 100% Sex 100% of individuals have available information				
	Female partner must not be diagnosed with generalized epilepsy	Sex Diagnostic code Date of diagnosis Presence of family linkage Type of linkage	*High*	Diagnostic codes: 100% Sex 100% of individuals have available information Family unit can be linked by deterministic linkage. 67% of livebirths linked to the mother can be linked to the father.	Deterministic linkage is used to obtain mother-child pairs, so deemed to be highly reliable. Father-child linkage was obtained using a deterministic approach, with direct and indirect linkage. Direct linkage was performed linking children with their male health cardholder or legal guardian, which had to comply with other criteria such as being 15y or older at the time of birth of the index child and their surnames should be compatible. Then we used household address combined with sex, age and surnames criteria, excluding households where more than one male were retrieved. This linkage is only possible for live births. So, only females experiencing a pregnancy ending in a live birth will be linked to the male partner.		This linkage is available from 2010 to 2024	https://www.eurolinkcat.eu/loadFile.aspx?filename=D2.4%20report%20submitted(2).pdf https://catalogues.ema.europa.eu/node/1077/quantitative-descriptors
	Male not have any medical condition that permanently prevents them from conception (i.e., infertility or any condition that makes a future conception impossible).	Sex Diagnostic code Date of diagnosis	*High*	Diagnostic codes: 100% Sex 100% of individuals have available information				
	Female partner must not have any medical condition that permanently prevents them from conception (i.e., infertility or any condition that makes a future conception impossible.)	Sex Diagnostic code Date of diagnosis Procedure codes Date of procedure Presence of family linkage Type of linkage	*High*	Diagnostic codes: 100% Sex 100% of individuals have available information Family unit can be linked by deterministic linkage. 67% of livebirths linked to the mother can be linked to the father.	Deterministic linkage is used to obtain mother-child pairs, so deemed to be highly reliable. Father-child linkage was obtained using a deterministic approach, with direct and indirect linkage. Direct linkage was performed linking children with their male health cardholder or legal guardian, which had to comply with other criteria such as being 15y or older at the time of birth of the index child and their surnames should be compatible. Then we used household address combined with sex, age and surnames criteria, excluding households where more than one male were retrieved. Only live births can be linked with the father. So, only females with a pregnancy ending in a live birth can be linked to the male partner.		This linkage is available from 2010 to 2024	https://www.eurolinkcat.eu/loadFile.aspx?filename=D2.4%20report%20submitted(2).pdf https://catalogues.ema.europa.eu/node/1077/quantitative-descriptors

	Female partner must not be pregnant at the time of inclusion.	Sex Diagnostic code Date of diagnosis Presence of family linkage Type of linkage	*High*	Diagnostic codes: 100% Sex 100% of individuals have available information Family unit can be linked by deterministic linkage. 67% of livebirths linked to the mother can be linked to the father.	Deterministic linkage is used to obtain mother-child pairs, so deemed to be highly reliable. Father-child linkage was obtained using a deterministic approach, with direct and indirect linkage. Direct linkage was performed linking children with their male health cardholder or legal guardian, which had to comply with other criteria such as being 15y or older at the time of birth of the index child and their surnames should be compatible. Then we used household address combined with sex, age and surnames criteria, excluding households where more than one male were retrieved. The method is based several on healthcare card holders, residency address, so its reliability should be interpreted with caution. Also, it is only possible for live births conceptions, so non-live conceptions cannot be linked with the father. So, only females experiencing a live birth conception will be possible to be linked to the male partner.		This linkage is available from 2010 to 2024	https://www.eurolinkcat.eu/loadFile.aspx?filename=D2.4%20report%20submitted(2).pdf https://catalogues.ema.europa.eu/node/1077/quantitative-descriptors
Treatment/exposure	Valproate use as monotherapy	Medication code Date of prescription/dispensing	*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 valproate prescription / dispensation)	Diagnoses and drugs follow UMLS ontologies.		
Comparator group (if applicable)	Active: levetiracetam use as monotherapy	Medication code Date of prescription/dispensing	*High*	Medication: ATC code (100%)	Active principle is the level of detail of medication (apparently enough to decipher lamotrigine or levetiracetam prescription / dispensation)	Diagnoses and drugs follow UMLS ontologies.		
Key endpoint(s)	Time to first ocurrence of either of: Composite of Autism, ADHD, congenital malformations, stillbirths, spontaneous abortions and post-birth death in offspring.	Diagnostic code Date of diagnosis Date of death Date of delivery Presence of family linkage	*High*	Diagnostic codes: 100% of individuals have available information. 100% of major anomalies is included (see link). A date of death is recorded for 100% of individuals who have died. There is a perinatal mortality registry, where neonatal deaths (until 28 days after birth) are registered and a mortality registry where all deaths are registered.	Minor anomalies are excluded according to EUROCAT criteria. Only livebirths are linked with the father, and thus stillbirths and spontaneous abortions cannot be determined. 67% of livebirths linked to the mother can be linked to the father.	Diagnoses and drugs follow UMLS ontologies. Comparison with EUROCAT classification has been performed. 98.9% of subjects match with congenital anormalities comparing to EUROCAT database	Some neurodevelopment disorders might be diagnosed after offspring reaches different ages.	"https://pmc.ncbi.nlm.nih.gov/articles/PMC10468043/pdf/pone.0290711.pdf https://eu-rd-platform.irc.ec.europa.eu/eurocat/eurocat-members/registries/Valencia-Region_en https://pubmed.ncbi.nlm.nih.gov/38507750/ https://pubmed.ncbi.nlm.nih.gov/38265792/
Confounders	**Mother (for neurodevelopmental disorders)**							
	Sociodemographic features: age, obesity, smoking	Date of birth Sex BMI (patient with BMI > 30) Smoking habits	*Low*	100% date of birth and sex. Smoking can be captured via ICD codes or a specific variable in VID tobacco use, BMI can be caputred via a BMI variable in VID or can be calculated with weight and height data				
	Diagnoses: substance abuse, alcohol abuse, affective disorder, schizophrenia (including schizotypal and delusional disorders), neurotic disorder, neurodevelopmental disorder, rubella, cytomegalovirus, diabetes & gestional diabetes	Diagnostic code Date of diagnosis	*Low*	Diagnostic codes: 100% of individuals have available information		Diagnoses and drugs follow UMLS ontologies.		
	Medications: any concomitant medications associated with valproate-indicated psychiatric conditions, any concomitant medications associated with neuropsychiatric effects	Medication code Date of prescription/dispensing	*Low*	Medication: ATC code 100% of individuals have available information, except for inpatient medication data which is not available in VID.		Diagnoses and drugs follow UMLS ontologies.		
	Mother (for congenital anomalies)							
	Sociodemographic features: age, obesity, smoking	Date of birth Sex BMI (patient with BMI > 30) Smoking habits	*Low*	Date of birth and sex: 100% of individuals have available information. Smoking can be captured via ICD codes or a specific variable in VID tobacco use, BMI can be caputred via a BMI variable in VID or can be calculated with weight and height data				
	Diagnoses: substance abuse, alcohol abuse, rubella, varicella, toxoplasmosis, herpes simplex virus, cytomegalovirus, diabetes & gestional diabetes, folate deficiency	Diagnostic code Date of diagnosis	*Low*	Diagnostic codes: 100% of individuals have available information		Diagnoses and drugs follow UMLS ontologies.		

