Study Protocol

P4-C4-002

DARWIN EU® - Enhanced Preparedness of the DARWIN EU® Network for regulatory questions in oncology – DARWIN EU® OncoNet

23/06/2026

Version 4.0

Authors: Talita Duarte Salles, Anton Barchuk, Melissa Leung, Yuqing Hu, Aniek Markus, Anum Zahra, Rana Jajou, Maxim Moinat, Nicholas Hunt, Guido van Leeuwen, Maarten van Kessel, Sumiya Abdirashid, Cesar Barboza, Ross Williams, Adam Black, Natasha Yefimenko, Sharon Geerlings

Public

CONTENTS

Study title	DARWIN EU® - Enhanced Preparedness of the DARWIN EU® Network for regulatory questions in oncology – DARWIN EU® OncoNet
Protocol version	V4.0
Date	23/06/2026
EUPAS number	EUPAS1000000824
Active Substance	Not applicable.
Medicinal Product	Not applicable.
Research question and Objectives	Research aim To enhance the DARWIN EU® Network in generating real world evidence (RWE) in oncology. Objectives The specific objectives of this study are: 1. To characterise and map oncology-related data into the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) for individuals with selected cancer types by: 1.1. Characterising individuals with selected cancer types in terms of age at the time of diagnosis, sex, performance status, cancer stage, cancer grade, clinical and laboratory measurements, and cancer treatments and procedures before and after data mappings. 1.2. Providing guidance and support in mapping data on tumour grade, stages (American Joint Committee on Cancer/Union for International Cancer Control (AJCC/UICC) Tumour, Node, Metastasis (TNM) classification), selected cancer treatments (treatment episodes, regimens, and information regarding cycles/sequencing), and biomarkers to the OMOP CDM. 1.3. Developing and implementing phenotypes for the definition of cancer subtypes, grades, and stages (AJCC/UICC TNM), and cancer-specific measurements (tumour-specific biomarkers) at and after cancer diagnosis in individuals with selected cancer types. 2. To develop and evaluate algorithms to identify oncological treatment episodes, regimens, and lines of therapy, and to describe regimen level treatment patterns received by patients with selected cancers. 3. To identify available data elements to describe treatment response, cancer progression and recurrence, and relevant time-to-event endpoints. 4. To create a database with aggregate counts related to oncology to support future feasibility assessments in the DARWIN EU® Data Network.
Countries of study	Croatia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Italy, Norway, Spain, Sweden, the Netherlands, the United Kingdom
Authors	Talita Duarte-Salles, t.duarte@darwin-eu.org; Anton Barchuk, a.barchuk@darwin-eu.org; Melissa Leung, m.leung@darwin-eu.org; Aniek Markus, a.markus@darwin-eu.org; Anum Zahra, a.zahra@darwin-eu.org; Natasha Yefimenko n.yefimenkonosova@darwin-eu.org; Yuqing Hu, y.hu@darwin-eu.org; Rana Jajou, r.jajou@darwin-eu.org; Maxim Moinat, m.moinat@darwin-eu.org; Nicholas Hunt, n.hunt@darwin-eu.org; Guido van Leeuwen, g.vanleeuwen@darwin-eu.org; Sumiya Sheikh Abdirashid s.sheikhabdirashid@darwin-eu.org Maarten van Kessel, m.vankessel@darwin-eu.org; Cesar Barboza, c.barboza@darwin-eu.org; Ross Williams, r.williams@darwin-eu.org; Sharon Geerlings, s.geerlings@darwin-eu.org.

LIST OF ABBREVIATIONS

Acronyms/terms	Description
A	Doxorubicin
AC	Doxorubicin and Cyclophosphamide (regimen)
ALK	Anaplastic Lymphoma Kinase
AJCC	American Joint Committee on Cancer
ATC	Anatomical Therapeutic Chemical
BRAF V600	B-Raf proto-oncogene V600 mutation
BRAF V600E	B-Raf Proto-Oncogene Valine-to-Glutamic Acid Substitution at Codon 600
BRCA1/2	Breast Cancer Susceptibility Genes 1 and 2
C	Cyclophosphamide
CC	Coordination centre
CDM	Common Data Model
CDW Bordeaux	Clinical Data Warehouse of Bordeaux University Hospital
CHUBX	Centre Hospitalier Universitaire de Bordeaux
CIPH	Croatian Institute of Public Health
CLDN18.2	Claudin 18.2
CPRD GOLD	Clinical Practice Research Datalink GOLD
CPT4	Current Procedural Terminology, 4th Edition
CRN	Cancer Registry Norway
DARWIN EU®	Data Analysis and Real-World Interrogation Network
DK-DHR	Danish Data Health Registries
DKMA	Danish Medicines Agency
Dist	Value distribution
DRE	Digital Research Environment
DTZ	Data Transfer Zone
EBB	Estonian Biobank
ECOG	Eastern Cooperative Oncology Group
ED	Emergency Department
EGFR	Epidermal Growth Factor Receptor
EHR	Electronic Health Records
EMA	European Medicines Agency
ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
ER/PR	Oestrogen Receptor/Progesterone Receptor
Erasmus MC	Erasmus Medical Center
ESMO	European Society for Medical Oncology
ESR1	Estrogen Receptor 1

Acronyms/terms	Description
EU	European Union
EUPAS	EU Post-Authorisation Studies Register
FGFR2	Fibroblast Growth Factor Receptor 2
FGFR3	Fibroblast Growth Factor Receptor 3
FinOMOP-ACI Varha	Auria Clinical Informatics
FinOMOP-HUS	Hospital District of Helsinki and Uusimaa
FinOMOP-TaUH Pirha	Tampere University Hospital patient cohort
FinOMOP-THL	Finnish Care Register for Health Care
GDPR	General Data Protection Regulation
GP	General Practitioner
HER2 (IHC 3+)	Human Epidermal Growth Factor Receptor 2 Overexpression Determined by Immunohistochemistry Score 3+
HER2/ERBB2	Human Epidermal Growth Factor Receptor 2 (Erb-B2 Receptor Tyrosine Kinase 2)
HRR Genes	Homologous Recombination Repair Genes
H12O	Hospital Universitario 12 de Octubre
HI-SPEED	Health Impact - Swedish Population Evidence Enabling Data-linkage
ICD	International Classification of Diseases
ICD-10-PCS	ICD-10 Procedure Coding System
ICD-O	International Classification of Diseases for Oncology
IKNL	Integraal Kankercentrum Nederland
IMASIS	Institut Municipal Assistència Sanitària Information System
InGef	Institute for Applied Health Research Berlin GmbH
InGef RDB	InGef Research Database
IPCI	Integrated Primary Care Information
IRB	Institutional Review Board
IQR	Interquartile range
KIT	KIT Proto-Oncogene Receptor Tyrosine Kinase (CD117)
KRAS G12C	Kirsten Rat Sarcoma Viral Oncogene Homolog Glycine-to-Cysteine Substitution at Codon 12
LDH	Lactate Dehydrogenase
LOINC	Logical Observation Identifiers Names and Codes
MET	Mesenchymal–Epithelial Transition Factor Receptor Tyrosine Kinase
MET Exon 14	Skipping, Mesenchymal–Epithelial Transition Factor Exon 14 Skipping Alteration
MMR	Mismatch Repair
MSI	Microsatellite Instability
NAJS	Croatian National Public Health Information System
NCR	Netherlands Cancer Registry
NIPH	Norwegian Institute of Public Health

Acronyms/terms	Description
NPV	Negative predictive value
NRAS	Neuroblastoma RAS Viral Oncogene Homolog
NTRK	Neurotrophic Tyrosine Receptor Kinase Gene Fusion
OGJ	Oesophagogastric Junction
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
OPS	Operation and Procedure codes
Oxford NDORMS	University of Oxford - Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences
PC	Person Count
PD-1	Programmed Cell Death Protein 1
PD-L1	Programmed Death-Ligand 1
PGH	Papageorgiou General Hospital;
PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha Gene
PSA	Prostate-Specific Antigen
PSMA	Prostate-Specific Membrane Antigen
PSMAR	Parc de Salut Mar Barcelona
POLIMI	Research Repository @Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico
PPV	Positive predictive value
RAS	Rat Sarcoma Viral Oncogene Family (Including KRAS, NRAS, and HRAS)
RC	Record count
RECIST	Response Evaluation Criteria in Solid Tumours
RET	Rearranged During Transfection Proto-Oncogene
ROS1	ROS Proto-Oncogene 1 Receptor Tyrosine Kinase
RWD	Real World Data
RW-DFS	Real-world disease-free survival
RWE	Real world evidence
RW-PFS	Real-world progression-free survival
RxNorm	Medical prescription normalised
SD	Standard Deviation
SIDIAP	The Information System for Research in Primary Care
SMPA-GU	Swedish Medical Products Agency - Gothenburg University
SNOMED	Systemised Nomenclature of Medicine
SU	Semmelweis University
SUCD	Semmelweis University Clinical Data
Synapse	Synapse Research Management Partners
TMB-High	High Tumour Mutational Burden

Acronyms/terms	Description
TNM classification	Tumour, Node, Metastasis classification
UICC	Union for International Cancer Control
UKBB	UK BioBank
TARTU	University of Tartu
WHO	World Health Organisation

1. TITLE

DARWIN EU® - Enhanced Preparedness of the DARWIN EU® Network for regulatory questions in oncology – DARWIN EU® OncoNet

2. DESCRIPTION OF THE STUDY TEAM

Study team role	Names	Organisation
Principal Investigators	Talita Duarte-Salles Anton Barchuk Melissa Leung Maxim Moinat Aniek Markus Rana Jajou	Erasmus MC
Co-Principal Investigators	Nicholas Hunt Yuqing Hu Guido van Leeuwen Anum Zahra Katia Verhamme	Erasmus MC
Data Scientists	Sumiya Sheikh Abdirashid Maarten van Kessel Cesar Barboza Cristiana Di Tullio Ger Inberg Adam Black Arianna Alamshahi Ross Williams	Erasmus MC
Network Operations	Anne van Winzum	The Hyve
Study Manager	Natasha Yefimenko Sharon Geerlings	Erasmus MC Synapse
Data source	**Names**	**Data Partner Organisation***
Croatian National Public Health Information System (NAJS)	Ivan Pristaš Marko Čavlina Anamaria Jurčević Antea Jezidžić Karlo Pintarić Jakov Vukovic	Croatian Institute of Public Health (CIPH)
Danish Data Health Registries (DK-DHR)	Elvira Bräuner Susanne Bruun Charlotte Wessel Skovlund	Danish Medicines Agency (DKMA)
Estonian Biobank (EBB)	Marek Oja	University of Tartu (TARTU)

	Raivo Kolde Ami Sild	
Hospital District of Helsinki and Uusimaa (FinOMOP-HUS)	Eric Fey Kimmo Porkka Marianna Niemi Lesonen Päivi Dimitrios Tsallos Valtteri Nieminen Johanna Niklander	FinOMOP
Tampere University Hospital patient cohort (FinOMOP-TaUH Pirha)	Sampo Kukkurainen Kati Kristiansson Hakkarainen Leena Harri Rantala	FinOMOP
Finnish Care Register for Health Care (FinOMOP-THL)	Tiina Wahlfors Laura Salonen Miia Ryhänen-Tompuri Toni Lehtonen Petteri Hovi Gustav Klingstedt Terhi Kilpi	FinOMOP
Auria Clinical Informatics (FinOMOP-ACI Varha)	Pia Tajanen-Doumbouya Virkki Arho Mikael Högerman Tommi Kauko	FinOMOP
Clinical Data Warehouse of Bordeaux University Hospital (CDW Bordeaux)	Romain Griffier Guillaume Verdy Loïc Ronflette	Centre Hospitalier Universitaire de Bordeaux (CHUBX)
InGef Research Database (InGef RDB)	Raeleesha Norris Annika Vivirito Alexander Harms Josephine Jacob Daniela Harnaß	Institute for Applied Health Research Berlin GmbH (InGef)
Papageorgiou General Hospital (PGH)	Achilleas Chytas Anastasia Farmaki Alexandros Rekkas Maria Bigaki Pantelis Natsiavas	Papageorgiou General Hospital (PGH)
Semmelweis University Clinical Data (SUCD)	Bagyura Zsolt István Kiss Loretta Zsuzsa (adjunktus)	Semmelweis University (SU)

	Ágota Mészáros	
Research Repository @Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico (POLIMI)	Gianluigi Galli Mauro Bucalo Vittoria Ramella	Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico (POLIMI)
Cancer Registry Norway (CRN)	Espen Enerly Yngvar Nilssen Simen Breivik Anna Skog	Norwegian Institute of Public Health (NIPH)
Hospital Universitario 12 de Octubre (H12O)	Juan Luis Cruz Bermudez Noelia Garcia Barrio Paula Rubio Mayo	Hospital Universitario 12 de Octubre (H12O)
Institut Municipal Assistència Sanitària Information System (IMASIS)	Angela Leis Juan Manuel Ramírez Anguita Miguel Angel Mayer Pujadas	Consorci Mar Parc de Salut de Barcelona (PSMAR) together with Fundació Institut Hospital del Mar d'investigacions Mèdiques
Health Impact - Swedish Population Evidence Enabling Data-linkage (HI-SPEED)	Fredrik Nyberg Huiqi Li Talbäck Mats Rickard Ljung Ballin Marcel	Swedish Medical Products Agency - Gothenburg University (SMPA-GU)
Netherlands Cancer Registry (NCR)	Jelle Evers Maaike van Swieten Peter Prinsen	Integraal Kankercentrum Nederland (IKNL)
UK BioBank (UKBB)	Antonella Delmestri Mandickel Kamtengeni Hezekiah Omulo	University of Oxford - Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences (Oxford NDORMS)

*Data partners do not have an investigator role. Data partners execute code at their data source, review, and approve their results.

