Item		Sub-item	Description	Origin of information	Maturity level grade	Maturity level criteria and definitions	Rationale
0	Data base identification	Country	Finland		*N/A*	*N/A*	*N/A*
		Data Access Provider	Finnish Social and Health Data Permit Authority Findata, with register maintainers (for national registers Finnish Institute of Health and Welfare (THL), Social Insurance Institution of Finland (Kela), Statistics Finland, DVV)	https://findata.fi/en/data/#what-data-are-available-via-findata//https://pubmed.ncbi.nlm.nih.gov/22899561/ // https://findata.fi/en/data/#what-data-are-available-via-findata		*N/A*	
		Organisation type	Government agency			*N/A*	
I	Rationale and scope for the RWD source creation	Primary purpose for which data are collected	Originally, these data have been collected for administrative purposes (since 1950s), but their use for research, regulatory, health policy and organisational planning is enabled by the act on secondary use of health and social data. Linkage across different registers is possible with personal identification numbers which are pseudonymised when the data are provided for research use	https://tehdas.eu/tehdas1/packages/package-4-outreach-engagement-and-sustainability/tehdas-country-visits/	*2*	*L1 if information is available as free text and/or online link(s)*	*Relevant for all DQ dimensions (reliability, extensiveness, coherence and timeliness) as it provides a general understandin g of the strengths and limitations of an RWD source.* *Knowing the triggers would ease the understanding of the content and motivations behind the data.*
		Criteria for the selection of the data being collected or integrated	All residents			*L2 if information is available using standardised templates to make information easy to digest and interpret (the EMA recommends to check this tool as reference: REQueST Tool and its vision paper [Internet]. EUnetHTA. 2019. Available from: 721 https://www.eunethta.eu/request-tool-and-its-vision-paper/.*	
		What triggers a record in the database	Birth, use of healthcare service, dispensing event, death	https://journal.fi/finjehew/article/view/146124/94799// https://pubmed.ncbi.nlm.nih.gov/34321928/		*L3 if the information is provided as Metadata (machine readable), including standrad formats, clear definitions and potentially some quality information*	
		Publications describing this RWD	https://pubmed.ncbi.nlm.nih.gov/34321928/				
II	Data collection or recording process	Description of data provider (geographical and oragnizational setting, nature of the data - reported by patients, HCP, etc)	The data on these national registers is recorded by health care providers. Dispensing events are recorded in pharmacies, healthcare (primary and specialised) by the healthcare providers. Death certificates are written by physicians.	https://pubmed.ncbi.nlm.nih.gov/34321928/ //https://journal.fi/finjehew/article/view/146124/94799/	*2*	*L1 if information is available as free text and/or online link(s)*	*Essential to understand extensiveness and to assess reliability (that can be affected by errors or biases in the collection process). Also, essential to evaluate SOP for data collection or recording practices that may impact coherence (e.g., where "curation at source" is involved and provide hard constraints for timeliness).*

		Standard Operating Procedures (SOPs) recording	The Finnish Population Information System is a computerised national register that contains basic information about Finnish citizens and foreign citizens residing in Finland on a permanent or temporary basis. The Kanta Services are a set of digital services that store citizens' social welfare and health care data. The use of this data makes it easier to manage affairs relating to your health and wellbeing. In addition, the data supports social welfare and health care providers in their decision making. The Kanta Services are a nationwide solution that cover all of Finland. Kanta is used by both public and private providers of health care and social welfare services, and by pharmacies. The statistics on causes of death are based on data derived from the death certificates that are complemented with data on deaths from the Population Information System of the Population Register. The statistics on causes of death include all deaths in Finland or abroad of persons permanently resident in Finland at the time of their death. Investigating the cause of death and the related procedures including the production of statistics and archiving of death certificates is based on the Act (1973/459) and Decree (1973/948) on the investigation of the cause of death.	https://dvv.fi/en/population-information-system https://stat.fi/meta/til/ksyyt_en.html https://thl.fi/en/research-and-development/thl-biobank/for researchers/application-process/access-to-national-register-data https://thl.fi/en/statistics-and-data/data-and-services/register-descriptions https://www.kanta.fi/en/research-and-knowledge-management Provided by DEAP		*L2 if information is available using standardised templates to make information easy to digest and interpret, and also standard vocabularies are available*	
		How SOPs are implemented and monitored	The Digital and Population Data Services Agency promotes the digitalisation of society, maintains population data, secures the availability of data, and provides services for the life events of its customers. For the Care register for Healthcare, every year, data suppliers (healthcare and social care units) are provided with a manual on the data content and on how to submit care notifications. More information is available in links in this and above sections.	https://dvv.fi/en/population-information-system https://stat.fi/meta/til/ksyyt_en.html https://thl.fi/en/research-and-development/thl-biobank/for researchers/application-process/access-to-national-register-data https://thl.fi/en/statistics-and-data/data-and-services/register-descriptions https://www.kanta.fi/en/research-and-knowledge-management Provided by DEAP			
		Key data elements captured (are they always recorded, are they optional, is there a planned coverage over time, ...)	Prescriptions (written and dispensed), inpatient and outpatient contatcs, including diagnoses, procedures and specialties, Cancer register, Cause of death register, Birth register, Laboratory database, Vaccinations, death dates and causes of death, migration, socioeconomic information (income, education) linkable via personal identification number. Different data is available since: - Hospitaisation data, cancer registry, medical births register, special reimbursements, since 1972, - Dispensed / reimbursed prescriptions since 1995, - All prescriptions since 2014 (with full coverage since 2017), - Secialized healthcare outpatient admissions since 1998, - Primary care since 2011, - Lab measurements since 2014.	https://pubmed.ncbi.nlm.nih.gov/34321928/ // https://journal.fi/finjehew/article/view/146124/94799/		*L3 if additionally SOPs specify KPIs to monitor*	
III	The selection of RWD sources and their	Criteria to accept or exclude a datasource	N/A		*N/A*	*L1 if information about selection criteria or DQ performance is available as free text and/or online link(s)*	*When data are provided by a data aggregator, ensure that all the available evidence related to systems and*

