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

		Software testing and software quality control in place	Requested to DEAP and unable to provide		N/A	*L2 if the hardware or software implementation complies with recognised quality standards that can be reported*	
		Measures to prevent accidental physical data alterations (e.g.: backups, redundant systems, checksums)	CPRD is oblied to complete an annual NHS Data Security and Protection Toolkit assessment to demonstrate that it meets the required standard for holding data securely. We are unsure of what this toolkit entails. information is broad and might be only available when you buy/contract the service.	https://www.cprd.com/safeguarding-patient-data https://www.dsptoolkit.nhs.uk/	*2*	*L3 NA*	
V	Data management and governance	Data management principles being followed (e.g., GCP, ISO, FAIR, etc)	Requested to DEAP and unable to provide		N/A	*L1 if information is available as free text and/or online link(s)*	*Data management and governance impact reliability, as well as all quality dimensions for metadata.*
		Data management processes in place (DQ controls, KPIs, SOPs, etc)	**Check:** the volume of data downloaded against that supplied data volumes are in the expected range all data elements received are of the correct type, length and format **Our range of validation and quality checks include:** Collection-level validation ensures integrity by checking that data received from practices contain only expected data files and ensures that all data elements are of the correct type, length and format. Duplicate records are identified and removed. Transformation-level validation checks for referential integrity between records ensure that there are no orphan records included in the database (for example, that all event records link to a patient). Research-quality-level validation covers the actual content of the data. CPRD provides a patient-level data quality metric in the form of a binary 'acceptability' flag. This is based on recording and internal consistency of key variables including date of birth, practice registration date and transfer out date. In addition to checks undertaken by the CPRD teams before the data is released, researchers using the data are advised to undertake study-specific checks themselves.	https://www.cprd.com/data-quality	*2*	*L2 if standard best practices are being used and a direct impact on DQ is reported. There are SOPs and dat amanagement processes that adhere to the standards. The representation of metadata follows FAIR standards*	
		Measures to prevent data alterations by unauthorised parties (cybersecurity)	Single study dataset licence – where a study dataset defined by an approved research application will be prepared by CPRD, and access granted to researchers via the CPRD Trusted Research Environment (TRE). As UU, they have a multistudy license; so data is extracted by UU themselves. The TRE is not used by UU at this moment; we use our own secure TRE for research purposes	https://www.cprd.com/cprd-safe-our-trusted-research-environment			
		Auditing and DQ imporvement procedures in place	Sensitive mortality data Operational management issues Data destruction Acces control Information transfer Risk managment Operational transfer	https://digital.nhs.uk/services/data-access-request-service-dars/data-sharing-audits/2021/post-audit-review-cprd		*L3 if data management and governance is implemented in the data platforms 'Digital Quality Measures' (DQMs) so that reports of performance and deviations are automated. Submitted metadata are generated "by design". Basically, if everything in L2 is automatised and generated by default*	
VI	Data manipulation steps	Frequency of data updates	GOLD: monthly; Aurum: Quarterly	https://catalogues.ema.europa.eu/node/976/data-flows-and-management	*2*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*Impacts reliability both in terms of accuracy (possible errors) and precision (i.e., the degree of approximatio n by which data represents reality). Essential to ensure traceability of information. Also impacts coherence and potentially timeliness.*
		Data transformations performed, data mapping steps, data cleaning	Requested to DEAP and unable to provide		N/A	*L2 if Tests performed follow some standard or shared set of tests, that can be re-used across RWD sources. Key performance indicators (KPIs) for data cleaning (e.g., data duplications, mislabelling, etc.) are provided. Data mapping tables and algorithms are described with a standard characterisation of their performance. Lists of standard test batteries used to detect loss of accuracy or precision are provided. All lineage information is provided as metadata associated to the dataset*	
		Information about loss of precision during data manipulation steps	Requested to DEAP and unable to provide			*L3 if information about data onboarding is directly provided by the platform, e.g.: * Transaction logs are available*	

		Lineage information (e.g., justifcation of data manipulation, track of changes and versions)	Each dataset has a digital object identifier (DOI) to trace specific database versions	https://www.cprd.com/digital-object-identifiers-dois-datasets	*2*	*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*	
VII	Data augmentation steps (e.g., imputation or linkage)	Is any augmentation happening in this datasource?	Patient-level data from consenting practices are linked via a trusted third party—the Health and Social Care Information Centre—to a range of other data sources. Established linkages include Hospital Episode Statistics (HES), covering Admitted Patient Care (APC), Accident & Emergency (A&E), and Outpatient (OP) data; Office for National Statistics (ONS) mortality records, including causes of death; and multiple deprivation indices such as the Index of Multiple Deprivation (IMD), Townsend index, Carstairs index, and Rural-Urban classification. Linkages also extend to disease registries, including the National Cancer Intelligence Network and tumour-level records from the National Cancer Data Repository (NCDR) submitted to ONS by the England Cancer Registries, as well as the Myocardial Ischaemia National Audit Project. Additional linkages are planned (see CPRD website), and researchers can request bespoke linkage for individual studies.	https://catalogues.ema.europa.eu/node/1026/data-flows-and-management https://www.cprd.com/cprd-linked-data https://pmc.ncbi.nlm.nih.gov/articles/PMC4521131/ https://www.cprd.com/sites/default/files/2022-02/Linkage%20Source%20Documentation_set22_1.20.pdf	*2*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*Data augmentation steps impact accuracy (reliability) and extensiveness. We consider here data transformations that produce new information subject to reliability issues: e.g.: imputation of missing values, or extraction of codes via natural language processing.*
		If yes, which are the methods applied	For linkage to the HES datasets, ONS Death, NCRAS, ICNARC and Mental Health data, the trusted third party use an eight-step process to match patients using some or all of the following: NHS number, date of birth, sex and postcode. It is explained in the attached link	https://www.cprd.com/sites/default/files/2022-02/Linkage%20Source%20Documentation_set22_1.20.pdf		*L2 if algorithms are published and their performance documented. Information on which values result from imputation is provided as part of the dataset (e.g., presented in metadata or a data dictionary)*	
		If yes, which algorithms and assumptions applied	It is explained in the attached link	https://www.cprd.com/sites/default/files/2022-02/Linkage%20Source%20Documentation_set22_1.20.pdf			
		If yes, which is the error rate when conducting the augmentation	Requested to DEAP and unable to provide		*N/A*	*L3 if an automatised process for data linkage/mapping exists*	
VIII	Known quality issues and independent QA assessment of the RWD source	Known DQ issues (e.g., poor overall completeness in Q3 2020 due to COVID-19)	A significant proportion of lab data lacking a normal range were missing units or had values inconsistent with units provided. A significant proportion of cases of hyperlipidemia or anemia will be missed ifthe investigator relies solely on diagnosis codes to select patients. Researchers should consider using available treatments, supporting codes, and lab data to supplement diagnosis codes and enhance case capture when studying anemia, diabetes and hyperlidiemia using CPRD. In rpevious articles, CPRD assumed that, for anemia, diabetes or hyperlimidemia, lab and prescription data were less likely than GP entered diagnosis codes to be missing or miscoded, as prescriptions must be entered into the electronic recordto be issued and lab data with a normal range are likely to be elec-tronically transferred from the laboratory. As CPRD has prescription data, it is unknown whether the patient took the prescription.	https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135 https://www.sciencedirect.com/science/article/pii/S2214623720300351?via%3Dihub#s0055	*1*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*Explicit description of known DQ issues, as well as external validation performed (all dimensions affected)*
		Validation studies and publications resulting from this EWD source	Useful publications on the quality of CPRD data for research	https://www.cprd.com/data-quality		*L2 if standard procedures are set for external/internal validation of the data* *L3 if the mechanism provided includes notification of automatically detected DQ issues*	
IX	The RWD source representation	Description of data model or models used (OMOP, FHIR, ...)	OMOP and CONCEPTION	https://catalogues.ema.europa.eu/node/1026/data-flows-and-management	*3*	*L1 if free-text information, links or publications are available reporting all the mentioned features* *L2 if the description refers to a model such as OMOP, I2B2, FHIR, others, or an extension of them. Data dictionaries are standard (and if non-standard, jusified* *L3 if a standard CDM is used, the datasource has been mapped to one or more than one CDM, and if data dictionaries are provided using standard formats that facilitate the mappings across different vocabularies and across languages*	*Descriptive of the intended coherence DQ of a dataset and its metadata.*
		Data ontology (dictionaries and vocabularies) being used, and if in standard formats that allow mapping across different languages (e.g., UMLS)	Medcodeid (unique code for the medical term selected by the GP), Prodcodeid (unique code for the treatment selected by the GP), Read (for diagnoses; from April 2018, Read codes are prospectively mapped to SNOMED CT codes by Vision), Snomed (added to clinical, immunisation, referral and test tables) Read Code (CPRD Gold) SNOMED (CPRD Aurum) Local EMIS® codes and ICD-10 for HES	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf https://www.cprd.com/sites/default/files/2024-08/CPRD%20Aurum%20Data%20Specification%20v3.5.pdf https://pmc.ncbi.nlm.nih.gov/articles/PMC4521131/ https://www.cprd.com/cprd-linked-data#HES%20Accident%20and%20Emergency%20data			
X	The RWD source declared Service Level Agreements (SLA)	Guaranteed frequency of updates and incident response time (e.g., corrections in case of errors)	Monthly		*1*	*L1 if free-text information and links are available reporting all the mentioned features*	*Descriptive of guaranteed timeliness and possible variations of extensivenes s/reliability provided.*

