Item		Sub-item	Description	Origin of information	Maturity level grade	Maturity level criteria and definitions	Rationale
0	Data base identification	Country	the Netherlands		*N/A*	*N/A*	*N/A*
		Data Access Provider	Netherlands Comprehensive Cancer Organisation (IKNL)			*N/A*	
		Organisation type	Quality institute for oncological and palliative research and practice. National, regional, or municipal public founding	https://iknl.nl/en/about-iknl		*N/A*	
I	Rationale and scope for the RWD source creation	Primary purpose for which data are collected	The main goal of the Netherlands Comprehensive Cancer Organisation (IKNL) is to reduce the impact of cancer, from the personal to the societal level. With the Netherlands Cancer Registry (NCR) as its core activity, IKNL enables health care professionals, researchers, policy makers and others to reflect on cancer and on palliative care. Together with care professionals, researchers, patients, and policy makers we translate data into valuable insights to improve oncological and palliative care.	https://iknl.nl/en/about-iknl	*1*	*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	The NCR compiles clinical data of all individuals newly diagnosed with cancer in the Netherlands. Hospital inpatient care, hospital outpatient care.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP		*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	Having performed a biopsy through PALGA (the national pathology database) Having a cancer diagnosis in LBZ **Event triggering registration of a person in the data source**: having performed a biopsy through PALGA (the national pathology database) and having a cancer diagnosis in LBZ **Event triggering de-registration of a person in the data source**: Persons are de-registered when they request this (they work with an opt-out system so everyone is included, and everyone can request to be taken out of the database), when they emigrate or die. **Event triggering creation of a record in the data source:** A group of data managers daily screen for new information of the patients registered in the data source. Persons (or actually tumors) are triggered as described in "event triggering registration of a person in the data source". We then register the relevant data for this person (tumor) after a set amount of time (typically 6-12 months after diagnosis). When the person develops another primary tumor, they go through the same process for that new tumor. Registration of patients is typically done about a year after the incidence date (the exact lag depends on the type of cancer). The vital status of patients is checked once per year.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP		*L3 if the information is provided as Metadata (machine readable), including standrad formats, clear definitions and potentially some quality information*	
		Publications describing this RWD	Not found PubMed, Google free search				

II	Data collection or recording process	Description of data provider (geographical and oragnizational setting, nature of the data - reported by patients, HCP, etc)	IKNL is a knowledge institute that is mostly government funded (by the Ministry of Health, Welfare and Sport)	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP	*1*	*L1 if information is available as free text and/or online link(s)* *L2 if information is available using standardised templates to make information easy to digest and interpret, and also standard vocabularies are available* *L3 if additionally SOPs specify KPIs to monitor*	*Essential to understand extensiveness and to assess reliability (that can be affected by errors or biases in the collection process). Also, essential to evaluate SOP for data collection or recording practices that may impact coherence (e.g., where "curation at source" is involved and provide hard constraints for timeliness).*
		Standard Operating Procedures (SOPs) recording	Data is collected by well-trained data managers using coding manuals. The data entry application performs checks on the data that is entered, automatic checks are done on the database, as well as manual checks of random samples. A group of data managers is responsible for data quality and researchers in the organization can flag potential quality issues.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP			
		How SOPs are implemented and monitored	Data managers receive regular training. See also 'SOPs recording'. SOPs are kept up to date by discussing the items in the IKNL tumor boards, they are implemented in daily practice by regular training of the data managers (there is a training program during onboarding and frequent data managers meetings thereafter, data managers are also informed about changes in the SOPs by email. The most recent version of the SOPs are available online and are used during registration. Quality checks also include cross checks by direct colleagues.	Provided by DEAP			
		Key data elements captured (are they always recorded, are they optional, is there a planned coverage over time, ...)	Disease information, rare diseases, prescriptions of medicines, indication for use, procedures, clinical measurements, patinet-reported outcomes, unique identifier persons, diagnostic code, medicinal product information, quality of life measurements, sociodemographic information (age, gender). These are items grouped by: patient, tumor and treatment. I removed the PROMs/QoL as they are only captured on project basis. Coverage over time: there is a planned delay of 9 months delay (as data managers only have to access the EHR once per patient to capture the primary treatment plan), but in practice this is 1 to 2 years. Note: NCR only covers the primary plan, there is no information on follow-up (like PFS or secondary treatment -> I see this is addressed in row 31). The day of death is known by linkage to CBS. TNM recorded	Provided by DEAP (Always recorded: overview on 240918-itemset-long.pdf Unfortunately only in Dutch.)			
