Fitness for purpose of data sources relevant for real-world data (RWD) studies on CAR-T cell therapy

FWC/EMA/2020/46/TDA/L5.07 – SC01

Objective 2

Protocol/Study Number	EMA/39169/2024
Protocol Date	18 July 2025
Version Number	1.0
Amendment	Not applicable
	Certara, Real World Evidence & Modelling Solutions Artak Khachatryan - Responsible Senior Investigator Nadia Quignot - Overall Project Coordinator Stephanie Read - Lead Epidemiologist Kazue Kikuchi - Project Manager **EMA** Carla Jonker (Carla.Jonker@ema.europa.eu) - Project Lead Alexandra Pacurariu (Alexandra.Pacurariu@ema.europa.eu) - Pharmacoepidemiologist

TABLE OF CONTENTS

List of tables

List of figures

LIST OF ABBREVIATIONS

Abbreviation	Definition
ATC	Anatomical Therapeutic Chemical
BMI	body mass index
CAR-T	chimeric antigen receptor T-cell
CIBMTR	Centre for International Blood and Marrow Transplant Research
CKS	cytokine release syndrome
Clinformatics	Optum® Clinformatics® Data Mart
DD	date
DESCAR-T	Dispositif d'Enregistrement et Suivi des patients traités par CAR-T cells
DLBCL	diffuse large B-cell lymphoma
EEA	European Economic Area
EHR	electronic health record
EMA	European Medicines Agency
EMR	electronic medical records
EU	European Union
GP	general practitioner
GUI	graphical user interface
HCPCS	Healthcare Common Procedure Coding System
HES	Hospital Episode Statistics
HTA	health technology assessment
IBM	International Business Machines Corporation
ICANS	immune effector cell-associated neurotoxicity syndrome
ICD	International Classification of Diseases
ICD-9	International Classification of Diseases, 9th Revision
ICD-10	International Classification of Diseases, 10th Revision
ICD-11	International Classification of Diseases, 10th Revision
ID	identity document
IQR	interquartile range
IRB	Institutional Review Board
MCL	mantle cell lymphoma
MM	month
NCRAS	National Cancer Registration and Analysis Service

Abbreviation	Definition
NDC	National Drug Code
OPCS	Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures
RW	real-world
RWD	real-world data
SNDS	Système National des Données de Santé
SPIFD	Structured Process to Identify Fitness-For-Purpose Data
US	United States
YYYY	year

RESPONSIBLE PARTIES

Name	Professional Title	Role in Study	Affiliation	Email Address
Carla Jonker	Scientific Administrator	EMA Project Lead	EMA	Carla.Jonker@ema.europa.eu
Alexandria Pacurariu	Scientific Administrator	Pharmaco-epidemiologist	EMA	Alexandra.Pacurariu@ema.europa.eu
Artak Khachatryan	Distinguished Scientist	Responsible Senior Investigator	Certara	Artak.Khachatran@certara.com
Nadia Quignot	Senior Director	Project Coordinator	Certara	Nadia.Qugnot@certara.com
Izabela Pieniazek	Senior Director	Senior Internal Expert (Literature Review)	Certara	Izabela.Pieniazek@certara.com
Stephanie Read	Associate Principal Scientist	Lead Epidemiologist	Certara	Stephanie.Read@certara.com
Giancarlo Pesce	Senior Technical Consultant	Epidemiologist	Certara	Giancarlo.Pesce@certara.com
Dmitry Gultyaev	Consultant	Senior Internal Expert (Data Repository Lead)	Certara	Dmitry.Gultyaev@certara.com
Kazue Kikuchi	Senior Analyst	Project Manager	Certara	Kazue.Kikuchi@certara.com