	onboarding *(Applies to RWD sources that integrate or repurpose other RWD sources)*	Is there a DQ assessment for data sources onboarded?	N/A		*N/A*	*L2 if a structure checklist and dataset version control are available* *L3 is only aspirational. N/A*	*processes potentially affecting DQ (extensiveness and reliability especially) can be followed. Provide information of impact on both reliability and evidence (as well as other dimensions if relative constraints are formulated in inclusion/excl usion (I/E) criteria)*
		If yes: does it follow any specific framework? Is there an assessment checklist? Are datasets versions traceable?	N/A		*N/A*		
IV	The data management infrastructure	List of systems used to manage the RWD (either for data collection, recording, processing, etc)	All the patient data recorded to the Kanta PDR are stored as XML docu-ments using Health Level Seven (HL7) Clinical Docu-ment Architecture Release 2 (CDA R2) format	https://academic.oup.com/jamia/article-abstract/13/1/30/781314?redirectedFrom=fulltext	*2*	*L1 if information is available as free text and/or online link(s)* *L2 if the hardware or software implementation complies with recognised quality standards that can be reported* *L3 N/A*	*Essential for reliability regarding data alterations resulting from system accidents, software errors or malicious intervention.*
		Software testing and software quality control in place	Requested to DEAP and unable to provide		*N/A*		
		Measures to prevent accidental physical data alterations (e.g.: backups, redundant systems, checksums)	The end users naturally employ backups and sum checks when data extractions are disposed to them. However, further information of this process at other stages is not available for the DEAP. As much as the data is originating from billing systems (healthcare contacts), there are internal audits in place but there is no consultable documentation on this.	Provided by DEAP	*1*		
V	Data management and governance	Data management principles being followed (e.g., GCP, ISO, FAIR, etc)	As per GDPR the data subjects have the legal right to request the correction of their information. Researchers / other data users can notify the permit authority on assumed errors in the data	Provided by DEAP	*1*	*L1 if information is available as free text and/or online link(s)* *L2 if standard best practices are being used and a direct impact on DQ is reported. There are SOPs and dat amanagement processes that adhere to the standards. The representation of metadata follows FAIR standards* *L3 if data management and governance is implemented in the data platforms 'Digital Quality Measures' (DQMs) so that reports of performance and deviations are automated. Submitted metadata are generated "by design". Basically, if everything in L2 is automatised and generated by default*	*Data management and governance impact reliability, as well as all quality dimensions for metadata.*
		Data management processes in place (DQ controls, KPIs, SOPs, etc)	All the patient data recorded to the Kanta PDR are stored as XML docu-ments using Health Level Seven (HL7) Clinical Docu-ment Architecture Release 2 (CDA R2) format	https://academic.oup.com/jamia/article-abstract/13/1/30/781314?redirectedFrom=fulltext	*2*		
		Measures to prevent data alterations by unauthorised parties (cybersecurity)	All data are analysed in audted remote use environments.	https://findata.fi/en/kapseli/regulation-on-secure-operating-environments/ Provided by DEAP	*2*		
		Auditing and DQ imporvement procedures in place	Medical record review is a common reference standard used in validation studies to confirm the presence or absence of a disease. Other types of reference standards include patient self-reports, physicians' reports, autopsy reports, and alternative data sources with presumably higher data quality (such as clinical quality data, laboratory data, and pathology data). In general, data in health and clinical quality databases have high validity and completeness, because they are collected prospectively with the aim of quality control and clinical care. Consultants within each medical field register the data in clinical quality databases, further increasing the accuracy of the data. Although also registered prospectively, the validity of clinical data in the administrative databases may vary considerably among databases and within each database. Government-initiated systematic validation of personal demographic data, hospital admission data, and overall diagnoses within different clinical specialties. Investigator-driven systematic validation of individual diagnoses, examinations, procedures, and surgery codes within a specific clinical specialty Investigator-driven ad hoc validation of study-specific variables, the most common type of validation study As stated under auditing documents, it should be possible to register which object locating systems produce data and, then, which applications consume this data. Periodic reviews (i.e. automatic reviews) of these log events can assist in detecting security issues, faulty allocation of rights, etc. Which systems produce data can be registered at low cost as a part of ensuring the persistence of location events. Registration of who uses the individual event data will lead to considerably larger overheads. The type of queries performed by the applications can be registered instead.	Provided by DEAP	*2*		