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	NA		*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/exclusion (I/E) criteria)*
		Is there a DQ assessment for data sources onboarded?	NA				
		If yes: does it follow any specific framework? Is there an assessment checklist? Are datasets versions traceable?	NA				
IV	The data management infrastructure	List of systems used to manage the RWD (either for data collection, recording, processing, etc)	Data is collected manually from the EMR systems of the hospitals by data managers and entered into a database. IKNL uses its own application for this (RANK). The changes in the database are loaded into a datawarehouse (DWH) every night. The DWH keept a history of the registration, performs transformations on the data, etc.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP	*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*	*Essential for reliability regarding data alterations resulting from system accidents, software errors or malicious intervention.*
		Software testing and software quality control in place	RANK is developed and maintained by the in-house Software Development department. They perform testing and quality control as well. The same applies to the DWH. This is based on a commercial application (from Microsoft).	Provided by DEAP			

		Measures to prevent accidental physical data alterations (e.g.: backups, redundant systems, checksums)	A history of the data is kept in a DWH. Backups are made as well each day.	Provided by DEAP		*L3 NA*	
V	Data management and governance	Data management principles being followed (e.g., GCP, ISO, FAIR, etc)	There are ongoing activities to make the NCR more FAIR. For example through introduction/use of (more) international standards (such as ATC, ICHI, SNOMED-CT).	Provided by DEAP	*1*	*L1 if information is available as free text and/or online link(s)*	*Data management and governance impact reliability, as well as all quality dimensions for metadata.*
		Data management processes in place (DQ controls, KPIs, SOPs, etc)	Data is collected by well-trained data managers using coding manuals. The data entry application performs checks on the data that is entered, automatic checks are done on the database, as well as manual checks of random samples. A group of data managers is responsible for data quality and researchers in the organization can flag potential quality issues.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP		*L2 if standard best practices are being used and a direct impact on DQ is reported. There are SOPs and dat amanagement processes that adhere to the standards. The representation of metadata follows FAIR standards*	
		Measures to prevent data alterations by unauthorised parties (cybersecurity)	IKNL has an IT department that is responsible for cyber security. There are also Information Security Officers that monitor this.	Provided by DEAP			
		Auditing and DQ imporvement procedures in place	IKNL is NEN-7510 certified. Quarterly internal audits are performed, as well as regular external audits. There is also a working group responsible for DQ. They perform checks on the data. Researchers can also signal potential DQ issues.	Provided by DEAP		*L3 if data management and governance is implemented in the data platforms 'Digital Quality Measures' (DQMs) so that reports of performance and deviations are automated. Submitted metadata are generated "by design". Basically, if everything in L2 is automatised and generated by default*	
VI	Data manipulation steps	Frequency of data updates	The data in the NCR (i.e. the data in the DWH) is updated daily (by overnight loading of the changes in the RANK database). Disease diagnosis: daily through PALGA (the national pathology database); remaining patients (those that do not receive a biopsy) are found by a yearly coupling with LBZ (https://www.dhd.nl/producten-diensten/registratie-data/ontdek-de-mogelijkheden-van-de-lbz). Other data: 6 months-1 year (depending on tumor type) after identification of a new primary tumor, a datamanager collects additional data around diagnosis and treatment from the EHRs of the hospitals where the patient was diagnosed and treated. The data in the OMOP-CDM is updated a few times per year.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP	*1*	*L1 if free-text information, links or publications are available reporting all the mentioned features*	*Impacts reliability both in terms of accuracy (possible errors) and precision (i.e., the degree of approximatio n by which data represents reality). Essential to ensure traceability of information. Also impacts coherence and potentially timeliness.*
		Data transformations performed, data mapping steps, data cleaning	Data is registered manually by highly trained data managers so data that is entered into the NCR is already of high qualitym so cleaning is not required. Data transformations performed in the DWH are mainly creation of new variables (derived from existing NCR variable) and handling of changes of variables (or definition of variables) over time. A team consisting of data base administrators and experts on the NCR data do this in close collaboration with the clinical experts (tumor teams).	Provided by DEAP		*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*	