1 SYNOPSIS

Study Title	Fitness for purpose of data sources relevant for real-world data (RWD) studies on CAR-T cell therapy
Rationale for Study	Chimeric antigen receptor T-cell (CAR-T) cell therapy has emerged as an important cellular immunotherapy for cancer treatment. Since 2018, six CAR-T cell therapies have been approved by the European Medicines Agency (EMA) (Abecma, Breyanzi, Carvykti, Kymriah, Tecartus and Yescarta) for the treatment of specific haematological malignancies. Non-interventional studies are important for supporting regulatory decisions especially when randomised controlled trials are unfeasible or unethical, which is often the case in the CAR-T cell area whose effectiveness and safety evidence mostly rely on single arm trials. The increasing availability of real-world data (RWD) offers important opportunities to inform the real-world (RW) effectiveness and safety of approved CAR-T cell therapies, as well as to provide benchmarking (context) to findings from single arm trials. However, to support regulatory decision-making, evidence generated from RWD studies needs to be reliable. A key determinant of the reliability of evidence derived from RWD studies is quality of the data used and their suitability (fitness-for-use) for the question of interest. This project aims to identify and assess the quality and fitness-for-use of existing European or United States (US)-based data sources for the conduct of non-interventional studies into CAR-T cell therapies. The outputs of this study will provide insight into the quality and relevance of such data, helping future feasibility assessments and data selection for non-interventional studies in CAR-T cell therapies as well as helping interpretation of existing or ongoing studies.
Objective(s)	To identify, describe and assess data quality of a set of selected data sources relevant to the conduct of CAR-T cell non-interventional studies,
Overview of Study Methodology	To address the research objective, the following multi-phased methodology will be applied. 1. Relevant data sources will be identified via a targeted literature review using EMBASE, MEDLINE, online search queries, and the HMA-EMA Catalogue of RWD sources. A selection will be made on high level criteria, to focus already on most relevant data sources. 2. A data repository will be developed in which the following metadata for each data source will be presented:

	a. General data source information (data source type, care setting, language, time span, which data dictionary is used, coding vocabularies used) b. Data source population (population captured, age groups, proportion of population captured from the target population [coverage], median observation time) c. Data flows and management (events triggering registration/de-registration, linkage to other data sources, data source refresh frequency) d. Data access requirements (requirements, timelines, access type, costs) e. Data elements captured (demographics, medical history, treatment history, CAR-T cell therapy administration information, outcomes) 3. The metadata of up to 50 data sources will be obtained from publicly available information and via direct contact with up to 15 data sources. 4. Up to 30 data sources will be included in a more in-depth assessment of data quality. Metrics for each of the 5 dimensions of data quality (reliability, extensiveness, coherence, relevance and timeliness) will be assessed. The outputs of the data quality assessment will be included in a separate section of the data repository document.
Milestones	Finalisation of Data Source List: 20 June 2025 Development of Data Repository: 11 July 2025 Publicly Available Data Capture: 14 August 2025 Detailed Data Quality Assessment: 10 October 2025 Finalisation of Data Repository & Delivery of Report (Deliverable 4): 15 Jan 2026

2 AMENDMENTS AND UPDATES

Not applicable.

3 MILESTONES

Milestone	Planned Completion Date
Finalisation of Data Source List	20 June 2025
Development of Data Repository	11 July 2025
Publicly Available Data Capture	14 August 2025
Detailed Data Quality Assessment	10 October 2025
Finalisation of Data Repository & Delivery of Report (*Deliverable 4)*	15 January 2026

4 BACKGROUND AND RATIONALE

Chimeric antigen receptor T-cell (CAR-T) therapy has emerged as an important cellular immunotherapy for cancer treatment involving the ex-vivo genetic manipulation of autologous T-cells and their reinfusing to initiate appropriate immune response and potential eradication of tumour cells. Since 2018, six CAR-T cell therapies have been approved by the European Medicines Agency (EMA) (Abecma, Breyanzi, Carvykti, Kymriah, Tecartus and Yescarta) for the treatment of specific haematological malignancies such as B-cell leukaemia, B-cell lymphoma, follicular lymphoma, multiple myeloma, and mantle cell lymphoma (MCL) in patients with relapsed or refractory forms. Initial regulatory approval of these therapies was primarily based upon the results of single-arm phase II trials, such as ZUMA-1 for Yescarta and JULIET for Kymriah.[1, 2] These trials demonstrated significant improved overall response rate with CAR-T cell therapy, alongside adverse effects including cytokine release syndrome (CKS), immune effector cell-associated neurotoxicity syndrome (ICANS), infections and tumour lysis syndrome.

Non-interventional studies play a critical role in informing regulatory decisions, particularly in situations where randomization is impractical or unethical, sample sizes are limited, or late-onset and rare outcomes or adverse events are expected. These challenges are frequently encountered in the assessment of CAR-T cell therapies. Most clinical studies in this area are limited to single-arm, uncontrolled trials, largely due to the small size of eligible patient populations and the ethical and practical difficulties of randomization, given the potential for cure in patients with otherwise limited treatment options.