		Processes and resources accompanying the data, such as documentation, training materials or help desk contact	Requested to DEAP and unable to provide		*N/A*	*L2 if details of established data processess by the provider are available*	
		Possibility to collect additional data if needed	Requested to DEAP and unable to provide			*L3 if SLA compliance is assessed and reported automatically*	
XI	The RWD source licensing and restrictions	Data use agreements that may limit data use or access (consent, limitations of use), accessibility policies, licensing constraints, standard policies of use, data retention	Access to CPRD data, including UK Primary Care Data, and linked data such as Hospital Episode Statistics, is subject to protocol approval via CPRD's Research Data Governance (RDG) Process.	https://www.cprd.com/data-access	*2*	*L1 if free-text information and links are available reporting all the mentioned features* *L2 if policies and licensing are standardised to a broad range of RWD* *L3 NA*	*Descriptive of aspects that can limit extensiveness and coherence in downstream data aggregations.*
XII	Feedback	Is there a data ecosystem in place so that quality assessment by data consumers can provide feedback to improve the data collection and production process, thus allowing a continuous monitoring and improvement of DQ?	A general email and address are available	https://www.cprd.com/contact	*1*	*L1 if a person of contact is provided for Q&A* *L2 if the contact provided allows tracking of issues and follow-up* *L3 if the mechanism provided includes notification of automatically detected DQ issues*	*Descriptive of feedback mechanisms in place to improve all aspects of DQ*

Item		Sub-item	Description	Origin of information	Maturity level grade	Maturity level criteria and definitions	Rationale
0	Data base identification	Country	Spain		*N/A*	*N/A*	*N/A*
		Data Access Provider	AEMPS (Spanish Agency for Medicines)			*N/A*	
		Organisation type	EU Institution/Body/Agency Not-for-profit Regulatory Authority			*N/A*	
I	Rationale and scope for the RWD source creation	Primary purpose for which data are collected	To serve as source of information for independent studies on drug safety and support of medicines regulation activities. BIFAP is a non-profit research project funded by the Spanish Agency for Medicines and Medical Devices (AEMPS). Information is collected by PCPs. Information on hospital discharge diagnoses is linked to patients included in BIFAP for a subset of periods and regions participating in the database. All information on prescriptions of medicines by the PCP is incorporated and linked by the PCP to a health problem (episode of care), and information on the dispensation of medicines at pharmacies is extracted from the e-prescription system that is widely implemented in Spain. The project started in 2001 and nine participant autonomous regions send their data to BIFAP every year. BIFAP includes anonymized clinical and prescription/dispensing data. From several regions, hospitalization data can be linked.		*2*	*L1 if information is available as free text and/or online link(s)*	*Relevant for all DQ dimensions (reliability, extensiveness, coherence and timeliness) as it provides a general understandin g of the strengths and limitations of an RWD source.* *Knowing the triggers would ease the understanding of the content and motivations behind the data.*
		Criteria for the selection of the data being collected or integrated	Data recorded by family doctors and primary care paediatricians in Electronic Medical Records (HCE-AP) provided by the Autonomous Communities (CCAA) that voluntarily participate through collaboration agreements.	https://catalogues.ema.europa.eu/node/955/quantitative-descriptors		*L2 if information is available using standardised templates to make information easy to digest and interpret (the EMA recommends to check this tool as reference: REQueST Tool and its vision paper [Internet]. EUnetHTA. 2019. Available from: 721 https://www.eunethta.eu/request-tool-and-its-vision-paper/.* *L3 if the information is provided as Metadata (machine readable), including standrad formats, clear definitions and potentially some quality information*	
		What triggers a record in the database	**Event triggering registration of a person in the data source, other:** Upon registration with a primary care physician within the Spanish NHS (=98,9% of the Spanish population) in the 9 out of the 17 Spanish regions that contribute data. **Event triggering de-registration of a person in the data source:** Death, Emigration **Event triggering creation of a record in the data source:** In every encounter with the general practitioner/paediatrician. Hospital admission and pharmacy dispensation will also trigger the creation of a record.	https://catalogues.ema.europa.eu/node/955/data-flows-and-management			
		Publications describing this RWD	https://pubmed.ncbi.nlm.nih.gov/32337840/ https://www.bifap.org/scientific-publications?lang=en				
II	Data collection or recording process	Description of data provider (geographical and oragnizational setting, nature of the data - reported by patients, HCP, etc)	Agencia Española de Medicamentos y Productos Sanitarios (Spanish Agency for Medicines and Medical Devices, AEMPS). It is the regulatory agency of the Spanish administration that oversees the quality, safety, efficacy and correct information of medicines and medical devices in Spain, as well as cosmetics, from their research to their use, in the interests of the protection and promotion of human health, animal health and the environment. The AEMPS also is the coordinating centre for the Spontaneous Reporting Scheme in Spain and also administer the database supporting this program (FEDRA).	https://catalogues.ema.europa.eu/institution/3331474	*2*	*L1 if information is available as free text and/or online link(s)*	*Essential to understand extensiveness and to assess reliability (that can be affected by errors or biases in the collection process). Also, essential to evaluate SOP for data collection or recording practices that may impact coherence (e.g., where "curation at source" is involved and provide hard constraints for timeliness).*
		Standard Operating Procedures (SOPs) recording	Yes , the document linked describes the flows and the different kinds of processing to which the data is subjected, the management of access to the data, the use of the data by virtue of the user type of the database, and the characteristics of the exploitation of the data.	https://www.bifap.org/data-governance?		*L2 if information is available using standardised templates to make information easy to digest and interpret, and also standard vocabularies are available*	
		How SOPs are implemented and monitored	Data Collection: Data is gathered from various sources, including the Autonomous Communities (CCAA) and AEMPS. Participants: CCAA, AEMPS, Information Systems Technicians. Data Cleansing and Structuring: The collected data is cleaned (errors and inconsistencies are removed) and structured to make it usable and organized. Participants: BIFAP Technicians at AEMPS. Standardized Reporting: A subset of the data is extracted and formatted into standardized reports for easier access and dissemination. Participants: AEMPS Technicians, Technicians from participating Autonomous Communities, Collaborating Physicians. Research Studies: The structured data is extracted for research purposes and used by researchers for analysis, with results being shared and disseminated.Participants: Researchers, BIFAP Technicians at AEMPS. Security Measures: Throughout the process, data is pseudonymized and protected with access controls and obfuscation procedures to ensure privacy and compliance. Participants: AEMPS Technicians. Dissemination: Research results are published in peer-reviewed journals, on the HMA EMA Catalogue and shared via the BIFAP website. Participants: Researchers, BIFAP Technicians at AEMPS.	https://www.bifap.org/data-governance?			