		Information about loss of precision during data manipulation steps	In general, there are no manipulation steps that cause loss of precision. There may be loss of precision (loss of information) in the creation of specific variables. However, the original variables are also part of the NCR.	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.*	
		Lineage information (e.g., justifcation of data manipulation, track of changes and versions)	A history of the DWH is maintained. Data manipulation that is performed by scripts has an accompanying justification.	Provided by DEAP			
VII	Data augmentation steps (e.g., imputation or linkage)	Is any augmentation happening in this datasource?	There are no data augmentation steps.	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,*
		If yes, which are the methods applied	N/A			*L2 if algorithms are published and their performance documented. Information on*	
		If yes, which algorithms and assumptions applied	N/A				

		If yes, which is the error rate when conducting the augmentation	N/A			*L3 if an automatised process for data linkage/mapping exists*	*or extraction of codes via natural language processing.*
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)	Cause of death not included. Comorbidity and cardiac events. Inclusion of only first line treatments. No data set has all information about a patient, you always need to make choices about what to collect. There is no cause of death and only first line treatment is registered. There is no registration of side effects and data about comorbidities is very incomplete. There are other minor quality issues in variables (sometimes variables are only registered in certain regions, or registration was nog mandatory in certain time periods so the data is incomplete). The department that handles data request knows these issues and they are communicated with the person that requests the data if this affects their research. There is no single overview of these DQ issues. NO information on recurrences	Provided by DEAP	*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	Possibility of data validation	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP		*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, ETL completed. IKNL uses its own data model for the NCR. Data deliveries to researchers are usually done as a csv file with accompanying data dictionary. Part of the NCR is also available in the OMOP-CDM. The data in the OMOP-CDM is updated a few times per year.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP	*3*	*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 (e.g., UMLS)	Prescription: ATC level 5, own vocabulary; Indication: ICD-O-3; Procedures vocabulary: own vocabulary; Diagnosis/medical event vocabulary: ICD-O-3. Starting: TNM	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP		*L2 if the description refers to a model such as OMOP, I2B2, FHIR, others, or an extension of them. Data dictionaries are standard (and if non-standard, justified why)* *L3 if a standard CDM is used, the datasource has been mapped to one or more than one CDM, and if data dictionaries are provided using standard formats that facilitate the mappings across different vocabularies and across languages*	
X	The RWD source declared Service Level Agreements (SLA)	Guaranteed frequency of updates and incident response time (e.g., corrections in case of errors)	Updates of the data are daily and occur in-house. Any issues with this can (and will) be handled immediately (during work hours).	Provided by DEAP	*2*	*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	Data deliveries are accompanied by a data dictionary. Data request are handled by a specialist at IKNL. They are available for additional questions about the data. There is also a general e-mail address for these questions.	Provided by DEAP		*L2 if details of established data processess by the provider are available*	
		Possibility to collect additional data if needed	Additional data can be collected by the data managers if there is additional funding available. There is a fee involved with this.	Provided by DEAP		*L3 if SLA compliance is assessed and reported automatically*	
XI	The RWD source licensing and restrictions	Data use agreements that may limit data use or access (consent, limitations of use), accessibility policies, licensing constraints, standard policies of use, data retention	Acces to data trough an application form https://iknl.nl/en/ncr/apply-for-data	https://iknl.nl/en/ncr-data	*1*	*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 group of data managers is responsible for data quality and researchers in the organization can flag potential quality issues.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP/ DEAP	*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)	NCR updates are daily. However, data is registered 6-12 months after diagnosis so there is a lag there. Vital status is indeed checked once per year.	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP/ DAP