3. ABSTRACT

Title

DARWIN EU® - Enhanced Preparedness of the DARWIN EU® Network for regulatory questions in oncology – DARWIN EU® OncoNet.

Rationale and background

There is a regulatory need to perform real-world data (RWD) studies in oncology. However, several DARWIN EU® studies were not feasible due to the absence of key variables, limited granularity of data, and analytical pipelines required to address specific regulatory questions. Some of these data elements are available in certain data sources (source data) within the DARWIN EU® Data Network but have not yet been mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).

This study focuses on improving data quality (with specific focus on extensiveness and reliability) related to oncology in the DARWIN EU® Data Network, as well as enhancing the analytical pipelines that will facilitate future oncology studies in DARWIN EU®. This initiative is important for advancing the use of RWD in cancer research by ensuring accurate, detailed data representation, and the availability of appropriate analytical tools.

Research question and Objectives

Research aim

To enhance the DARWIN EU® Network in generating real world evidence (RWE) in oncology.

Objectives

The specific objectives of this study are:

1. To characterise and map oncology-related data into the OMOP CDM for individuals with selected cancer types by:
    1.1. Characterising individuals with selected cancer types in terms of age at the time of diagnosis, sex, performance status, cancer stage, cancer grade, clinical and laboratory measurements, and cancer treatments and procedures before and after data mappings.
    1.2. Providing guidance and support in mapping data on tumour grade, stages (American Joint Committee on Cancer/ Union for International Cancer Control (AJCC/UICC) Tumour, Node, Metastasis (TNM) classification), selected cancer treatments (treatment episodes, regimens, and information regarding cycles/sequencing), and biomarkers to the OMOP CDM.
    1.3. Developing and implementing phenotypes for the definition of cancer subtypes, grades, and stages (AJCC/UICC TNM), and cancer-specific measurements (tumour-specific biomarkers) at and after cancer diagnosis in individuals with selected cancer types.
2. To develop and evaluate algorithms to identify oncological treatment episodes, regimens, and lines of therapy, and to describe regimen level treatment patterns received by patients with selected cancers.
3. To identify available data elements to describe treatment response, cancer progression and recurrence, and relevant time-to-event endpoints.
4. To create a database with aggregate counts related to oncology to support future feasibility assessments in the DARWIN EU® Data Network.

Methods

Retrospective cohort studies will be conducted using routinely collected health data from 18 data sources from 14 countries across Europe.

These data sources are diverse and cover primary care, secondary and specialised care (including data collected for hospital outpatient and inpatient visits), biobanks, cancer registries, medical claims, prescription, and mortality data.

Population

The study population will include all individuals with a diagnosis of one of the selected types of cancer (breast, lung, prostate, colorectal, skin melanoma, bladder, oesophageal cancer), date of the first diagnosis during the study period (01/01/2010 to the most recent available data), and aged ≥18 years at the date of cancer diagnosis.

Analysis

Clinical characteristics, including measurements, conditions, observations, procedures, and performance status, will be characterised for cancer type across different time windows spanning the period before cancer diagnosis and follow-up after diagnosis.

Cancer stage and grade will be described by cancer type across different time windows around the index date, including before diagnosis, at diagnosis, and after diagnosis. Cancer-specific biomarkers, including molecular genetics markers, specified for each cancer type, will be included.

Treatment patterns, including radiotherapy and surgical procedures, will be described by cancer type across different time windows before and after the index date. Treatments will be selected based on the most recent cancer-specific treatment guidelines.

Additionally, for Objective 3, we will assess whether we can assess cancer progression and recurrence, and relevant time-to-event outcomes (disease-free survival and progression-free survival) using variables describing metastatic disease, palliative and symptomatic care, and others.

In Objective 4, a tool will be created to summarise aggregated counts of relevant variables for the cohort of patients with the selected cancer types to facilitate further feasibility exercises for potential future cancer studies.

4. AMENDMENTS AND UPDATES

None.

5. MILESTONES

Study milestones and deliverables	Planned dates[1]
Draft Protocol	15/04/2026
Final Protocol	10/06/2026
Creation of mapping guidelines[2]	01/07/2026
Last mapped data to be included to the study report[3]	04/09/2026
Draft Report	30/10/2026
Final Report	27/11/2026
Feasibility tool	27/11/2026

[1]Planned dates are dependent on obtaining approvals from the internal review boards of the data sources and deliverables review rounds.
[2]Timelines for the finalisation of mapping guidelines will depend on the number of variables needed to be included in the guideline, which is pending agreement between the CC and EMA.
[3]Timelines for finalisation of mappings will depend on data partners. Data mapped after this date might not be included in the study report.

6. RATIONALE AND BACKGROUND

There is a regulatory need to perform real-world data (RWD) studies in oncology.[1,2] However, several DARWIN EU® feasibility assessments have concluded that studies were not feasible or lacked a wider range of data sources and/or European country representativeness due to the absence of key cancer-related variables and/or analytical pipelines required to address specific regulatory questions within acceptable timeframes. Some of these data may be available in certain data sources within the DARWIN EU® Data Network but have not yet been mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).

Among these crucial variables is cancer staging, which provides insight into the extent, spread, and severity of cancer. Staging determines how far the cancer has spread, commonly using the Tumour, Node, Metastasis (TNM) classification system for solid tumours, evaluating primary tumour size and local invasion (T), regional lymph node involvement (N), and the presence of distant metastasis (M).[3,4] Once the T, N, and M are determined, they are combined, and an overall stage of 0, I, II, III, IV is assigned. This information is essential for determining prognosis, guiding treatment decisions, and estimating patient survival rates. However, this information is not always included in observational data sources and/or prioritised for standardisation to the OMOP CDM.

Cancer treatment is a central component of cancer care and encompasses a wide range of modalities, including surgery, radiotherapy, chemotherapy, targeted therapy, immunotherapy, and hormonotherapy.[5] These modalities are often used in combination rather than in isolation,[6] reflecting the complexity of modern oncology care. It is therefore important to study cancer treatment in a way that captures these multimodal therapeutic patterns. In particular, examining treatment at the regimen level is essential for understanding how specific combinations and sequences of care influence effectiveness, toxicity, resistance, and long-term outcomes. This level of granularity is especially important for evaluating comparative effectiveness, safety, adherence, and treatment pathways in real-world data, and for generating evidence that can better inform clinical decision making, guideline development, and precision oncology.

Recurrence and progression are commonly used terms in oncology, particularly in clinical trials,[7–9] but have more recently also been actively used in RWD research.[10–12]

The definitions of these concepts are context-dependent and must be explicitly specified in each study, particularly given their use in time-to-event analyses. Progression-free survival and disease-free survival have been defined in variable ways, even within similar disease settings.[13] Heterogeneity in the definitions and operationalisation of recurrence and progression may complicate RWD analyses, interpretation of clinical outcomes, and regulatory decisions.

In oncology clinical trials, treatment response in solid tumours is most commonly defined using the Response Evaluation Criteria in Solid Tumours (RECIST), a framework for standardised response assessment.[14]

Recurrence refers to the reappearance of cancer after a period when the patient was considered disease-free. The term is used to describe the regrowth of cancer following surgery or definitive locoregional therapy.[15,16] Progression describes the increase in tumour burden or the appearance of new lesions in a patient with known, measurable disease. Progression can occur during active treatment or surveillance and is the primary endpoint for assessing treatment failure.

Recurrence and progression of solid cancers in RWD studies are typically identified in electronic health records (EHRs), registries, and claims data using algorithmic approaches or manual abstraction. Manual chart review remains the gold standard for the ascertainment of recurrence/progression.[17] Recurrence has also been identified using algorithmic approaches, typically flagging new metastatic diagnoses, restaging, pathology codes, or the initiation of cancer-directed therapy after a defined disease-free interval.[18–20]

This study focuses on improving data quality and availability related to oncology in the DARWIN EU® Data Network, as well as improving the analytical pipelines that will facilitate future oncology studies in DARWIN EU®. This initiative is important for advancing the use of RWD in cancer research by ensuring accurate, detailed data representation and the availability of appropriate analytical tools.

7. RESEARCH QUESTION AND OBJECTIVES

Research aim

To enhance the DARWIN EU® Network in generating real-world evidence (RWE) in oncology.

Objectives

The specific objectives of this study are:

1. To characterise and map oncology-related data into the OMOP CDM for individuals with selected cancer types by:
    1.1. Characterising individuals with selected cancer types in terms of age at the time of diagnosis, sex, performance status, cancer stage, cancer grade, clinical and laboratory measurements, and cancer treatments and procedures before and after data mappings.
    1.2. Providing guidance and support in mapping data on tumour grade, stages (American Joint Committee on Cancer/Union for International Cancer Control (AJCC/UICC) Tumour, Node, Metastasis (TNM) classification), selected cancer treatments (treatment episodes, regimens, and information regarding cycles/sequencing), and biomarkers to the OMOP CDM.
    1.3. Developing and implementing phenotypes for the definition of cancer subtypes, grades, and stages (AJCC/UICC TNM), and cancer-specific measurements (tumour-specific biomarkers) at and after cancer diagnosis in individuals with selected cancer types.

2. To develop and evaluate algorithms to identify oncological treatment episodes, regimens, and lines of therapy, and to describe regimen-level treatment patterns received by patients with selected cancers.
3. To identify available data elements to describe treatment response, cancer progression and recurrence, and relevant time-to-event endpoints.
4. To create a database with aggregate counts related to oncology to support future feasibility assessments in the DARWIN EU® Data Network.

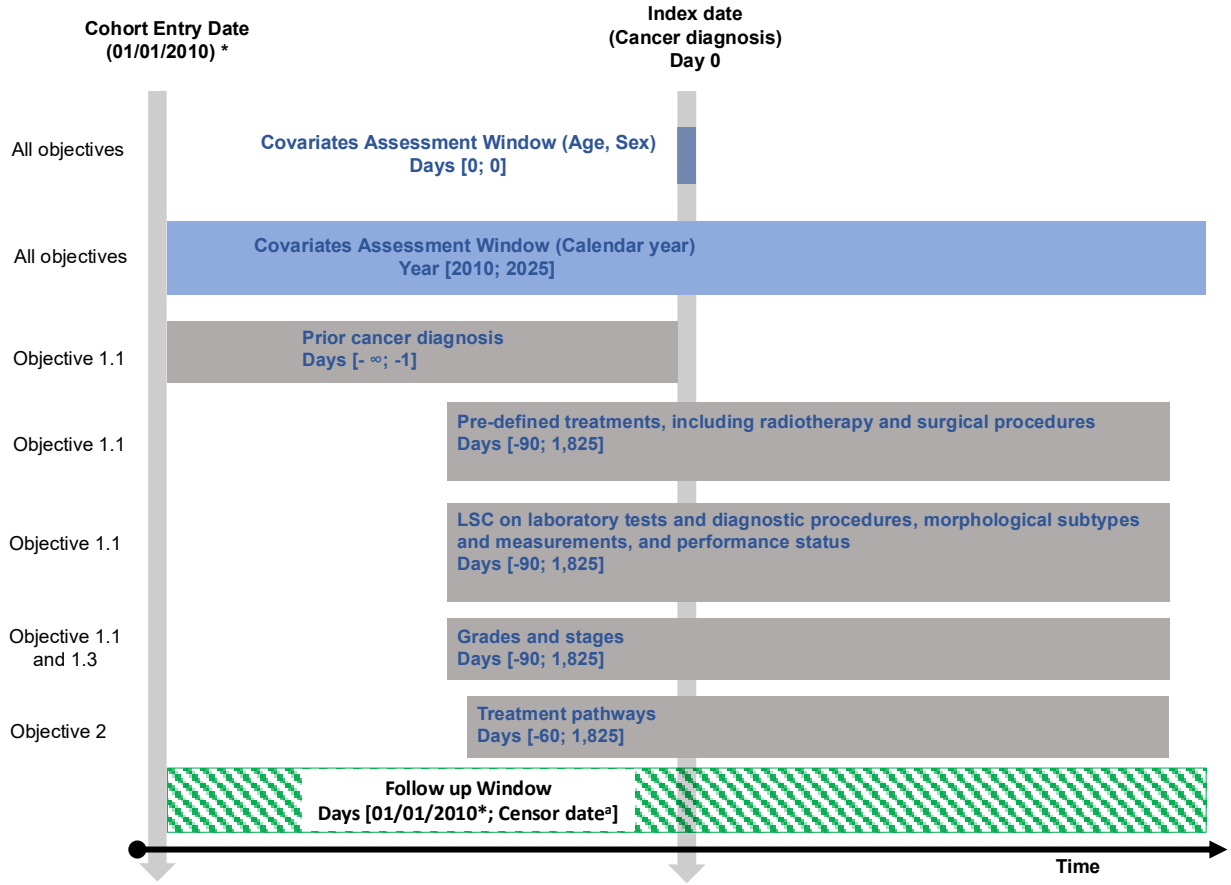
8. RESEARCH METHODS

8.1. Study design

Cohort studies will be conducted using routinely collected health data from 18 data sources from 14 countries across Europe.