VI	Data manipulation steps	Frequency of data updates	Monthly causes of death annually	https://catalogues.ema.europa.eu/node/1094/data-flows-and-management Provided by DEAP	*2*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*Impacts reliability both in terms of accuracy (possible errors) and precision (i.e., the degree of approximatio n by which data represents reality). Essential to ensure traceability of information. Also impacts coherence and potentially timeliness.*
		Data transformations performed, data mapping steps, data cleaning	Values or missings are NOT imputed. Internal quality checks are conducted, but they are not acknowledged. Data validation procedures vary accorss registers. Variable names might be changed, especially for old data.	Provided by DEAP	*1*	*L2 if Tests performed follow some standard or shared set of tests, that can be re-used across RWD sources. Key performance indicators (KPIs) for data cleaning (e.g., data duplications, mislabelling, etc.) are provided. Data mapping tables and algorithms are described with a standard characterisation of their performance. Lists of standard test batteries used to detect loss of accuracy or precision are provided. All lineage information is*	
		Information about loss of precision during data manipulation steps	Requested to DEAP and unable to provide		*N/A*	*L3 if information about data onboarding is directly provided by the platform, e.g.: * Transaction logs are available including deviations and actions that required manual intervention Actual data transformation code is accessible and verifiable. Quality checks and KPIs reported are automatically generated by the data platform (e.g.: unit testing) Lineage information is automatically generated by the processing platform*	
		Lineage information (e.g., justifcation of data manipulation, track of changes and versions)	Requested to DEAP and unable to provide		*N/A*		
VII	Data augmentation steps (e.g., imputation or linkage)	Is any augmentation happening in this datasource?	Family relationship linkage with different registers, linkage to FinnGen biobank data possible	Provided by DEAP	*1*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*Data augmentation steps impact accuracy (reliability) and extensiveness. We consider here data transformations that produce new information subject to reliability issues: e.g.: imputation of missing values, or extraction of codes via natural language processing.*
		If yes, which are the methods applied	Pseudonymised personal identification number (deterministic)	Provided by DEAP		*L2 if algorithms are published and their performance documented. Information on which values result from imputation is provided as part of the dataset (e.g.,*	
		If yes, which algorithms and assumptions applied	Direct for personal information, family relation linkage for paternity	Provided by DEAP			
		If yes, which is the error rate when conducting the augmentation	Not available	Provided by DEAP	*N/A*	*L3 if an automatised process for data linkage/mapping exists*	
VIII	Known quality issues and independent QA assessment of the RWD source	Known DQ issues (e.g., poor overall completeness in Q3 2020 due to COVID-19)	Individuals declared legally dead are not included in the number of deceased in the statistics on causes of death. The statistics will lack the cause of death for some 100 to 400 individuals every year, because they cannot be provided with a death certificate. In 2023, the statistics lacked a death certificate in 284 deaths, corresponding to 0.5 per cent of all deaths. The data pertaining to causes of death are produced annually and completed by the end of the following year. More data quality characteristics are described in the link.	https://stat.fi/en/statistics/documentation/ksyyt#Accuracy,%20reliability%20and%20timeliness https://stat.fi/en/statistical-data	*2*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*Explicit description of known DQ issues, as well as external validation performed (all dimensions affected)*
		Validation studies and publications resulting from this EWD source	In general, there is ongoing work on data quality and structuring of data across Finland. Stakeholders • Registry data at THL are validated and checked through basic automated checks, manual checks, comparison to previous years and feedback loops. More automated checks are being developed.	https://tehdas.eu/tehdas1/packages/package-4-outreach-engagement-and-sustainability/tehdas-country-visits/		*L2 if standard procedures are set for external/internal validation of the data* *L3 if the mechanism provided includes notification of automatically detected DQ issues*	
IX	The RWD source representation	Description of data model or models used (OMOP, FHIR, ...)	Conception, OMOP	Provided by DEAP	*2*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*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	Diagnoses: ICD-10, ICD-9, ICPC Drugs: ATC Procedures: NOMESCO	https://academic.oup.com/jamia/article-abstract/13/1/30/781314?redirectedFrom=fulltext https://journal.fi/finjehew/article/view/146124		*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)*	

		(e.g., UMLS)		https://zenodo.org/records/13384860		*L3 if a standard CDM is used, the datasource has been mapped to one or more than one CDM, and if data dictionaries are provided using standard formats that facilitate the mappings across different vocabularies and across languages*	
X	The RWD source declared Service Level Agreements (SLA)	Guaranteed frequency of updates and incident response time (e.g., corrections in case of errors)	Requested to DEAP and unable to provide		*N/A*	*L1 if free-text information and links are available reporting all the mentioned features*	*Descriptive of guaranteed timeliness and possible variations of extensivenes s/reliability provided.*
		Processes and resources accompanying the data, such as documentation, training materials or help desk contact	Requested to DEAP and unable to provide			*L2 if details of established data processess by the provider are available*	
		Possibility to collect additional data if needed	It is possible to collect additional data. For example, when data is extracted from the database but more personal information is warranted, person identification number cvould be provided to link the extracted information, get the individuals' consent, and link especific surveyed information to the actual data.	Provided by DEAP	*2*	*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	Permissions can be applied via the Nationalpermit authority Findata. Individual data can be analysed only within aufdited remote use envionment. Suggested link to cell E40-41: https://findata.fi/en/	https://findata.fi/en/data/#what-data-are-available-via-findata https://thl.fi/en/statistics-and-data/data-and-services/research-use-and-data-permits	*2*	*L1 if free-text information and links are available reporting all the mentioned features* *L2 if policies and licensing are standardised to a broad range of RWD* *L3 N/A*	*Descriptive of aspects that can limit extensiveness and coherence in downstream data aggregations.*
XII	Feedback	Is there a data ecosystem in place so that quality assessment by data consumers can provide feedback to improve the data collection and production process, thus allowing a continuous monitoring and improvement of DQ?	Requested to DEAP and unable to provide		*N/A*	*L1 if a person of contact is provided for Q&A* *L2 if the contact provided allows tracking of issues and follow-up* *L3 if the mechanism provided includes notification of automatically detected DQ issues*	*Descriptive of feedback mechanisms in place to improve all aspects of DQ*