		Key data elements captured (are they always recorded, are they optional, is there a planned coverage over time, ...)	**(HCE-AP) data** **PATIENT DATA:** patient pseudonym unique identifier, gender, date of birth, date of entry into the database, date of deletion from database, cause of deletion (includes patient's death), status in database **VISITS:** date **CONSTRAINTS:** diagnosis with dates, code, and descriptor **HISTORY:** diagnosis with dates, code, and descriptor **GENERAL PATIENT DATA:** includes data such as tobacco, alcohol, blood pressure,body mass index, etc. The date of collection and type are recorded. **EPISODES OR DIAGNOSIS:** descriptor with date, code. **COMMENTS ASSOCIATED WITH THE EPISODE:** date, observations **INTER-VISITS:** dates, medical specialty, motivation, results **VACCINES:** date, code, antigens. **ANALYTICS:** request and result dates, type, determination, value, units, ranges **PRESCRIPTIONS OF PRIMARY HEALTHCARE:** dates, type, drug code, number of containers, dosage. **HCE-AP** **HOSPITAL DISCHARGE DIAGNOSES:** admission reasons recorded and coded by the RAE-CMBD system: includes dates of admission and discharge, type of discharge, primary and secondary diagnoses at hospital discharge. **DATA CONCERNING ELECTRONIC DISPENSATIONS OF MEDICINAL PRODUCTS PRESCRIBED IN PRIMARY HEALTHCARE SETTINGS:** identification of the dispensed drug code, dates, type, number of containers, dosage.	https://www.bifap.org/data-governance?		*L3 if additionally SOPs specify KPIs to monitor*	
III	The selection of RWD sources and their onboarding *(Applies to RWD sources that integrate or repurpose other RWD sources)*	Criteria to accept or exclude a datasource	BIFAP draws on the data provided by the autonomous communities that voluntarily participate through collaboration agreements. These agreements detail the commitments of the Autonomous Community and the Spanish Agency of Medicines and Medical Devices, as well as other operational aspects related to data processing, monitoring commissions, etc. Each autonomous community gathers data in each own database before sending them to BIFAP. Private partners are not included.	https://bifap.aemps.es/autonomous-communities?lang=en	*2*	*L1 if information about selection criteria or DQ performance is available as free text and/or online link(s)*	*When data are provided by a data aggregator, ensure that all the available evidence related to systems and processes potentially affecting DQ (extensiveness and reliability especially) can be followed. Provide information of impact on both reliability and evidence (as well as other dimensions if relative constraints are formulated in inclusion/exclusion (I/E) criteria)*
		Is there a DQ assessment for data sources onboarded?	Each Autonomous Community has the right to find the basic structure and collection of data. When this data is extracted and pseudonymised , it is sent to BIFAP. There this data is cleaned and standardized	https://pubmed.ncbi.nlm.nih.gov/32337840/ https://bifap.aemps.es/autonomous-communities?lang=en		*L2 if a structure checklist and dataset version control are available*	
		If yes: does it follow any specific framework? Is there an assessment checklist? Are	No, it follows an internal structure (BIFAP Common Data Model)	https://pubmed.ncbi.nlm.nih.gov/32337840/		*L3 is only aspirational. NA*	
IV	The data management infrastructure	List of systems used to manage the RWD (either for data collection, recording, processing,	"Comprehensive Study Management (GIE) software, different operating modules are made available to the researcher"	https://www.bifap.org/data-governance?lang=en	*2*	*L1 if information is available as free text and/or online link(s)*	*Essential for reliability regarding data alterations resulting from system accidents, software errors or malicious intervention.*
		Software testing and software quality control in place	Requested to DEAP and unable to provide	https://www.bifap.org/data-governance?lang=en	*N/A*	*L2 if the hardware or software implementation complies with recognised quality standards that can be reported*	
		Measures to prevent accidental physical data alterations (e.g.: backups, redundant systems, checksums)	The transfer of data from the Autonomous Communities to the BIFAP Data Processing Centre (CPD) in the AEMPS is carried out via an SFTP server enabled by the AEMPS for this purpose where each IT manager in the Autonomous Community has a protected folder to store the data. To maintain absolute confidentiality and privacy on the pseudonymised data files to which they have access, which they may not copy or use for any purpose other than the study, nor disclose or assign to anyone outside the research team, even for conservation purposes. This commitment shall be maintained upon completion of the study. As an additional measure, data storage is performed on encrypted hard disks in order to avoid any opening of the files in the event of disk theft.	https://www.bifap.org/data-governance?lang=en	*2*	*L3 N/A*	
V	Data management and governance	Data management principles being followed (e.g., GCP, ISO, FAIR, etc)	Data protection regulations in force in Spain: General Data Protection Regulation, GDPR. Royal Decree LOPDDD	https://www.bifap.org/data-governance?lang=en	*3*	*L1 if information is available as free text and/or online link(s)*	*Data management and governance impact reliability, as well as all quality dimensions for metadata.*

		Data management processes in place (DQ controls, KPIs, SOPs, etc)	Minimum quality for research: - Invalid or not coded gender - Invalid date of birth (before 1800 or after follow-up start date) - Age over 115 years at end of follow-up date - Tracking start date on or after tracking end date - No patient clinical records - HCE-AP with "inactive" record without administrative deletion record. - Existence of data in the HCE-AP with a date prior to the start of follow-up or after the end of follow-up date. - Administrative deletion due to transfer to another primary care quota Identifyind data: ensure effective anonymisation Data on prescribed or dispensed medicinal products: purged and addpated for research use Data related to clinical and diagnostic events: structuring of the diagnostic coding is carried out Free text information is used in BIFAP for better event characterisation, event validation, or event identification that is not properly encoded. EMRs received during the annual extraction procedure are reviewed and the following checks and actions take place: Dates of start of follow-up and end of follow-up are defined for each patient. Only the information dated within this period is used for research. These dates are defined based on quality of clinical and administrative data available in the EMR. If needed, the first registered date of death (from administrative or clinical data) is defined as the end of follow-up. KPI: numeouse publications >75% of electronic prescription dispensation records in 2018 44% of patients with linkage to hospital discharged data in 2018 8.1 million patients in 2018			*L2 if standard best practices are being used and a direct impact on DQ is reported. There are SOPs and dat amanagement processes that adhere to the standards. The representation of metadata follows FAIR standards*	
		Measures to prevent data alterations by unauthorised parties (cybersecurity)	Physical and logical data security measures to prevent re-identification and access by unauthorized third parties are summarised below: a) The BIFAP database is subject to all physical access controls applied by the Ministry of Health under the National Security Scheme b) As an additional measure, data storage is performed on encrypted hard disks in order to avoid any opening of the files in the event of disk theft. c) Access to the computers where the database is accessed requires a password. Only 4 people who are currently part of the BIFAP's IT Unit can access as password users. d) Access to equipment is only possible at AEMPS headquarters. Access to the AEMPS headquarters is recorded at check-in and check-out. e) The computer security measures applied at the software level are subject to the security criteria applied by the Ministry of Health in all its information systems. f) All activity performed by users accessing BIFAP data is monitored by generating Programming Reports: with each output file the application keeps a report on the activity carried out by the user, in order to track and make it possible to trace the queries made in BIFAP with the tools, as well as the delivery of Structured Data Files for statistical	https://www.bifap.org/data-governance?lang=en			
		Auditing and DQ imporvement procedures in place	An DPIA has been carried out using the ASSI-GDPR Tool (Information Systems Security Audit for compliance with the General Data Protection Regulation).	https://www.bifap.org/data-governance?lang=en (Annexes)		*L3 if data management and governance is implemented in the data platforms 'Digital Quality Measures' (DQMs) so that reports of performance and deviations are automated. Submitted metadata are generated "by design". Basically, if everything in L2 is automatised and generated by default*	
VI	Data manipulation steps	Frequency of data updates	Every 1 year	Provided by DEAP	*1*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*Impacts reliability both in terms of accuracy (possible errors) and precision (i.e., the degree of approximatio n by which data represents reality). Essential to ensure traceability of information. Also impacts coherence and potentially timeliness.*
		Data transformations performed, data mapping steps, data cleaning	Requested to DEAP and unable to provide		*N/A*	*L2 if Tests performed follow some standard or shared set of tests, that can be re-used across RWD sources. Key performance indicators (KPIs) for data cleaning (e.g., data duplications, mislabelling, etc.) are provided. Data mapping tables and algorithms are described with a standard characterisation of their performance. Lists of standard test batteries used to detect loss of accuracy or precision are provided. All lineage information is provided as metadata*	
		Information about loss of precision during data manipulation steps	Requested to DEAP and unable to provide			*L3 if information about data onboarding is directly provided by the platform, e.g.: * Transaction logs are available including deviations and actions that required manual intervention Actual data transformation code is accessible and verifiable. Quality checks and KPIs reported are automatically generated by the data platform (e.g.: unit testing) Lineage information is automatically generated by the processing platform*	
		Lineage information (e.g., justifcation of data manipulation, track of changes and versions)	Requested to DEAP and unable to provide				