A characterisation study will be conducted to address Objectives 1.1, 1.3, 2, and 3, assessing individual characteristics, cancer-related characteristics, measurements and procedures, treatment use and its grouping (into regimens and lines of therapy), and response and recurrence after treatment. The graphical depiction of the study design is shown in **Figure 1**.

To address Objective 1.2, we will support OMOP mapping for those data sources that have available cancer-specific information in their source data but remain unmapped. This data identification will be based on initial characterisation (Objective 1.1) together with results from two surveys shared with data partners at the feasibility assessment stage on data availability and mapping, to identify missing information.



* In some data sources, data availability begins after 2010
a. Earliest date of: Loss to follow-up, death, end of the observation period, or study end date

Figure 1. Graphical depiction of the study design.

8.2. Follow-up

Follow-up will start on the date of the first cancer diagnosis.

Follow-up will end on the earliest of loss to follow-up, death, or the end of the observation period (the latest available data). For Objective 2 on cancer treatments, follow-up will also end at time of a new cancer diagnosis.

8.3. Study population with inclusion and exclusion criteria

The study population will include all individuals who meet the eligibility criteria at study entry.

Eligibility criteria are:

- Diagnosis of one of the selected types of cancer:
    - Breast cancer (in women)
    - Lung cancer
    - Prostate cancer
    - Colorectal cancer
    - Skin melanoma
    - Bladder cancer

    - Oesophageal cancer
- Date of the first diagnosis of a selected type of cancer during the study period
- Age ≥18 years at the date of cancer diagnosis

The concepts linked to the corresponding Systematized Nomenclature of Medicine (SNOMED) and International Classification of Diseases, 10th Revision (ICD-10) codes listed in **Table S1** in **Annex IV**, along with their descendants, will be used to build the preliminary concept sets for the definition of cancer diagnoses. Concepts based on the International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) codes will be used in data sources that cover cancer registries. Oesophageal cancer will also include oesophagogastric junction (OGJ) carcinoma, which would also be reported separately.

8.4. Study setting and data sources

This study will be conducted using routinely collected data from 18 data sources representing different settings (secondary care (including hospital registries), claims and electronic records, biobanks, cancer registries) in the DARWIN EU® Network of data partners from 14 European countries. All data were a priori mapped to the OMOP CDM.

For each data source, Table 1 shows the health care setting, type of data, number of active individuals, and calendar period covered. All data sources will contribute to Objective 1.1, and the contribution to other objectives will be determined by the fitness for use for those objectives.

Table 1. Data sources characteristics.

Name of Data source	Health Care setting	Number of active individuals	Calendar period covered	Biobank	Claims	EHR	Registry nationwide/ regional	Cancer registry
EBB	Primary care GPs and specialists, community pharmacists, secondary care, hospital inpatient care	212k	2004–2024	✓	✓	✓	✓	
CDW Bordeaux	Secondary care, hospital inpatient care	2.41M	2005–2025	✓	✓	✓		
UKBB	Primary care GPs and specialists, secondary care specialists, hospital inpatient care	502k	2006–2025	✓		✓	✓	
InGef RDB	Primary care GPs and specialists, community pharmacists, secondary care, hospital inpatient care	10.7M	2016–2025		✓			
SUCD	Secondary care specialists, hospital inpatient care	2.49M	2010–2025		✓	✓	✓	
PGH	Primary care GPs, secondary care, hospital inpatient care	1.49M	1999–2025		✓	✓		
NAJS	Primary care GPs, secondary care, hospital inpatient care	4.79M	2014–2025		✓	✓	✓	✓
DK-DHR	Community pharmacists, secondary care, hospital inpatient care	8.59M	1995–2024		✓	✓	✓	✓
H12O	Secondary care, hospital inpatient care	1.87M	2015–2024			✓	✓	
FinOMOP-THL	Primary care GPs and specialists, secondary care, hospital inpatient care	6.62M	2011–2024			✓	✓	
FinOMOP-ACI Varha	Secondary care, hospital inpatient care	224K	2002–2024			✓		
FinOMOP-HUS	Secondary care, hospital inpatient care	2.92M	2004–2024			✓		

Name of Data source	Health Care setting	Number of active individuals	Calendar period covered	Biobank	Claims	EHR	Registry nationwide/ regional	Cancer registry
FinOMOP-TaUH Pirha	Secondary care, hospital inpatient care	862k	2007–2025			✓		
IMASIS	Secondary care, hospital inpatient care	1.79M	1990–2025			✓		
POLIMI	Secondary care specialists, hospital inpatient care	1.72M	2010–2024			✓		
HI-SPEED	Primary care GPs, secondary care specialists, hospital inpatient care	12.1M	2015–2025				✓	
CRN	Secondary care, hospital inpatient care	1.25M	1953–2024				✓	✓
NCR	Secondary care, hospital inpatient care	2.76M	1993–2024					✓

CDW Bordeaux = Clinical Data Warehouse of Bordeaux University Hospital; CRN = Cancer Registry Norway; DK-DHR = Danish Data Health Registries; EBB = Estonian Biobank; EHR = electronic health records; FinOMOP-ACI Varha = Auria Clinical Informatics; FinOMOP-HUS = Hospital District of Helsinki and Uusimaa; FinOMOP-TaUH Pirha = Tampere University Hospital patient cohort; FinOMOP-THL = Finnish Care Register for Health Care; GP = general practitioner; H12O = Hospital Universitario 12 de Octubre; HI-SPEED = Health Impact - Swedish Population Evidence Enabling Data-linkage; IMASIS = Institut Municipal Assistència Sanitària Information System; InGef RDB = InGef Research Database; k = thousand; M = million; NAJS = Croatian National Public Health Information System; NCR = Netherlands Cancer Registry; PGH = Papageorgiou General Hospital; POLIMI = Research Repository @Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico; SUCD = Semmelweis University Clinical Data; UKBB = UK BioBank.

Data sources selection

As part of the feasibility assessment, we conducted a survey among the DARWIN EU® Network to identify suitable data partners for this oncology study, focusing on key variables and their presence in the source and/or mapped data. Of the 30 data partners in the network (as of July 2025), 18 were invited to participate in the study, as they met the required settings (registry, inpatient care, and/or claims), had a data source recency/lookback period available, and (at least partially) covered the key variables. The selection was inclusive, in order to also include data sources that are not yet completely fit for purpose but can become so by being part of the project and by improving the data quality. Data sources from the network that mainly collect primary healthcare data (e.g. Integrated Primary Care Information (IPCI), Clinical Practice Research Datalink (CPRD) GOLD, The Information System for Research in Primary Care (SIDIAP)) were not included in the study, as it is likely that they do not capture important cancer-related variables (stage, biomarkers, treatment) mainly recorded in specialised care settings.

See **Annex I** for full descriptions of the data sources and Annex II for a full fitness-for-use assessment.

Table 2 provides an overview of how each data source that will contribute to the various study objectives.

Table 2. Data sources participation per study objective.

Study Objectives	NAJS	DK-DHR	EBB	FinOMOP-ACI Varha	FinOMOP-HUS	FinOMOP-TaUH Pirha	FinOMOP-THL	CDW Bordeaux	InGef RDB	PGH	SUCD	POLIMI	CRN	H12O	IMASIS	HI-SPEED	NCR	UKBB
1.1 Characterisation	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
1.2 Mapping*	x	x	x	x	x	x		x		x	x		x	x			x	
1.3 Stages and measurements	x	x	x	x	x	x					x		x	x	x	x	x	x
2 Treatment	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
3 Response		x			x						x		x			x	x	
4 Feasibility tool	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

CDW Bordeaux = Clinical Data Warehouse of Bordeaux University Hospital; CRN = Cancer Registry Norway; DK-DHR = Danish Data Health Registries; EBB = Estonian Biobank; EHR = electronic health records; FinOMOP-ACI Varha = Auria Clinical Informatics; FinOMOP-HUS = Hospital District of Helsinki and Uusimaa; FinOMOP-TaUH Pirha = Tampere University Hospital patient cohort; FinOMOP-THL = Finnish Care Register for Health Care; GP = general practitioner; H12O = Hospital Universitario 12 de Octubre; HI-SPEED = Health Impact - Swedish Population Evidence Enabling Data-linkage; IMASIS = Institut Municipal Assistència Sanitària Information System; InGef RDB = InGef Research Database; k = thousand; M = million; NAJS = Croatian National Public Health Information System; NCR = Netherlands Cancer Registry; PGH = Papageorgiou General Hospital; POLIMI = Research Repository @Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico; SUCD = Semmelweis University Clinical Data; UKBB = UK BioBank.
*These data sources did not report having relevant information available in their source requiring mapping. However, revisions to their current mappings may be needed to ensure consistency in mapping across data sources of the network.

8.5. Study period

The study period will extend from 01/01/2010 up to the most recent available data for each contributing data source. It should be noted that, in some data sources, data availability begins after 2010 (see **Table 1**).

8.6. Variables

Objective 1

1. Individual characteristics will be assessed at the cancer diagnosis date (index date):
    - Sex, categorised into female and male.
    - Age, categorised into 18–65 years and ≥65 years.
    - Performance status, assessed using the Eastern Cooperative Oncology Group (ECOG) score or an equivalent (0 to 4).
2. Prior cancer diagnosis will be assessed any time before the date of diagnosis for the cancer of interest, i.e. [-Inf; -1] days relative to the index date.
3. Cancer-related characteristics and relevant cancer subtypes will be assessed around the index date [-90 days, +90 days] (in the case of multiple records, the latest in this period is reported). Cancer subtypes will be defined primarily by histology and molecular markers. The subtypes will be further refined based on clinical input and data readiness.

    If feasible, some characteristics on this list (e.g. staging, performance status) will also be assessed in 30-day periods during the time windows 90 days before and 1825 days after the index date:

- Subtypes will be described for lung, oesophageal, and breast cancers (**Table 3**); the possibility of describing other subtypes will be explored after reviewing the large-scale characterisation.
- Cancer stage (described using AJCC/UICC Tumour, Node, Metastasis (TNM) classification system; 6th, 7th, and 8th Editions).[3,4,21]
- Cancer grade (1 to 4).
- Cancer-specific biomarkers, including molecular genetics markers.[22,23] A preliminary list of biomarkers is specified for each cancer type in **Table 4**.

4. Cancer medical treatments, radiotherapy, and surgery:

   A predefined list of drugs and procedures will be selected from the relevant therapeutic subgroup of the Anatomical Therapeutic Chemical Classification System (ATC code L01, Antineoplastic agents and other as applicable), aligned with but not limited to the most recent cancer-specific treatment guidelines selected by clinical experts, as it may not cover historical or off-label use.

   Drugs and concepts are available in **Table S2** in **Annex IV**. Radiation therapy and surgical procedures will also be described using SNOMED and ICD-10-PCS codes (see **Table S3** in **Annex IV**).

Table 3. Morphological subtypes for lung, oesophageal, and breast cancers (draft classification).