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

IV	The data management infrastructure	List of systems used to manage the RWD (either for data collection, recording, processing, etc)	EMIS Web® electronic patient record system software for CPRD Aurum Vision® software for CPRD GOLD (From April 2018, Read codes are prospectively mapped to SNOMED CT codes by Vision)	https://www.cprd.com/primary-care-data-public-health-research https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135 https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf	*2*	*L1 if information is available as free text and/or online link(s)* *L2 if the hardware or software implementation complies with recognised quality standards that can be reported* *L3 NA*	*Essential for reliability regarding data alterations resulting from system accidents, software errors or malicious intervention.*
		Software testing and software quality control in place	Requested to DEAP and unable to provide		*N/A*		
		Measures to prevent accidental physical data alterations (e.g.: backups, redundant systems, checksums)	CPRD is oblied to complete an annual NHS Data Security and Protection Toolkit assessment to demonstrate that it meets the required standard for holding data securely. We are unsure of what this toolkit entails. information is broad and might be only available when you buy/contract the service.	https://www.cprd.com/safeguarding-patient-data https://www.dsptoolkit.nhs.uk/	*2*		
V	Data management and governance	Data management principles being followed (e.g., GCP, ISO, FAIR, etc)	Requested to DEAP and unable to provide		*N/A*	*L1 if information is available as free text and/or online link(s)* *L2 if standard best practices are being used and a direct impact on DQ is reported. There are SOPs and dat amanagement processes that adhere to the standards. The representation of metadata follows FAIR standards* *L3 if data management and governance is implemented in the data platforms 'Digital Quality Measures' (DQMs) so that reports of performance and deviations are automated. Submitted metadata are generated "by design". Basically, if everything in L2 is automatised and generated by default*	*Data management and governance impact reliability, as well as all quality dimensions for metadata.*
		Data management processes in place (DQ controls, KPIs, SOPs, etc)	**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*		
		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			
VI	Data manipulation steps	Frequency of data updates	GOLD: monthly; Aurum: Quarterly	https://catalogues.ema.europa.eu/node/976/data-flows-and-management	*2*	*L1 if free-text information, links or publications are available reporting all the mentioned features* *L2 if Tests performed follow some standard or shared set of tests, that can be re-used across RWD sources. Key performance indicators (KPIs) for data cleaning (e.g., data duplications, mislabelling, etc.) are provided. Data mapping tables and algorithms are described with a standard characterisation of their performance. Lists of standard test batteries used to detect loss of accuracy or precision are provided. All lineage information is provided as metadata associated to the dataset* *L3 if information about data onboarding is directly provided by the platform, e.g.: * Transaction logs are available*	*Impacts reliability both in terms of accuracy (possible errors) and precision (i.e., the degree of approximatio n by which data represents reality). Essential to ensure traceability of information. Also impacts coherence and potentially timeliness.*
		Data transformations performed, data mapping steps, data cleaning	Requested to DEAP and unable to provide		*N/A*		
		Information about loss of precision during data manipulation steps	Requested to DEAP and unable to provide				

		Lineage information (e.g., justifcation of data manipulation, track of changes and versions)	Each dataset has a digital object identifier (DOI) to trace specific database versions	https://www.cprd.com/digital-object-identifiers-dois-datasets	*2*	*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*

