Item		Sub-item	Description	Origin of information	Maturity level grade	Maturity level criteria and definitions	Rationale
0	Data base identification	Country	SPAIN		*N/A*	*N/A*	*N/A*
		Data Access Provider	The data owner of the Valencia Health System Integrated Database (VID) is the Valencia regional government. Reseach teams at Fisabio, such as the HSRP Unit, can be considered as providers, as they are granted access to VID data on a project-basis, after 1. ethics committee approval of a research protocol and 2. data commission approval of the data extraction.	https://catalogues.ema.europa.eu/node/1077/administrative-details		*N/A*	
		Organisation type	EU Institution/Body/Agency Not-for-profit			*N/A*	
I	Rationale and scope for the RWD source creation	Primary purpose for which data are collected	Primary purpose of data collection is registry of daily clinical practice in the public healthcare system of the Valencia region. Data can also be used for research under the conditions explained above. Data in VID require expert data analyst review and manipulation, as well as data linkage procedures, before being ready for research purposes.	https://catalogues.ema.europa.eu/institution/3331380; HSRP Unit	*2*	*L1 if information is available as free text and/or online link(s)*	*Relevant for all DQ dimensions (reliability, extensiveness, coherence and timeliness) as it provides a general understandin g of the strengths and limitations of an RWD source.* *Knowing the triggers would ease the understanding of the content and motivations behind the data.*
		Criteria for the selection of the data being collected or integrated	Data are sourced of all general population covered by the universal public health care system in the Health Department of the Valencia Regional Government All data available in the VID (see IJE publication)	https://academic.oup.com/ije/article/49/3/740/5707448 https://catalogues.ema.europa.eu/node/1077/quantitative-descriptors		*L2 if information is available using standardised templates to make information easy to digest and interpret (the EMA recommends to check this tool as reference: REQueST Tool and its vision paper [Internet]. EUnetHTA. 2019. Available from: 721 https://www.eunethta.eu/request-tool-and-its-vision-paper/.*	
		Data prompts	**Event triggering registration of a person in the data source, other:** Any contact with the health system triggers the registration of the person **Event triggering de-registration of a person in the data source:** Death, Emigration, Insurance coverage end **Event triggering creation of a record in the data source:** Most of them are created when a contact with the health system is produced. In other cases, such as pharmacy data, when the prescription is created or when the dispensing is produced. Each table of the data source has their own triggers	https://catalogues.ema.europa.eu/node/1077/data-flows-and-management		*L3 if the information is provided as Metadata (machine readable), including standrad formats, clear definitions and potentially some quality information*	
		Publications describing this RWD	https://academic.oup.com/ije/article/49/3/740/5707448				
II	Data collection or recording process	Description of data provider (geographical and oragnizational setting, nature of the data - reported by patients, HCP, etc)	The Valencia Health System Integrated Database (VID) is a set of multiple, public, population-wide electronic databases for the Valencia Region, the fourth most populated Spanish region, with ~5 million inhabitants. VID provides exhaustive longitudinal information including sociodemographic and administrative data (sex, age, nationality, etc.), clinical (diagnoses, procedures, diagnostic tests, imaging, etc.), pharmaceutical (prescription, dispensation) and healthcare utilization data from hospital care, emergency departments, specialized care (including mental and obstetrics care), primary care and other public health services.	https://academic.oup.com/ije/article/49/3/740/5707448 https://catalogues.ema.europa.eu/node/1077/quantitative-descriptors	*2*	*L1 if information is available as free text and/or online link(s)*	*Essential to understand extensiveness and to assess reliability (that can be affected by errors or biases in the collection process). Also, essential to evaluate SOP for data collection or recording practices that may impact coherence (e.g., where "curation at source" is involved and provide hard constraints for timeliness).*
		Standard Operating Procedures (SOPs) recording	Data owner (Valencia government) and data users (Fisabio) comply with the regulations on the protection of personal data and the guarantee of digital rights, specifically Organic Law 3/2018 of December 5, on the Protection of Personal Data and the Guarantee of Digital Rights, and Regulation (EU) 2016/679 of the European Parliament and of the Council of April 27, 2016 (General Data Protection Regulation – GDPR). Likewise, both organisations are obliged to implement the necessary technical and organizational measures to ensure the security and integrity of personal data and to prevent its alteration, loss, or unauthorized processing or access.	https://www.san.gva.es/es/web/investigacio/solicitud-datos-sia-gaia https://fisabio.san.gva.es/es/registro-de-actividades-de-tratamiento-de-datos/ Provided by DEAP		*L2 if information is available using standardised templates to make information easy to digest and interpret, and also standard vocabularies are available*	
		How SOPs are implemented and monitored	Data extraction: data is gathered from various databases within the VID system and extracted based on a previously approved research protocol by an Institutional Review Board (IRB). Participants: IT technicians from the Valencia Region Health Department. Data cleansing and linkage: extracted data is reviewed (consistency and quality checks) before database linkage. Participants: Data analysts from the Health Services and Policy Research (HSRP) Unit at Fisabio. Security measures: throughout the process, data is pseudonymized and protected with access controls and obfuscation procedures to ensure privacy and regulatory compliance. Participants: IT technicians from the Valencia Region Health Department. Dissemination: research findings are published in peer-reviewed journals, included in the HMA EMA Catalogue, and disseminated via various social media platforms. Participants: Researchers from the HSRP Unit at Fisabio.	Provided by DEAP	*N/A*		