Cancer	Cancer subtype group	Cancer subtypes (preliminary list)
Lung cancer	Non-small cell lung cancer (NSCLC)	Adenocarcinoma (SCLC)
		Squamous cell carcinoma
		Large cell carcinoma
	Small cell carcinoma	Small cell carcinoma
	Other	Carcinoid
Oesophageal cancer	Based on histology	Oesophageal Squamous Cell Carcinoma
		Oesophageal Adenocarcinoma
	Based on location	Oesophagogastric junction (OGJ)
		Higher oesophageal cancer
Breast cancer	Based on morphology	Infiltrating ductal carcinoma
		Infiltrating lobular carcinoma
		Mixed ductal/lobular carcinoma
	Based on receptor status	Luminal A
		Luminal B
		Human Epidermal Growth Factor Receptor 2 (HER2) positive
		Triple negative

Table 4. Preliminary list of biomarkers to be described in the study.

Biomarker	Relevance (cancer type)
FGFR3 / FGFR2	Bladder
ER/PR	Breast
PIK3CA mutations	Breast

Biomarker	Relevance (cancer type)
BRCA1/2 mutation status	Breast
ESR1	Breast
PD-1/PDL-1	Breast
HER2/ER receptor status	Breast
Ki-67	Breast
RAS	Colorectal
EGFR mutations	Lung
ALK rearrangements	Lung
ROS1 fusions	Lung
KRAS G12C	Lung
MET exon 14 skipping	Lung
MET	Lung
RET fusions	Lung
BRAF V600	Melanoma
NRAS	Melanoma
KIT	Melanoma
LDH	Melanoma
Claudin 18.2 (CLDN18.2)	Oesophageal
PSA	Prostate
HRR genes	Prostate
PSMA	Prostate
HER2 / ERBB2	Tumor agnostic, relevant for breast, oesophageal, colorectal
MSI/MMR	Tumour agnostic, relevant for colorectal, oesophageal, prostate, bladder
PD-L1 expression	Tumour agnostic, relevant for, lung, bladder, oesophageal
BRAF V600E point mutation	Tumour agnostic, relevant for melanoma, lung, colorectal
TMB-high	Tumour agnostic, relevant for melanoma, lung, bladder, colorectal, oesophageal
NTRK	Tumour agnostic, relevant for lung, colorectal
HER2 (IHC 3+)	Tumour agnostic, relevant for breast, lung, colorectal, bladder, oesophageal

ALK, anaplastic lymphoma kinase; BRAF V600, B-Raf proto-oncogene V600 mutation; BRAF V600E, B-Raf proto-oncogene valine-to-glutamic acid substitution at codon 600; BRCA1/2, breast cancer susceptibility genes 1 and 2; CLDN18.2, claudin 18.2; EGFR, epidermal growth factor receptor; ER/PR, oestrogen receptor/progesterone receptor; ESR1, oestrogen receptor 1; FGFR2, fibroblast growth factor receptor 2; FGFR3, fibroblast growth factor receptor 3; HER2 (IHC 3+), human epidermal growth factor receptor 2 overexpression determined by immunohistochemistry score 3+; HER2/ERBB2, human epidermal growth factor receptor 2 (erb-b2 receptor tyrosine kinase 2); HRR genes, homologous recombination repair genes; KIT, KIT proto-oncogene receptor tyrosine kinase (CD117); KRAS G12C, Kirsten rat sarcoma viral oncogene homolog glycine-to-cysteine substitution at codon 12; LDH, lactate dehydrogenase; MET, mesenchymal–epithelial transition factor receptor tyrosine kinase; MET exon 14 skipping, mesenchymal–epithelial transition factor exon 14 skipping alteration; MMR, mismatch repair; MSI, microsatellite instability; NRAS, neuroblastoma RAS viral oncogene homolog; NTRK, neurotrophic tyrosine receptor kinase gene fusion; PD-1, programmed cell death protein 1; PD-L1, programmed death-ligand 1; PIK3CA, phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha gene; PSA, prostate-specific antigen; PSMA, prostate-specific membrane antigen; RAS, rat sarcoma viral oncogene family (including KRAS, NRAS, and HRAS); RET, rearranged during transfection proto-oncogene; ROS1, ROS proto-oncogene 1 receptor tyrosine kinase; TMB-high, high tumour mutational burden.

Objective 2

Treatment regimens mainly include systemic therapies, such as chemotherapy, targeted therapy, hormone therapy, and immunotherapy, as well as guideline-recommended combinations of these therapies. Regimens may broadly follow either a cyclic administration pattern, characterised by predefined treatment cycles, or a continuous administration pattern, characterised by uninterrupted daily treatment until disease progression, unacceptable toxicity, planned treatment completion, or treatment discontinuation.

Considering that treatment regimens in real-world data are often less standardised than those described in clinical guidelines, some flexibility is needed for regimen identification, analysis, and interpretation in this study. Therefore, the extraction procedure will allow for variation in how these regimens are present in the data. The number of cycles will also be captured, if available in the data source.

For example, as shown in **Figure 2**, for the guideline-recommended Doxorubicin and Cyclophosphamide (AC)-Paclitaxel regimen, if prescriptions of doxorubicin and cyclophosphamide are identified matching the expected cyclic pattern in a patient with breast cancer, the patient will be classified as having received the AC regimen. If repeated cycles of paclitaxel are subsequently identified after AC, the patient will be interpreted as having received the AC-Paclitaxel regimen. For tamoxifen regimen, if continuous prescription of tamoxifen is identified in a patient with breast cancer, the patient will be classified as having received the tamoxifen regimen.

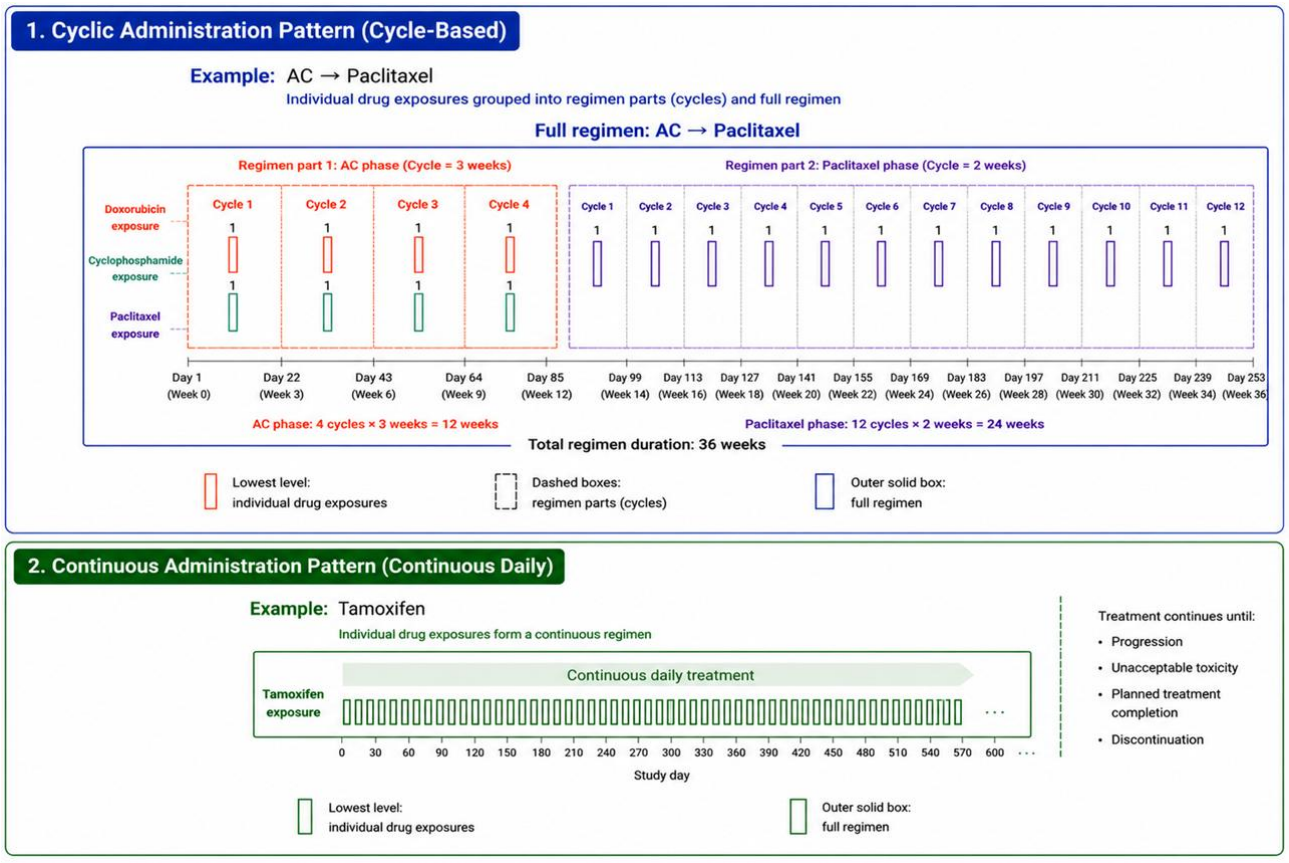


Figure 2. Example of treatment regimen for cyclic and continuous administration patterns.

A=Doxorubicin; C= Cyclophosphamide.

We will define the main regimens per cancer type for regimen extraction following HemOnc definitions (https://hemonc.org) and will prioritise the most important regimens per cancer type for testing based on

clinical guidelines (e.g. European Society for Medical Oncology (ESMO), https://www.esmo.org/guidelines). Both lists will be reviewed by clinical experts.

Lines of therapy will be defined as the sequence of treatment regimen used per cancer type.

Objective 3

Treatment response (complete response, partial response, stable disease, and progression) in solid tumours can be most accurately defined using RECIST-related variables; however, in the absence of such variables, which is likely in RWD, we will explore the potential to identify cancer progression in cancer RWD studies using surrogate variables.

The information on the availability of concepts describing response based on RECIST and surrogate variables will be reviewed under Objectives 1.1, 1.3, and 2. We will pre-define a surrogate (or proxy variable) based on information obtained in these objectives.

In the absence of RECIST, we need to define cancer progression as disease worsening during or immediately after active treatment, without a preceding disease-free interval, resulting in stopping any cancer therapy or switching to a subsequent line of systemic therapy, metastatic disease during an active treatment episode, or upstaging during or immediately after therapy.

Cancer recurrence will be defined as the reappearance of cancer and cancer-related clinical activity after a defined disease-free interval following completion of initial curative-intent treatment. It may also be defined using specific variables or inferred from surrogate signals (e.g. new treatment initiation (which needs to be distinguished from change of therapy due to tolerability without progression and progression), metastatic disease, or palliative care use).

The following additional surrogate variables and their combinations will be explored to identify progression and recurrence in, with a focus on timing (after initial cancer diagnosis and after initial 1st line treatment, which will be identified in Objective 2) (**Table 5**):

- change in cancer stage (upstaging will be specifically assessed over time from the index date based on results from Objective 1);
- information about metastatic disease or carcinomatosis (for non-metastatic cases at the baseline) (see **Table S4** in **Annex IV**);
- new cancer treatment initiated (surgical procedures, the initiation of a new treatment regimen line, or the initiation of radiotherapy);
- any other characteristics and variables considered relevant based on the review in Objective 1.1.

Table 5. Variables related to the definition of progression and recurrence.

Surrogate variables	Definition	Vocabularies
Metastatic disease	New coding of distant or regional metastasis after the disease-free interval	ICD-10, SNOMED
New staging (upstaging)	A coded upstaging event (higher T, N, or M category, or higher stage group)	Cancer Modifier vocabulary: AJCC/UICC stage concepts
New cancer-directed systemic therapy	Dispensing or administration of antineoplastic, targeted, or endocrine therapy. Needs to be distinguished from change of therapy due to tolerability without progression	ATC, RxNorm, HemOnc
New radiotherapy	New course of radiotherapy not consistent with adjuvant schedule	SNOMED procedure codes for external beam radiotherapy, brachytherapy, stereotactic radiosurgery
New cancer surgery	Surgical procedure for cancer treatment, not consistent with combination therapy	SNOMED procedure codes for resection. Specific codes for metastasectomy, debulking, salvage nephrectomy / hepatectomy
Death from cancer (cause of death)	Death certificate with cancer as underlying or contributing cause of death	Cause of death ICD-10

RECIST = Response Evaluation Criteria in Solid Tumours; SNOMED = Systemised Nomenclature of Medicine; T, N, or M = Tumour, Node, or Metastasis; AJCC = American Joint Committee on Cancer; UICC = Union for International Cancer Control; ATC = Anatomical Therapeutic Chemical; RxNorm = Medical prescription normalised.

Table 6 presents the definitions of the endpoints. For each endpoint, we specify the eligible population (defined by tumour stage and treatment intent), the events that trigger the endpoint, the competing risks, the censoring rules, the appropriate statistical estimator, and any specific time-window rules.