Dimension	Sub-dimension	Metrics	Description	Origin of information
Timeliness	Currency	How often is the database updated (i.e., frequency of updates)	Varies between registers (monthly-annual)	Provided by DEAP by consulting register maintainers
		The time gap between the latest available data and date when data is delivered to user. (i.e., how up-to-date data are when it reach the user)	From a few months to 2 years, depending on the needed.	Provided by DEAP
		The time elapsed from when a user requests the data to when they actually receive it	4-12 months	Provided by DEAP
		Median time (years) between first and last available records for unique individuals	Specific number of years is unknown but data subjects are followed from birth/immigration to death/emigration.	Provided by DEAP
Extensiveness	Coverage	Percentage of a target population present in a database	>99%	Although these are nationwide data, it is possible to deny the use of personal information for reasons outlined in the act on secondary use of health and social data. Until now, the number of these persons has been small, less than 1% of population.
	Completeness	% of subjects in the data with a recorded birth date	100%	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
		% of subjects in the data, irrespective of vital status, that have a recorded date of death	A date of death is recorded for 100% of individuals who are known to have died	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
		% of subjects in the data with a record of sex	100%	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
		% of subjects in the data who had an event with a code for the event	99.97%	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
		% of subjects in the data who had a prescription/dispensing with a recorded code for the medicine	99.99%	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
		% of subjects in the data who got vaccinated with a recorded code for the vaccine	A register of vaccination with a code for the vaccine is recorded for 98.67% of individuals who are known to have been vaccinated	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
Reliability	Accuracy	The population distribution in the data source aligns with that of the country	As the source includes population at a national level, this assessment is not applicable.	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
		Records of diagnostics, exposures or medical observations that do not agree with common expectations and knowledge or feasible ranges (e.g., pregnancy records in males, a human with 4 arms, systolic pressure higher than 250mmHg, etc)	<0.5%	Provided by DEAP. Checks of diagnosis data in our previous studies accross different therapeutic areas
		Records of healthcare events (diagnoses, prescriptions, admissions, etc) with logical inconsistencies (e.g., and admission occurs after death)	<1%	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
		Variables that are based in imputation, derivation or inference (e.g., end of treatment date is derived from treatment start date and treatment cycle length)	End of treatment episodes derived based on dispensing date and dispensed amount	
	Precision	Exposures codes precision level, including medicines and vaccines (e.g., active principle, therapeutic group, ...)	Active principle (ATC level 5 codes)	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
		Precision of date of birth (e.g., day, month, year)	Month, year (Day can be obtained if justified by the research question	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
		Precision of date of death (e.g., day, month, year)	Day, month, year	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
		Precision of date of the event/diagnosis (e.g., day, month, year)	Day, month, year	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
		Precision of date of the exposure (e.g., day, month, year)	Day, month, year (dispensing date). End date must be calculated based on data)	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
	Traceability	Provenance of event records	Primary care, specialised health care (inpatient and outpatient); intensive care unit (limited)	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
		Provenance of medicines/vaccines records	Both prescribed and dispensed included, vaccinations also from primary care records	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
Coherence	Format coherence	For dates, formatting constraint being followed	Character, length 8: ddmmyyyy	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
		For sex, formatting constraint being followed	0 Unknown 1 Man ☐ 2 Woman 3 Not known / cannot be defined☐ 9 undefined	https://www.jidonline.org/article/S0022-202X(18)30110-6/fulltext
	Relational coherence	% of records with the Person ID in the PERSONS table	100%	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017
	Semantic coherence	For EVENTS definitions, codelists/data dictionaries being employed according to external standards	ICD-10 (https://koodistopalvelu.kanta.fi/codeserver/pages/classification-view-page.xhtml?classificationKey=23&versionKey=58), ICPC (https://koodistopalvelu.kanta.fi/codeserver/pages/classification-view-page.xhtml?classificationKey=210&versionKey=282)	https://koodistopalvelu.kanta.fi/codeserver/pages/classification-view-page.xhtml?classificationKey=210&versionKey=282
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	ATC, nordic product number (vnr)	
	Uniqueness	Number of records flagged as potential duplicates	0%	Provided by DEAP. Checks on our 50% random sample of Finnish population in 2017

Dimension	Sub-dimension	Metrics	Description	Origin of information
Timeliness	Currency	How often is the database updated (i.e., frequency of updates)	GOLD: monthly; Aurum: quaterly	https://academic.oup.com/ije/article/44/3/827/632531
		The time gap between the latest available data and date when data is delivered to user. (i.e., how up-to-date data are when it reach the user)	1 month plus lag of delivery for CPRD GOLD, and 3 months plus lag of delivery for CPRD Aurum	Provided by DEAP
		The time elapsed from when a user requests the data to when they actually receive it	Requested to DEAP and unable to provide	
		Median time (years) between first and last available records for unique individuals	5.89 years	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors
Extensiveness	Coverage	Percentage of a target population present in a database	CPRD-GOLD 2,894,922 current acceptable patients (i.e. registered at currently contributing practices that use Vision software, excluding transferred out, deceased patients and those flagged by CPRD as not acceptable for clinical research for data quality issues) equal to **4.32%** based on the UK population estimates of 67,026,300 from the Office of National Statistics (July 2024). CPRD-AURUM 16,585,135 Current acceptable patients (i.e. registered at currently contributing practices2, excluding transferred out and deceased patients) equal to **24.27%** percentage UK population coverage (67,026,300) (september 2024).	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors https://www.cprd.com/doi/cprd-gold-november-2024-dataset https://www.cprd.com/doi/cprd-aurum-september-2024-dataset https://jech.bmj.com/content/76/10/880
	Completness	% of subjects in the data with a recorded birth date	Percentage not provided (only year of birth available)	
		% of subjects in the data, irrespective of vital status, that have a recorded date of death	A date of death is recorded for 100% of individuals who are known to have died	https://zenodo.org/records/13384860
		% of subjects in the data with a record of sex	100%	https://zenodo.org/records/13384860
		% of subjects in the data who had an event with a code for the event	100% (86% of the emergency room setting)	https://zenodo.org/records/13384860 https://www.cprd.com/cprd-linked-
		% of subjects in the data who had a prescription/dispensing with a recorded code for the medicine	100%	https://zenodo.org/records/13384860
		% of subjects in the data who got vaccinated with a recorded code for the vaccine	A register of vaccination with a code for the vaccine is recorded for 100% of individuals who are known to have been vaccinated	https://zenodo.org/records/13384860
		Others: BMI	BMI completeness increased over calendar timefrom 37% in 1990–1994 to 77% in 2005–2011, was higher among female and increased with age	https://bmjopen.bmj.com/content/3/9/e003389 https://www.sciencedirect.com/science/article/pii/S2666776224001534#appsec1
Reliability	Accuracy	The population distribution in the data source aligns with that of the country	Population distribution as expected based on the statistics of the general population of England. Previous literature acknowledges some potential overrepresentation of minority ethnic groups. There is a study ongoing in regards to CPRD representativeness (see link). Active population size by ageband: -Paediatric Population (< 18 years): 519902 (13.1%) -Children (2 to < 12 years): 287819 (8.3%) -Adolescents (12 to < 18 years): 200949 (5.1%) -Adults (18 to < 46 years): 1061418 (26.7%) -Adults (46 to < 65 years): 725924 (18.3%) -Elderly (≥ 65 years): 587470 (14.8%) -Adults (65 to < 75 years): 303212 (7.6%) -Adults (75 to < 85 years): 205960 (5.2%) -Adults (85 years and over): 78298 (2.0%)	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors https://jech.bmj.com/content/76/10/880 https://pophealthmetrics.biomedcentral.com/articles/10.1186/s12963-023-00302-0 https://www.cprd.com/approved-studies/representativeness-clinical-practice-research-datalink-cprd-primary-care-databases
		Records of diagnostics, exposures or medical observations that do not agree with common expectations and knowledge or feasible ranges (e.g., pregnancy records in males, a human with 4 arms, systolic pressure higher than 250mmHg, etc)	A data cleaning procedure is performed to avoid inconsistencies and other unfeasible data (see link) Rate of adherence among metformin new users is lower than rates determined in previous UK studies Nearly all patients who had elevated HbA1c labs or hypoglycemic treatments also had a type 2 diabetes diagnosis code Completeness for hyper-cholesterolemia and anemia diagnoses is modest even when the presence of treatments and lab results indicated the conditions were likely present (51%-59% and 58%-70%, respectively)	https://www.cprd.com/sites/default/files/2023-02/CPRD%20Aurum%20Glossary%20Terms%20v2.pdf https://www.sciencedirect.com/science/article/pii/S2214623720300351?via%3Dihub#s0055 https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.5135
		Records of healthcare events (diagnoses, prescriptions, admissions, etc) with logical inconsistencies (e.g., and admission occurs after death)	Data values after death: 0% (from DEAP experience, some event dates may occur after censoring) Date values before birth: 0.02%	https://zenodo.org/records/13384860 https://www.cprd.com/sites/default/files/2023-02/CPRD%20Aurum%20Glossary%20Terms%20v2.pdf
		Variables that are based in imputation, derivation or inference (e.g., end of treatment date is derived from treatment start date and treatment cycle length)	Mother-baby id, pregnancy, ethnicity	https://onlinelibrary.wiley.com/doi/10.1002/pds.5135 https://www.cprd.com/cprd-algorithm-derived-data
	Precision	Exposures codes precision level, including medicines and vaccines (e.g., active principle, therapeutic group, ...)	Active principle (ATC level 5 codes)	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
		Precision of date of birth (e.g., day, month, year)	Year (Month/year only for children)	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
		Precision of date of death (e.g., day, month, year)	Day, month, year	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
		Precision of date of the event/diagnosis (e.g., day, month, year)	Day, month, year	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf
		Precision of date of the exposure (e.g., day, month, year)	Day, month, year	https://www.cprd.com/sites/default/files/2024-08/CPRD%20GOLD%20Full%20Data%20Specification%20v2.6.pdf