The increasing availability of real-world data (RWD) offers important opportunities to inform the real-world (RW) effectiveness and safety of approved CAR-T cell therapies, as well as to provide benchmarking (context) to findings from single arm trials. To support regulatory decision-making, evidence generated from RWD studies needs to be reliable. A key determinant of the reliability of evidence derived from RWD studies is quality of the data used and their suitability (fitness-for-use) for the question of interest. Assessment of fitness-for-use includes evaluating whether the data source satisfies criteria relating to the reliability, extensiveness, coherence, timeliness and relevance of the data.[3]

The overarching aim of this project is to identify and assess the quality and fitness-for-use of existing European or United States (US)-based data sources for the conduct of non-interventional studies into CAR-T cell therapies. The outputs of this study will provide insight into the quality and relevance of such data, helping future feasibility assessments and data selection for non-interventional studies in CAR-T cell therapies as well as helping interpretation of results from finalised or ongoing studies.

5 STUDY OBJECTIVES

To identify, describe and assess data quality of a set of selected data sources relevant to the conduct of CAR-T cell non-interventional studies. To address this aim, the following three objectives were set:

Objective 1: To provide a comprehensive overview of the healthcare systems where CAR-T cell therapy is approved and effectively used/prescribed in the different European Union (EU) member states/European Economic Area (EEA) countries and the US[4]

Objective 2: To identify and provide information on a set of available data sources relevant for the conduct of CAR-T cell non-interventional studies, describe their metadata and assess their data quality and option(s) for data access.

Objective 3: To assess whether the data sources identified as part of Objective 2 are fit-for-use for conducting CAR-T cell non-interventional studies for the below three theoretical (regulatory relevant) scenarios (hereafter "scenarios")

- As part of this assessment, to describe underlying CAR-T cell treated population(s), including indications, data on alternative treatments, average follow-up, and selected outcome(s) data, in connection with three scenarios

- To identify expected data and methodological limitations based on these scenarios

This study protocol outlines the methodology for addressing **objective 2.**

6 STUDY METHODOLOGY

6.1 Overview of Methodology

The methodological approach to Objective 2 is summarised in Figure 1 and an overview of the methods is provided in subsequent sections.