	Father (for neurodevelopmental disorders)							
	Sociodemographic features: age, calendar year of conception of offspring	Date of birth Sex Date of conception/Date of delivery	*Low*	Available in VID. In regards to conception date, it is calcualted using gestational age at birth estimated by ultrasound (available in most live births), or otherwise using date of LMP.	Given that gestational age as estimated by ultrasound is highly reliable and it is available in most of livebirths, date of conception can be considered reliable in VID.	VID has experience with running pregnancy algorithms. VID has assembled a pregnancy cohort (PREGVAL), including over 500.000 pregnancies.		https://link.springer.com/article/10.1007/s10654-025-01260-7 https://github.com/ARS-toscana/ConcePTIONAlgorithmPregnancies
	Diagnoses: substance abuse, alcohol abuse, affective disorder, schizophrenia (including schizotypal and delusional disorders), neurotic disorder, neurodevelopmental disorder	Diagnostic code Date of diagnosis	*Low*	Diagnostic codes: 100% of individuals have available information		Diagnoses and drugs follow UMLS ontologies.		
	Medications: any concomitant medications associated with valproate-indicated psychiatric conditions, any concomitant medications associated with neuropsychiatric adverse effects	Medication code Date of prescription/dispensing	*Low*	Medication: ATC code 100% of individuals have available information, except for inpatient medication data, not available in VID.		Diagnoses and drugs follow UMLS ontologies.		
	Father (for congenital anomalies)							
	Sociodemographic features: age, calendar year of conception of offspring	Date of birth Sex Date of conception/Date of delivery	*Low*	Available in VID. In regards to conception date, it might be derived from the LMP, date of typical trimestral birth control tests, or date of delivery.	Conception date might be derived from the LMP, date of typical trimestral birth control tests, or date of delivery. Reliability may depend on the inferences that need to be made depending on the information available.	VID has experience with running with pregnancy algorithms ongoing validation.		https://github.com/ARS-toscana/ConcePTIONAlgorithmPregnancies
	Offspring (for neurodevelopmental disorders)							
	Sociodemographic features: sex	Sex	*Low*	Available in VID				
	Diagnoses: foetal alcohol syndrome, fragile X syndrome, congenital cytomegalovirus, lejeune/cri du chat syndrome, tuberous sclerosis	Diagnostic code Date of diagnosis	*Low*	Diagnostic codes: 100% of individuals have available information		Diagnoses and drugs follow UMLS ontologies.		
	Offspring (for congenital anomalies)							
	Diagnoses: foetal alcohol syndrome, congenital rubella, congenital varicella, congenital cytomegalovirus, congenital herpes syndrome, congenital toxoplasmosis	Diagnostic code Date of diagnosis	*Low*	Diagnostic codes: 100% of individuals have available information		Diagnoses and drugs follow UMLS ontologies.		
Intercurrent events	Treatment discontinuation less or more than 3 months prior to conception	Medication code Date of prescription/dispensing Date of discontinuation Treatment duration Date of conception/Date of delivery	*Low*	Medication: ATC code 100% of individuals have available information, except for inpatient medication data, not available in VID. Duration of exposure may be estimated using dispensation data and prescription information on dosing schedule.	Prescription and dispensation are only indirect date of stopping intervention and may suffer imprecisions (a patient may decide not to take usual medication despite persistente of prescription and dispensation)	Diagnoses and drugs follow UMLS ontologies.		
	No conception	Lack of: Date of conception/Date of delivery Pregnancy diagnosis Male-female partner linkage	*High*	Linkage of males with a partner when they have not conceived is not possible since such linkage is made through livebirths.	Deterministic linkage is used to obtain mother-child pairs, so deemed to be highly reliable. Father-child linkage was obtained using a deterministic approach, with direct and indirect linkage. Direct linkage was performed linking children with their male health cardholder or legal guardian, which had to comply with other criteria such as being 15y or older at the time of birth of the index child and their surnames should be compatible. Then we used household address combined with sex, age and surnames criteria, excluding households where more than one male were retrieved. It is only possible for live births to be linked with the father. So, only females experiencing a pregnancy ending in live birth conception can be linked to the male partner.		This linkage is available from 2010 to 2024	https://www.eurolinkcat.eu/loadFile.aspx?filename=D2.4%20report%20submitted(2).pdf https://catalogues.ema.europa.eu/node/1077/quantitative-descriptors
	Conception within 3 months post-treatment initiation	Medication code Date of prescription/dispensing Date of discontinuation Treatment duration Date of conception/Date of delivery	*Low*	Medication: ATC code 100% of individuals have available information, except for inpatient medication data, not available in VID. Duration of exposure may be estimated using dispensation data and prescription information on dosing schedule.		Diagnoses and drugs follow UMLS ontologies.		