	Traceability	Provenance of event records	Primary care medical records, Emergency room, Intensive care unit, Hospitalisation (ER/ICU, HOSP only through linked data. UU only has access to HES admited patient care)	https://cataloques.ema.europa.eu/node/1026/administrative-details
		Provenance of medicines/vaccines records	Primary care medical records (Prescription medicines, No dispensing medicines)	https://cataloques.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		**Comparison of single-device vilanterol/fluticasone furoate with other inhaled corticosteroid-long-acting beta agonist single-device combinations in the risk of pneumonia in adolescents with asthma**						
Design elements	**Operationalization of definitions**	**Data elements for valid capture of variables**	**Criticality of the quality of the element**	**Extensiveness assessment (if applicable)**	**Reliability assessment (if applicable)**	**Coherence assessment (if applicable)**	**Timeliness assessment (if applicable)**	**Origin of information**
Study population	**Inclusion criteria**							
	12-17 at treatment initiation (adolescents)	Date of birth (years)	*High*	100% have date of birth	Only year is available, this may impact precision.		As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	https://catalogues.ema.europa.eu/node/1026/quantitative-descriptors
	Asthma diagnosis	Diagnostic code Date of diagnostic	*High*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				https://bmjopen.bmj.com/content/7/8/e017474.abstract
	Step-up from ICS alone to ICS+LABA	Medication code Date of prescription/dispensing Duration of treatment	*High*	100% medication code 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication	As ATC codes are not available, a mappint to ATC will potentially be needed to extract study drugs information. The DEAP informed they already have the mapping ready, so this should not be a problem.		
	Exclusion criteria							
	Previous pneumonia diagnosis within the previous year	Diagnostic code Date of diagnostic	*High*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room			As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	
	LABA treatment within the previous year	Medication code Date of prescription/dispensing	*High*	100%	Only prescription date is available	As ATC codes are not available, a mappint to ATC will potentially be needed to extract study drugs information. The DEAP informed they already have the mapping ready, so this should not be a problem.	As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	