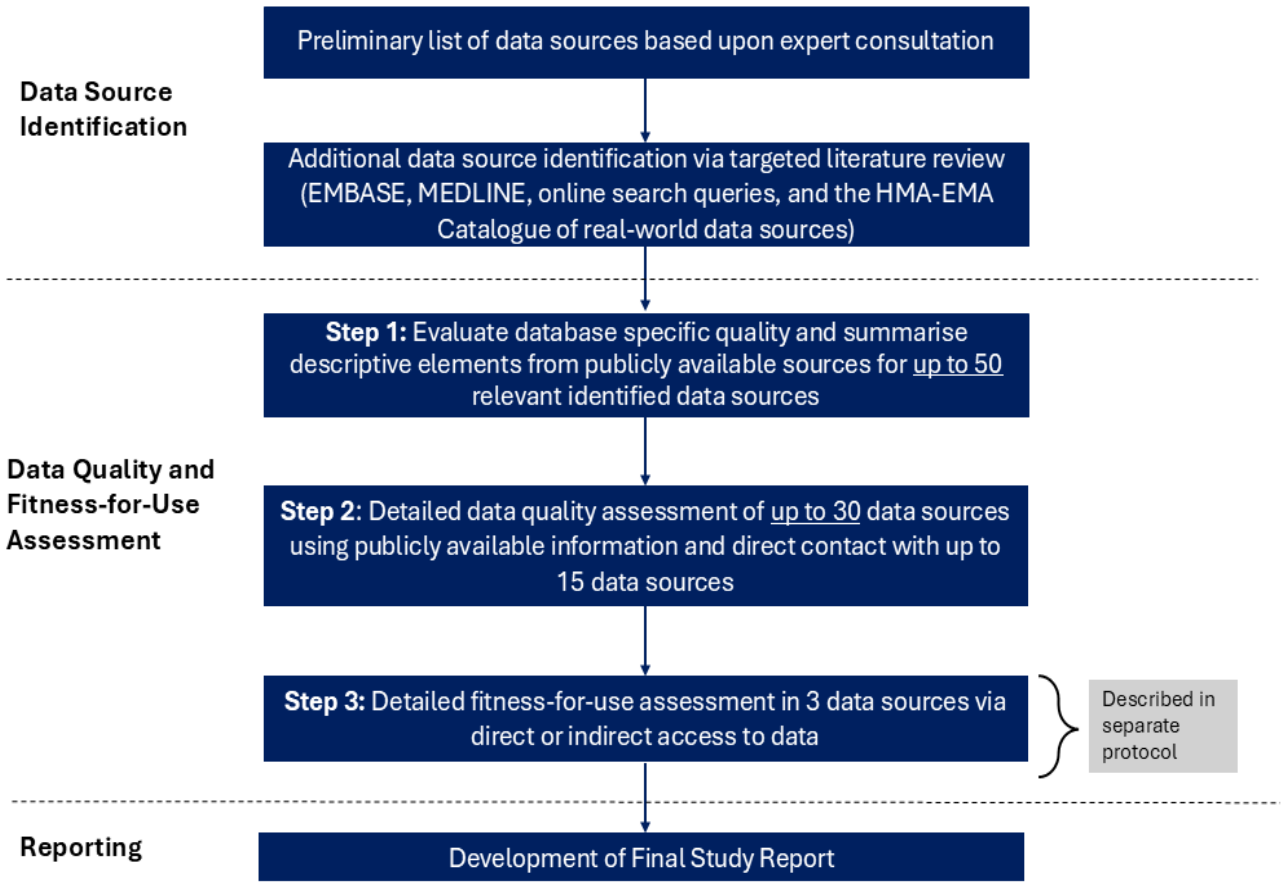

Figure 1. Overview of methodological approach to addressing Objective 2
Abbreviation: EMA = European Medicines Agency.

6.2 Data Source Identification & Screening

To address the objectives, relevant eligible data sources that may be used for non-interventional studies into CAR-T cell therapies need to be identified. Eligible data sources should meet the following criteria:

- Capture CAR-T cell therapy administration
- Person-level data
- Retrospective data sources
- US and Europe only

Table 1 presents a preliminary list of potentially relevant data sources to be included in the data repository based upon expert consultation.

Table 1. Preliminary list of potentially relevant data sources

Data Source	Country	Healthcare Setting Covered	Type of Data
European Society for Blood and Marrow Transplantation CAR-T cell Registry	Across Europe	Transplant centres	Registry
Danish national population-based registries	Denmark	Primary care and secondary care	Collection of registries
NCRAS & linked HES	England	Primary care and secondary care	Cancer registry and EHRs
French Health Insurance database SNDS	France	Primary care and secondary care	Administrative claims
Unicancer	France	18 French Cancer Centres	EHRs
DESCAR-T Registry	France	CAR-T treatment centres	Registry
German Registry for Stem Cell Transplantation	Germany	Adult transplant centres	Registry
Therapie Monitor/TriNetx	Germany/Spain	Secondary care	Registry
Dutch CAR-T Tumorboard	Netherlands	CAR-T treatment centres	Registry
Norwegian national population-based registries	Norway	Primary care and secondary care	Collection of registries
GELTAMO-GETH	Spain	CAR-T treatment centres	Registry
Swedish national population-based registries	Sweden	Primary care and secondary care	Collection of registries
US Flatiron Health Research Database	US	Cancer clinics	EHRs
Panalgo EHR/EMR Data	US	Hospitals, primary-care units, or specialists	EHRs
TriNetX research Network	US	Hospitals, primary-care units, or specialists	EHRs
Inovalon closed claims database	US	Medical and pharmacy visits	Claims database
Medicare Fee-for-Service database	US	Inpatient & Outpatient	Claims database
IBM MarketScan Commercial and Medicare databases	US	Inpatient & Outpatient	Claims database
Clinformatics	US	Inpatient & Outpatient	Claims database
PINC AI Healthcare Database	US	Inpatient & Outpatient	Claims database
CIBMTR's data registry (CIBMTR database)	US	Transplantation centres	Registry

Abbreviations: CAR-T = chimeric antigen receptor T-cell; CIBMTR = Centre for International Blood and Marrow Transplant Research; Clinformatics = Optum® Clinformatics® Data Mart; DESCAR-T cell = Dispositif d'Enregistrement et Suivi des patients traités par CAR-T cells; DM1 = Myotonic Dystrophy Type 1; EHR = electronic health record; EMR = electronic

medical records; HES = Hospital Episode Statistics; IBM = International Business Machines Corporation; NCRAS = National Cancer Registration and Analysis Service; SD = standard deviation; SNDS = Système National des Données de Santé; US = United States.

The preliminary list of data sources presented in Table 1 will be supplemented by data sources identified through a literature search to include a maximum of 50 data sources. The structured literature search of the PubMed and Embase (via the Ovid platform) electronic databases will be performed based on a search strategy developed using keywords, synonyms, and strings adequate to identify relevant data sources. The timeframe of the last five years will be considered to align with the first CAR-T cell therapy approvals. No restrictions to languages will be implemented to enable sensitive search strategy development and identification of relevant data sources. In addition, supplementary online search queries as well as the review of HMA-EMA Catalogue of RWD sources will be undertaken.