		Key data elements captured (are they always recorded, are they optional, is there a planned coverage over time, ...)	Disease information (information about COVID-19 test in RedMIVA), Rare diseases, Pregnancy and/or neonates, Hospital admission/or discharge, ICU admission, Cause of death, Prescriptions of medicines, Dispensing of medicines, Contraception measures covered by public health system, Indication for use, Medical devices, Administration of vaccines, Procedures, Clinical measurements, Healthcare provider, Patient-generated data, Units of healthcare utilisation, Unique identifier for persons, Diagnostic codes, Medicinal product information (Active ingredient(s), ATC and product codes, INN, Presentation, Dosage regime, Formulation, Package size, Strength, Prescribed duration), Lifestyle factors (alcohol use, tobacco use), Sociodemographic information (age, country of origin, deprivation index, gender, health area, socioeconomic status) Included databanks: - SIP (basic information of VHS coverage and sociodemographic data) - ABUCASIS (ambulatory medical record): includes GAIA (prescription and dispensation) and SIA (diagnoses, history, lab results, lifestyle habits) - ORION (hospital medical record): includes MBDS (diagnoses and procedures, discharge and admission data) and AED (triage data, diagnoses, tests and procedures in public emergency rooms) - CRC (corporate info: physician information and geographical and functional organization of health services) - RedMIVA (results of microbiological analysis) - SIV (vaccination information: type, manufacturer, batch number, number of doses, location and administration date, adverse reactions related to vaccines, rejected vaccinations and, if applicable, risk groups) - CIS (cancer information: incidence, prevalence, tumour site and tumour type) - SIER-CV (epidemiological information on rare diseases: incidence, prevalence, patient characteristics, geographical distribution...): includes the Congenital Anomalies Registry (prevalence of congenital anomalies in the region and the exposure to teratogen agents) - BIMCV (a digital biobank of medical images) All databanks that composes VID (01_SIP, 02_PCV, 03_CEX, 04_MBDS, 05_AED, 06_DIAGNOSES, 07_GAIA, 08_SIV, 09_MDR, 10_PMR, 11_EOS, 12_TESTS, 13_CONG and 14_REDMIVA) have a linking ID number that identifies uniquely each person.	https://catalogues.ema.europa.eu/node/1077/data-flows-and-management#darwin-data-source-linkage https://academic.oup.com/ije/article/49/3/740/5707448 https://catalogues.ema.europa.eu/node/1077/quantitative-descriptors	2	*L3 if additionally SOPs specify KPIs to monitor*	
III	The selection of RWD sources and their onboarding *(Applies to RWD sources that integrate or repurpose other RWD sources)*	Criteria to accept or exclude a datasource	N/A		N/A	*L1 if information about selection criteria or DQ performance is available as free text and/or online link(s)* *L2 if a structure checklist and dataset version control are available* *L3 is only aspirational. N/A*	*When data are provided by a data aggregator, ensure that all the available evidence related to systems and processes potentially affecting DQ (extensiveness and reliability especially) can be followed. Provide information of impact on both reliability and evidence (as well as other dimensions if relative constraints are formulated in inclusion/excl usion (I/E) criteria)*
		Is there a DQ assessment for data sources onboarded?	N/A				
		If yes: does it follow any specific framework? Is there an assessment checklist? Are datasets versions traceable?	N/A				
IV	The data management infrastructure	List of systems used to manage the RWD (either for data collection, recording, processing, etc)	The IJE publication provides full detail of the different systems used to gather RWD in VID.	https://academic.oup.com/ije/article/49/3/740/5707448	*1*	*L1 if information is available as free text and/or online link(s)* *L2 if the hardware or software implementation complies with recognised quality standards that can be reported* *L3 N/A*	*Essential for reliability regarding data alterations resulting from system accidents, software errors or malicious intervention.*
		Software testing and software quality control in place	The HSRP Unit in Fisabio as DAP has measures in place to ensure full traceability and data protection for data handled by the research team. Measures adopted by the data owner (Valencia regional government) are regulated by european and national regulations about data privacy and data protection.	Provided by DEAP			
		Measures to prevent accidental physical data alterations (e.g.: backups, redundant systems, checksums)	The HSRP Unit in Fisabio as data provider has measures in place to ensure full traceability and data protection for data handled by the research team. Measures adopted by the data owner (Valencia regional government) are regulated by european and national regulations about data privacy and data protection.	Provided by DEAP			
V	Data management and governance	Data management principles being followed (e.g., GCP, ISO, FAIR, etc)	Data owner is subjet to european and national regulations on data management.	Provided by DEAP	*2*	*L1 if information is available as free text and/or online link(s)* *L2 if standard best practices are being used and a direct impact on DQ is reported. There are SOPs and dat amanagement processes that adhere to the standards. The representation of metadata follows FAIR standards* *L3 if data management and governance is implemented in the data platforms 'Digital Quality Measures' (DQMs) so that reports of performance and deviations are automated. Submitted metadata are generated "by design". Basically, if everything in L2 is automatised and generated by default*	*Data management and governance impact reliability, as well as all quality dimensions for metadata.*
		Data management processes in place (DQ controls, KPIs, SOPs, etc)	Only in some cases, such as the MBDS and the AED records, are data subject to a consolidation and quality check process before data are available for research	https://academic.oup.com/ije/article/49/3/740/5707448			
		Measures to prevent data alterations by unauthorised parties (cybersecurity)	Data owner implements all safety regulations imposed by legal mandate to ensure data safety. With regard to data managed by HSRP Unit, these are stored in a secure server permanently, and access is tightly restricted only to data analysts and senior researers within the team that work in the project.	https://www.san.gva.es/ca/web/sanidad/clien			
		Auditing and DQ imporvement procedures in place	In February 2025, a public body (Oficina Autonómica de Auditoria e Inspección Sanitaria de la Comunidad Valenciana) was created to audit Valencian Health System. No repots have been published yet.	https://www.gva.es/es/inicio/atencion_ciudadano/buscadores/departamentos/detalle_departamentos?id_dept=28011			
VI	Data manipulation steps	Frequency of data updates	Monthly data updating	https://catalogues.ema.europa.eu/node/1077/data-flows-and-management	*1*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*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*