	Long-term hospitalisation within the previous year	Date of admission Date of discharge Duration of hospitalisation	*High*	Not available for the core data source; available with data augmentation (linkage) only Known issues in CPRD that could affect extensiveness include: Provisional HES data are monthly publications of HES data (these data may be incomplete or contain errors for which no adjustments have yet been made by HES), It is also probable that clinical data are not complete, which may affect the last two months of any given period.	Known issues in CPRD that could affect reliability include: Unfinished episodes, at the end of the fiscal year ("month 13"), the annual data is refreshed and known data quality issues are corrected, prior to locking the annual published data, HRG files provided to CPRD with high level of null values for variable hes_yr compared to Set22	Hospitalisations refer to the total period of inpatient hospital stay from admission to discharge. When a hospitalisation spans the end of the HES year, it is artificially modelled as two hospitalisations. Known issues in CPRD that concern coherence include: • Invalid/missing dates depicted as 15/10/1582 or 01/01/1600 • Episodes where admission date precedes the epistart date • Explicit duplicate records which vary only by unique episode identifier (epikey) • Maternity records may have inconsistencies which need to be considered when using the data • Counts produced from provisional data are likely to be lower than those generated for the same period in the final dataset. There may also be errors due to coding inconsistencies that have not yet been investigated and corrected.	Unfinished episodes might be found at the end of the fiscal year ("month 13") As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	
Treatment/exposure	Vilanterol-flucatisone furoate combination	Medication code Date of prescription/dispensing	*High*	100% medication code 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication	As ATC codes are not available, a mappint to ATC will potentially be needed to extract study drugs information. The DEAP informed they already have the mapping ready, so this should not be a problem.		
Comparator group (if applicable)	Other ICS-LABA combinations	Medication code Date of prescription/dispensing	*High*	100% medication code 100% date of prescription (only prescription is available)		As ATC codes are not available, a mappint to ATC will potentially be needed to extract study drugs information. The DEAP informed they already have the mapping ready, so this should not be a problem.		
Key endpoint(s)	Time to first ocurrence of pneumonia	Diagnostic code Date of diagnostic	*High*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				
Confounders	sex	Sex	*Low*	>99%. Sex categories in CPRD include unknown and indeterminate sex, but are never included in data extractions (<1% of records without sex information are excluded); they are extremely rare.				
	age	Date of birth (years)	*Low*	100% have date of birth	Only year is available, this may impact precision.			
	season	Date of diagnostic Date of prescription/dispensing	*Low*	100% Date of diagnostic 100% Date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
	calendar year	Date of diagnostic Date of prescription/dispensing	*Low*	100% Date of diagnostic 100% Date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
	age at 1st asthma diagnosis	Date of birth Date of diagnostic Diagnostic code	*Low*	100% have date of birth	Only year is available, this may impact precision.			
	diabetes	Diagnostic codes Date of diagnostic	*Low*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				
	rheumatological diseases	Diagnostic codes Date of diagnostic	*Low*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				

malignancies	Diagnostic codes Date of diagnostic	*Low*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				
cardiovascular diseases	Diagnostic codes Date of diagnostic	*Low*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				
n of hospital days (any cause)	Date of admission Date of discharge Duration of hospitalisation	*Low*					
n of outpatient visits	Date of visit Setting of visit	*Low*					
Down/intellectual disabilities/ congenital malformations	Diagnostic codes Date of diagnostic	*Low*	Diagnostic codes available for 100% of patients 100% Date of diagnostic Diagnostic codes are available for 86% subjects in attending emergency room				
psychotropics (or specific groups)	Medication code Date of prescription/dispensing	*Low*	100% medication code 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
asthma exacerbations	Medication code Date of prescription/dispensing Diagnostic code Date of diagnosis	*Low*	100% medication code 100% date of prescription Admission or emergency room diagnostic code only available in HES (not disposable for the current study, would need data augmentation, i.e., linkage).	Prescription of a medication does not guarantee the subject collected and took the medication Exacerbations leading to hospital admission or emergency room might be needed for a more reliable capturing of exacerbation.			
oral corticosteroids	Medication code Date of prescription/dispensing	*Low*	100% medication code 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
biologicals	Medication code Date of prescription/dispensing	*Low*	100% medication code 100% date of prescription If these medications are dispensed/prescribed inhospital, availability is unknown	Prescription of a medication does not guarantee the subject collected and took the medication			
SABA	Medication code Date of prescription/dispensing	*Low*	100% medication code 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
leukotriene receptor antagonists	Medication code Date of prescription/dispensing	*Low*	100% medication code 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
antibiotic use	Medication code Date of prescription/dispensing	*Low*	100% medication code 100% date of prescription (only prescription is available) If used inhospital, information might not available	Prescription of a medication does not guarantee the subject collected and took the medication			
pneumococcal vaccination	Medication code Date of prescription/dispensing	*Low*	Vaccine codes available for 100% of patients 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
influenza vaccination	Medication code Date of prescription/dispensing	*Low*	Vaccine codes available for 100% of patients 100% date of prescription (only prescription is available)	Prescription of a medication does not guarantee the subject collected and took the medication			
death	Date of death	*Low*	98.2% of deaths in the Office of National Statistics data are recorded in the CPRD GOLD primary care data, while agreement onthe exact date of death increased over time to 78.0% in 2013.			As data is updated monthly, is to bear in mind that information extracted will be at least 1 month old. Median time between first and last records for unique active individuals is ~13 years. For all individuals is ~6 years.	https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.4747

