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)* *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 standard 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	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/			
		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			
		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)* *L2 if information is available using standardised templates to make information easy to digest and interpret, and also standard vocabularies are available* *L3 if additionally SOPs specify KPIs to monitor*	*Essential to understand extensiveness and to assess reliability (that can be affected by errors or biases in the collection process). Also, essential to evaluate SOP for data collection or recording practices that may impact coherence (e.g., where "curation at source" is involved and provide hard constraints for timeliness).*
		Standard Operating Procedures (SOPs) recording	The 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			
		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			
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)* *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	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			

		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. 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		*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	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			*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://catalogues.ema.europa.eu/node/991/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)	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			
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	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-	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*

Item		Sub-item	Description	Origin of information	Maturity level grade	Maturity level criteria and definitions
0	Data base identification	Country	Spain	https://www.sidiap.org/index.php/ca/	*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://catalogues.ema.europa.eu/institution/50154	*2*	*L1 if information is available as free text and/or online link(s)*
		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://catalogues.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)*
		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://catalogues.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*	
		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)*

		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 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/dyac068		*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*
		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*
		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://catalogues.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*
		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		*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, ...)	TrineTX, ConcepTION, OMOP	https://catalogues.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*
		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://catalogues.ema.europa.eu/node/1019/data-flows-and-management		*L2 if the description refers to a model such as OMOP, I2B2, FHIR, others, or an extension of them. Data dictionaries are standard (and if non-standard, justified why)* *L3 if a standard CDM is used, the datasource has been mapped to one or more than one CDM, and if data dictionaries are provided using standard formats that facilitate the mappings across different vocabularies and across languages*
X	The RWD source declared Service Level Agreements (SLA)	Guaranteed frequency of updates and incident response time (e.g., corrections in case of errors)	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*
		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		*L2 if details of established data processess by the provider are available*
		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		*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	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*
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*

Rationale
N/A
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.*
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).
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)
Essential for reliability regarding data alterations resulting from system accidents, software errors or malicious intervention.

Data management and governance impact reliability, as well as all quality dimensions for metadata.		
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 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.		

Explicit description of known DQ issues, as well as external validation performed (all dimensions affected)
Descriptive of the intended coherence DQ of a dataset and its metadata.
Descriptive of guaranteed timeliness and possible variations of extensivenes s/reliability provided.
Descriptive of aspects that can limit extensiveness and coherence in downstream data aggregations.
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)	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	

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

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.				

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

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
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. [1]	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. [2]	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

REFERENCES

[1] 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.

[2] Ibáñez L, Sabaté M, Vidal X, Ballarin E, Rottenkolber M, Schmiedl S, Heeke A, Huerta C, Martin 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.