Each of the records identified during the search will be screened against predefined eligibility criteria (see Table 2 below). During initial screening, titles and abstracts will be reviewed, and records clearly not relevant or unclear will be excluded. A single record review will be performed during the screening of titles and abstracts.

Relevant publications will be retrieved for full-text screening. A single review of the full text will be performed against the same eligibility criteria.

Table 2. Inclusion criteria for RWD data sources mapping

Category	Inclusion criteria
Interventions	CAR-T cell therapies in general or certain types (tisagenlecleucel, axicabtagene ciloleucel, brexucabtagene autoleucel, idecabtagene vicleucel, lisocabtagene maraleucel, ciltacabtagene autoleucel)
Type of data source	Registries, claims databases, EMR or electronic, health records
Type of study	All excluding randomised control trials.
Regions	EU member states/EEA countries and the US
Language	No restriction
Time frame	Last 5 years to align with first CAR-T cell therapy approvals
Bibliographic databases/other sources	PubMed Embase (via Ovid) Online search queries (e.g., Google) HMA-EMA Catalogue of RWD sources

Abbreviations: CAR-T = chimeric antigen receptor T-cell; EEA = European Economic Area; EMR = electronic medical records; EU = European Union; RWD = real-world data; US = United States.

The inclusion and exclusion of data sources identified in the literature search will be documented. The number of data sources will be restricted to 50 sources for inclusion in the data repository. The choice of the most appropriate data sources will be limited to those meeting the following criteria:

- Captures day-level CAR-T cell administration data;
- Captures at least 1 safety of effectiveness outcome relevant to CAR-T cell therapy.

6.3 Database-Specific Quality and Descriptive Elements

A data repository will be developed in Excel in a searchable format that will enable data sources to be filtered according to features of interest (e.g., disease, geography). The proposed features of the repository will be discussed with the EMA prior to implementation.

The repository will be programmed in Microsoft Excel with a focus on clean and straightforward graphical user interface ("GUI") navigation to efficiently display essential data source information and key parameters, ensuring ease of use.

The foundational data quality and descriptive elements outlined in Table 3 will be reported in the data repository for each of the up to 50 data sources. This list of metrics was developed and adapted from the list of metadata for the HMA-EMA Catalogues of RWD sources and studies.[5]

Table 3. Foundational data quality and descriptive metrics [5]

Element	Format
General Data Source Information	
Type of Data Source	Registry/Claims/EHRs
Data Source Countries (where data originates)	<free text>
Data Access Provider Name	<free text>
Data Access Provider Organization Type (e.g., academic group, private companies, public institutions/organizations)	<free text>
Data Access Provider Contact	<free text>
Data Source Language(s)	<free text>
Regulatory/HTA acceptance	Yes/No
Registered in HMA-EMA catalogue	Yes/No
Publication or website characterizing data source	Yes/No
If applicable, publication citation or website address characterizing data source information	<free text>
Data Source Time Span - First Collection Date	Date (DD/MM/YYYY)

Element	Format
Data Source Time Span - Last Collection Date	Date (DD/MM/YYYY) or Ongoing
Data Source Time Lag (i.e. delay between collection of data at its source and time it is available for extraction/analysis)	None/>0 to <3 months/≥3 to 6 months/≥6 to 12 months/≥12 to 18 months/≥24 months
Data Source Refresh Frequency	None/Monthly/Quarterly/Every 6 months/Yearly/Unknown
Availability of Data Dictionary	Yes – publicly available/ Yes, at-request/No
Treatment Vocabulary	ATC/NDC/Free text/Other/Unknown
Diagnosis Vocabulary	ICD-9/ICD-10/ICD-11/Read codes/Free text/Other/Unknown/MedDRA
Procedures Vocabulary	ICD-9/ICD-10/ICD-11/OPCS/HCPCS/Free text/Other/Unknown
Verification and/or Monitoring Activities (Whether data owners conduct monitoring activities to avoid the capture/collection of incorrect data (i.e. plausibility ranges).	<free text>
Data Source Validation/Quality Checks (i.e., Whether the data owners have conducted validation exercises to confirm data reliability once data are received (i.e., published validation studies)	<free text>
Data Provenance Documentation (i.e., whether there is documentation detailing the data origin, transformations, management and cleaning processes	Yes/No
Linkage to other data sources	Yes/No/Unknown
Names of linked data source, if applicable	<free text>
Identifier for data linkage, if applicable (e.g., social security number)	<free text>
Data Access Requirements	
Data Access Requirements (e.g., approval from scientific committee) and time	<free text>
Patient-Level Data Access Authorized Parties (i.e., no restrictions, academics only, no pharmaceutical companies)	<free text>
Typical timeline for acquiring access to data (including any ethical/IRB/other approvals, as needed)	< 3 months/≥3 months to < 6 months/≥6 months to < 12 months/≥12 months /Unknown
Data access type	<multiple may be selected> Provision of data extract/Remote access/Analysis in-house and reporting results/Other/Unknown
Fee associated with data access	Yes/No/Unknown