		Data transformations performed, data mapping steps, data cleaning	MBDS and AED records are reviewed by data coding specialists to improve coding accuracy and data quality. In the context of a research project, in the case of HRSP Unit all steps with regard to data cleaning etc are tracked and stored safely in the internal server.	https://academic.oup.com/ije/article/49/3/740/5707448		*L2 if Tests performed follow some standard or shared set of tests, that can be re-used across RWD sources. Key performance indicators (KPIs) for data cleaning (e.g., data duplications, mislabelling, etc.) are provided. Data mapping tables and algorithms are described with a standard characterisation of their performance. Lists of standard test batteries used to detect loss of accuracy or precision are provided. All lineage information is provided as metadata associated to the dataset*	*impacts coherence and potentially timeliness.*
		Information about loss of precision during data manipulation steps	In the context of data for a specific research project, HSRP Unit has measures in place to ensure full traceability and data protection.	Provided by DEAP		*L3 if information about data onboarding is directly provided by the platform, e.g.: * Transaction logs are available including deviations and actions that required manual intervention Actual data transformation code is accessible and verifiable. Quality checks and KPIs reported are automatically generated by the data platform (e.g.: unit testing) Lineage information is automatically generated by the processing platform*	
		Lineage information (e.g., justifcation of data manipulation, track of changes and versions)	In the context of data for a specific research project, HSRP Unit has measures in place to ensure full traceability and data protection.	Provided by DEAP			
VII	Data augmentation steps (e.g., imputation or linkage)	Is any augmentation happening in this datasource?	All the data tables above stated belong to the core VID database. Father-child linkage is perfomed.	https://catalogues.ema.europa.eu/node/1077/data-flows-and-management#darwin-data-source-linkage Provided by DEAP	*N/A*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*Data augmentation steps impact accuracy (reliability) and extensiveness. We consider here data transformations that produce new information subject to reliability issues: e.g.: imputation of missing values, or extraction of codes via natural language processing.*
		If yes, which are the methods applied	Probabilistic	Provided by DEAP		*L2 if algorithms are published and their performance documented. Information on which values result from imputation is provided as part of the dataset (e.g., presented in metadata or a data dictionary)* *L3 if an automatised process for data linkage/mapping exists*	
		If yes, which algorithms and assumptions applied	Father-child or family unit linkage is based on residency address.	Provided by DEAP			
		If yes, which is the error rate when conducting the augmentation	Unknown	Provided by DEAP			
VIII	Known quality issues and independent QA assessment of the RWD source	Known DQ issues (e.g., poor overall completeness in Q3 2020 due to COVID-19)	Incompleteness of early data from AED records or coding reliability of diagnostic information in the EMR Different datasets cover different periods (ABUCASIS from 2009, ORION from 2008, AED reliable since 2017, RedMIVA from 2008, SIV reliable since 2005, CIS from 2004, SIER-CV reliable since 2012) Lacking data on in-hospital pharmaceutical prescription (pending to be integrated as part of the ORION information system)	https://academic.oup.com/ije/article/49/3/740/5707448	*2*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*Explicit description of known DQ issues, as well as external validation performed (all dimensions affected)*