VII	Data augmentation steps (e.g., imputation or linkage)	Is any augmentation happening in this datasource?	Linked data sources into BIFAP common data model: •EMRs from Primary Care •Diagnosis Tests of Covid-19 (SARS-CoV2 positive test results) during the COVID pandemic •Hospital Diagnosis at in patients discharge in a subset of the participating regions and calendar periods •Medicines Dispensed at Community Pharmacies for the total BIFAP population •Vaccines Covid-19 administered, National Registry, for the total BIFAP population •All vaccines administered, National Registry, for the total BIFAP population •Causes of Death recorded in the national registry •Hospital Pharmacies dispensing Data in a subset of the participating regions	https://catalogues.ema.europa.eu/node/955/data-flows-and-management	2	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*Data augmentation steps impact accuracy (reliability) and extensiveness. We consider here data transformations that produce new information subject to reliability issues: e.g.: imputation of missing values, or extraction of codes via natural language processing.*
		If yes, which are the methods applied	Linkagae strategy: deterministic			*L2 if algorithms are published and their performance documented. Information on which values result from imputation is provided as part of the dataset (e.g., presented in metadata or a data dictionary)*	
		If yes, which algorithms and assumptions applied	None	Provided by DEAP			
		If yes, which is the error rate when conducting the augmentation	Unknown	Provided by DEAP		*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)	Not aware of any issue	Provided by DEAP	2	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*Explicit description of known DQ issues, as well as external validation performed (all dimensions affected)*
		Validation studies and publications resulting from this RWD source	https://www.bifap.org/scientific-publications?lang=en Validation of digestive cancer: https://www.mdpi.com/2077-0383/13/2/361 COVID-19 validation: https://pmc.ncbi.nlm.nih.gov/articles/PMC11102056/			*L2 if standard procedures are set for external/internal validation of the data* *L3 if the mechanism provided includes notification of automatically detected DQ issues*	
IX	The RWD source representation	Description of data model or models used (OMOP, FHIR, ...)	OMOP, CONCEPTION, BIFAP DATA MODEL	https://catalogues.ema.europa.eu/node/955/data-elements-collected	3	*L1 if free-text information, links or publications are available reporting all the mentioned features* *L2 if the description refers to a model such as OMOP, I2B2, FHIR, others, or an extension of them. Data dictionaries are standard (and if non-standard, jusified why)* *L3 if a standard CDM is used, the datasource has been mapped to one or more than one CDM, and if data dictionaries are provided using standard formats that facilitate the mappings across different vocabularies and across languages*	*Descriptive of the intended coherence DQ of a dataset and its metadata.*
		Data ontology (dictionaries and vocabularies) being used, and if in standard formats that allow mapping across different languages (e.g., UMLS)	Cause of death: ICD-10-CM / ICD-9-CM Indication: SNOMED CT Procedures: ICD-10-CM / ICD-9-CM / SNOMED CT Diagnosis / medical event: ICD-10-CM / ICD-9-CM / SNOMED CT / Not coded (Free text) Prescriptions of medicines: ATC / SNOMED Dispensing vocabulary: ATC / SNOMED	https://catalogues.ema.europa.eu/node/955/data-elements-collected ICPC: https://pubmed.ncbi.nlm.nih.gov/32337840/			
X	The RWD source declared Service Level Agreements (SLA)	Guaranteed frequency of updates and incident response time (e.g.	Requested to DEAP and unable to provide		N/A	*L1 if free-text information and links are available reporting all the mentioned features*	*Descriptive of guaranteed timeliness and possible variations of extensivenes s/reliability provided.*
		Processes and resources accompanying the data, such as documentation, training materials or help desk contact	For researchers registered on the BIFAP website, there are training courses online (Introduction to BIFAP data; Medication data in BIFAP; Diagnosis data in BIFAP; Creation of variables in BIFAP) before using BIFAP data, as well as aggregated data and a list of available standard variables to use in research studies. Governance and auditing documents are public, and an automatic online process is in place to submit studies for evaluation by the Scientific Commitee. Help desk contact available.	Provided by DEAP	2	*L2 if details of established data processess by the provider are available*	
		Possibility to collect additional data if needed	Possible to request reextractions, but not found if collection of new data. Additional data from linked sources are under evaluation (Hospital Pharmacies dispensing Data; Specialist Prescriptions; Link mother-child; etc).	Provided by DEAP		*L3 if SLA compliance is assessed and reported automatically*	
XI	The RWD source licensing and restrictions	Data use agreements that may limit data use or access (consent, limitations of use), accessibility policies, licensing constraints, standard policies of use, data retention	AEMPS Administrator All other functions of the AEMPS Autonomous Communities: Members of the Advisory Committee and staff authorised by each member of the BIFAP Advisory Committee in the collaborating Autonomous Communities. Collaborating physicians Registered Researchers: Healthcare professionals or health research professionals	https://www.bifap.org/learn-more?lang=en	2	*L1 if free-text information and links are available reporting all the mentioned features* *L2 if policies and licensing are standardised to a broad range of RWD* *L3 N/A*	*Descriptive of aspects that can limit extensiveness and coherence in downstream data aggregations.*
XII	Feedback	Is there a data ecosystem in place so that quality assessment by data consumers can provide feedback to improve the data collection and production	A general email and phone number are available. Review of clinical histories for case confirmation must be conducted at the AEMPS offices. Consequently, feedback from researchers/data consumers who validate recorded outcomes is frequent facilitated through personal communication and collaboration.	https://www.bifap.org/ Provided by DEAP	2	*L1 if a person of contact is provided for Q&A* *L2 if the contact provided allows tracking of issues and follow-up* *L3 if the mechanism provided includes notification of automatically detected DQ issues*	*Descriptive of feedback mechanisms in place to improve all aspects of DQ*