Element	Format
Data Source Population	
Population size	<number of people included in data source >
Active population size (e.g., patients still contributing data to data source at time of assessment)	<number of active patients at time of data extraction> / Unknown
CAR-T cell population size	<number of CAR-T cell therapy recipients included in data source>
Data Source Representativeness	
Care Setting for Data Source	<Multiple may be selected> Hospital inpatient care/Hospital outpatient care/ Primary care-GP, community pharmacist level/Primary care-specialist level (e.g. paediatricians)/Secondary care-specialist level (ambulatory)/Other/Unknown
Population captured by data source	All people in region or country/Only people with specific disease/Only recipients of specific therapy/Only people registered at specific healthcare provider
Event triggering registration	<Multiple may be selected> Disease diagnosis/Healthcare registration/Insurance coverage start/Start of treatment/Other/Unknown
Event triggering de-registration	<Multiple may be selected> Death/Emigration/End of treatment/Insurance coverage end/Loss to follow-up/Site or practice de-registration/Other/Unknown
Specific disease or treatment administered (if data source captures only patients with a specific disease or recipients of a treatment)	<free text>
Captures conditions other than haematological malignancies	Yes/No
Population age group-estimated minimum age	<number> years/no limit/Unknown
Population age group-estimated maximum age	<number> years/no limit/Unknown
Estimated percentage of source population covered by data source	0 to 20%/21 to 40%/41 to 60%/61 to 80%/81 to 100%/Unknown
Observation time	<median and IQR> years/Unknown
Key Data Elements Captured	
Demographics	Yes/No
Diagnoses	Yes/No
Diagnosis date	Yes/No

Element	Format
Symptoms / Signs	Yes – via diagnostic codes only/Yes – via clinical notes or structured variable/Yes – via diagnostic codes and clinical notes or structured variable/No
Treatments	Yes/No
Treatment dates	Yes/No
CAR-T cell administrations	Yes/No
Test results	Yes/No
Test results dates	Yes/No
Procedures	Yes/No
Procedures dates	Yes/No
Lifestyle Characteristics/Behaviour (e.g., smoking/alcohol consumption/exercise)	Yes/No
Height and weight (or BMI)	Yes/No
Referrals	Yes/No
Referral dates	Yes/No
Hospital admissions	Yes/No
Hospital admission dates	Yes/No
Outpatient visits	Yes – all/ Yes – hospital outpatients only/Yes – primary care visits / No
Outpatient visits	Yes/No
Patient-reported outcomes	Yes/No
Medical Costs	Yes/No
Deaths	Yes/No
Death Date	Yes/No
Cause of death	Yes/No

Abbreviations: ATC = Anatomical Therapeutic Chemical; BMI = body mass index; CAR-T = chimeric antigen receptor T-cell; DD = date; EHR = electronic health record; GP = general practitioner; HCPCS = Healthcare Common Procedure Coding System; HTA = health technology assessment;ICD-9 = International Classification of Diseases, 9th Revision; ICD-10 = International Classification of Diseases, 10th Revision; ICD-11 = International Classification of Diseases, 11th Revision; IQR = interquartile range; IRB = Institutional Review Board; MM = month; NDC = National Drug Code; OPCS = Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures; YYYY = year.

6.4 Detailed Data Quality Assessment

A total of (up to) 30 data sources will be included in the data quality assessment. These will be the most relevant data sources from the initial list. They will be selected based on a series of ranked minimal inclusion criteria, following the Structured Process to Identify Fitness-For-Purpose Data (SPIFD) data feasibility assessment framework (Table 4).[6, 7] This will be an iterative process whereby each next lower ranked criterion will be considered, as needed, until the list of data sources is restricted to 30.