		Validation studies and publications resulting from this RWD source	There are many publications using VID data.			*L2 if standard procedures are set for external/internal validation of the data* *L3 if the mechanism provided includes notification of automatically detected DQ issues*	
IX	The RWD source representation	Description of data model or models used (OMOP, FHIR, ...)	ConcePTION, OMOP	https://catalogues.ema.europa.eu/sites/default/files/cdm-etl-spec/0_3_VID_Catalogue_RTL_specifications	*3*	*L1 if free-text information, links or publications are available reporting all the mentioned features* *L2 if the description refers to a model such as OMOP, I2B2, FHIR, others, or an extension of them. Data dictionaries are standard (and if non-standard, jusified why)* *L3 if a standard CDM is used, the datasource has been mapped to one or more than one CDM, and if data dictionaries are provided using standard formats that facilitate the mappings across different vocabularies and across languages*	*Descriptive of the intended coherence DQ of a dataset and its metadata.*
		Data ontology (dictionaries and vocabularies) being used, and if in standard formats that allow mapping across different languages (e.g., UMLS)	Cause of death: ICD-10-CM (ICD10-ES or Spanish clinical modification) Indication: ICD-10-CM / ICD-9-CM Procedures: ICD-10-CM / ICD-9-CM Diagnosis / medical event vocabulary: ICD-10-CM / ICD-9-CM Prescriptions of medications: ATC Dispensing of medicines: ATC Medicinal product: ATC / Other Oncology: ICD-O3	https://catalogues.ema.europa.eu/sites/default/files/cdm-etl-spec/0_3_VID_Catalogue_RTL_specifications_0.pdf			
X	The RWD source declared Service Level Agreements (SLA)	Guaranteed frequency of updates and incident response time (e.g., corrections in case of errors)	Defined on a reserch contract-basis, corrections in the case of errors is a common procedure.	*Provided by DEAP*	*1*	*L1 if free-text information and links are available reporting all the mentioned features*	*Descriptive of guaranteed timeliness and possible variations of extensivenes s/reliability provided.*
		Processes and resources accompanying the data, such as documentation, training materials or help desk contact	N/A	*N/A*	*N/A*	*L2 if details of established data processess by the provider are available*	
		Possibility to collect additional data if needed	VID data is linkable to a set of databases, as defined in the IJE reference.	*Provided by DEAP*	*1*	*L3 if SLA compliance is assessed and reported automatically*	
XI	The RWD source licensing and restrictions	Data use agreements that may limit data use or access (consent, limitations of use), accessibility policies, licensing constraints, standard policies of use, data retention	Ethics approval by an accredited ethical research committee is required to access the data for research purposes. After that, the regional data commission has to aprove the data extraction. Access to data for researchers has no financial cost but is covered by research ethics and authorization processes.	https://academic.oup.com/ije/article/49/3/740/5707448	*2*	*L1 if free-text information and links are available reporting all the mentioned features* *L2 if policies and licensing are standardised to a broad range of RWD* *L3 NA*	*Descriptive of aspects that can limit extensiveness and coherence in downstream data aggregations.*
XII	Feedback	Is there a data ecosystem in place so that quality assessment by data consumers can provide feedback to improve the data collection and production process, thus allowing a continuous monitoring and improvement of DQ?	No.		*N/A*	*L1 if a person of contact is provided for Q&A* *L2 if the contact provided allows tracking of issues and follow-up* *L3 if the mechanism provided includes notification of automatically detected DQ issues*	*Descriptive of feedback mechanisms in place to improve all aspects of DQ*