	Still birth, spontaneous abortion or congenital malformation	Diagnostic code Date of diagnosis Linkage father-child	*High*	Diagnostic codes: 100% of individuals have available information Around 67% of livebirths are linked to a father. All congenital malformation codes are available for livebirths. Stillbirths and spontaeous abortions can not be linked to the fathers.	Deterministic linkage is used to obtain mother-child pairs, so deemed to be highly reliable. Father-child linkage was obtained using a deterministic approach, with direct and indirect linkage. Direct linkage was performed linking children with their male health cardholder or legal guardian, which had to comply with other criteria such as being 15y or older at the time of birth of the index child and their surnames should be compatible. Then we used household address combined with sex, age and surnames criteria, excluding households where more than one male were retrieved. The method is based several on healthcare card holders, residency address, so its reliability should be interpreted with caution. Also, it is only possible for live births conceptions, so non-live conceptions (i.e., spontaneous abortions, still births) cannot be linked with the father. So, only females experiencing a live birth conception will be possible to be linked to the male partner.	Diagnoses and drugs follow UMLS ontologies.	This linkage is available from 2010 to 2024	
	Switch to another anti-seizure drug prior to conception	Medication code Date of prescription/dispensing Date of discontinuation Treatment duration Date of conception/Date of delivery	*Low*	"Switching" may be operationalised using prescription and dispensing data. Inpatient medication data is not available in VID.		Diagnoses and drugs follow UMLS ontologies.		
	Death of offspring after birth	Date of death Date of delivery/birth Linkage father-child	*High*	A date of death is recorded for 100% of individuals who are known to have died. Around 67% of livebirths are linked to a father. All congenital malformations are available for livebirths. There is a perinatal mortality registry, where neonatal deaths (until 28 days after birth) are registered and a mortality registry where all deaths are registered.	Deterministic linkage is used to obtain mother-child pairs, so deemed to be highly reliable. Father-child linkage was obtained using a deterministic approach, with direct and indirect linkage. Direct linkage was performed linking children with their male health cardholder or legal guardian, which had to comply with other criteria such as being 15y or older at the time of birth of the index child and their surnames should be compatible. Then we used household address combined with sex, age and surnames criteria, excluding households where more than one male were retrieved. The method is based several on healthcare card holders, residency address, so its reliability should be interpreted with caution. Also, it is only possible for live births conceptions, so non-live conceptions (i.e., spontaneous abortions, still births) cannot be linked with the father. So, only females experiencing a live birth conception will be possible to be linked to the male partner.			
Follow-up time needed per patient in the study	15 years	15 years (including recruitment and follow-up)	*High*	Children linked to their fathers can be followed from 2010 to 2025, entailing a maximum follow-up of 15 years. One year of look-back will be available.			Children linked to their fathers can be followed from 2010 to 2025, entailing a maximum follow-up of 15 years. One year of look-back will be available.	
Minimum time in the data source for lookback assessment	3 months (males)	3 months	*Low*	Available in VID.			Children linked to their fathers can be followed from 2010 to 2025, entailing a maximum follow-up of 15 years. One year of look-back will be available.	

	Estimated sample size: Approx. 2,574 children			Considering that the Valencia Integrated Database (VID) covers approximately 98% of the 5 million inhabitants of the Valencia region—representing 10.7% of the Spanish population and around 1% of the European population—with an annual birth cohort of approximately 48,000 newborns. Considering that I) the incidence of epilepsy is of 37,7 cases every 100,000 inhabitants, II) in 2023 levetiracetam represented the 21% of DHD of antiepileptic medicines and valproate the 13%, III) the 15% of no conception and IV) that linkage with the father is possible for the 67% of livebirths, the sample size might be challenging to achieve.				https://app.powerbi.com/view?r=eyJrIjoiNjY1NzVhZjAtYWNmNS00ZTllLTgyNDEtN2E3MGQ5ZTNkZTNmIiwidCI6IjJkM2I1MGUwLTZlZjQtNGViYy05MjQ2LTdkMWNiYjc3MDg5YyIsImMiOjh9 Quintana M, Sánchez-López J, Mazuela G, Santamarina E, Abraira L, Fonseca E, Seijo I, Álvarez-Sabin J, Toledo M. Incidence and mortality in adults with epilepsy in northern Spain. Acta Neurol Scand. 2021 Jan;143(1):27-33. doi: 10.1111/ane.13349. Epub 2020 Oct 13. PMID: 32969054. Villanueva V, Carreño M, Gil-Nagel A, Serrano-Castro PJ, Serratosa JM, Toledo M, Álvarez-Barón E, Gil A, Subías-Labazuy S. Identifying key unmet needs and value drivers in the treatment of focal-onset seizures (FOS) in patients with drug-resistant epilepsy (DRE) in Spain through Multi-Criteria Decision Analysis (MCDA). Epilepsy Behav. 2021 Sep;122:108222. doi: 10.1016/j.yebeh.2021.108222. Epub 2021 Aug 6. PMID: 34371462. https://www.ncbi.nlm.nih.gov/books/NBK546611/

Step 3. CPRD-C9

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 enzyme levels. We may find such codes in the control group (as liver enzyme levels will not be tested) but not in the intervention group. So, in short, in the intervention group we will mainly see small increases in enzyme levels, in the control group we will see ICD-10 codes for liver injuries, but we cannot compare these easily. To bear in mind, there are safety measures for tolvaptan users; treatment is stopped before patients reach these levels of increases in liver enzymes.			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.				