Table 6. Definition of endpoints related to progression and recurrence.

Endpoint	Intent	Outcome events	Competing events	Estimator	Censoring
Disease-free survival	Curative	Local or regional recurrence/regrowth, New distant metastases, Death from any cause.	None	Kaplan-Meier	End of follow-up.
Recurrence proportion	Curative	Local recurrence/regrowth, Regional recurrence/regrowth, New distant metastases, death from cancer of interest	Death from any non-cancer cause	Cumulative incidence	End of follow-up
Progression-free survival	Palliative	Disease progression, New distant metastases, death from any cause.	None	Kaplan-Meier	End of follow-up.
Progression proportion	Palliative	Disease progression, New distant metastases, death from cancer of interest	Death from any non-cancer cause	Cumulative incidence	End of follow-up

Objective 4

All cancer-related variables mentioned above will be described.

8.7. Study size

No sample size has been calculated, as this is a descriptive disease epidemiology study which will not test a specific hypothesis. Thus, the sample size is driven by the availability of data for individuals with selected cancers. The preliminary feasibility assessment of the expected number of persons counted for individuals with the cancer type of interest in the data sources is presented in **Table 7**.

Table 7. Number of descendant person counts per data source for general concepts, describing the cancer type of interest.

	Cancer type						
	Malignant melanoma of skin	Malignant tumour of urinary bladder	Malignant tumour of large intestine	Malignant tumour of breast	Malignant tumour of prostate	Malignant tumour of oesophagus	Malignant neoplasm of lower respiratory tract
CDW Bordeaux	9k	5.2k	13.5k	9.2k	12.5k	2.9k	14.8k
CRN	63.4k	32.1k	181k	150k	181k	12.1k	119k
DK-DHR	57.3k	38.3k	137k	158k	118k	18.8k	129k
EBB	1.2k	700	2.5k	3.6k	3.1k	200	1.1k
FinOMOP-ACI Varha	3.6k	3.7k	8.5k	14.6k	14.3k	1.1k	5.3k
FinOMOP-HUS	11.9k	8.1k	20.7k	37.5k	30.4k	2.7k	15k
FinOMOP-THL	29.4k	25.2k	61.7k	109k	105k	6.7k	43.4k
FinOMOP-TaUH Pirha	3.4k	3k	8.5k	14.5k	16.3k	1.3k	6.4k
H12O	500	2.1k	19.5k	12.4k	8.3k	200	6.7k
HI-SPEED	36.1k	60.2k	133k	185k	224k	10k	52k
IMASIS	1.9k	6k	10.4k	13.6k	9k	1.1k	7k
InGef RDB	12.1k	32.1k	57.5k	87.2k	65.9k	8.9k	47.4k
NAJS	22k	26.7k	71.3k	75.9k	51.6k	4.1k	40.4k
PGH	900	3k	4.7k	6.6k	4k	300	200
POLIMI	700	1.7k	2.2k	3.7k	2k	200	2.7k
SUCD	2.8k	6.3k	16.3k	14.5k	8.3k	2.2k	2.6k
NCR	141k	89.7k	366k	422k	321k	55.3k	364k
UKBB	3k	5.2k	11.4k	23.5k	18.2k	2.1k	6.4k

CDW Bordeaux = Clinical Data Warehouse of Bordeaux University Hospital; CRN = Cancer Registry Norway; DK-DHR = Danish Data Health Registries; EBB = Estonian Biobank; EHR = electronic health records; FinOMOP-ACI Varha = Auria Clinical Informatics; FinOMOP-HUS = Hospital District of Helsinki and Uusimaa; FinOMOP-TaUH Pirha = Tampere University Hospital patient cohort; FinOMOP-THL = Finnish Care Register for Health Care; GP = general practitioner; H12O = Hospital Universitario 12 de Octubre; HI-SPEED = Health Impact - Swedish Population Evidence Enabling Data-linkage; IMASIS = Institut Municipal Assistència Sanitària Information System; InGef RDB = InGef Research Database; k = thousand; M = million; NAJS = Croatian National Public Health Information System; NCR = Netherlands Cancer Registry; PGH = Papageorgiou General Hospital; POLIMI = Research Repository @Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico; SUCD = Semmelweis University Clinical Data; UKBB = UK BioBank.

8.8. Analysis

8.8.1. Federated network analyses

All analyses will be conducted separately for each data source and will be carried out in a federated manner, allowing analyses to be run locally without sharing individuals' data.

Before sharing the study package, test runs of the analytics will be performed on a subset of the data sources, and quality control checks will be performed. After all the tests are passed (see **Annex III.**), the final package will be released in a version-controlled study repository for execution against all the participating data sources.

8.8.2. Data privacy protection

The data partners will locally execute the analytics against the OMOP CDM in R Studio and review and approve the default aggregated results. They will then be made available to the Principal Investigators and study team in a secure online repository of DTZ (Data Transfer Zone). All results will be locked and timestamped for reproducibility and transparency. The study results of all data sources will be checked, after which they will be made available to the team, and the Study Dissemination Phase can start. All analyses will be conducted separately for each data source and will be carried out in a federated manner, allowing analyses to be run locally without sharing patient-level data. Cell counts <5 will be suppressed when reporting results to comply with the data source's privacy protection regulations.

8.8.3. Statistical model specification and assumptions of the analytical approach considered

Objective 1.1

The prevalence of the variables describing individuals with cancer will be reported per cancer type at or around the index date.

Prior cancer diagnosis will be described per cancer type during the time window [-Inf; -1 days] relative to the index date.

The prevalence of cancer subtypes will be described per cancer type at the time of cancer diagnosis (based on morphological codes that define cancer diagnosis).

The prevalence of the other cancer-related characteristics (i.e. cancer stage and grade, cancer-specific biomarkers, performance status, cancer treatments, and laboratory tests and procedures) will be described per cancer type around the index date (-90 days; 90 days after, using the latest record), and the use of those characteristics will also be assessed per 30-day period in the time window [-90; 1,825 days] relative to index date. This will inform other objectives, including the best time window to use in Objective 1.3 for defining cancer stage and grade, the misclassification of the cancer diagnosis date, and Objective 3.

Measurements, conditions, observations, and procedures will be described through large-scale characterisation per cancer type 90 days prior to the index date (cancer diagnosis), at the index date, and up to 365 days after the index date.

For each data source, we will calculate counts and percentages for the categorical variables described above, as well as mean, standard deviation (SD), median, and interquartile range (IQR) for continuous variables. Diagnostic test results presented as continuous variables will be categorised based on clinically relevant thresholds, if needed.

The characterisation study will be conducted using the following R packages: *PatientProfiles*[25], *CohortCharacteristics*[26], and *DrugUtilisation*[28].

This objective will run across all cancers and all data partners.

Objective 1.2

We will support OMOP mapping for data sources that contain unmapped cancer-specific information in their source data or information not mapped according to conventions. The focus will be on the following oncology variables: tumour stage and grade, treatment regimens and procedures, performance status, and tumour-specific biomarkers. The result from the already taken feasibility survey, together with initial characterisation (Objective 1.1), will identify information already available in each participating data source and inform how complete the data already is, as well as how it is currently captured (domain, CDM table).

The first step in this objective will be to review the Observational Health Data Sciences and Informatics (OHDSI) oncology working group [29] guidelines and assess their applicability to the DARWIN EU® Data Network.[30] This task will be conducted by two OMOP mapping experts. The second step will be to review these guidelines together with the clinical team involved in this project, as well as the European Medicines Agency (EMA) team, in order to align on the prioritisation of mappings. The third step will be to develop DARWIN EU®-specific mapping guidance and to support data partners in their implementation. Meetings will be organised with data partners to provide guidance on the mapping convention of their oncology data and support in mapping.

A post-mapping characterisation will be conducted by re-running Objective 1.1 in those data sources that have mapped new information. Finally, the results of mapping will then be assessed through reporting of measurements and treatments, both before and after the data mappings from source data to the OMOP CDM, and the mean change in % will be reported.

All data partners are expected to complete mapping to the best of their ability within the timelines of the project.

Objective 1.3

This objective will focus on more detailed phenotype development of cancer subtypes, grades, and stages (AJCC/UICC TNM), and cancer-specific measurements (tumour-specific biomarkers) at and after diagnosis in individuals with the selected cancer types.

Cancer stages and grades: Cancer stage (defined as stage I to IV, depending on the cancer type) will be identified based on AJCC/UICC TNM staging. Morphological cancer grade (defined as grade 1 to 4, depending on the cancer type) will also be described, with additional focus on Gleason grading for prostate cancer.

Results from Objective 1.1 will show the proportion of cancer patients with a record of cancer stage and grade around the index date (-90;+90), but also in different 30-day time windows, 90 days before, and 1825 days after the cancer diagnosis. These results will help refine the optimal time window for identifying cancer stage and grade.

Both clinical and pathological stages will be identified for selected cancers. Cancer stage will be determined based on aggregated stage concepts (stage I, II, III, and IV); if these are not available or incomplete, stage will be derived using individual TNM components registered in the data source and aggregated according to the TNM staging guidelines specific to each cancer type.

Additionally, the mean time between cancer diagnosis and the first cancer stage or grade will be reported in days.

Cancer biomarkers: these will be the biomarkers identified in Objective 1.1. As part of this objective, we will develop definitions to distinguish relevant cancer subtypes based on clinical and biological characteristics.

Objective 1.3 will be completed in two stages. First, the current mapping of cancer morphological subtypes, stage, grade, and tumour-specific biomarkers will be assessed as part of Objective 1.1 to determine the extent and granularity of information available across the data sources. The results of this initial run will

inform the mapping activities under Objective 1.2, guiding the prioritisation of mapping these cancer-specific variables in data sources that contain this information but have not yet been mapped.

This initial run will also guide the development of the study code needed to phenotype these variables across the participating data sources. This study code will first be implemented in data sources where the variables of interest are already mapped to the OMOP CDM, and subsequently in data sources that have completed mapping activities based on the guidelines provided by Objective 1.2.

We will report counts and percentages for the variables mentioned above.

R packages *PatientProfiles*[25] and *CohortCharacteristics*[26] will be used in this objective.

This objective will run across all cancers and all data sources, with the exception of subtypes, which will be defined only for selected cancers.

Objective 2

This objective will enable the analysis of oncology regimen-based treatment patterns (see **Figure 3**).

Figure 3. Flow for construction of regimen-level treatment pathways.

We will develop methods to 1) extract regimens from the available data and 2) support data-driven analysis of treatment pathways that reflect lines of therapy using the extracted regimens. Finally, we will apply these two steps to perform a descriptive characterisation of regimen-level treatment pathways for breast and lung cancer.

For regimen extraction, we will explore using R package *Artemis*[31] as the primary extraction layer to identify and standardise treatment regimens from longitudinal clinical data by grouping drugs into coherent regimen entities. Where *Artemis* does not fully capture more complex real-world scenarios, we will explore complementing this with custom code.

For pathway visualisation, we aim to treat regimens as units within patient journeys. We will explore options to visualise this either based on direct output from *Artemis* and the sequences of regimens, or by integrating it into the R package *TreatmentPatterns*[27], treating each regimen as a step in the pattern.

Where elements are not yet fully defined, we will explore and refine this setup, with the goal of using *Artemis* for robust regimen extraction and *TreatmentPatterns* for intuitive pathway visualisation, supported by targeted custom development, where necessary.

The methodological development will start with breast and lung cancers as work examples, and application will be extended to the other cancers in an iterative way.

Further details on both steps are described below.

1. Regimen Extraction Procedure and Testing

Procedure

Oncology treatments are recorded heterogeneously across the DARWIN EU® Data Network, with data coming from different populations and settings and are collected at different levels of granularity. To

support regimen-level analyses, we aim to accommodate these variations in our methodology. **Figure 4** shows the high-level approach for extracting a regimen based on what is available in the data source:

- For data partners who have indicated the presence of regimen in source data, we will first verify its origin and suitability for use. When it aligns with the regimen definition in this study, we will map it using the HemOnc vocabulary)[32]. However, we anticipate that regimen information will not be available for most data sources.
- For data partners who only have ingredient-level information, the use of *Artemis* for extracting regimen from ingredient-level data will be explored. As input, *Artemis* takes a list of regimens and individual ingredient-level drug exposures. Drug exposures can be filtered to ensure that only treatment-relevant ingredients are used in subsequent steps. *Artemis* uses a modified temporal Smith-Waterman algorithm to find local alignment regions between patient drug records and sequences of drug exposures from predefined regimens. This alignment step can produce multiple candidate regimens; these candidates are processed such that only one most likely candidate remains. *Artemis* also includes functionality to define regimen eras, combining regimens that are the same and occur in short succession into a single era, while keeping distinct regimens in separate eras.
- For data partners who have indicated both ingredient-level and regimen information in the source data, we will use mapped regimens, if available in any data source, as ground truth to validate the package.