Dimension	Sub-dimension	Metrics	Description	Origin of information
Timeliness	Currency	How often is the database updated (i.e., frequency of updates)	Once a year	https://www.bifap.org/data-governance?lang=en
		The time gap between the latest available data and date when data is delivered to user. (i.e., how up-to-date data are when it reach the user)	At least 6 months plus the lag of delivery	Provided by DEAP
		The time elapsed from when a user requests the data to when they actually receive it	5 to 6 months (including 1-2 months for Institutional Rewvie Board's approval)	Provided by DEAP
		Median time (years) between first and last available records for unique individuals	10 years	https://catalogues.ema.europa.eu/node/955/quantitative-descriptors
Extensiveness	Coverage	Percentage of a target population present in a database	98.9% of the Spanish population is registered in the Spanish NHS 36% of the total Spanish population in the NHS (46.6 Million) 95% of the population of the nine included regions Up to 2018, 57.6% active patients of the participating regions (14 million)	https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.5006 AEMPS Internal statistics
	Completness	% of subjects in the data with a recorded birth date	100%	https://es.slideshare.net/slideshow/protecto-bifap/13799690
		% of subjects in the data, irrespective of vital status, that have a recorded date of death	Among the patients with a recorded administrative death in BIFAP, 33.6% had a death date that matched the National Death Registry, and 84.8% were recorded within the same 30-day period.	https://zenodo.org/records/13384860 Validation against Death National Registry among people aged 35-80 years, partially provided in Abstract #1375. Volume33, IssueS2. Supplement: Abstracts of ISPEs 2024, 40th international conference, 24–28 August 2024, Germany. November 2024. e5891;
		% of subjects in the data with a record of sex	100%	https://zenodo.org/records/13384860
		% of subjects in the data who had an event with a code for the event	100%	https://zenodo.org/records/13384860
		% of subjects in the data who had a prescription/dispensing with a recorded code for the medicine	ATC code (100%), MPID (100%)	https://zenodo.org/records/13384860
		% of subjects in the data who got vaccinated with a recorded code for the vaccine	A register of vaccination with a code for the vaccine is recorded for 100% of individuals who are known to have been vaccinated	https://zenodo.org/records/13384860 ADVANCE database characterisation and fit for purpose assessment for multi-country studies on the coverage, benefits and risks of pertussis vaccinations. Sturkenboom M et al. Vaccine. 2020 Dec 22;38 Suppl 2:B8-B21. doi: 10.1016/j.vaccine.2020.01.100. Epub 2020 Feb 12. PMID: 32061385. Age-specific vaccination coverage estimates for influenza, human papillomavirus and measles containing vaccines from seven population-based healthcare databases from four EU countries - The ADVANCE project. Braeye T et al. Vaccine. 2020 Apr 3;38(16):3243-3254. doi: 10.1016/j.vaccine.2020.02.082. Epub 2020 Mar 12. PMID: 32171573
Reliability	Accuracy	The population distribution in the data source aligns with that of the country	Up to the end of 2018, BIFAP includes data from 7566 PCPs (6419 general practitioners and 1147 pediatricians). This yields a total of 12 million (2.3 million pediatric) patients for studies. Out of them, 8 million contains up-to-date information the year 2018 ("active" patients), representing 57.6% of the participating regions (14 million) and 17% of the total Spanish population (46.6 million). Now, all these counts have increased. Population age distribution are aligned with a developed country demographics reported by the National Statistics Institute (INE), although the yough adults are slightly less represented in BIFAP than country population due to the less frequent health seeking behaviour. Active population:□ Paediatric Population (< 18 years):2664591 (11.2%) Neonate: 81731 (0.3%) Infants and toddlers (28 days – 23 months):116259 (0.5%) Children (2 to < 12 years): 1410787 (5.9%) Adolescents (12 to < 18 years):1055814 (4.5%) Adults (18 to < 46 years):5930964 (25.0%) Adults (46 to < 65 years):5089694 (21.4%) Elderly (≥ 65 years):3698925 (15.6%) Adults (65 to < 75 years):1784108 (7.5%) Adults (75 to < 85 years):1215680 (5.1%) Adults (85 years and over): 699137 (2.9%)	https://catalogues.ema.europa.eu/node/955/quantitative-descriptors Maciá-Martínez MA, Gil M, Huerta C, Martín-Merino E, Álvarez A, Bryant V, Montero D; BIFAP Team. Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria (BIFAP): A data resource for pharmacoepidemiology in Spain. Pharmacoepidemiol Drug Saf. 2020 Oct;29(10):1236-1245. doi: 10.1002/pds.5006. Epub 2020 Apr 26. PMID: 32337840. https://www.ine.es/
		Records of diagnostics, exposures or medical observations that do not agree with common expectations and knowledge or feasible ranges (e.g., pregnancy records in males, a human with 4 arms, systolic pressure higher than 250mmHg, etc)	Internal quality checks include criteria to identify and eliminate implausible or clearly erroneous data.	Provided by DEAP
		Records of healthcare events (diagnoses, prescriptions, admissions, etc) with logical inconsistencies (e.g., and admission occurs after death)	Date values outside observation periods: 0-8.7% Date values before birth: 0% Date values after death: 0%	https://zenodo.org/records/13384860
		Variables that are based in imputation, derivation or inference (e.g., end of treatment date is derived from treatment start date and treatment cycle length)	Pregnancy	https://pubmed.ncbi.nlm.nih.gov/31749191/
	Precision	Exposures codes precision level, including medicines and vaccines (e.g., active principle, therapeutic group, ...)	Active principle (ATC level 5 codes)	https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2023.1207976/full https://pmc.ncbi.nlm.nih.gov/articles/PMC5993167/
		Precision of date of birth (e.g., day, month, year)	Day, month and year (in house), but for researching pourpose only the year of birth or death are obtained and the month-year of birth is considered for research on babies/children. Date of birth can serve as the source of indirect re-identification.	https://zenodo.org/records/13384860
		Precision of date of death (e.g., day, month, year)	day, month and year. For researching pourpose only the year of death are obtained. Date of dearth can serve as the source of indirect re-identification.	https://www.bifap.org/data-governance?lang=en
		Precision of date of the event/diagnosis (e.g., day, month, year)	Day, month, year	Provided by DEAP
		Precision of date of the exposure (e.g., day, month, year)	Day, month, year	Provided by DEAP
	Traceability	Provenance of event records	Primary care medical records/Hospital discharge records	https://catalogues.ema.europa.eu/node/955/administrative-details

		Provenance of medicines/vaccines records	Medicines record from Pharmacy dispensing records and Vaccination records from vaccines administrations in every public healthcare setting (for most of participating regions)	https://catalogues.ema.europa.eu/node/955/administrative-details
Coherence	Format coherence	For dates, formatting constraint being followed	Character, length 8: yyyymmdd	Provided by DEAP
		For sex, formatting constraint being followed	M (male), F (female)	Provided by DEAP
	Relational coherence	% of records with the Person ID in the PERSONS table	100%	https://zenodo.org/records/13384860
	Semantic coherence - to determine whether the database uses a standardised dictionary	For EVENTS definitions, codelists/data dictionaries being employed according to external standards	ICD-10-CM ICD-9-CM Not coded (Free text) SNOMED CT	https://catalogues.ema.europa.eu/node/955/data-elements-collected
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	ATC SNOMED National drug code	https://catalogues.ema.europa.eu/node/955/data-elements-collected
	Uniqueness	Number of records flagged as potential duplicates	Requested to DEAP and unable to provide	

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

Coherence	Format coherence	For dates, formatting constraint being followed	Date of birth: MM/YY Other dates: DD/MM/YYYY (Death, events/diagnosis/exposure) Character, length 5 or 10	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
		For sex, formatting constraint being followed	Mapping: Lookup SEX Type: INTEGER, Format:1, 1 (male) 2 (female) 3 (indeterminate) 4 (unknown)	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
	Relational coherence	% of records with the Person ID in the PERSONS table	98.2-100%	https://zenodo.org/records/13384860
	Semantic coherence - to determine whether the database uses a standardised dictionary	For EVENTS definitions, codelists/data dictionaries being employed according to external standards	Read Code (CPRD Gold): these are used for diagnoses; from April 2018, Read codes are prospectively mapped to SNOMED CT codes SNOMED (CPRD Aurum) Local EMIS® codesICD-10 for HES Medcodeid (unique code for the medical term selected by the GP)	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf https://www.cprd.com/sites/default/files/2024-08/CPRD%20Aurum%20Data%20Specification%20v3.5.pdf
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	Prodcodeid (unique code for the treatment selected by the GP), SNOMED for some immunisations No ATC codes available in the raw data but ATC for active substances link is available at the Utrecht University	https://zenodo.org/records/13384860
	Uniqueness	Number of records flagged as potential duplicates	Requested to DEAP and unable to provide	