Step 3. NCR-10

Scientific research question		**Clinical Benefit of Capecitabine with Oxaliplatin (CapOx) plus Bevacizumab versus CapOx only in patients with Metastatic Colorectal Cancer**						
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**							
	Histologically confirmed mCRC diagnosis in the last year prior to randomization	Pathology results	*High*	100% of individuals have available information	Once year all cancer dx are reviewed to identify cancer patients that did not have a biopsy and pathology finding.		Since 1989 OMOP-CDM since 1992. Daily updates	
	Age > or = 18y	Date of birth	*High*	100% of individuals have available information	As the format is not known, precision can not be evaluated. Low impact in the study?		Since 1989 OMOP-CDM since 1992. Daily updates	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP
	ECOG < or = 1	ECOG score	*High*	Recorded 15% missing	ECOG is dependent on the eye of the beholder.			
	Not felt to be amenable to curative resection	Pathology results	*High*	100% of individuals have available information	Once year all cancer dx are reviewed to identify cancer patients that did not have a biopsy and pathology finding.		Since 1989 OMOP-CDM since 1992. Daily updates	
	Life expectancy longer than 3 months	Pathology results	*High*	100% of individuals have available information	Once year all cancer dx are reviewed to identify cancer patients that did not have a biopsy and pathology finding.		Since 1989 OMOP-CDM since 1992. Daily updates	
	No prior systemic therapy for mCRC or previous treatment with oxaliplatin or bevacizumab	Medication code Date of prescription/dispensing	*High*	Only first line treatment				
	Adequate hematologic/ clotting, hepatic and renal function	Laboratory tests	*High*	Unknown missingness				
	Exclusion criteria							
	Pregnant or breastfeeding women	Pregnancy/breastfeeding status	*High*	Not registered for colon cancer patients				Provided by DEAP
Treatment/exposure	Bevacizumab (7.5 mg/kg IV, on day 1 of a 3-week cycle) + Capecitabine-Oxaliplatin regimen (IV/3wk).	Medication code Date of prescription/dispensing	*High*	Prescription first line treatment, dose not registered				DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP
Comparator group (if applicable)	Capecitabine-Oxaliplatin regimen (IV/3wk).	Medication code Date of prescription/dispensing	*High*	Prescription first line treatment, dose not registered				
Key endpoint(s)	Progression Free survival (PFS)	Date of treatment initiation Date of progression (imaging) Date of death	*High*	A date of death is recorded for 100% of individuals who are known to have died	Vital status checked once per year. As the date of death is registered it will be possible to calculate. PFS is not directly provided, although an algorithm using prognostic markers has been used in this database to predict PFS, being included in published papers (see column I)		Vital status checked once per year	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). https://onlinelibrary.wiley.com/doi/10.1002/cam4.6223 Provided by DEAP
Confounders	Age > or = 18y	Date of birth	*Low*	100% of individuals have available information				
	Sex	Sex	*Low*	100% of individuals have available information				
	ECOG performance statust score	ECOG score	*Low*	100% of individuals have available information				
Intercurrent events	Treatment discontinuation	Medication code Treatment end date	*Low*	100% of individuals have available information				DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP
	Partial discontinuation: capecitabine	Medication code Capecitabine end date Oxaliplatin end date	*Low*	100% of individuals have available information				DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP
	Treatment switch	Medication code Date of prescription/dispensing Date of discontinuation Treatment duration	*Low*	Only first line treatment	As only first line treatment is recorded it won't be posible to diferentiate descontinuation than switch			
	Local treatment	Date of procedure Procedure code	*Low*	Procedures available, cancer related surgery might be picked if a specific code is available				
Follow-up time needed per patient in the study	48 weeks	48 weeks	*High*				The median lenght of follow-up per patient is approximately 9 months	
Minimum time in the data source for lookback assessment	1 week	1 week	*Low*				The median lenght of follow-up per patient is approximately 9 months	

	Estimated sample size: Approx 440 participants			Considering that the Netherlands Cancer Registry (NCR) recorded 22,192 patients aged ≥70 years with metastatic colon cancer between 2005 and 2020—of whom 23% received targeted therapy—the target sample size is anticipated to be reached.				