Dimension	Sub-dimension	Metrics	Description	Origin of information
Timeliness	Currency	How often is the database updated (i.e., frequency of updates)	Data banks are updated daily according to clinical practice. MBDS and AED are updated every 6 months	https://zenodo.org/records/13384860
		The time gap between the latest available data and date when data is delivered to user. (i.e., how up-to-date data are when it reach the user)	1 day to 6 months (depending on the data bank) plus the lag of delivery	Provided by DEAP
		The time elapsed from when a user requests the data to when they actually receive it	Between 8 and 14 months	Provided by DEAP
		Median time (years) between first and last available records for unique individuals	12 years	https://catalogues.ema.europa.eu/node/1077/quantitative-descriptors
Extensiveness	Coverage	Percentage of a target population present in a database	Aproximately 98% of the 5 million inhabitants of the region of Valencia, with an annual birth cohort of 48000 newborns, representing 10.7% of the Spanish population and around 1% of the European population	https://academic.oup.com/ije/article/49/3/740/5707448 https://zenodo.org/records/13384860
	Completness	% of subjects in the data with a recorded birth date	100%	
		% of subjects in the data, irrespective of vital status, that have a recorded date of death	A date of death is recorded for 100% of individuals who are known to have died	https://zenodo.org/records/13384860
		% of subjects in the data with a record of sex	100%	https://zenodo.org/records/13384860
		% of subjects in the data who had an event with a code for the event	100%	https://zenodo.org/records/13384860
		% of subjects in the data who had a prescription/dispensing with a recorded code for the medicine	ATC code (100%), MPID (100%)	https://zenodo.org/records/13384860
		% of subjects in the data who got vaccinated with a recorded code for the vaccine	From the total of individuals known to have been vaccinated, 100% had the vaccine batch number recorded and 100% had the vaccine type available	https://zenodo.org/records/13384860
Reliability	Accuracy	The population distribution in the data source aligns with that of the country	Population age distribution are aligned with a developed country demographics reported by the National Statistics Institute (INE). To bear in mind, information about people with no contact with the healthcare system or attending the private health sector is not represented. Active population size: Paediatric Population (< 18 years): 461000 (9.6%) Preterm newborn infants (0 – 27 days): 2700 (0.1%) Term newborn infants (0 – 27 days): 33000 (0.7%) Infants and toddlers (28 days – 23 months): 99000 (2.1%) Children (2 to < 12 years): 521000 (10.9%) Adolescents (12 to < 18 years): 263000 (5.5%) Adults (18 to < 46 years): 679000 (14.2%) Adults (46 to < 65 years): 748000 (15.6%) Elderly (≥ 65 years): 996000 (20.8%) Adults (65 to < 75 years): 517000 (10.8%) Adults (75 to < 85 years): 332000 (6.9%) Adults (85 years and over): 147000 (3.1%)	https://zenodo.org/records/13384860 https://www.ine.es/ García-Sempere A, Orrico-Sánchez A, Muñoz-Quiles C, Hurtado I, Peiró S, Sanfélix-Gimeno G, Diez-Domingo J. Data Resource Profile: The Valencia Health System Integrated Database (VID). Int J Epidemiol. 2020 Jun 1;49(3):740-741e. doi: 10.1093/ije/dyz266. PMID: 31977043; PMCID: PMC7394961.
		Records of diagnostics, exposures or medical observations that do not agree with common expectations and knowledge or feasible ranges (e.g., pregnancy records in males, a human with 4 arms, systolic pressure higher than 250mmHg, etc)	Requested to DEAP and unable to provide	
		Records of healthcare events (diagnoses, prescriptions, admissions, etc) with logical inconsistencies (e.g., and admission occurs after death)	Data values before birth: 0-0.1% Data values after death: 0-0%	https://zenodo.org/records/13384860
		Variables that are based in imputation, derivation or inference (e.g., end of treatment date is derived from treatment start date and treatment cycle length)	In VID no imputation, derivation or inference is performed unless required for a specific project.	https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2023.1207976/full Provided by DEAP
	Precision	Exposures codes precision level, including medicines and vaccines (e.g., active principle, therapeutic group, ...)	Active principle (ATC level 5) and national product codes are available.	Provided by DEAP
		Precision of date of birth (e.g., day, month, year)	Day, month and year	Provided by DEAP
		Precision of date of death (e.g., day, month, year)	Day, month and year	Provided by DEAP
		Precision of date of the event/diagnosis (e.g., day, month, year)	Day, month and year	https://zenodo.org/records/13384860
		Precision of date of the exposure (e.g., day, month, year)	Day, month and year	Provided by DEAP
	Traceability	Provenance of event records	Primary care, Emergency, Hospital, Specialist and ICU	https://zenodo.org/records/13384860
		Provenance of medicines/vaccines records	Prescription, dispensation, vaccine information system	https://catalogues.ema.europa.eu/node/1077/administrative-details
Coherence	Format coherence	For dates, formatting constraint being followed	yyyy/mm/dd Uncertain variable format and length.	Provided by DEAP
		For sex, formatting constraint being followed	STRING 1 character, M (male), F (female)	Provided by DEAP
	Relational coherence	% of records with the Person ID in the PERSONS table	100%. This is controlled at extraction. Data must have all Person IDs in their persons table to be used for a study.	https://zenodo.org/records/13384860
	Semantic coherence - to determine	For EVENTS definitions, codelists/data dictionaries being employed according to external standards	ICD -10-CM (The ICD-10-CM used is the ICD10-ES (Spanish clinical modification), ICD-9-CM,	Provided by DEAP
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	ATC code (100%), MPID (0%) MIPD (100%)	Provided by DEAP
	Uniqueness	Number of records flagged as potential duplicates	Requested to DEAP and unable to provide	

