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*
		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://cataloques.ema.europa.eu/node/1026/data-flows-and-management https://www.cprd.com/cprd-linked-data https://pmc.ncbi.nlm.nih.gov/articles/PMC4521131/ https://www.cprd.com/sites/default/files/2022-02/Linkage%20Source%20Documentation_set22_1.20.pdf	*2*	*L1 if free-text information, links or publications are available reporting all the mentioned features* *L2 if algorithms are published and their performance documented. Information on which values result from imputation is provided as part of the dataset (e.g., presented in metadata or a data dictionary)* *L3 if an automatised process for data linkage/mapping exists*	*Data augmentation steps impact accuracy (reliability) and extensiveness. We consider here data transformations that produce new information subject to reliability issues: e.g.: imputation of missing values, or extraction of codes via natural language processing.*
		If yes, which are the methods applied	For linkage to the HES datasets, ONS Death, NCRAS, ICNARC and Mental Health data, the trusted third party use an eight-step process to match patients using some or all of the following: NHS number, date of birth, sex and postcode. It is explained in the attached link	https://www.cprd.com/sites/default/files/2022-02/Linkage%20Source%20Documentation_set22_1.20.pdf			
		If yes, which algorithms and assumptions applied	It is explained in the attached link	https://www.cprd.com/sites/default/files/2022-02/Linkage%20Source%20Documentation_set22_1.20.pdf			
		If yes, which is the error rate when conducting the augmentation	Requested to DEAP and unable to provide		*N/A*		
VIII	Known quality issues and independent QA assessment of the RWD source	Known DQ issues (e.g., poor overall completeness in Q3 2020 due to COVID-19)	A significant proportion of lab data lacking a normal range were missing units or had values inconsistent with units provided. A significant proportion of cases of hyperlipidemia or anemia will be missed ifthe investigator relies solely on diagnosis codes to select patients. Researchers should consider using available treatments, supporting codes, and lab data to supplement diagnosis codes and enhance case capture when studying anemia, diabetes and hyperlidiemia using CPRD. In rpevious articles, CPRD assumed that, for anemia, diabetes or hyperlimidemia, lab and prescription data were less likely than GP entered diagnosis codes to be missing or miscoded, as prescriptions must be entered into the electronic recordto be issued and lab data with a normal range are likely to be elec-tronically transferred from the laboratory. As CPRD has prescription data, it is unknown whether the patient took the prescription.	https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135 https://www.sciencedirect.com/science/article/pii/S2214623720300351?via%3Dihub#s0055	*1*	*L1 if free-text information, links or publications are available reporting all the mentioned features* *L2 if standard procedures are set for external/internal validation of the data* *L3 if the mechanism provided includes notification of automatically detected DQ issues*	*Explicit description of known DQ issues, as well as external validation performed (all dimensions affected)*
		Validation studies and publications resulting from this EWD source	Useful publications on the quality of CPRD data for research	https://www.cprd.com/data-quality			
IX	The RWD source representation	Description of data model or models used (OMOP, FHIR, ...)	OMOP and CONCEPTION	https://cataloques.ema.europa.eu/node/1026/data-flows-and-management	*3*	*L1 if free-text information, links or publications are available reporting all the mentioned features* *L2 if the description refers to a model such as OMOP, I2B2, FHIR, others, or an extension of them. Data dictionaries are standard (and if non-standard, jusified* *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 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*	*Descriptive of feedback mechanisms in place to improve all aspects of DQ*

	production process, thus allowing a continuous monitoring and improvement of DQ?				*L3 if the mechanism provided includes notification of automatically detected DQ issues*	

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/997/administrative-details	*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)* *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: REQuesST Tool and its vision paper [Internet]. EUnetHTA. 2019. Available from: 721 https://www.eunethta.eu/request-tool-and-its-vision-paper/.* *L3 if the information is provided as Metadata (machine readable), including standrad formats, clear definitions and potentially some quality information*	*Relevant for all DQ dimensions (reliability, extensiveness, coherence and timeliness) as it provides a general understandin g of the strengths and limitations of an RWD source.* *Knowing the triggers would ease the understanding of the content and motivations behind the data.*
		Criteria for the selection of the data being collected or integrated	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			
		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			*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	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.				
		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 GP□Diagnoses 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 pharmacies□Type 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 hospitalization□Type 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 specialists□Date 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.b□Diagnoses 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.c□Information on mothers, eg: ▪Maternal age	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*	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*	*When data are provided by a data aggregator, ensure that all the available evidence related to systems and processes potentially affecting DQ*
		Is there a DQ assessment for data sources onboarded?	N/A				

	integrate or repurpose other RWD sources)	If yes: does it follow any specific framework? Is there an assessment checklist? Are datasets versions traceable?	N/A			*L3 is only aspirational. NA*	*(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)*
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* *L3 NA*	*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.				
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-fulltext-article-CLEP	*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* *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)*	*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	General Practitioner (GP) Database (in-house)Diagnoses and symptoms: ICPC Drug prescriptions: WHO ATC Classification System Out-patient Pharmacy Database (in-house)Dispensings:	https://pmc.ncbi.nlm.nih.gov/articles/PMC7196787/			

		across different languages (e.g., UMLS)	National product classification WHO ATC Classification System In-patient Pharmacy Database (in-house)Dispensings: National product classification WHO ATC Classification System Clinical Laboratory Database (in-house)WCIA Coding System LDINC (partly) Hospital Database (external)Diagnoses: WHO ICD Procedures: DHD registration system for procedures Medication: Dutch classification system Pathology Registry (external)Diagnosis codes are a combination of diagnostic terms (localization, acquisition technique, abnormality) and related to the SNOMED coding system. Cancer Registry (external)Tumor staging: TNM-classification Tumor site and morphology: WHO ICD-O			*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* *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	An epidemiologist from the PHARMO Institute can be made available to support data	Provided by DEAP			
		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/			
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 -Correct the problem -Prevent (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*

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

	Relational coherence	% of records with the Person ID in the PERSONS table	98.2-100%	https://zenodo.org/records/13384860
	Semantic coherence - to determine whether the database uses a standardised dictionary	For EVENTS definitions, codelists/data dictionaries being employed according to external standards	Read Code (CPRD Gold): these are used for diagnoses; from April 2018, Read codes are prospectively mapped to SNOMED CT codes SNOMED (CPRD Aurum) Local EMIS® codesICD-10 for HES Medcodeid (unique code for the medical term selected by the GP)	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf https://www.cprd.com/sites/default/files/2024-08/CPRD%20Aurum%20Data%20Specification%20v3.5.pdf
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	Prodcodeid (unique code for the treatment selected by the GP), SNOMED for some immunisations No ATC codes available in the raw data but ATC for active substances link is available at the Utrecht University	https://zenodo.org/records/13384860
	Uniqueness	Number of records flagged as potential duplicates	Requested to DEAP and unable to provide	

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 https://catalogues.ema.europa.eu/node/997/administrative-details
		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

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.				

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.				

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
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). [1]	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.

REFERENCES

[1] https://www.gipdatabank.nl/