FINAL FEASIBILITY ASSESSMENT

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 criticality 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 criticality 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.
2 (Nivolumab plus ipilimumab versus pembrolizumab in patients with advanced non-small-cell lung cancer)	**NCR**	The estimated sample consists of approximately 244 participants (1:1 ratio of nivolumap + ipilimumab + chemotherapy and pembrolizumab + chemotherapy, thus 122 per group). NCR includes 5,000 patients with stage IV NSCLC and 1,000 with stage III NSCLC. Since 2021, ~100 patients have been treated with nivolumab + ipilimumab, compared to 3,000 patients receiving pembrolizumab. The sample size for pembrolizumab is adequate, while the size for nivolumab + ipilimumab could be limited.	Yes	Elements with high criticality are in their majority available, but some of them have limitations. The time elapsed from when a user requests the data to when they actually receive it is 2 months. Data recency is ~12 months before extraction, reasonably enough for the research question. Sample size is achievable.	·Potentially major: The median lenght of follow-up per patient is approximately 9 months ·Potentially major: 15% ECOG missing ·Potentially major: Difficult to detect previous systemic anti-cancer treatment, autoimmune disease or severe infectious disease (e.g., HIV) ·Minor: Only prescription of first line of treatment (if stage changes a treatment is considered as a new first line); it won't be posible to diferentiate discontinuation from switch. ·Minor: TNM reliable, but 6% have missing the specific stage ·Minor: Some cancer patients do not have a biopsy and pathology, but might be picked by diagnostic code ·Minor: Data is registered 6-12 months after diagnosis so there is a lag	Although the median follow-up time in the NCR is 9 months, this includes patients with all types of cancer with different survival durations. However, this variation is likely non-differential, meaning it is not expected to bias the results in favour of or against any particular cancer group. If the patients included in the study have a longer survival time, the registry will allow for the follow-up required by protocol. Missing ECOG data may prevent us from including some subjects. Previous cancer or anti-cancer treatments can be detected from the patient's previous records in the registry. The history of autoimmune disease or severe infections cannot be detected, but we believe that this fact is already implicit in the physician's decision to treat the patient. This should be taken into account in the interpretation of the results.
3 (Dapagliflozin and Major Adverse Cardiovascular Events in Type 2 Diabetes)	**BIFAP**	With an approximate estimated sample size of 13,341 (based on a 1:1 ratio of dapagliflozin and DPP-4i), and considering that BIFAP includes data from approximately 14 million inhabitants (up to 2018), the target sample size is anticipated to be reached. Furthermore, experimental exposure is expected to occur frequently (13.5 DHD in 2023) [1].	Yes	Elements with high criticality are available and fairly reliable. 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 1-4 months. Sample size is achievable.	·Minor: Hospital information in BIFAP includes dates of admission and discharge, type of discharge, primary and secondary diagnoses at hospital discharge. So, an acute cardiovascular event will only be picked if it constituted one of the main reasons for admission. ·Minor: Mortality data is updated with a one-year delay relative to the present time. For research purposes only the year of death is available. ·Minor: In mortality, a non-random pattern of missingness (MNAR) was observed due to incomplete or inaccurate recording of the cause of death, with a tendency to preferentially register cardiovascular-related deaths. A non-random pattern of association between missingness and MACE was seen. GPs do not have a complete registry of deaths and, particularly, there is not an appropriate recording of the cause of death. Consequently, adjustments using statistical methods for MNAR should be considered in the TTE protocol. ·Minor: Discontinuation date is not available, but calculated by dispentation date+number of packages+posology if written by doctor, if not, calculated by algorithm. ·Minor: Drug use is not linked to a specific indication. ·Minor: Smoking status may be biased, as the criterion is 'current use or use within one year prior to randomization; therefore, patients who smoked before this period would be classified as non-smokers.	As in-hospital cardiovascular events might not be fully captured, some underestimation of outcomes may exist. However, as these are usually severe and with chronic repercussions, we expect primary care setting will capture them even with some delay. As mortality data is delayed and only the year of death is available, this can impact precision and the time sequence of outcomes.
	CPRD	With an approximate estimated sample size of 13,341 (based on a 1:1 ratio of dapagliflozin and DPP-4i), and considering that CPRD includes data from approximately 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	Elements with high criticality are available and fairly reliable. Data recency of 3 months before extraction, reasonably enough for the research question. Sample size is achievable.	·Minor: Dispensing is not available, only prescription. ·Minor: Treatment discontinuation is not readily available but inferred from prescription duration. ·Minor: Diagnostic codes are available for 86% subjects attending emergency room. ·Minor: Diabetes mellitus without type specification occurs frequently as well; usually insulin in monotherapy is used to assess T1D.	As this database only has prescription data, it is unknown if patients took the prescription or if they discontinued it. However, treatment duration is available, from which this data may be estimated. Diagnostic codes are reported to be available for 86% of subjects in the emergency room; however, the missing cases we expect to capture them from hospitalization records or primary care records, since the severity of this disease may justify an admission and/or the follow-up with the GP, or change of baseline treatment. As diabetes type is frequently not specified, insulin in monotherapy might be used as a proxy to detect T1D cases. Very low misclassification of indication is extected since insulin in monotherapy is not used for other infications rather than T1D.
4 (Rivaroxaban and risk of major gastrointestinal bleeding in elderly patients with non-valvular atrial fibrillation)	**DNR**	With an approximate estimated sample size of 45,493 individuals (based on a 1:1 ratio between treatment arms, with 22,747 participants in each), and considering that the Danish population includes approximately 5.9 million inhabitants (as of 2023), the target sample size is anticipated to be reached. Furthermore, previous literature reports 46 675 patients with non-valvular atrial fibrillation (NVAF) claimed a prescription of anticoagulation between 2011 and 2014 in Denmark. [2]	Yes	Elements with high criticality seem available but reliability is unknown. Data recency of 2-3 years old, depending on the datasets needed, reasonably enough for the research question. The time elapsed from when a user requests the data to when they actually receive it is 3-6 months (if the cohort is already extracted and used within a specific approved purpose for other projects, no lag in delivery). Sample size is achievable.	·Minor: Treatment duration and discontinuation needs to be estimated by means date of last medication acquisition. ·Minor: If the prescribed daily dose is not recorded, the defined daily dose (as defined by WHO) can be used as a proxy of consumption. ·Minor: Some standard UMLS dictionaries are available, such as SNOMED, RxNorm or ATC; but some mapping might be needed for ICD10DA codes and procedure codes.	As exact treatment duration is not available, depending on the method to estimate it we may under or overestimate exposure episodes.
	SIDIAP	With an approximate estimated sample size of 45,493 individuals (based on a 1:1 ratio between treatment arms, with 22,747 participants in each), and considering that SIDIAP includes data from approximately 5.8 million inhabitants, with 11,962 patients with non-valvular atrial fibrillation (NVAF) claimed a prescription of anticoagulation between 2011 and 2014 identified in previous literature, the target sample size is anticipated to be reached. [3]	Yes	Elements with high criticality are available and fairly reliability. Data recency of 8-9 months, dreasonably enough for the research question. The time elapsed from when a user requests the data to when they actually receive it is 2-3 months. Sample size is achievable.	·Potentially major: Inability of capturing the cause of death in the database. Cause of death might need to be inferred since it is not recorded in the data source. ·Minor: Treatment duration and discontinuation needs to be estimated.	Since cause of death is not available, there is a risk of outcome misclassification. However, this limitation could be mitigated by inferring the likely cause of death based on diagnostic information recorded near the time of death. Additionally, because exact treatment duration is not available, estimates of exposure episodes may be under- or overestimated depending on the method used to approximate treatment duration
5 (Vilanterol/fluticasone furoate in the risk of pneumonia in adolescents with asthma)	**CPRD**	With an approximate estimated sample size of 26,750 individuals (based on a 1:1 ratio between treatment arms, with 13,375 participants in each), and considering that CPRD includes data from approximately 4.4 million inhabitants (as of 2014), the target cohort size is anticipated to be achievable. In the UK, asthma affects approximately 7.2 million individuals—about 8% of the population—with up to 5.4 million currently receiving any treatment. This translates to approximately 281,600 subjects. Teeneger population in CPRD is 200,949, and the prevalence of asthma in UK teenegers is 15.2%, this translates in 30,544 asmathic teenegers in CPRD. The target sample size is anticipated to be reached. [4,5,6]	Yes	Elements with high criticality are available and fairly reliable. Data recency of 3 months before extraction, reasonably enough for the research question. Sample size is achievable.	·Potentially major: Hospitalisation data is not available for the core data source. ·Potentially major: Not registered date of death. ·Minor: Dispensing is not available, only prescription. ·Minor: Diagnostic codes are available for 86% subjects attending emergency room.	The DEAP is able to perform data augmentation (linkage) to retrieve hospital admission and discharge diagnoses to detect the outcomes of this case-study. For exact date of death or the cause of death data augmentation (linkage) is needed to the ONS death registration data. As this database only have prescription data, it is unknown if patients took the prescription, and so, if they discontinued it. However, treatment duration is available, from which this data may be estimated. Diagnostic codes are reported to be available for 86% of subjects in the emergency room; however, the missing cases we expect to capture them from hospitalization records or primary care records, since the severity of this disease may justify an admission and/or the follow-up with the GP, or change of baseline treatment.