		The time gap between the latest available data and date when data is delivered to user. (i.e., how up-to-date data are when it reach the user)	1 to 2 years, as data managers only have access to EHR once per patient to capture the primary treatment plan	Provided by DEAP
		The time elapsed from when a user requests the data to when they actually receive it	~2 months	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP/ DAP
		Median time (years) between first and last available records for unique individuals	0.7 years	https://catalogues.ema.europa.eu/node/952/quantitative-descriptors
Extensiveness	Coverage	Percentage of a target population present in a database	>95% coverage of the total population in The Netherlands. >= 18 y Population size: 3,677,269	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP
	Completness	% of subjects in the data with a recorded birth date	100%	Provided by DEAP
		% of subjects in the data, irrespective of vital status, that have a recorded date of death	A date of death is recorded for 100% of individuals who are known to have died	Provided by DEAP
		% of subjects in the data with a record of sex	100%	Provided by DEAP
		% of subjects in the data who had an event with a code for the event	100%	Provided by DEAP
		% of subjects in the data who had a prescription/dispensing with a recorded code for the medicine	99.27% of registered chemotherapies have an ATC code. 0.73% of registered chemotherapies are either coded as "intensive chemotherapy" (the majority, mainly for hematology) or "trial medication" (for all cancer patients in the past 5 years)	Provided by DEAP
		% of subjects in the data who got vaccinated with a recorded code for the vaccine	None	Provided by DEAP
Reliability	Accuracy	The population distribution in the data source aligns with that of the country	As NCR is a disease registry, it reflects only the population affected by colorectal cancer, rather than the general population of the country. Yearly participants to the registry: -2019: 1,574,506 -2020: 1,338,052 -2021: 1,632,493	https://iknl.nl/getmedia/046b1a45-b673-4117-9320-9fd8e1823bd6/Monitor-darmkanker-2021-UK-definitieve-versie.pdf https://www.rivm.nl/sites/default/files/2021-10/Monitor_bevolkingsonderzoek_darmkanker_2020_eng.pdf
		Records of diagnostics, exposures or medical observations that do not agree with common expectations and knowledge or feasible ranges (e.g., pregnancy records in males, a human with 4 arms, systolic pressure higher than 250mmHg, etc)	Requested to DEAP and unable to provide	
		Records of healthcare events (diagnoses, prescriptions, admissions, etc) with logical inconsistencies (e.g., and admission occurs after death)	Requested to DEAP and unable to provide	
		Variables that are based in imputation, derivation or inference (e.g., end of treatment date is derived from treatment start date and treatment cycle length)	Requested to DEAP and unable to provide	
	Precision	Exposures codes precision level, including medicines and vaccines (e.g., active principle, therapeutic group, ...)	Active principle (ATC level 5 codes)	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP
		Precision of date of birth (e.g., day, month, year)	Age at diagnosis, the NCR contains date of birth (day, month, year), but this is generally not shared in a data request, instead age at diagnosis is shared, for example	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP/ DAP
		Precision of date of death (e.g., day, month, year)	Day, month, year	Provided by DEAP
		Precision of date of the event/diagnosis (e.g., day, month, year)	Day, month, year; but usually not shared in a data request, instead interval since diagnosis (or a different interval) is shared	Provided by DEAP
		Precision of date of the exposure (e.g., day, month, year)	Day, month, year	Provided by DEAP
	Traceability	Provenance of event records	EMR. Death, so that is from CBS (it is retrieved from the EHR as well if known, but will be checked during linkage with CBS)	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP
		Provenance of medicines/vaccines records	EMR	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP
Coherence	Format coherence	For dates, formatting constraint being followed	Requested to DEAP and unable to provide	
		For sex, formatting constraint being followed	Requested to DEAP and unable to provide	
	Relational coherence	% of records with the Person ID in the PERSONS table	100%	Provided by DEAP
	Semantic coherence to determine whether the database uses a	For EVENTS definitions, codelists/data dictionaries being employed according to external standards	Indication: ICD-O; Procedures vocabulary: own vocabulary; Diagnosis/medical event vocabulary: ICD-O Stage: TNM	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP/ DAP
		For EXPOSURES, codelists/data dictionaries being employed according to external standards	Prescription: ATC level 5, own vocabulary;	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DAP