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

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

	Sociodemographic features: age, obesity, smoking	Date of birth Sex BMI (patient with BMI > 30) Smoking habits	*Low*	Date of birth and sex: 100% of individuals have available information. Smoking can be captured via ICD codes or a specific variable in VID tobacco use, BMI can be caputred via a BMI variable in VID or can be calculated with weight and height data				
	Diagnoses: substance abuse, alcohol abuse, rubella, varicella, toxoplasmosis, herpes simplex virus, cytomegalovirus, diabetes & gestional diabetes, folate deficiency	Diagnostic code Date of diagnosis	*Low*	Diagnostic codes: 100% of individuals have available information		Diagnoses and drugs follow UMLS ontologies.		
	Father (for neurodevelopmental disorders)							
	Sociodemographic features: age, calendar year of conception of offspring	Date of birth Sex Date of conception/Date of delivery	*Low*	Available in VID. In regards to conception date, it is calcualted using gestational age at birth estimated by ultrasound (available in most live births), or otherwise using date of LMP.	Given that gestational age as estimated by ultrasound is highly reliable and it is available in most of livebirths, date of conception can be considered reliable in VID.	VID has experience with running pregnancy algorithms. VID has assembled a pregnancy cohort (PREGVAL), including over 500.000 pregnancies.		https://link.springer.com/article/10.1007/s10654-025-01260-7 https://github.com/ARS-toscana/ConcePTIONAlgorithmPregnancies
	Diagnoses: substance abuse, alcohol abuse, affective disorder, schizophrenia (including schizotypal and delusional disorders), neurotic disorder, neurodevelopmental disorder	Diagnostic code Date of diagnosis	*Low*	Diagnostic codes: 100% of individuals have available information		Diagnoses and drugs follow UMLS ontologies.		
	Medications: any concomitant medications associated with valproate-indicated psychiatric conditions, any concomitant medications associated with neuropsychiatric adverse effects	Medication code Date of prescription/dispensing	*Low*	Medication: ATC code 100% of individuals have available information, except for inpatient medication data, not available in VID.		Diagnoses and drugs follow UMLS ontologies.		
	Father (for congenital anomalies)							
	Sociodemographic features: age, calendar year of conception of offspring	Date of birth Sex Date of conception/Date of delivery	*Low*	Available in VID. In regards to conception date, it might be derived from the LMP, date of typical trimestral birth control tests, or date of delivery.	Conception date might be derived from the LMP, date of typical trimestral birth control tests, or date of delivery. Reliability may depend on the inferences that need to be made depending on the information available.	VID has experience with running with pregnancy algorithms ongoing validation.		https://github.com/ARS-toscana/ConcePTIONAlgorithmPregnancies
	Offspring (for neurodevelopmental disorders)							
	Sociodemographic features: sex	Sex	*Low*	Available in VID				
	Diagnoses: foetal alcohol syndrome, fragile X syndrome, congenital cytomegalovirus, lejeune/cri du chat syndrome, tuberous sclerosis	Diagnostic code Date of diagnosis	*Low*	Diagnostic codes: 100% of individuals have available information		Diagnoses and drugs follow UMLS ontologies.		
	Offspring (for congenital anomalies)							
	Diagnoses: foetal alcohol syndrome, congenital rubella, congenital varicella, congenital cytomegalovirus, congenital herpes syndrome, congenital toxoplasmosis	Diagnostic code Date of diagnosis	*Low*	Diagnostic codes: 100% of individuals have available information		Diagnoses and drugs follow UMLS ontologies.		
Intercurrent events	Treatment discontinuation less or more than 3 months prior to conception	Medication code Date of prescription/dispensing Date of discontinuation Treatment duration Date of conception/Date of delivery	*Low*	Medication: ATC code 100% of individuals have available information, except for inpatient medication data, not available in VID. Duration of exposure may be estimated using dispensation data and prescription information on dosing schedule.	Prescription and dispensation are only indirect date of stopping intervention and may suffer imprecisions (a patient may decide not to take usual medication despite persistente of prescription and dispensation)	Diagnoses and drugs follow UMLS ontologies.		