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

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

Confounders	Age at index	Date of birth (month/years)	*Low*	100% have date of birth available				Rodríguez-Miguel A, Fernández-Fernández B, Ortiz A, Gil M, Rodríguez-Martín S, Ruiz-Hurtado G, Fernández-Antón E, Ruilope LM, de Abajo FJ. Glucose-Lowering Drugs and Primary Prevention of Chronic Kidney Disease in Type 2 Diabetes Patients: A Real-World Primary Care Study. Pharmaceuticals (Basel). 2024 Sep 29;17(10):1299. doi: 10.3390/ph17101299. PMID: 39458940; PMCID: PMC11510410. Mar Martín-Pérez et al. Multiple vertebral fractures after antiosteoporotic medications discontinuation: A comparative study to evaluate the potential rebound effect of denosumab PMID: 39521365 DOI: 10.1016/j.bone.2024.117325 Quesada JA, Orozco-Beltran D. Analysis of missing data in electronic health records of people with diabetes in primary care in Spain: A population-based cohort study. Int J Med Inform. 2025 Feb;194:105722. doi: 10.1016/j.ijmedinf.2024.105722. Epub 2024 Nov 23. PMID: 39586146. Martín-Merino E et al. Cessation rate of anti-osteoporosis treatments and risk factors in Spanish primary care settings: a population-based cohort analysis Arch Osteoporos (2017) 12:1-12. DOI 10.1007/s11657-017-0331-6
	Gender: female or male	Sex	*Low*	100% have date of birth, sex, diagnostic codes and drug codes available.				
	Frailty	Diagnostic code (ICD-10 or equivalent)	*Low*		In a published study in specific regions of Spain, 26.2% of 855 were found to be frail by using Fried criteria. Other studies report ~10%. Institucionalized patients went up to 68.8%. This can be taken as reference to benchmark to.			Rivas-Ruiz F, Machón M, Contreras-Fernández E, Vrotsou K, Padilla-Ruiz M, Díez Ruiz AI, de Mesa Berenguer Y, Vergara I; Group GIFEA. Prevalence of frailty among community-dwelling elderly persons in Spain and factors associated with it. Eur J Gen Pract. 2019 Oct;25(4):190-196. doi: 10.1080/13814788.2019.1635113. Epub 2019 Oct 22. PMID: 31637940; PMCID: PMC6853242. Jürschik P, Nunin C, Botigué T, Escobar MA, Lavedán A, Viladrosa M. Prevalence of frailty and factors associated with frailty in the elderly population of Lleida, Spain: the FRALLE survey. Arch Gerontol Geriatr. 2012 Nov-Dec;55(3):625-31. doi: 10.1016/j.archger.2012.07.002. Epub 2012 Jul 31. PMID: 22857807. González-Vaca J, de la Rica-Escuín M, Silva-Iglesias M, Arjonilla-García MD, Varela-Pérez R, Oliver-Carbonell JL, Abizanda P. Frailty in INstitutionalized older adults from ALbacete. The FINAL Study: rationale, design, methodology, prevalence and attributes. Maturitas. 2014 Jan;77(1):78-84. doi: 10.1016/j.maturitas.2013.10.005. Epub 2013 Oct 16. PMID: 24189222.
	Obesity: defined as a separate diagnosis and/or BMI greater than or equal to 30.	BMI or weight and height Diagnostic code (ICD-10 or equivalent)	*Low*	In a published study on 515,701 T2D patients 312,383 were fount to be obese by using BMI > or = 30kg/m^2. Another study reported 35.4% missingness of BMI in diabetic patients.				Rodríguez-Miguel A, Fernández-Fernández B, Ortiz A, Gil M, Rodríguez-Martín S, Ruiz-Hurtado G, Fernández-Antón E, Ruilope LM, de Abajo FJ. Glucose-Lowering Drugs and Primary Prevention of Chronic Kidney Disease in Type 2 Diabetes Patients: A Real-World Primary Care Study. Pharmaceuticals (Basel). 2024 Sep 29;17(10):1299. doi: 10.3390/ph17101299. PMID: 39458940; PMCID: PMC11510410. Quesada JA, Orozco-Beltran D. Analysis of missing data in electronic health records of people with diabetes in primary care in Spain: A population-based cohort study. Int J Med Inform. 2025 Feb;194:105722. doi: 10.1016/j.ijmedinf.2024.105722. Epub 2024 Nov 23. PMID: 39586146.
	Heart transplant	Diagnostic code (ICD-10 or equivalent) Procedure code	*Low*					
	Microvascular complications: mono-/polyneuropathy, eye complications, Diabetic foot/Peripheral angiopathy, nephropathy, Diabetes with several-/unspecified complications	Diagnostic code (ICD-10 or equivalent)	*Low*					
	Severe hypoglycemia	Diagnostic code (ICD-10 or equivalent) Laboratory values	*Low*	In previous studies on patients with diabetes in BIFAP, The proportion of individuals with at least one missing value was 76.0%. Regarding diabetes control measures, 10.8% of records had missing glycated hemoglobin values, and 21.4% had missing basal blood glucose values.				Quesada JA, Orozco-Beltran D. Analysis of missing data in electronic health records of people with diabetes in primary care in Spain: A population-based cohort study. Int J Med Inform. 2025 Feb;194:105722. doi: 10.1016/j.ijmedinf.2024.105722. Epub 2024 Nov 23. PMID: 39586146.
	Keto-/lactate acidosis	Diagnostic code (ICD-10 or equivalent) Laboratory values	*Low*					
	Lower limb amputations	Diagnostic code (ICD-10 or equivalent) or procedure	*Low*					
	Chronic obstructive pulmonary disease (COPD)	Diagnostic code (ICD-10 or equivalent) Medication code and date (as proxies)	*Low*	100% medication codes available.	COPD frenquency and comparison between diabetic and non-diabetic patients has beed assessed and reported in a paper referred in column 'I'			Arias Fernández, L., Pardo Seco, J., Cebey-López, M. et al. Differences between diabetic and non-diabetic patients with community-acquired pneumonia in primary care in Spain. BMC Infect Dis 19, 973 (2019). https://doi.org/10.1186/s12879-019-4534-x

Cancer	Diagnostic code (ICD-10 or equivalent) Medication code and date (as proxies)	*Low*	A previous study reported 0 missings on cancer recording in diabetic patients. 100% medication codes available.	Validation studies have been performed on validation of colorectal cancer diagnosis in BIFAP, so these are deemed to eb reliable (PPV>92% and NPV 100%). Also, an algorithm to detect digestive cancer has been developed and validated.			Quesada JA, Orozco-Beltran D. Analysis of missing data in electronic health records of people with diabetes in primary care in Spain: A population-based cohort study. Int J Med Inform. 2025 Feb;194:105722. doi: 10.1016/j.ijmedinf.2024.105722. Epub 2024 Nov 23. PMID: 39586146. Gil M, Rodríguez-Miguel A, Montoya-Catalá H, González-González R, Álvarez-Gutiérrez A, Rodríguez-Martín S, García-Rodríguez LA, de Abajo FJ. Validation study of colorectal cancer diagnosis in the Spanish primary care database, BIFAP. Pharmacoepidemiol Drug Saf. 2019 Feb;28(2):209-216. doi: 10.1002/pds.4686. Epub 2018 Dec 12. PMID: 30548462. Fernández-Antón E, Rodríguez-Miguel A, Gil M, Castellano-López A, de Abajo FJ. Development and Validation of Case-Finding Algorithms for Digestive Cancer in the Spanish Healthcare Database BIFAP. J Clin Med. 2024 Jan 9;13(2):361. doi: 10.3390/jcm13020361. PMID: 38256495; PMCID: PMC10816118.
Major organ specific bleeding	Diagnostic code (ICD-10 or equivalent) Procedure code	*Low*					
Bariatric surgery	Diagnostic code (ICD-10 or equivalent) or procedure	*Low*					
Chronic kidney disease (CKD) stages 1-4	Diagnostic code (ICD-10 or equivalent) Laboratory values	*Low*	Cases whose first abnormal record was albuminuria/proteinuria or a diagnosis of CKD and without a record of eGFR could not be classified. Neither albuminuria nor proteinuria were exhaustively recorded, which may lead to a certain under-recording of CKD in its early stages. Imaging or hisological findings are not recorded.	In previous studies using BIFAP, they classified CKD using the KDIGO criteria (except imaging or histological findings). Also, they defined as "chronic kidney insufficiency" cases of CKD at stages from G3a to G5. The following information was concluded from a previous study using BIFAP, which may help to assess the plausability of our results: "Regarding the epidemiology of CKD in primary care and among patients with type 2 diabetes, we found the following: (1) over the study period, the incidence rate of CKD was stable overall; (2) in patients older than 70 years, the incidence rate was higher in females than in males; and (3) the factors more strongly associated with incident CKD were the antecedents of gout, hyperuricemia, hyperkalemia, hypertension, heart failure, hyperparathyroidism, and prior isolated abnormal values of eGFR or proteinuria/albuminuria." In T2D patients, levels of serum creatinine or eGFR were recorded in the database for each subject every year, on average.	Around 10% of cases were detected by albuminuria/proteinuria or a recorded diagnosis of CKD but with no data on eGFR or creatinine at the index date. Thus, these two variables seem not to be concordantly recorded.	Levels of serum creatinine or eGFR were recorded in the database for each subject every year, on average	Rodríguez-Miguel A, Fernández-Fernández B, Ortiz A, Gil M, Rodríguez-Martín S, Ruiz-Hurtado G, Fernández-Antón E, Ruilope LM, de Abajo FJ. Glucose-Lowering Drugs and Primary Prevention of Chronic Kidney Disease in Type 2 Diabetes Patients: A Real-World Primary Care Study. Pharmaceuticals (Basel). 2024 Sep 29;17(10):1299. doi: 10.3390/ph17101299. PMID: 39458940; PMCID: PMC11510410.
End stage kidney disease (CKD stage 5)	Diagnostic code (ICD-10 or equivalent) Laboratory values Procedure codes (i.e., dialysis)	*Low*	Cases whose first abnormal record was albuminuria/proteinuria or a diagnosis of CKD and without a record of eGFR could not be classified. Neither albuminuria nor proteinuria were exhaustively recorded, which may lead to a certain under-recording of CKD in its early stages. Imaging or hisological findings are not recorded.	In previous studies using BIFAP, they classified CKD using the KDIGO criteria (except imaging or histological findings). Also, they defined as "chronic kidney insufficiency" cases of CKD at stages from G3a to G5. The following information was concluded from a previous study using BIFAP, which may help to assess the plausability of our results: "Regarding the epidemiology of CKD in primary care and among patients with type 2 diabetes, we found the following: (1) over the study period, the incidence rate of CKD was stable overall; (2) in patients older than 70 years, the incidence rate was higher in females than in males; and (3) the factors more strongly associated with incident CKD were the antecedents of gout, hyperuricemia, hyperkalemia, hypertension, heart failure, hyperparathyroidism, and prior isolated abnormal values of eGFR or proteinuria/albuminuria." In T2D patients, levels of serum creatinine or eGFR were recorded in the database for each subject every year, on average.	Around 10% of cases were detected by albuminuria/proteinuria or a recorded diagnosis of CKD but with no data on eGFR or creatinine at the index date. Thus, these two variables seem not to be concordantly recorded.	Levels of serum creatinine or eGFR were recorded in the database for each subject every year, on average	Rodríguez-Miguel A, Fernández-Fernández B, Ortiz A, Gil M, Rodríguez-Martín S, Ruiz-Hurtado G, Fernández-Antón E, Ruilope LM, de Abajo FJ. Glucose-Lowering Drugs and Primary Prevention of Chronic Kidney Disease in Type 2 Diabetes Patients: A Real-World Primary Care Study. Pharmaceuticals (Basel). 2024 Sep 29;17(10):1299. doi: 10.3390/ph17101299. PMID: 39458940; PMCID: PMC11510410.