	Uniqueness	Number of records flagged as potential duplicates	NO but If one patients gets two different tumors, they would get two entries in the NCR and therefore there are more records in the source data. These two tumors will be connected to the same single patient.	DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP

Scientific research question		**Clinical Benefit of Capecitabine with Oxaliplatin (CapOx) plus Bevacizumab versus CapOx only in patients with Metastatic Colorectal Cancer**						
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**							
	Histologically confirmed mCRC diagnosis in the last year prior to randomization	Pathology results	*High*	100% of individuals have available information	Once year all cancer dx are reviewed to identify cancer patients that did not have a biopsy and pathology finding.		Since 1989 OMOP-CDM since 1992. Daily updates	
	Age > or = 18y	Date of birth	*High*	100% of individuals have available information	As the format is not known, precision can not be evaluated. Low impact in the study?		Since 1989 OMOP-CDM since 1992. Daily updates	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP
	ECOG < or = 1	ECOG score	*High*	Recorded 15% missing	ECOG is dependent on the eye of the beholder.			
	Not felt to be amenable to curative resection	Pathology results	*High*	100% of individuals have available information	Once year all cancer dx are reviewed to identify cancer patients that did not have a biopsy and pathology finding.		Since 1989 OMOP-CDM since 1992. Daily updates	
	Life expectancy longer than 3 months	Pathology results	*High*	100% of individuals have available information	Once year all cancer dx are reviewed to identify cancer patients that did not have a biopsy and pathology finding.		Since 1989 OMOP-CDM since 1992. Daily updates	
	No prior systemic therapy for mCRC or previous treatment with oxaliplatin or bevacizumab	Medication code Date of prescription/dispensing	*High*	Only first line treatment				
	Adequate hematologic/ clotting, hepatic and renal function	Laboratory tests	*High*	Unknown missingness				
	Exclusion criteria							
	Pregnant or breastfeeding women	Pregnancy/breastfeeding status	*High*	Not registered for colon cancer patients				Provided by DEAP
Treatment/exposure	Bevacizumab (7.5 mg/kg IV, on day 1 of a 3-week cycle) + Capecitabine-Oxaliplatin regimen (IV/3wk).	Medication code Date of prescription/dispensing	*High*	Prescription first line treatment, dose not registered				DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP
Comparator group (if applicable)	Capecitabine-Oxaliplatin regimen (IV/3wk).	Medication code Date of prescription/dispensing	*High*	Prescription first line treatment, dose not registered				
Key endpoint(s)	Progression Free survival (PFS)	Date of treatment initiation Date of progression (imaging) Date of death	*High*	A date of death is recorded for 100% of individuals who are known to have died	Vital status checked once per year. As the date of death is registered it will be possible to calculate. PFS is not directly provided, although an algorithm using prognostic markers has been used in this database to predict PFS, being included in published papers (see column I)		Vital status checked once per year	DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). https://onlinelibrary.wiley.com/doi/10.1002/cam4.6223 Provided by DEAP
Confounders	Age > or = 18y	Date of birth	*Low*	100% of individuals have available information				
	Sex	Sex	*Low*	100% of individuals have available information				
	ECOG performance statust score	ECOG score	*Low*	100% of individuals have available information				
Intercurrent events	Treatment discontinuation	Medication code Treatment end date	*Low*	100% of individuals have available information				DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP
	Partial discontinuation: capecitabine	Medication code Capecitabine end date Oxaliplatin end date	*Low*	100% of individuals have available information				DARWIN:"2024_IKNL.xlsx (for the EMA catalogue) and DARWIN.pdf (for the onboarding as a data partner). Provided by DEAP
	Treatment switch	Medication code Date of prescription/dispensing Date of discontinuation Treatment duration	*Low*	Only first line treatment	As only first line treatment is recorded it won't be posible to diferentiate descontinuation than switch			
	Local treatment	Date of procedure Procedure code	*Low*	Procedures available, cancer related surgery might be picked if a specific code is available				
Follow-up time needed per patient in the study	48 weeks	48 weeks	*High*				The median lenght of follow-up per patient is approximately 9 months	
Minimum time in the data source for lookback assessment	1 week	1 week	*Low*				The median lenght of follow-up per patient is approximately 9 months	

	Estimated sample size: Approx 440 participants			Considering that the Netherlands Cancer Registry (NCR) recorded 22,192 patients aged ≥70 years with metastatic colon cancer between 2005 and 2020—of whom 23% received targeted therapy—the target sample size is anticipated to be reached.				