A list of regimens of interest will be created, as mentioned in **Section 8.6**.

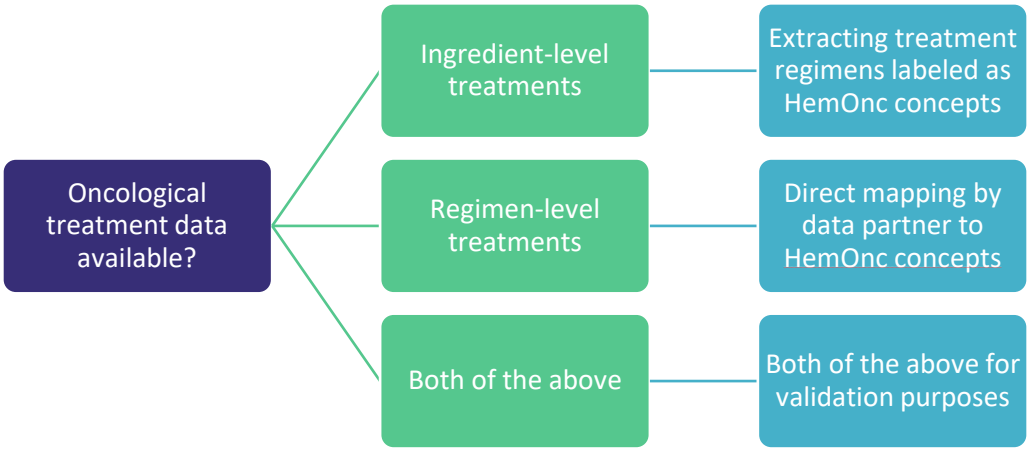


Figure 4. Strategy to convert to regimen-level data based on data availability.

Testing

To evaluate the algorithm, we will perform extensive validation with various synthetic test cases to confirm its functional correctness (e.g. monotherapy regimens, regimens with multiple drugs, regimen discontinuation) and run *Artemis* across a selection of data sources (based on settings and data source set-ups) to assess operational performance.

For test cases, we will develop synthetic patients aligned to predefined regimens to benchmark extraction accuracy. We will then introduce incrementally deviating scenarios (e.g. missing treatments, timing shifts, overlaps, incomplete regimen) to assess robustness and identify failure points, informing clear rules for when extraction is considered valid. Additionally, we will generate non-conformant patients to ensure the approach avoids false-positive regimen assignments.

Key questions to investigate include how the algorithm: 1) handles missing or incomplete information, 2) accounts for variations between RWD and regimen definitions in clinical guidelines, and 3) deals with ambiguous or overlapping regimens. Detected issues will be optimised within the package.

If *Artemis* is insufficient, we will provide suggestions for an alternative solution to be explored in follow-up studies. We will provide short- and long-term recommendations for the regimen extraction procedure to guide and support future oncology studies. Finally, we will run the algorithm to derive regimens for breast and lung cancers across the selected data partners. Furthermore, we will include metrics in this run to evaluate the quality of the derived regimen for each data source. These evaluation metrics will be co-designed in collaboration with EMA and other oncology experts.

2. Regimen-Level Treatment Pathway Characterisation

To perform pathway characterisation, we will first assess the current limitations of *TreatmentPatterns* in oncology research based on previous use cases and studies. We will then adjust the package, as needed, to analyse treatment pathways at the regimen level. Furthermore, the study team will work on additional identified enhancements (e.g. supporting the choice of study parameters, incorporating treatment duration), prioritised based on impact and feasibility and informed by evaluation of other tools, such as *Artemis*. We will test the added features and provide user documentation to support future oncology studies.

Finally, we will use the above to perform a descriptive characterisation of regimen-level treatment pathways. The number and proportion of cancer patients receiving each regimen will be described from 90 days prior to the index date up to 5 years after the index date. For cyclic regimens, the number of cycles will be reported, where available in the data. For continuous regimens, the duration of regimen use will be reported, where available in the data. Treatment sequences from 90 days prior to 5 years after the index date will also be assessed and visualised using sunburst and Sankey plots per cancer. Analyses will be stratified by sex, relevant age groups per cancer type, and disease stage.

Objective 3

The study code for Objective 3 will be run after the additional mapping (Objective 1.2) and definition of lines of therapy (Objective 2) are completed.

We will describe treatment response and cancer progression captured using RECIST-related variables, if available.

We will use three groups of surrogate variables to describe progression and recurrence: upstaging and metastatic disease, new cancer treatment, and supportive care (**Section 8.6**). We will review and refine a set of candidate variables representing potential progression triggers after running Objective 1 and weigh them against available literature and clinical relevance.[33,34]

The variables are cancer-specific; therefore, this exercise will be done for lung cancer and breast cancer first, as working examples, and then extended to all other cancers.

We will then identify the percentage of patients with each cancer type who meet the selected criteria for progression and recurrence within consecutive 30-day periods after the start of first-line therapy.

Validation

Only if explicit information on progression or recurrence (e.g. RECIST) is available in any of the data sources, we will validate the use of surrogate variables in that specific data source. The following metrics should be reported for each outcome:

- Sensitivity: proportion of true recurrences or progressions captured by the surrogate algorithm;
- Specificity: proportion of true non-events correctly classified as such;

- Negative predictive value (NPV): proportion of algorithm-identified non-events confirmed by the reference standard;
- Positive predictive value (PPV): proportion of algorithm-identified recurrences or progressions confirmed by the reference standard.

Definitions of outcomes

Real-world progression-free survival (RW-PFS) will be defined as the time from the diagnosis date to cancer progression or death from any cause. An alternative definition will be applied in the competing risk scenario: it will use death from any cause as a competing event. If cause of death information is available, death from cancer will be considered as an event, as well as progression.

Real-world disease-free survival (RW-DFS) is defined as the time from the date of diagnosis to recurrence of disease or death from any cause. It will typically only be measured after surgery with neoadjuvant/adjuvant treatment for selected cancer types. A competing risk scenario will also be applied, similar to RW-PFS.

Results will be compared to the literature and reviewed by clinical experts.

In general, we will follow a framework that partially aligns with the consensus presented by Vissers et al.[12] for defining and operationalising time-to-event endpoints in RWD oncology studies.

The work will start with breast and lung cancers as work examples and will extend to the other cancers in an iterative way.

Objective 4

Aggregated counts from the variables described in Objectives 1 to 3 will be made available in a dashboard for the DARWIN EU® Coordination Centre (CC) personnel. This dashboard will inform future feasibility requests by showing which data sources have the relevant variables for the potential new oncology study for specific cancer cohorts.

The aggregated counts of information recorded among cancer patients will be collected from data partners upon run and re-run of the code of Objectives 1 to 3.

The aggregated counts on the new mapped information could be collected from data partners at each release of their OMOP CDM data after execution of the *DashboardExport* package[35]. This package is run by all data partners in the DARWIN EU® Data Network during onboarding or OMOP CDM updates, which means that all data partners in DARWIN EU® will generate these aggregates upon a new data release, ensuring full coverage of the network for future feasibility requests. After execution of the *DashboardExport* package, aggregated counts will be added to the DARWIN EU® Portal (a data source catalogue containing metadata of each of the DARWIN EU® data sources, which is used internally to inform feasibility assessments).

The type of aggregate counts included in the DARWIN EU® Portal will include:

- Record count (RC): Counts representing the total number of occurrences of a specific variable.
- Person count (PC): Counts representing the number of unique individuals who have a record of a specific variable.
- Value distribution (Dist): Set of numerical statistics: mean, standard deviation, minimum, maximum, median, 1st, and 3rd quartile.

When reporting, a minimum cell count of 5 will be used, and results will be rounded up to the nearest 100.

This information will be supplemented by a qualitative assessment of the data source's fitness for purpose for oncology studies, with regard to setting, representativeness, completeness of cancer-related information, and accuracy of death information.

8.8.4. Output

The output of this project is complex and contingent upon the progress of the mapping process, as well as other contributing factors. A draft proposal for the output will be stored in a separate supplementary file. The final deliverable may be revised at a later stage, depending on the scope, quality, and availability of the data.

8.9. Evidence synthesis

Results from all analyses described will be presented separately for each data source. No meta-analysis of results will be conducted.

9. STRENGTHS AND LIMITATIONS

This is one of the first comprehensive initiatives to improve the conduct of oncology studies using RWD, potentially benefiting studies conducted outside of the DARWIN EU® Data Network. Improving data representativeness and analytical tools for oncology research will enable the conduct of large-scale, complex studies. This study will focus on improving data representativeness across the network and on developing new tools for future use.

Some limitations should be mentioned. The study will be informed by routinely collected healthcare data, so data quality issues must be considered. We expect that identification of cancer patients will be possible across all participating data sources; however, the granularity of the data will vary by data source type and setting. Moreover, the availability of a history of comorbidities and concomitant medications prior to cancer diagnosis, which is necessary for patient characterisation, may not be available in cancer registries and might be limited in secondary care data sources.

Another limitation concerns how certain variables are currently mapped to the OMOP CDM; for example, the mapping of drug exposure duration may vary across data sources for the same drugs. The development of guidelines in this study should address the mapping of new variables and the limitations of heterogeneous mapping, which could be improved in upcoming releases of CDM across the network. However, a lack of access to the source document will limit our ability to assess the quality of the data.

We provide guidelines and support for the new or updated mappings (Objective 1.2). However, the mappings will be implemented by the data partners themselves, with support from the DARWIN EU® CC, creating a dependency on their internal timelines for new CDM releases. This risk can be mitigated by setting clear goals for the new releases and monitoring progress per individual data partner. Besides Teams, the meetings with data partners will also be used to communicate the timelines and give data partners the opportunity to voice any concerns.

The lack of explicit information on treatment response (RECIST) and recurrence might limit our ability to validate algorithms in Objective 3. In the future, as the DARWIN EU® Data Network expands and the volume of data collected across individual data sources increases, validation may be feasible. Additionally, limited information on progression limits the ability to estimate progression-free survival.

Another limitation is that this study focuses on solid tumours in adults, and the results may not be directly applicable to haematological malignancies and childhood cancers.

Additionally, the results estimated in this study will reflect only the populations represented in the included data sources. Countries in Eastern Europe are underrepresented in the DARWIN EU® Data Network, which might affect the generalisability of the results.

Finally, many of the tools and R packages used in DARWIN EU® studies were not developed to specifically address questions in oncology and require additional adaptation.

10. REFERENCES

1. Saesen R, Lacombe D, Huys I. Real-world data in oncology: a questionnaire-based analysis of the academic research landscape examining the policies and experiences of the cancer cooperative groups. *ESMO Open*. 2023;8(2):100878. doi:10.1016/j.esmoop.2023.100878
2. Ramsey SD, Onar-Thomas A, Wheeler SB. Real-World Database Studies in Oncology: A Call for Standards. *JCO*. 2024;42(9):977-980. doi:10.1200/JCO.23.02399
3. Brierley J, Eycken E van, Rous BA, Giuliani M, O'Sullivan B, eds. *TNM Classification of Malignant Tumours*. Ninth edition. Wiley; 2025.
4. Brierley J, Gospodarowicz MK, Wittekind C. *TNM Classification of Malignant Tumours*. 8th ed. John Wiley & Sons, Incorporated; 2017.
5. DeVita VT, Lawrence TS, Rosenberg SA, eds. *Devita, Hellman, and Rosenberg's Cancer: Principles and Practice of Oncology*. Twelfth edition. Wolters Kluwer; 2023.
6. Mokhtari RB, Homayouni TS, Baluch N, et al. Combination therapy in combating cancer. *Oncotarget*. 2017;8(23):38022-38043. doi:10.18632/oncotarget.16723
7. Hudis CA, Barlow WE, Costantino JP, et al. Proposal for Standardized Definitions for Efficacy End Points in Adjuvant Breast Cancer Trials: The STEEP System. *Journal of Clinical Oncology*. 2007;25(15):2127-2132. doi:10.1200/JCO.2006.10.3523
8. Oxnard GR, Morris MJ, Hodi FS, et al. When Progressive Disease Does Not Mean Treatment Failure: Reconsidering the Criteria for Progression. *JNCI Journal of the National Cancer Institute*. 2012;104(20):1534-1541. doi:10.1093/jnci/djs353
9. Gourgou-Bourgade S, Cameron D, Poortmans P, et al. Guidelines for time-to-event end point definitions in breast cancer trials: results of the DATECAN initiative (Definition for the Assessment of Time-to-event Endpoints in CANcer trials). *Annals of Oncology*. 2015;26(5):873-879. doi:10.1093/annonc/mdv106
10. Digkas E, Tabiim AJ, Smith D, Valachis A. Randomized Versus Real-World Evidence on the Efficacy and Toxicity of Checkpoint Inhibitors in Cancer in Patients with Advanced Non-small Cell Lung Cancer or Melanoma: A Meta-analysis. *Targeted Oncology*. 2022;17(5):507-515. doi:10.1007/s11523-022-00901-1
11. Lu Y, Langerman SS, McCain E, et al. Response- and Progression-Based End Points in Trial and Observational Cohorts of Patients With NSCLC. *JAMA Network Open*. 2024;7(5):e249286. doi:10.1001/jamanetworkopen.2024.9286
12. Vissers PAJ, Elferink MAG, Bijlsma MJ, et al. Recommendations and definitions for time-to-event endpoints using real-world data in oncology. *ESMO Real World Data and Digital Oncology*. 2025;10:100649. doi:10.1016/j.esmorw.2025.100649
13. Lim AM, McDowell L, Hurt C, et al. Assessment of endpoint definitions in curative-intent trials for mucosal head and neck squamous cell carcinomas: Head and Neck Cancer International Group consensus recommendations. *The Lancet Oncology*. 2024;25(7):e318-e330. doi:10.1016/S1470-2045(24)00067-6
14. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*. 2009;45(2):228-247. doi:10.1016/j.ejca.2008.10.026