	All separate GLD (glucose-lowering drugs): biguanides (metformin), sulfonylurea, sulfonamides, alfa glucoside inhibitors, thiazolidinediones, other blood glucose lowering drugs, insulin	Medication codes	*Low*	100% have date of birth, sex, diagnostic codes and drug codes available.				
	Drugs to prevent CVD: I) Antihypertensives (Angiotensin-converting-enzyme inhibitors, Angiotensin receptor blockers, Beta-blockers, Low-/high ceiling diuretics, Aldosterone antagonists, Thiazide diuretics); II) Ca channel blockers; III) Digitoxin/digoxin, IV) Antiarrhythmics (flekanide, amiodarone); V) Statins; VI) Anticoagulants (Warfarin); and VII) Antiplatelet agents (Low dose acetylsalicylic acid, Receptor P2Y12 antagonists, Other antiplatelets)	Medication codes	*Low*	100% have date of birth, sex, diagnostic codes and drug codes available.				
	Corticosteroids	Medication codes	*Low*					
	Weigh loss drug	Medication codes	*Low*					
Intercurrent events	Treatment discontinuation	Date of drug discontinuation	*High*		In BIFAP, the information regarding the date of dispensation, the number of dispensed packages and the number of doses per package is recorded. However, duration of each dispensed package might be only calculated if the doctor wrote prescription instructions (i.e. posology). If it is not available, there is a multistep algorithm to derive the duration of the package dispensed, based on the date of dispensation at the pharmacy, the distance between subsequent prescriptions, the mode or median of prescription duration in the data set, or imputing 30 days. Consequently, treatment discontinuation can be determined whenever algorithms are based on these data points.			
	Treatment switch	Date of drug discontinuation Date of drug start	*High*		In BIFAP, the information regarding the date of dispensation, the number of dispensed packages and the number of doses per package is recorded. However, duration of each dispensed package might be only calculated if the doctor wrote prescription instructions (i.e. posology). If it is not available, there is a multistep algorithm to derive the duration of the package dispensed, based on the date of dispensation at the pharmacy, the distance between subsequent prescriptions, the mode or median of prescription duration in the data set, or imputing 30 days. Consequently, treatment discontinuation can be determined whenever algorithms are based on these data points.			
	Addition of another antihyperglycemic therapy	Medication codes	*High*	100% have date of birth, sex, diagnostic codes and drug codes available.				
	Non-CV death	Diagnostic code Date of death	*High*		Depending on the instance utilized, BIFAP may or may not have the cause of death data linked (i.e., the National Registry of Mortality). In instances where such data is linked, the cause of death will be recorded using diagnostic codes (ICD-10 or equivalent).Mortality data is updated with a one-year delay relative to the present time. For research purposes only the year of death is available. This can impact precision and the time sequence of outcomes.			

Follow-up time needed per patient in the study	5 years	6 years (including recruitment and follow-up)	*High*	The median time between first and last available records for any individual is 10 years. For active individuals, 12 years. Thus, presumably 6 years of follow-up time will be available.			The median time between first and last available records for any individual is 10 years. For active individuals, 12 years. Seems the needed timeliness will be met.	https://cataloques.ema.europa.eu/node/955/quantitative-descriptors
Minimum time in the data source for lookback assessment	1 year	1 year of lookback	*High*				Data is updated once a year so we should be able to have this information for the year prior randomisation. However, the delay to data delivery should be accounted for. The median lenght of follow-up per patient is 10 years.	https://cataloques.ema.europa.eu/node/955/quantitative-descriptors
Others	Rescue medications: acute use of insulin or sulfonylureas	Medication codes Treatment duration	*Low*	In BIFAP, the information regarding the date of dispensation, the number of dispensed packages and the number of doses per package is recorded. However, duration of each dispensed package might be only calculated if the doctor wrote prescription instructions (i.e. posology). If it is not available, there is a multistep algorithm to derive it, to calculate the duration of the package dispensed, based on the date of dispensation at the pharmacy, the distance between subsequent prescriptions, the mode or median of prescription duration in the data set, or imputing 30 days.	Depending on what is available, the accuracy of the duration measurement can be impacted.			

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

Scientific research question		**Dapagliflozin and Major Adverse Cardiovascular Events in Type 2 Diabetes**						
Design elements	**Operationalization of definitions**	**Data elements for valid capture of variables**	**Criticality of the quality of the element**	**Extensiveness assessment (if applicable)**	**Reliability assessment (if applicable)**	**Coherence assessment (if applicable)**	**Timeliness assessment (if applicable)**	**Origin of information**
Study population	**Inclusion criteria**							
	Adult >40 years of age at the time of eligibility screening	Date of birth (years)	*High*	100% have date of birth	Only year is available, this may impact precision.		As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years	https://cataloques.ema.europa.eu/node/1026/quantitative-descriptors
	Diagnosis of type diabetes 2	Diagnostic code	*High*	Diagnostic codes available for 100% of patients	A diagnosis code of type 2 diabetes is likely to be correct where present (correctness 99%). From DEAP experience, diabetes mellitus without type specification occurs frequently as well; usually insulin in monotherapy is used to assess T1D and NIADS for, T2D.	Nearly all patients who had elevated HbA1c labs or hypoglycemic treatments also had a type 2 diabetes diagnosis code (concordance >90%). In CPRD seems lab values and prescriptions are less likely to be missing than diagnoses.		CPRD Aurum database: Assessment of data quality andcompleteness of three important comorbidities, including diabetes: https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135 *"Among 37 502 patients in CPRD Aurum, correctness of type 2 diabetes,hyperlipidemia, and anemia diagnoses was high (99%, 93%, and 97%, respectively).Completeness was only high for type 2 diabetes (94%-98%);"*
	Established ACVD as a history or diagnosis of ischemic heart disease, ischemic cerebrovascular disease or peripheral arterial disease during the 1 year eligibility	Diagnostic code	*High*	Diagnostic codes available for 100% of patients			As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years	https://cataloques.ema.europa.eu/node/1026/quantitative-descriptors
	High ACVD risk defined as no established ACVD, age ≥ 55 years in men and ≥ 60 in women and one orf more of the following: - History or diagnosis of dyslipidemia - Current lipid lowering therapy - History or diagnosis of hypertension - Current anti-hypertensive medication use prescribed for blood pressure lowering - Current tobacco use or within 1 year prior to randomisation	Sex Date of birth Diagnostic code (ICD-10 or equivalent) Smoking habits Prescription/dispensing data Indication linked to drug use	*High*	Smoking present for 89.7% of records. Only prescription medicines (100%) Not especific linked indication to drug use; but codes Dx are used as a aproximation			As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years	https://cataloques.ema.europa.eu/node/1026/quantitative-descriptors
	Exclusion criteria							
	Treatment with SGLT2i or DPP-4i in the last year prior randomisation	Prescription/dispensing data	High	Only prescription medicines (100%)		As ATC codes are not available, a mapping to ATC will potentially be needed to extract study drugs information.	As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years	https://cataloques.ema.europa.eu/node/1026/quantitative-descriptors
	Treatment with pioglitazone or rosiglitazone treatment in the last year prior randomisation	Prescription/dispensing data	*High*	Only prescription medicines (100%)			As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years	https://cataloques.ema.europa.eu/node/1026/quantitative-descriptors
	Acute cardiovascular event in the last year prior randomisation	Diagnostic code (ICD-10 or equivalent) Emergency room and/or hospitalisation diagnoses	*High*	Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects attending emergency room			As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years	Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC) https://academic.oup.com/ije/article/46/4/1093/3072145?login=true
	Diagnosis Type 1 diabetes any time before randomisation	Diagnostic code (ICD-10 or equivalent)	*High*	Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects attending emergency room	From DEAP experience, diabetes mellitus without type specification occurs frequently as well; usually insulin in monotherapy is used to assess T1D and NIADS for, T2D.		As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years	https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135