Table 4. Minimal criteria required for inclusion in the detailed data quality assessment

Design/Data Source Element	Minimal criteria required for inclusion in data quality assessment	Rank
Data source contact	Data source willing to collaborate on data quality assessment	5
Eligible population sample size	Contains data for at least 30 recipients for CAR-T cell therapy	4
Treatment/Exposures	Day level CAR-T cell therapy administration data	1
Comparator group (if applicable)	Day level standard of care or other CAR-T cell therapy administration data	3
Safety or Effectiveness Outcomes	SAFETY: Inpatient or outpatient diagnostic records or codes for safety outcomes of interest (CKS, neurotoxicity, infections, neutropenia, secondary primary malignancies) EFFECTIVENESS: Records documenting either disease relapse, progression, recurrence, or death	2
Length of follow-up	3 months minimum	7
Provision of actual dates for treatments and events**	Actual dates available	6

Abbreviations: CAR-T = chimeric antigen receptor T-cell; CKS = cytokine release syndrome.

* Not ranked since these criteria were part of initial inclusion criteria of selected data sources

** No imputed or off-set dates

A data quality assessment will be conducted primarily using publicly available information. In addition, direct contact with data owners will be attempted to fill the remaining gaps. Where possible, quality checks will be run directly on the data.

The data quality assessment will involve the assessment of metrics for each of the 5 dimensions of data quality according to the EMA Data Quality Framework (reliability, relevance, extensiveness, coherence and timeliness).[3]

1. **Data Reliability:** This dimension describes how closely the data reflect what it is meant to be.
2. **Data Extensiveness:** This dimension explores whether the data are sufficient with regard to data coverage and completeness. Elements of data coverage are captured as part of phase 3 of this study.
3. **Data Coherence:** This dimension addresses whether the data are analysable and is largely dependent on foundational determinants of the data source.
4. **Timeliness:** This dimension covers whether the data is available at the right time. Assessment of this dimension includes an assessment of the frequency of data updates, time lags and time to data access. These metrics are already assessed and captured in the data repository as part of phase 3 of the study.
5. **High level relevance metrics for CAR-T related questions:** This dimension measures the extent to which the data source captures the key data elements critical to address CAR-T specific research question(s) (exposures, outcomes, covariates).

Details of each metric to be assessed are presented in Table 5. Results for each metric will be reported along with the reference source of the information.

Table 5. Metrics for detailed data quality assessment

Dimension	Sub-dimension	Metrics
Extensiveness	Coverage	Number and percentage of patients with CAR-T cell therapy in the country available in the data source
	Completeness	Number and percentage of patients with complete age data
		Number and percentage of patients with complete sex data
		Completeness of mortality data <free text–qualitative assessment>
		Completeness of diagnoses <free text–qualitative assessment>
		Completeness of treatment data <free text–qualitative assessment>
		Completeness of clinical status data (e.g., treatment response, progression, remission) <free text-qualitative assessment>
		Number and percentage of patients with complete CAR-T cell therapy administration date
		Number and percentage of CAR-T cell recipients with a CAR-T cell therapy indication (e.g., multiple myeloma)
		Overall completeness (based on published studies)
		Measures taken to improve the completeness of data (e.g., prerequisites for reimbursements to data collectors, automated data collection methods) <free text– qualitative assessment>
Reliability	Accuracy	Plausibility of population min-max age
		Plausibility of population min-max age at diagnosis of CAR-T cell indication (e.g., multiple myeloma)
		Plausibility of min-max age at CAR-T cell administration
		Presence of logical inconsistencies in death date (e.g., diagnoses occurring after date of death)
		Presence of logical inconsistencies in birth date (e.g., diagnoses occurring before date of birth)
		Other flags for accuracy concerns for any variable <free text-qualitative assessment>
	Precision	Treatment type precision level (e.g., active principle, therapeutic group)
		Diagnostic codes precision level (e.g., ICD code decimal places)
		Treatment response precision level (e.g., algorithm, proxy, unstructured clinical notes, structured variable, not captured)

Dimension	Sub-dimension	Metrics
		Progression precision level (e.g., algorithm, proxy, unstructured clinical notes, structured variable, not captured)
		Recurrence precision level (e.g., algorithm, proxy, unstructured clinical notes, structured variable, not captured)
		Refractory status precision level (e.g., algorithm, proxy, unstructured clinical notes, structured variable, not captured)
		Precision of date of birth (e.g., day, month, year)
		Precision of date of death (e.g., day, month, year)
		Precision of date of the diagnoses (e.g., day, month, year)
		Precision of date of the CAR-T exposure (e.g., day, month, year)
		Other flags for precision concerns for any variable <free text-qualitative assessment>
Coherence	Relational coherence	Number and percentage of patients without internal identifier (e.g., person ID) to enable linkage with internal databases
	Semantic coherence	Consistency in reporting of diagnoses (i.e., ICD-10 codes all reported to same number of decimal places) <free text qualitative assessment>
	Uniqueness	Number of records flagged as potential duplicates <free text qualitative assessment>
Relevance	Availability and completeness of key variables for CAR-T study	Availability and qualitative assessment of the completeness of following variables: • Diagnosis of CAR-T cell therapy indication • Diagnosis date of underlying condition of CAR-T therapy indication • Performance status • Performance status assessment date • Transplantation • Transplantation date • Resistance to prior treatments • Date of CAR-T cell administrations/infusions • Whether CAR-T cell therapy was done as part of compassionate use or trial (investigational drug product) • Treatment response • Disease progression • Disease recurrence/relapse