15. Mahvi DA, Liu R, Grinstaff MW, Colson YL, Raut CP. Local Cancer Recurrence: The Realities, Challenges, and Opportunities for New Therapies. *CA A Cancer J Clinicians*. 2018;68(6):488-505. doi:10.3322/caac.21498

16. Apolo AB, Milowsky MI, Kim L, et al. Eligibility and Radiologic Assessment in Adjuvant Clinical Trials in Bladder Cancer. *JAMA Oncol*. 2019;5(12):1790. doi:10.1001/jamaoncol.2019.4114

17. Griffith SD, Miksad RA, Calkins G, et al. Characterizing the Feasibility and Performance of Real-World Tumor Progression End Points and Their Association With Overall Survival in a Large Advanced Non–Small-Cell Lung Cancer Data Set. *JCO Clinical Cancer Informatics*. 2019;(3):1-13. doi:10.1200/CCI.19.00013

18. Chubak J, Yu O, Pocobelli G, et al. Administrative Data Algorithms to Identify Second Breast Cancer Events Following Early-Stage Invasive Breast Cancer. *JNCI: Journal of the National Cancer Institute*. 2012;104(12):931-940. doi:10.1093/jnci/djs233

19. Nors J, Iversen LH, Erichsen R, Gotschalck KA, Andersen CL. Incidence of Recurrence and Time to Recurrence in Stage I to III Colorectal Cancer: A Nationwide Danish Cohort Study. *JAMA Oncol*. 2024;10(1):54. doi:10.1001/jamaoncol.2023.5098

20. Hassett MJ, Uno H, Cronin AM, Carroll NM, Hornbrook MC, Ritzwoller D. Detecting Lung and Colorectal Cancer Recurrence Using Structured Clinical/Administrative Data to Enable Outcomes Research and Population Health Management. *Medical Care*. 2017;55(12):e88-e98. doi:10.1097/MLR.0000000000000404

21. Amin MB, Greene FL, Edge SB, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA A Cancer J Clinicians*. 2017;67(2):93-99. doi:10.3322/caac.21388

22. Van De Haar J, Roepman P, Andre F, et al. ESMO Recommendations on clinical reporting of genomic test results for solid cancers. *Annals of Oncology*. 2024;35(11):954-967. doi:10.1016/j.annonc.2024.06.018

23. Zhou Y, Tao L, Qiu J, et al. Tumor biomarkers for diagnosis, prognosis and targeted therapy. *Sig Transduct Target Ther*. 2024;9(1):132. doi:10.1038/s41392-024-01823-2

24. Warner JL, Dymshyts D, Reich CG, et al. HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. *Journal of Biomedical Informatics*. 2019;96:103239. doi:10.1016/j.jbi.2019.103239

25. Catala M, Guo Y, Du M, et al. PatientProfiles: Identify Characteristics of Patients in the OMOP Common Data Model. Published online March 5, 2025. Accessed April 8, 2025. https://cran.r-project.org/web/packages/PatientProfiles/index.html

26. Catala M, Guo Y, Lopez-Guell K, Burn E, Mercade-Besora N, Alcalde M. *CohortCharacteristics: Summarise and Visualise Characteristics of Patients in the OMOP CDM*. https://darwin-eu.github.io/CohortCharacteristics/

27.. Markus AF, Verhamme KMC, Kors JA, Rijnbeek PR. TreatmentPatterns: An R package to facilitate the standardized development and analysis of treatment patterns across disease domains. *Computer Methods and Programs in Biomedicine*. 2022;225:107081. doi:10.1016/j.cmpb.2022.107081

28. Burkard T, López-Güell K, Gorbachev A, et al. Calculating daily dose in the OBSERVATIONAL MEDICAL OUTCOMES PARTNERSHIP COMMON DATA MODEL. *Pharmacoepidemiology and Drug*. 2024;33(6):e5809. doi:10.1002/pds.5809

29. OHDSI Oncology Working Group. Oncology Conventions. Accessed February 12, 2026. https://ohdsi.github.io/OncologyWG/conventions.html

30. Belenkaya R, Gurley MJ, Golozar A, et al. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. *JCO Clinical Cancer Informatics*. 2021;(5):12-20. doi:10.1200/CCI.20.00079

31. Golozar A, Lawrence-Archer L, Zack T, Warner JL. Introducing ARTEMIS: Advanced Regimen Detec8on Using an Adapted Smith- Waterman Algorithm. In. 2023 [cited 2026 Oct 6]. p. 3. Available from: https://www.ohdsi.org/wp-content/uploads/2023/10/Golozar-Asieh_Introducing-ARTEMIS-Advanced-Regimen-Detection_2023Symposium-Asieh-Golozar.pdf

32. Warner JL, Dymshyts D, Reich CG, et al. HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. *Journal of Biomedical Informatics*. 2019;96:103239. doi:10.1016/j.jbi.2019.103239

33. Joshi V, Adelstein B, Schaffer A, et al. Validating a proxy for disease progression in metastatic cancer patients using prescribing and dispensing data. *Asia-Pac J Clncl Oncology*. 2017;13(5). doi:10.1111/ajco.12602

34. Nordstrom BL, Simeone JC, Malley KG, et al. Validation of Claims Algorithms for Progression to Metastatic Cancer in Patients with Breast, Non-small Cell Lung, and Colorectal Cancer. *Front Oncol*. 2016;6. doi:10.3389/fonc.2016.00018

35. darwin-eu/DashboardExport [R] [Internet]. Darwin EU® Public Code Repository; 2026 [cited 2026 Jun 10]. Available from: https://github.com/darwin-eu/DashboardExport

11. ANNEXES

ANNEX I. Description of data sources

Croatia: Croatian National Public Health Information System (NAJS)

#	Section	Description
1	Data source identification and country	NAJS (Croatian National Public Health Information System) Croatia
2	Data partner information section	Croatian Institute of Public Health (CIPH) Department of Data Science and Analytics
3	Coverage and timespan	Data collection since: 1998 Extent: Nation-wide. Geographic coverage covers whole Croatia, with various levels of resolution for different registries. Current estimates for the population in Croatia will be available at: https://podaci.dzs.hr/hr/podaci/stanovnistvo/procjena-stanovnistva/ for each year. The total and active person count in the NAJS data is larger than the current population of Croatia. This explained by: a) the person table included deceased and all previously insured people and b) there is no information about insurance ending, c) healthcare is also used by people with dual citizenship from neighbouring countries It is known that a lot of people emigrated (300k-400k) and weren't included in the last population census but still are in the NAJS database. There is also an influx of immigrant workers that are insured and registered but weren't included in the census.
4	Healthcare setting / type of data	Primary care – General Practitioner, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. Primary care – gps, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. For both inpatient and outpatient setting diagnoses, medication, procedures, and measurements are captured. The year of availability of information depends on the setting • 2014-2025 for biochemical lab tests in primary care from EHR patients records (measurements with results) • 2015-2025 for primary care data from EHR patient records (conditions, procedures, and drug prescriptions) • 2015- 2024 for inpatient hospital data from EHR administrative records (conditions, procedures, measurements without results and drug administrations) • 2016-2025 for health risk assessment data entered by GPs (measurements with results - height, weight…) • 2016-2022 for secondary conciliatory care data from EHR administrative records (conditions, procedures, measurements without results and drug administrations) • 2016-2022 for emergency care data from EHR patient records (conditions) • 2017-2025 for hospital records from registry data (conditions and procedures) • 2020-2025 for vaccination data from EHR patient records
5	Data collection process	Inpatient hospital billing systems, and Other. Data is entered by clinicians at healthcare contact, then combined by CIPH into the NAJS database and integrated with registries for public health purposes.
6	General representativeness	The data is collected from the evidence of public health records collected for public health purposes, as the majority of health care in Croatia is public and under single health insurance provider. Personal details are collected to a better extent for insured individuals compared to uninsured patients, who are excluded in the ETL process.
7	Data content /source coding	Medication prescriptions are recorded with ATC codes with an additional 3-digit code denoting the package. Diagnoses with ICD10 codes (Australian modification). Procedures with local source codes. Lab results with local source codes.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Records from 2017 include insured patients with reliable IDs. Uninsured patients do not have

#	Section	Description
		reliable IDs. For example, if a patient changed her status from insured to uninsured, or vice versa, she could be counted several times, as could tracking records from before 2017 and after. By using the unique personal identifier for Croatian citizens, it can be checked and verified.
9	Quality control (data source specific)	There is a network of registry personnel (leaders, administrators, coders, sources) working on data coverage and other quality dimensions. An analytical team routinely checks for erroneous entries in hospital records, removing double entries, false dates, and overlapping stays. Entries without enough data or with obviously erroneous dates from primary care analysis are being excluded.
10	Linkage	The national death registry is updated yearly, with one year lag, but the fact of someone's death (just the date) is updated daily, without the cause of death or any other additional details. Primary care is updated weekly and hospital level care monthly. Specific registries are included in NAJS (e.g. diabetes registry), where inclusion criteria vary across these registries.
11	Vital status	NAJS is linked to the national death registry.
12	Limitations	Hospital data is available from 2017 onwards. This is often used as start of data collection, while laboratory and GP data is captured before that (since 2014 and 2015 respectively). Drug duration is often not available and set to 1 day for administration and 30 days for prescription. Hospital discharge summaries are currently not captured in NAJS. Hospital drug administration data is less reliable than prescription data from primary care, with some drugs (monoclonal antibodies / precision medicine drugs) that require additional approval not being recorded at all.
13	Main references	No main reference provided.
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111155 Website: https://www.hzjz.hr/nacionalni-javnozdravstveni-informacijski-sustav-najs/

Denmark: Danish Data Health Registries (DK-DHR)

#	Section	Description
1	Data source identification and country	DK-DHR (Danish Data Health Registries) Denmark
2	Data partner information section	Danish Medicines Agency (DKMA) Data Analytics Centre (DAC)
3	Coverage and timespan	Data collection since: 1995 Extent: Nation-wide. The data is representative of the entire Danish population.
4	Healthcare setting / type of data	Community pharmacists, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. The following data elements are collected: diagnosis (including rare diseases and pregnancy data), hospital admissions, discharge and ICU data, Cause of death, Drug prescriptions, dispensing, vaccination and contraception, Procedures (surgical and non-surgical hospital), and Sociodemographic information (sex and age only).
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Registries, and Other. All causes of deaths, all retrieved drug prescriptions, all records of vaccinations, all hospital inpatient and outpatients contacts including disease diagnoses and hospital surgical and non-surgical procedures, histologically confirmed incident cancers, laboratory test results for the entire Danish population from 1/1/1995 onwards.
6	General representativeness	The data is representative of the entire Danish population. Healthcare is free in Denmark, so we do not expect any bias in data collection based on socio-economic status.