Treatment/exposure	Dapagliflozin	Medication codes	*High*	100% of individuals have available information		Nearly all patients who had elevated HbA1c labs or hypoglycemic treatments also had a type 2 diabetes diagnosis code. As ATC codes are not available, a mapping to ATC will potentially be needed to extract study drugs information. The DEAP informed they already have the mapping ready, so this should not be a problem.		https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135
Comparator group (if applicable)	DPP4i	Medication codes	*High*	100% of individuals have available information		As ATC codes are not available, a mappint to ATC will potentially be needed to extract study drugs information. The DEAP informed they already have the mapping ready, so this should not be a problem.		
Key endpoint(s)	Time to first MACE (non-fatal MI, stroke, all-cause death)	Date of death Date of diagnosis	*High*	8% of the whole population (irrespective of vital status) has a date of death recorded; 100% of death people have a date of death Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects in attending emergency room				
Confounders	Age at index	Date of birth (month/years)	*Low*	100% have date of birth	Only year is available, this may impact precision.		As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years	https://cataloques.ema.europa.eu/node/1026/quantitative-descriptors
	Gender: female or male	Sex	*Low*	>99%. Sex categories in CPRD include unknown and indeterminate sex, but are never included in data extractions (<1% of records without sex information are excluded); they are extremely rare.				
	Frailty	Diagnostic code (ICD-10 or equivalent)	*Low*	In previous studies the Charlson comorbidity index has been used, as well as dementia. In CPRD also eFI index is available (see link), among others.				https://www.sciencedirect.com/science/article/pii/S2214623720300351?via%3Dihub#s0055 https://www.cprd.com/approved-studies/exploring-role-electronic-frailty-index-efi-using-routine-primary-care-electronic
	Obesity: defined as a separate diagnosis and/or BMI greater than or equal to 30.	BMI or weight and height Diagnostic code (ICD-10 or equivalent)	*Low*	In previous studies it seems obesity has been defined by BMI and also as diagnose. Both strategies might be considered to capture obesity.		A significant proportion of cases of hyperlipidemia will be missed if the investigator relies solely on diagnosis codes to select patients.		https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135 https://www.sciencedirect.com/science/article/pii/S2214623720300351?via%3Dihub#s0055
	Heart transplant	Diagnostic code (ICD-10 or equivalent) Procedure code	*Low*	Diagnostic codes available for 100% of patients				
	Microvascular complications: mono-/polyneuropathy, eye complications, Diabetic foot/Peripheral angiopathy, nephropathy, Diabetes with several-/unspecified complications	Diagnostic code (ICD-10 or equivalent)	*Low*	Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects attending emergency room				
	Severe hypoglycemia	Diagnostic code (ICD-10 or equivalent) Laboratory values	*Low*	Diagnostic codes available for 100% of patients Diagnostic codes are available for 86% subjects attending emergency room	A significant proportion of lab data lacking a normal range were missing units.	A significant proportion of lab data lacking a normal range had values inconsistent with units provided.		https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135
	Keto-/lactate acidosis	Diagnostic code (ICD-10 or equivalent) Laboratory values	*Low*	DX codes 100% Laboratory values	A significant proportion of lab data lacking a normal range were missing units.	A significant proportion of lab data lacking a normal range had values inconsistent with units provided.		https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135
	Lower limb amputations	Diagnostic code (ICD-10 or equivalent) or procedure code	*Low*	Diagnostic codes available for 100% of patients				
	Chronic obstructive pulmonary disease (COPD)	Diagnostic code (ICD-10 or equivalent) Medication code and date (as proxies)	*Low*	Diagnostic codes available for 100% of patients Drug codes are available for 100% of patients Diagnostic codes are available for 86% subjects in attending emergency room				

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

Follow-up time needed per patient in the study	5 years	6 years (including recruitment and follow-up)	*High*				As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors
Minimum time in the data source for lookback assessment	1 year	1 year of lookback	*High*				As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors

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

Case study	RWD source	Sample size estimation form the hypotetical trial protocol	Feasibility assessment *(yes/yes, with limitations/no)*	Rationale for the feasibility assessment	Limitations identified during the feasibility assessment and categorisation	Description of potential impact of the identified limitations on the study results
3 (Dapagliflozin and Major Adverse Cardiovascular Events in Type 2 Diabetes)	**BIFAP**	With an approximate estimated sample size of 13,341 (based on a 1:1 ratio of dapagliflozin and DPP-4i), and considering that BIFAP includes data from approximately 14 million inhabitants (up to 2018), the target sample size is anticipated to be reached. Furthermore, experimental exposure is expected to occur frequently (13.5 DHD in 2023) [1].	Yes	Elements with high criticality are available and fairly reliable. Data recency of 6 months before extraction, reasonably enough for the research question. The time elapsed from when a user requests the data to when they actually receive it is 1-4 months. Sample size is achievable.	·Minor: Hospital information in BIFAP includes dates of admission and discharge, type of discharge, primary and secondary diagnoses at hospital discharge. So, an acute cardiovascular event will only be picked if it constituted one of the main reasons for admission. ·Minor: Mortality data is updated with a one-year delay relative to the present time. For research purposes only the year of death is available. ·Minor: In mortality, a non-random pattern of missingness (MNAR) was observed due to incomplete or inaccurate recording of the cause of death, with a tendency to preferentially register cardiovascular-related deaths. A non-random pattern of association between missingness and MACE was seen. GPs do not have a complete registry of deaths and, particularly, there is not an appropriate recording of the cause of death. Consequently, adjustments using statistical methods for MNAR should be considered in the TTE protocol. ·Minor: Discontinuation date is not available, but calculated by dispentation date+number of packages+posology if written by doctor, if not, calculated by algorithm. ·Minor: Drug use is not linked to a specific indication. ·Minor: Smoking status may be biased, as the criterion is 'current use or use within one year prior to randomization; therefore, patients who smoked before this period would be classified as non-smokers.	As in-hospital cardiovascular events might not be fully captured, some underestimation of outcomes may exist. However, as these are usually severe and with chronic repercussions, we expect primary care setting will capture them even with some delay. As mortality data is delayed and only the year of death is available, this can impact precision and the time sequence of outcomes.
	CPRD	With an approximate estimated sample size of 13,341 (based on a 1:1 ratio of dapagliflozin and DPP-4i), and considering that CPRD includes data from approximately 4.4 million inhabitants (as of 2014), the target sample size is anticipated to be reached. Furthermore, experimental exposure is expected to occur frequently.	Yes	Elements with high criticality are available and fairly reliable. Data recency of 3 months before extraction, reasonably enough for the research question. Sample size is achievable.	·Minor: Dispensing is not available, only prescription. ·Minor: Treatment discontinuation is not readily available but inferred from prescription duration. ·Minor: Diagnostic codes are available for 86% subjects attending emergency room. ·Minor: Diabetes mellitus without type specification occurs frequently as well; usually insulin in monotherapy is used to assess T1D.	As this database only has prescription data, it is unknown if patients took the prescription or if they discontinued it. However, treatment duration is available, from which this data may be estimated. Diagnostic codes are reported to be available for 86% of subjects in the emergency room; however, the missing cases we expect to capture them from hospitalization records or primary care records, since the severity of this disease

REFERENCES

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