Intercurrent events	Treatment discontinuation	Medication code Date of prescription/dispensing Duration of treatment	*High*	Prescription information available 100% medication codes and date of prescription	In CPRD information on treatment duration of the prescription is available. This may help defining drug exposure-related variables. Prescription of a medication does not guarantee the subject collected and took the medication			
	Switch to another ICS+LABA (i.e., budesonide/formoterol, or salmeterol/fluticasone propionate)	Medication code Date of prescription/dispensing Duration of treatment	*High*	Prescription information available 100% medication codes and date of prescription	In CPRD information on treatment duration of the prescription is available. This may help defining drug exposure-related variables.			
	Oral corticosteroids use	Medication code Date of prescription/dispensing	*Low*	Prescription information available 100% medication codes and date of prescription				
	Add on SABA, LAMA, leukotriene receptor antagonist, biologics	Medication code Date of prescription/dispensing Duration of treatment	*Low*	Prescription information available 100% medication codes and date of prescription	In CPRD information on treatment duration of the prescription is available. This may help defining drug exposure-related variables.Prescription of a medication does not guarantee the subject collected and took the medication			
	Rescue medications	Medication code Date of prescription/dispensing	*Low*	Prescription information available 100% medication codes and date of prescription				
	Non-pneumonia death	Date of death Diagnostic code Date of diagnosis	*High*	8% of the whole population (irrespective of vital status) has a date of death recorded; 100% of death people have a date of death Diagnostic codes available for 100% of patients 100% Date of diagnostic Where the exact date of death or the cause is importantin a CPRD study, it may be advisable to include the individually linked national ONS death registration data				https://onlinelibrary.wiley.com/doi/epdf/10.1002/pds.4747
Follow-up time needed per patient in the study	up to a maximum of 2 years after the randomisation day	Date of exit of the database	*High*				The median lenght of follow-up per patient is approximately 6 years and 13 years for active individuals	
Minimum time in the data source for lookback assessment	1 year	1 year	*High*				The median lenght of follow-up per patient is approximately 6 years and 13 years for active individuals	

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

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

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

	Non-pneumonia death	Date of death Diagnostic code Date of diagnosis	*High*	Death and cause of death are available. From previous publications, it seems a cause of death register exists.				https://stat.fi/en/statistics/ksyyt , and description of the individual-level data : https://aineistokatalogi.fi/catalog/studies/778c33bf-aceb-423f-89d9-e5abb5a0585c
Follow-up time needed per patient in the study	Up to a maximum of 2 years after the randomisation day	Date of exit of the database	*High*				Population registers have a long history in Finland, with population information having been registered since the 1530s.	
Minimum time in the data source for lookback assessment	1 year	1 year	*High*					

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

Case study	RWD source	Sample size estimation form the hypotetical trial protocol	Feasibility assessment *(yes/yes, with limitations/no)*	Rationale for the feasibility assessmernt
5 (Vilanterol/fluticasone furoate in the risk of pneumonia in adolescents with asthma)	**CPRD**	With an approximate estimated sample size of 26,750 individuals (based on a 1:1 ratio between treatment arms, with 13,375 participants in each), and considering that CPRD includes data from approximately 4.4 million inhabitants (as of 2014), the target cohort size is anticipated to be achievable. In the UK, asthma affects approximately 7.2 million individuals—about 8% of the population—with up to 5.4 million currently receiving any treatment. This translates to approximately 281,600 subjects. Teeneger population in CPRD is 200,949, and the prevalence of asthma in UK teenegers is 15.2%, this translates in 30,544 asmathic teenegers in CPRD. The target sample size is anticipated to be reached. [1,2,3]	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.
	Finnish registers	With an approximate estimated sample size of 26,750 individuals (based on a 1:1 ratio between treatment arms, with 13,375 participants in each), and considering that Finland has a population exceeding 5 million (380,000 individuals aged 12-17y, 10.5% with asthma). Despite there is up to 18,293 users of vilanterol-fluticasone furoate among the 233,261 patients with chronic asthma or similar chronic obstructive pulmonary diseases as of 2021, there is no specific data available for teenage users. Since the number of teenage users is expected to be smaller than the target sample size, we anticipate that reaching the desired sample may not be feasible. [4,5]	Yes, with limitations on sample size acquisition	Elements with high criticality are available and fairly reliable. Data recency is variable depending on the data sets used, but for the current case of a few months are reasonably enough for the research question. The time elapsed from when a user requests the data to when they actually receive it is 4-12 months. **Sample size might be difficult to be reached**.

REFERENCES

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

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

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

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

[5] Gehrt L, Vahlkvist S, Petersen TH, Englund H, Nieminen H, Laake I, Kofoed PE, Feiring B, Benn CS, Trogstad L, Sørup S. Trends in childhood asthma in Denmark, Finland, Norway and Sweden. Acta Paediatr. 2025

Limitations identified during the feasibility assessment and categorisation	Description of potential impact of the identified limitations on the study results
·Potentially major: Hospitalisation data is not available for the core data source. ·Potentially major: Not registered date of death. ·Minor: Dispensing is not available, only prescription. ·Minor: Diagnostic codes are available for 86% subjects attending emergency room.	The DEAP is able to perform data augmentation (linkage) to retrieve hospital admission and discharge diagnoses to detect the outcomes of this case-study. For exact date of death or the cause of death data augmentation (linkage) is needed to the ONS death registration data. As this database only have prescription data, it is unknown if patients took the prescription, and so, if they discontinued it. However, treatment duration is available, from which this data may be estimated. Diagnostic codes are reported to be available for 86% of subjects in the emergency room; however, the missing cases we expect to capture them from hospitalization records or primary care records, since the severity of this disease may justify an admission and/or the follow-up with the GP, or change of baseline treatment.
·Potentially major: limited sample size. ·Minor: No inhospital drug use available. ·Minor: Duration of hospitalisation is not readily available. ·Minor: Exacerbations inhospital are not available. ·Minor: Duration of treatment and end of treatment are not directly available.	Limited sample size in Finnish registers might lead to underpowered analyses for this population. This can be mitigated by meta-analysing the results together with CPRD. Duration of hospitalisation can be calculated from the admission and discharge dates. Inhospital exacerbations can be inferred from admission, discharge and primary care diagnoses that are readily available. End of treatment episodes may be derived based on dispensing date and dispensed amount.

; Jan 13. doi: 10.1111/apa.17573. Epub ahead of print. PMID: 39803879.