No conception	Lack of: Date of conception/Date of delivery Pregnancy diagnosis Male-female partner linkage	*High*	Linkage of males with a partner when they have not conceived is not possible since such linkage is made through livebirths.	Deterministic linkage is used to obtain mother-child pairs, so deemed to be highly reliable. Father-child linkage was obtained using a deterministic approach, with direct and indirect linkage. Direct linkage was performed linking children with their male health cardholder or legal guardian, which had to comply with other criteria such as being 15y or older at the time of birth of the index child and their surnames should be compatible. Then we used household address combined with sex, age and surnames criteria, excluding households where more than one male were retrieved. It is only possible for live births to be linked with the father. So, only females experiencing a pregnancy ending in live birth conception can be linked to the male partner.		This linkage is available from 2010 to 2024	https://www.eurolinkcat.eu/loadFile.aspx?filename=D2.4%20report%20submitted(2).pdf https://catalogues.ema.europa.eu/node/1077/quantitative-descriptors
Conception within 3 months post-treatment initiation	Medication code Date of prescription/dispensing Date of discontinuation Treatment duration Date of conception/Date of delivery	*Low*	Medication: ATC code 100% of individuals have available information, except for inpatient medication data, not available in VID. Duration of exposure may be estimated using dispensation data and prescription information on dosing schedule.		Diagnoses and drugs follow UMLS ontologies.		
Still birth, spontaneous abortion or congenital malformation	Diagnostic code Date of diagnosis Linkage father-child	*High*	Diagnostic codes: 100% of individuals have available information Around 67% of livebirths are linked to a father. All congenital malformation codes are available for livebirths. Stillbirths and spontaeous abortions can not be linked to the fathers.	Deterministic linkage is used to obtain mother-child pairs, so deemed to be highly reliable. Father-child linkage was obtained using a deterministic approach, with direct and indirect linkage. Direct linkage was performed linking children with their male health cardholder or legal guardian, which had to comply with other criteria such as being 15y or older at the time of birth of the index child and their surnames should be compatible. Then we used household address combined with sex, age and surnames criteria, excluding households where more than one male were retrieved. The method is based several on healthcare card holders, residency address, so its reliability should be interpreted with caution. Also, it is only possible for live births conceptions, so non-live conceptions (i.e., spontaneous abortions, still births) cannot be linked with the father. So, only females	Diagnoses and drugs follow UMLS ontologies.	This linkage is available from 2010 to 2024	
Switch to another anti-seizure drug prior to conception	Medication code Date of prescription/dispensing Date of discontinuation Treatment duration Date of conception/Date of delivery	*Low*	"Switching" may be operationalised using prescription and dispensing data. Inpatient medication data is not available in VID.		Diagnoses and drugs follow UMLS ontologies.		