	Finnish registers	With an approximate estimated sample size of 26,750 individuals (based on a 1:1 ratio between treatment arms, with 13,375 participants in each), and considering that Finland has a population exceeding 5 million (380,000 individuals aged 12-17y, 10.5% with asthma)- Despite there is up to 18,293 users of vilanterol-fluticasone furoate among the 233,261 patients with chronic asthma or similar chronic obstructive pulmonary diseases as of 2021, there is no specific data available for teenage users. Since the number of teenage users is expected to be smaller than the target sample size, we anticipate that reaching the desired sample may not be feasible. [7,8]	Yes, with limitations on sample size acquisition	Elements with high criticality are available and fairly reliable. Data recency is variable depending on the data sets used, but for the current case of a few months are reasonably enough for the research question. The time elapsed from when a user requests the data to when they actually receive it is 4-12 months. **Sample size might be difficult to be reached**.	·Potentially major: limited sample size. ·Minor: No inhospital drug use available. ·Minor: Duration of hospitalisation is not readily available. ·Minor: Exacerbations inhospital are not available. ·Minor: Duration of treatment and end of treatment are not directly available.	Limited sample size in Finnish registers might lead to underpowered analyses for this population. This can be mitigated by meta-analysing the results together with CPRD. Duration of hospitalisation can be calculated from the admission and discharge dates. Inhospital exacerbations can be inferred from admission, discharge and primary care diagnoses that are readily available. End of treatment episodes may be derived based on dispensing date and dispensed amount.
6 (Sacubitril/valsartan in the risk of angioedema)	**CPRD**	With an approximate estimated sample size of 30,784 (based on a 1:1 ratio of stopping current ACEi and starting Sacubitril/Valsartan versus continuing on ACEi), and considering that CPRD includes data from approximately 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	Elements with high criticality are available and fairly reliable. Data recency of 3 months before extraction, reasonably enough for the research question. Sample size is achievable.	·Minor: Dispensing is not available, only prescription. ·Minor:Diagnostic codes are available for 86% subjects attending emergency room. ·Minor: Treatment discontinuation not directly available but can be assessed using standard adherence calculation methods.	As this database only has prescription data, it is unknown if patients took the prescription, and so, if they discontinued it. However, treatment duration is available, from which this data may be estimated. Diagnostic codes are reported to be available for 86% of subjects in the emergency room; however, the missing cases we expect to capture them from hospitalization records or primary care records, since the severity of this disease may justify an admission and/or the follow-up with the GP, or change of baseline treatment.
	PHARMO	With an approximate estimated sample size of 30,784 (based on a 1:1 ratio of stopping current ACEi and starting Sacubitril/Valsartan versus continuing on ACEi), and considering that Pharmo includes data from 40% of the Dutch population, the target sample size is anticipated to be reached. Furthermore, experimental exposure is expected to occur frequently (50,102 Sacubitril/Valsartan users and 1,099,000 ACEi users recorded in the Netherlands in 2023). [9]	Yes	Elements with high criticality are available, and fairly reliable. Sample size is achivable. Data are available with an approximately 1-year lag depending on the databases required.	·Minor: 70-100% completness in most of the variables.	No major impact expected.
7 (Paternal exposure to valproate and the risk of neurodevelopment disorders in offspring)	**VID**	With an approximate cohort estimated sample of 2,574 children (based on a 1:1 ratio between exposure groups), and considering that the Valencia Integrated Database (VID) covers approximately 98% of the 5 million inhabitants of the Valencia region—representing 10.7% of the Spanish population and around 1% of the European population—with an annual birth cohort of approximately 48,000 newborns, considering that that I) the incidence of epilepsy is of 37,7 cases every 100,000 inhabitants, II) in 2023 levetiracetam represented the 21% of DHD of antiepileptic medicines and valproate the 13%, III) the 15% of no conception and IV) that linkage with the father is possible for the 67% of livebirths, the **sample size might be challenging to achieve**. In 2023, valproate use was estimated at 1.7 DHD (defined daily doses per 1,000 inhabitants per day). [1][10-13]	Yes, with limitations on sample size and elements of high criticality.	Most elements with high criticality are available and fairly reliable. However, the approach or methods to tackle **intention to conceive, no conception, female partner-related selection criteria and some outcomes (i.e., stillbirth, spontaneous abortions)** should be accounted for. 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 usually from 6 to 12 months, but in this case VID will have the data available at the time the analysis is expected. **Sample size is can be challenging to achieve**, so the precision it enables should be recalculated. Children linked to their fathers can be followed from 2010 to 2025, entailing a maximum follow-up of 15 years. One year of look-back will be available.	