Dimension	Sub-dimension	Metrics
		Availability of granular clinical data (disease subtype- e.g., MCL, DLBCL; biomarkers (e.g. CD19 status); performance status, comorbidities) <free text – qualitative assessment)
	Follow-up duration	Median and range of follow-up post CAR-T cell therapy administration

Abbreviations: CAR-T cell = chimeric antigen receptor T-cell; DLBCL = diffuse large B-cell lymphoma; ICD = International Classification of Diseases; ICD-10 = International Classification of Diseases, 10th Revision; ID = identity document; MCL = mantle cell lymphoma.

Given that some data points captured in the data repository are dynamic and may evolve with time (e.g., patient counts), the data repository will be updated in December 2025-January 2026 prior to final report completion. The following metadata will be updated for the shortlisted (up to) 30 data sources:

- Data Source time span-last collection date
- Population age group-estimated minimum age
- Population age group-estimated maximum age
- Estimated percentage of population covered by data source
- Estimated Population Size (CAR-T recipients)
- Linkage to other data sources
- Names of linked data source, if applicable

Please note, no further data sources will be added to the data repository at data repository update.

7 LIMITATIONS OF RESEARCH METHODS

Missing metadata may be a key limitation of this study. The study will rely upon published information and direct contact with data owners to capture the required metadata for inclusion in the data repository. However, some of the metadata may not be available in published sources and some data owners may not be responsive to requests for collaboration. This may lead to missing data in the data repository. Missing data will be clearly marked in the data repository. A more comprehensive data quality assessment of selected data sources will be conducted as part of objective 3 of this project.

A further limitation of this study is that some of the metadata may become outdated. However, plans are in place to update selected metadata prior to the final report in 2026. Also, the developed data repository could present a useful framework for future updates.

8 COMMUNICATION OF STUDY RESULTS

The results of this study will be provided in a final study report, documenting the findings for objectives 2 and 3 of the overall project. The section for objective 2 will describe the overall methodologic details and provide an overview of identified data sources (including a flow-chart diagram detailing, for example, the total number of data sources identified, total number of data sources selected following screening, reasons for exclusion from data quality assessment) as well as findings from the data quality assessment. The finalised data repository will be embedded in this document.

9 REFERENCES

1. Neelapu SS, Locke FL, Bartlett NL, Lekakis LJ, Miklos DB, Jacobson CA, et al. Axicabtagene Ciloleucel CAR T-Cell Therapy in Refractory Large B-Cell Lymphoma. *New England Journal of Medicine*. 377(26), 2531-2544 (2017)

2. Schuster SJ, Bishop MR, Tam CS, Waller EK, Borchmann P, McGuirk JP, et al. Tisagenlecleucel in Adult Relapsed or Refractory Diffuse Large B-Cell Lymphoma. *New England Journal of Medicine*. 380(1), 45-56 (2019)

3. European Medicines Agency. Data Quality Framework for EU medicines regulation. (2023); *Available from:* https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf.

4. Certara, ROC26 – Fitness for purpose of data sources relevant for real-world data (RWD) studies on CAR-T cell therapy – FWC/EMA/2020/46/TDA/L5.07 – SC01. Study Report, Objective 1. 2025.

5. European Medicines Agency, List of metadata for the HMA-EMA Catalogues of realworld data sources and studies 2024.

6. Gatto NM, Campbell UB, Rubinstein E, Jaksa A, Mattox P, Mo J, et al. The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework. *Clinical Pharmacology & Therapeutics*. 111(1), 122-134 (2022)

7. Gatto NM, Vititoe SE, Rubinstein E, Reynolds RF and Campbell UB. A Structured Process to Identify Fit-for-Purpose Study Design and Data to Generate Valid and Transparent Real-World Evidence for Regulatory Uses. *Clinical Pharmacology & Therapeutics*. 113(6), 1235-1239 (2023)