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

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

Case study	RWD source	Sample size estimation form the hypotetical trial protocol	Feasibility assessment *(yes/yes, with limitations/no)*	Rationale for the feasibility assessmernt	Limitations identified during the feasibility assessment and categorisation	Description of potential impact of the identified limitations on the study results
7 (Paternal exposure to valproate and the risk of neurodevelopment disorders in offspring)	**VID**	With an approximate cohort estimated sample of 2,574 children (based on a 1:1 ratio between exposure groups), and considering that the Valencia Integrated Database (VID) covers approximately 98% of the 5 million inhabitants of the Valencia region—representing 10.7% of the Spanish population and around 1% of the European population—with an annual birth cohort of approximately 48,000 newborns, considering that that I) the incidence of epilepsy is of 37,7 cases every 100,000 inhabitants, II) in 2023 levetiracetam represented the 21% of DHD of antiepileptic medicines and valproate the 13%, III) the 15% of no conception and IV) that linkage with the father is possible for the 67% of livebirths, the **sample size might be challenging to achieve**. In 2023, valproate use was estimated at 1.7 DHD (defined daily doses per 1,000 inhabitants per day). [1-5]	Yes, with limitations on~~,~~ sample size and elements of high criticality.	Most elements with high criticality are available and fairly reliable. However, the approach or methods to tackle **intention to conceive, no conception, female partner-related selection criteria and some outcomes (i.e., stillbirth, spontaneous abortions)** should be accounted for. Data recency of 6 months before extraction, reasonably enough for the research question. The time elapsed from when a user requests the data to when they actually receive it is usually from 6 to 12 months, but in this case VID will have the data available at the time the analysis is expected. **Sample size is can be challenging to achieve**, so the precision it enables should be recalculated. Children linked to their fathers can be followed from 2010 to 2025, entailing a maximum follow-up of 15 years. One year of look-back will be available.	·Potentially major: Intention to conceive, or no conception cannot be fairly assessed in RWD sources in general. Proxies might be considered. Also, as the father-child linkage is performed through livebirths, stillbirth and spontaneous abortions cannot be assessed for offsprings with fathers receiving the exposures of interest. Female partner-related selection criteria (such as not having a diagnosis of epilepsy), will only be possible for fathers who had livebirths. Approaches or methods to tackle these should be accounted for. ·Potentially major: As the sample size is challenging to achieve, statistical precision should be recalculated. ·Minor: Father-child linkage is obtained using a deterministic approach-the 67% of livebirths linked to the mother. Reliability should be interpreted with caution. ·Minor: Children linked to their fathers can be followed from 2010 to 2025, entailing a maximum follow-up of 15 years. One year of look-back will be available. ·Minor: Inpatient medication not available. ·Minor: Duration of exposure may be estimated using dispensation data and prescription information on dosing schedule. ·Minor: Prescription and dispensation need to be used to calculate indirectly the date of stopping or switching intervention and may suffer imprecisions. ·Minor: If the particular lifestyle interventions for this research question have a codelist, they might be captured. ·Minor: The criteria "Female partner must not be diagnosed with generalised epilepsy" cannot be assessed at the time of randomisation. This can be assessed when having the newborn.	Some misclassification of father-child linkage is possible, and the linkage is available for 67% of livebirths. Cases of spontanoues abortion or stillbirth cannot be linked with the father, which may underestimate the key endpoint or misclassify the outcome. Partner-related information will be assessed retrospectively once linked by means the liveborn. We do not expect a relevant impact in this regard. Only females experiencing a live birth conception will be possible to be linked to the male partner. Additionally, intention to conceive is not direcvtly available in RWD sources. So, items like "Participants must have a female partner with which they intend to conceive" or "No conception" are not available. Children linked to their fathers can be followed from 2010 to 2025, entailing a maximum follow-up of 15 years. The impact is expected to be low as neurodevelopmental disorders usually happens (from 18 months for autism, and from 4-5 years for ADHD) and this can be handled with the appropriate analysis (e.g., survival analysis). Sample size may not achieve sufficient precision for regulatory purposes, but reflects the reality of using RWD sources. Precision can be lower than calculated initially.

REFERENCES

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

[2] Quintana M, Sánchez-López J, Mazuela G, Santamarina E, Abraira L, Fonseca E, Seijo I, Álvarez-Sabin J, Toledo M. Incidence and mortality in adults with epilepsy in northern Spain. Acta Neurol Scand. 2021 Jan;143(1):27-33. doi: 10.1111/ane.13349. Epub 2020 Oct 13. PMID: 32969054.

[3] Villanueva V, Carreño M, Gil-Nagel A, Serrano-Castro PJ, Serratosa JM, Toledo M, Álvarez-Barón E, Gil A, Subías-Labazuy S. Identifying key unmet needs and value drivers in the treatment of focal-onset seizures (FOS) in patients with drug-resistant epilepsy (DRE) in Spain through Multi-Criteria Decision Analysis (MCDA). Epilepsy Behav. 2021 Sep;122:108222. doi: 10.1016/j.yebeh.2021.108222. Epub 2021 Aug 6. PMID: 34371462.

[4] https://www.ncbi.nlm.nih.gov/books/NBK546611/

[5] https://app.powerbi.com/view?r=eyJrIjoiNjY1NzVhZjAtYWNmNS00ZTllLTgyNDEtN2E3MGQ5ZTNkZTNmIiwidCI6IjJkM2I1MGUwLTZlZjQtNGViYy05MjQ2LTdkMWNiYjc3MDg5YyIsImMiOjh9