Case study	RWD source	Sample size estimation form the hypotetical trial protocol	Feasibility assessment *(yes/yes, with limitations/no)*	Rationale for the feasibility assessmernt	Limitations identified during the feasibility assessment and categorisation	Description of potential impact of the identified limitations on the study results
10 (Capecitabine with Oxaliplatin (CapOx) plus Bevacizumab versus CapOx in patients with Metastatic Colorectal Cancer)	**NCR**	With an approximate estimated sample size of 440 individuals (based on a 1:1 ratio between treatment arms, comparing CAPOX plus bevacizumab versus CAPOX alone), and considering that the Netherlands Cancer Registry (NCR) recorded 22,192 patients aged ≥70 years with metastatic colon cancer between 2005 and 2020—of whom 23% received targeted therapy—the target sample size is anticipated to be reached. [1]	Yes, with limitations on a design element	Elements with high criticality are available and fairly reliable, **with reservations regarding a design element endpoint**. The time elapsed from when a user requests the data to when they actually receive it is 2 months. Data recency is ~12 months before extraction, reasonably enough for the research question. Sample size is achievable.	·Potentially major: Progression free survival (key endpoint) is not directly provided, although an algorithm using prognostic markers has been used in this database to predict PFS. ·Potentially major: ECOG is 15% missing. ·Potentially major: The median lenght of follow-up per patient is approximately 9 months. ·Minor: Some cancer patients do not have a biopsy and pathology, but might be picked by diagnostic code. ·Minor: Only prescription of first line of treatment is available, but cancer stage changes mean a new first treatment line is started; so, we will be able to identify previous treatments. ·Minor: Data is registered 6-12 months after diagnosis so there is a lag. ·Minor: Imaging information to assess progression-free survival is not available, only death is captured ·Minor: Procedure codes are available, but cancer-related surgery might only be picked if a specific code is available.	Although PFS is not directly available, a previously developed algorithm using prognostic markers has been applied in this database to estimate PFS. Missing ECOG data may prevent us from including certain subjects. Although the median follow-up time in the NCR is 9 months, this includes patients with all types of cancer with different survival durations. However, this variation is likely non-differential, meaning it is not expected to bias the results in favour of or against any particular cancer group. If the patients included in the study have a longer survival time, the registry will allow for the follow-up required by protocol.

REFERENCES

[1] Baltussen JC, de Glas NA, Liefers GJ, Slingerland M, Speetjens FM, van den Bos F, Cloos-van Balen M, Verschoor AJ, Jochems A, Spierings LEAMM, Holterhues C, van Gerven LA, Mooijaart SP, Portielje JEA, Derks MGM. Time trends in treatment patterns and survival of older patients with synchronous metastatic colorectal cancer in the Netherlands: A population-based study. Int J Cancer. 2023 May 15;152(10):2043-2051. doi: 10.1002/ijc.34422. Epub 2023 Jan 13. PMID: 36620951.