·Potentially major: Intention to conceive, or no conception cannot be fairly assessed in RWD sources in general. Proxies might be considered. Also, as the father-child linkage is performed through livebirths, stillbirth and spontaneous abortions cannot be assessed for offsprings with fathers receiving the exposures of interest. Female partner-related selection criteria (such as not having a diagnosis of epilepsy), will only be possible for fathers who had livebirths. Approaches or methods to tackle these should be accounted for. ·Potentially major: As the sample size is challenging to achieve, statistical precision should be recalculated. ·Minor: Father-child linkage is obtained using a deterministic approach-the 67% of livebirths linked to the mother. Reliability should be interpreted with caution. ·Minor: Children linked to their fathers can be followed from 2010 to 2025, entailing a maximum follow-up of 15 years. One year of look-back will be available. ·Minor: Inpatient medication not available. ·Minor: Duration of exposure may be estimated using dispensation data and prescription information on dosing schedule. ·Minor: Prescription and dispensation need to be used to calculate indirectly the date of stopping or switching intervention and may suffer imprecisions. ·Minor: If the particular lifestyle interventions for this research question have a codelist, they might be captured. ·Minor: The criteria "Female partner must not be diagnosed with generalised epilepsy" cannot be assessed at the time of randomisation. This can be assessed when having the newborn.	Some misclassification of father-child linkage is possible, and the linkage is available for 67% of livebirths. Cases of spontanoues abortion or stillbirth cannot be linked with the father, which may underestimate the key endpoint or misclassify the outcome. Partner-related information will be assessed retrospectively once linked by means the liveborn. We do not expect a relevant impact in this regard. Only females experiencing a live birth conception will be possible to be linked to the male partner. Additionally, intention to conceive is not direcvtly available in RWD sources. So, items like "Participants must have a female partner with which they intend to conceive" or "No conception" are not available. Children linked to their fathers can be followed from 2010 to 2025, entailing a maximum follow-up of 15 years. The impact is expected to be low as neurodevelopmental disorders usually happens (from 18 months for autism, and from 4-5 years for ADHD) and this can be handled with the appropriate analysis (e.g., survival analysis). Sample size may not achieve sufficient precision for regulatory purposes, but reflects the reality of using RWD sources. Precision can be lower than calculated initially.
8 (Nirsevimab against RSV-respiratory tract infection in infants)	**PEDIANET**	With an approximate estimated sample size of 7,408 individuals, and considering that PEDIANET includes data on 24,572 toddlers aged between 28 days and 23 months, and ~ 30K birth every year.	Yes	Elements with high criticality are available and seem fairly reliable. 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 unknown. Sample size can be reached.	·Potentially major: For those patients born since 2024, some unknown maternal RSV immunisation is expected. ·Minor: The total number of bronchiolitis tested for pathogens is unknown. ·Minor: Unknown exact missingness of gestational age at birth. ·Minor: Duration of treatment is not available.	Some underdetection of maternal RSV immunisation is expected. However, maternal RSV inmunization is rare in Italy since it is not reimbursed apart from some specific local health units. Since duration of immunosupressive therapy is not available, it will be determined using a proxy based on diagnosis needing corticosteroides for more than 2 weeks and drugs prescriptions Previous studies using the same database to analyze pediatric lower respiratory tract infections (LRTIs) did not report the total number of bronchiolitis cases tested for pathogens. This omission may result in an underestimation of the incidence of RSV-associated LRTIs.
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. [14,15]	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.
10 (Capecitabine with Oxaliplatin (CapOx) plus Bevacizumab versus CapOx in patients with Metastatic Colorectal Cancer)	**NCR**	With an approximate estimated sample size of 440 individuals (based on a 1:1 ratio between treatment arms, comparing CAPOX plus bevacizumab versus CAPOX alone), and considering that the Netherlands Cancer Registry (NCR) recorded 22,192 patients aged ≥70 years with metastatic colon cancer between 2005 and 2020—of whom 23% received targeted therapy—the target sample size is anticipated to be reached. [16]	Yes, with limitations on a design element	Elements with high criticality are available and fairly reliable, **with reservations regarding a design element endpoint**. The time elapsed from when a user requests the data to when they actually receive it is 2 months. Data recency is ~12 months before extraction, reasonably enough for the research question. Sample size is achievable.	·Potentially major: Progression free survival (key endpoint) is not directly provided, although an algorithm using prognostic markers has been used in this database to predict PFS. ·Potentially major: ECOG is 15% missing. ·Potentially major: The median lenght of follow-up per patient is approximately 9 months. ·Minor: Some cancer patients do not have a biopsy and pathology, but might be picked by diagnostic code. ·Minor: Only prescription of first line of treatment is available, but cancer stage changes mean a new first treatment line is started; so, we will be able to identify previous treatments. ·Minor: Data is registered 6-12 months after diagnosis so there is a lag. ·Minor: Imaging information to assess progression-free survival is not available, only death is captured ·Minor: Procedure codes are available, but cancer-related surgery might only be picked if a specific code is available.	Although PFS is not directly available, a previously developed algorithm using prognostic markers has been applied in this database to estimate PFS. Missing ECOG data may prevent us from including certain subjects. Although the median follow-up time in the NCR is 9 months, this includes patients with all types of cancer with different survival durations. However, this variation is likely non-differential, meaning it is not expected to bias the results in favour of or against any particular cancer group. If the patients included in the study have a longer survival time, the registry will allow for the follow-up required by protocol.

REFERENCES

[1] https://www.aemps.gob.es/medicamentos-de-uso-humano/observatorio-de-uso-de-medicamentos/informes/?lang=ca

[2] Sørensen R, Jamie Nielsen B, Langtved Pallisgaard J, Ji-Young Lee C, Torp-Pedersen C. Adherence with oral anticoagulation in non-valvular atrial fibrillation: a comparison of vitamin K antagonists and non-vitamin K antagonists. Eur Heart J Cardiovasc Pharmacother. 2017 Jul 1;3(3):151-156. doi: 10.1093/ehjcvp/pvw048. PMID: 28158553.

[3] Ibáñez L, Sabaté M, Vidal X, Ballarín E, Rottenkolber M, Schmiedl S, Heeke A, Huerta C, Martín Merino E, Montero D, Leon-Muñoz LM, Gasse C, Moore N, Droz C, Lassalle R, Aakjaer M, Andersen M, De Bruin ML, Groenwold R, van den Ham HA, Souverein P, Klungel O, Gardarsdottir H. Incidence of direct oral anticoagulant use in patients with nonvalvular atrial fibrillation and characteristics of users in 6 European countries (2008-2015): A cross-national drug utilization study. Br J Clin Pharmacol. 2019 Nov;85(11):2524-2539. doi: 10.1111/bcp.14071. Epub 2019 Sep 4. PMID: 31318059; PMCID: PMC6848911.

[4] https://www.asthmaandlung.org.uk/conditions/asthma/what-asthma

[5] https://cks.nice.org.uk/topics/asthma/background-information/prevalence/

[6] Couriel J. Asthma in adolescence. Paediatr Respir Rev. 2003 Mar;4(1):47-54. doi: 10.1016/s1526-0542(02)00309-3. PMID: 12615032.

[7] https://www.julkari.fi/bitstream/handle/10024/145777/Finnish_statistics_on_medicines_2021.pdf?sequence=5&isAllowed=y

[8] Gehrt L, Vahlkvist S, Petersen TH, Englund H, Nieminen H, Laake I, Kofoed PE, Feiring B, Benn CS, Trogstad L, Sørup S. Trends in childhood asthma in Denmark, Finland, Norway and Sweden. Acta Paediatr. 2025 Jan 13. doi: 10.1111/apa.17573. Epub ahead of print. PMID: 39803879.

[9] https://www.gipdatabank.nl/
[10] Quintana M, Sánchez-López J, Mazuela G, Santamarina E, Abraira L, Fonseca E, Seijo I, Álvarez-Sabín J, Toledo M. Incidence and mortality in adults with epilepsy in northern Spain. Acta Neurol Scand. 2021 Jan;143(1):27-33. doi: 10.1111/ane.13349. Epub 2020 Oct 13. PMID: 32969054.
[11] Villanueva V, Carreño M, Gil-Nagel A, Serrano-Castro PJ, Serratosa JM, Toledo M, Álvarez-Barón E, Gil A, Subías-Labazuy S. Identifying key unmet needs and value drivers in the treatment of focal-onset seizures (FOS) in patients with drug-resistant epilepsy (DRE) in Spain through Multi-Criteria Decision Analysis (MCDA). Epilepsy Behav. 2021 Sep;122:108222. doi: 10.1016/j.yebeh.2021.108222. Epub 2021 Aug 6. PMID: 34371462.
[12] https://www.ncbi.nlm.nih.gov/books/NBK546611/
[13] https://app.powerbi.com/view?r=eyJrIjoiNjY1NzVhZjAtYWNmNS00ZTllLTgyNDEtN2E3MGQ5ZTNkZTNmIiwidCI6IjJkM2I1MGUwLTZlZjQtNGViYy05MjQ2LTdkMWNiYjc3MDg5YyIsImMiOjh9
[14] https://www.ncbi.nlm.nih.gov/books/NBK532934/
[15] 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.
[16] Baltussen JC, de Glas NA, Liefers GJ, Slingerland M, Speetjens FM, van den Bos F, Cloos-van Balen M, Verschoor AJ, Jochems A, Spierings LEAMM, Holterhues C, van Gerven LA, Mooijaart SP, Portielje JEA, Derks MGM. Time trends in treatment patterns and survival of older patients with synchronous metastatic colorectal cancer in the Netherlands: A population-based study. Int J Cancer. 2023 May 15;152(10):2043-2051. doi: 10.1002/ijc.34422. Epub 2023 Jan 13. PMID: 36620951.