#	Section	Description
7	Data content /source coding	Diagnoses and causes of death are collected using the ICD-10 vocabulary. ATC and RxNorm are used for Drugs. SNOMED codes are used for Procedures.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients have unique identifiers used to link datasets.
9	Quality control (data source specific)	The data we have received relating to nationwide Danish Health Data registries offer an opportunity for large-scale, population-based studies with several advantages 1) Their large size improves the precision of estimates and enables the study of rare exposures and outcomes with long-term latency, 2) Inclusion of nearly all individuals in the target population ensures that the data reflect routine clinical care and all clinical segments of the source population, 3) Data are collected independently of each research study, thus minimising certain types of bias, e.g., non-response, and the influence from attention to the research question on the diagnostic process. Before the source data is sent to us, the Danish Health Data Authority runs and does comprehensive checks of the registry table data validity of the variables, breaks in data, changes in variable coding, missingness, etc. We perform checks of missingness/completeness in relation to requested variables. In essence, we are receiving a dump of a mirror of the data that is controlled by the SDS. The documentation performed by SDS is available online, in Danish primarily https://www.esundhed.dk/Dokumentation (all variables), but also in English https://sundhedsdatastyrelsen.dk/da/english/health_data_and_registers/national_health_registers
10	Linkage	There is no linkage in this data source.
11	Vital status	The Cause of Death registry (DAR) is used, the cause of death is collected using ICD-10 codes.
12	Limitations	DK-DHR has the following limitations, which may be relevant confounders for certain complex DARWIN EU studies: - We lack information on key socio-economic status (SES) factors, such as occupation, education, and income. These variables may be important for analysis in some studies. - We only have complete data on lifestyle factors (such as smoking status and weight) for pregnant women. - We have no information on patient contacts in primary care (visits to the GP). Consequently, the incidence of chronic diseases like Type 2 Diabetes (T2D) and asthma must be determined using drug prescriptions as a proxy. Stillborn children will not have any records in our CDM. This means that e.g. birth length of stillborns is not recorded.
13	Main references	Schmidt M, Schmidt SAJ,Adelborg K,Sundbøll J,Laugesen K,Ehrenstein V,Sørensen HT "The Danish health care system and epidemiological research: from health care contacts to database records." Clinical epidemiology (2019): 31372058
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111217 Website: https://sundhedsdatastyrelsen.dk/da/english/health_data_and_registers/healthdatadenmark

Estonia: Estonian Biobank (EBB)

#	Section	Description
1	Data source identification and country	EBB (Estonian Biobank) Estonia
2	Data partner information section	University of Tartu (TARTU) Institute of Computer Science
3	Coverage and timespan	Data collection since: 2004 Extent: Nation-wide. EBB is a nation-wide database containing records from 2004 onwards. Estonian population-based cohort size of 211,800 participants (01/01/2024) aged 18 years and older recruited at GP offices, private practices, and hospitals or in the recruitment offices of the Estonian Genome Center.

#	Section	Description
4	Healthcare setting / type of data	Primary care – General Practitioner, and community pharmacists, and primary care specialists (e.g. paediatricians), and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. Healthcare providers must report the data to national registries. The data includes claims, prescriptions, EHR and two registries - the cause of death and cancer registry. The claims data covers diagnosis, procedures and services performed. Claims data contains information about services paid by the Estonian Health Insurance Fund and not the services paid by the person themself. The prescription data contains information prescribed and dispensed prescription drugs. It does not contain information about the over-the-counter drugs. The EHR data includes diagnosis, results of the procedures and lab measurements and case summaries. The cause of death and cancer registry data is checked by the registry before it is entered to the registry. Also, Estonian Biobank, has collected genetic and questionnaire data when the person gives consent. Questionnaire data includes personal data (e.g. nationality), genealogical data (family history of medical conditions), educational history, lifestyle data (e.g. smoking, alcohol consumption etc).
5	Data collection process	Insurance/administrative claims, and Outpatient electronic health records, and Inpatient hospital electronic health records, and Registries, and Biobank, and Other. Data is retrieved by Estonian Biobank once a year from national registries. The insurance claims are requested from Estonian Health Insurance Fund. The inpatient and outpatient electronic health records are requested from National Health Information System. The cancer registry and cause of death registry information is requested from The National Institute for Health Development. The data is sent to the national registry by the healthcare providers.
6	General representativeness	The age, sex, and geographical distribution closely reflect those of the Estonian adult population and encompass approximately 20% of adult population. Female participants are over-represented in EBB. Older people tend to participate less frequently, however, all age groups are well represented.
7	Data content /source coding	All participants have undergone a standardized health assessment, including provision of blood samples for purification of DNA, white blood cells, and plasma, and completed a questionnaire covering various health-related topics, such as lifestyle, diet, and clinical diagnoses. Diseases and health problems are recorded as ICD-10 codes and prescribed medicine according to the ATC classification and local package codes. Procedures and services are coded with NOMESCO classifier and local service codes. Lab measurements are coded with local codes and in later years with LOINC codes. In cancer registry also ICD-O-3 codes are used. The data is available with different granularity. For claims data we only now that procedure or service is done. The result of the procedure or measurement is recorded in EHR. Usually, the result is in unstructured text (except lab measurements) and needs specific pipelines for extraction.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. There is one national identifier that allows linking together all encounters across databases.
9	Quality control (data source specific)	The quality control procedures in the Estonian Biobank aim to remove the most obvious mistakes in the data, misspellings, impossible dates, duplicates. Before performing the ETL, several problems are fixed on the source data. Since the ETL procedures are used for a number of different datasets (from the same national sources), we have a growing number of pre-processing steps that correspond to the issues we have discovered previously in the data, such as checking for the presence of critical values, harmonizing date and unit of measurement formats, checking the validity of certain entries against classifiers, etc.
10	Linkage	Follow-up data are available via linkage with national health-related registries and via re-examination of participants. Furthermore, electronic health records are updated for phenotypic outcome information every year. The EBB database is regularly linked with national registries, hospital databases, and the databases of the Estonian Health Insurance Fund (EHIF) and the National Health Information System (NHIS).

#	Section	Description
11	Vital status	Vital status (death date and causes of death) is obtained from the Causes of Death Registry.
12	Limitations	No database-specific limitations documented. General limitations for the data type applicable.
13	Main references	Milani L, Alver M,Laur S,Reisberg S,Haller T,Aasmets O,Abner E,Alavere H,Allik A,Annilo T,Fischer K,Hofmeister R,Hudjashov G,Jõeloo M,Kals M,Karo-Astover L,Kasela S,Kolde A,Krebs K,Krigul KL,Kronberg J,Kruusmaa K,Kukuškina V,Kõiv K,Lehto K,Leitsalu L,Lind S,Luitva LB,Läll K,Lüll K,Metsalu K,Metspalu M,Mõttus R,Nelis M,Nikopensius T,Nurm M,Nõukas M,Oja M,Org E,Palover M,Palta P,Pankratov V,Pantiukh K,Pervjakova N,Pujol-Gualdo N,Reigo A,Reimann E,Smit S,Rogozina D,Särg D,Taba N,Talvik HA,Teder-Laving M,Tõnisson N,Vaht M,Vainik U,Võsa U,Yelmen B,Esko T,Kolde R,Mägi R,Vilo J,Laisk T,Metspalu A "The Estonian Biobank's journey from biobanking to personalized medicine." Nature communications (2025): 40188112
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111114 Website: https://genomics.ut.ee/en/content/estonian-biobank

Finland: Auria Clinical Informatics (FinOMOP-ACI Varha)

#	Section	Description
1	Data source identification and country	FinOMOP-ACI Varha (Auria Clinical Informatics) Varsinais-Suomi, Finland
2	Data partner information section	FinOMOP Hospital District of Southwest Finland
3	Coverage and timespan	Data collection since: 2004 Extent: Regional. The data covers the patient register at the Hospital District of Southwest Finland (HDSF), containing Turku University Hospital, which is one of the five university hospitals in Finland. The region of Southwest Finland is the third largest of the 21 wellbeing counties in Finland.
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. The data covers the public specialist health care and most emergency health care in the area of Southwest Finland (Varsinais-Suomi) for all demographic groups. The most relevant data domains are patients, visits, inpatient episodes, diagnoses, laboratory results, procedures, medication, pathology, radiology, radiotherapy, chemotherapy, obstetrics, and narrative patient reports,
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Inpatient hospital billing systems, and Biobank. Data is entered by clinicians at point of contact.
6	General representativeness	The data includes all residents of Southwest Finland that have needed specialist health care or emergency health care.
7	Data content /source coding	Source data is coded in ICD10, and several local coding systems: MIKROBI_FIN_VSSHP (microbes list), NCSP_MODIFIER_FIN (procedure modifiers from the Nomesco NCSP vocabulary), ICD-10-FIN (Finnish modification of ICD-10), PATO_DGN_FIN (Snomed II –based pathology organ-diagnosis pairs), Hilmo_eala (Finnish national medical specialty classification), UNIT_FIN (measurement units list found in data), NCSP-FIN (Finnish modification of Nomesco NCSP procedure classification), LAB-FIN (Finnish national lab test classification), LAB-FIN-VSSHP (local additions to Finnish national lab test classification), ATC. Available in prescribed medication data but not yet brought into the CDM: Vnr (Nordic medicine product identifiers https://www.laaketietokeskus.fi/en/pharmaceutical-information/vnr-services)
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Yes, but only briefly and only if the patient is unidentifiable upon arrival to the hospital. However,

#	Section	Description
		the differing ID numbers are always harmonized to the correct single ID number as soon as the patient's identity is determined.
9	Quality control (data source specific)	The source databases are operational EHR systems. We bring the data from them to the ACI analytical data warehouse. We monitor the ETL processes and get automated alerts if data is not accumulated at an expected rate. We have separate warehouse environments for development, test, and production. Raw data is never used for research. We have a data refinery pipeline that includes the following steps: 1) convert raw data to relational format (PostgreSQL), 2) contact clinicians and nurses who are experts in the specific data source, 3) understand the important concepts and locate them in source data, 4) form preliminary views on top of raw data that include the essential data and makes it easier to interpret, 5) calculate preliminary statistics and basic info from the view, and have clinical experts evaluate the correctness and compare to existing reports taken directly from possible reporting tools in the data source or other similar data sources, 6) make necessary modifications to views and repeat clinician evaluations until results are satisfactory, 7) document all relevant information to view/column comments and ACI Wiki, 8) train data users to use new views, 9) continually improve the views based on feedback from data users.
10	Linkage	No known linkages.
11	Vital status	Source for vital status unknown.
12	Limitations	No database-specific limitations documented. General limitations for the data type applicable.
13	Main references	No main reference provided.
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111111 Website: https://www.auria.fi/tietopalvelu/en/index.html

Finland: Hospital District of Helsinki and Uusimaa (FinOMOP-HUS)

#	Section	Description
1	Data source identification and country	FinOMOP-HUS (Hospital District of Helsinki and Uusimaa) Uusimaa, Finland
2	Data partner information section	FinOMOP HUS Tietohallinto (Information Management)
3	Coverage and timespan	Data collection since: 2004 Extent: Regional. HUS is responsible for specialized healthcare in Finland's Uusimaa region and treatment of many rare and severe diseases that are nationally centralized to HUS.
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. All visits, examinations, laboratory test, procedures, and treatments are recoded in the HUS IT systems and integrated into the data lake. The data lake stores decades of clinical information in digital format, and data from both past and current source systems are available. Systems providing data into the data lake include: CGI Uranus and Epic Apotti (EHR: visits, diagnoses, medication, etc.), Opera (procedure records), Kemokur and Beacon (cancer-specific medications), Marela (hospital pharmacy), Multilab/Mylab+ (laboratory system), Qpati/Mylab+ (pathology records system). MerlotMedi (prehospital emergency care), Disease specific quality registers (disease specific). Obstetrix (birth register).
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Other. All visits, all procedures, and all given treatments have been recorded systematically in the electronic format. We use more than hundred different operational IT systems.

#	Section	Description
6	General representativeness	The data reflects patients treated at the hospital and is thus not a generic population.
7	Data content /source coding	Drugs: ATC, generic name, brand name + dosage and strength information in the data + local route mapping Nordic product number (VNR) Procedures: Nomesco NCSP Genetic: special mappings to OMOP Genomics Diagnosis: ICD10 + Finnish adaptation "ICD10fi", ICD9fi, ICD8fi Laboratory tests: National laboratory test numbers (LABfi kuntaliitto) + local HUS variants National microbe listings, local "additional information" -coding Various observations during care: LOINC and local coding Pathology: SNOMED2
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. All patients are identified with a unique, national social security number, which is a permanent person identifier. Based on this, the derived patient identifiers (pseudonyms) in the HUS data lake also are unique.
9	Quality control (data source specific)	The HUS data lake and IT management is ISO 13485:2016 and ISO 9001:2015 certified and has capabilities in developing, validating, and CE-marking medical devices and software. Data passes through three layers of validations and transformations (bronze, silver, cold i.e. medallion lakehouse architecture) before being stored in a layout optimized for efficient analytics. The data is monitored for data freshness and data anomaly. OMOP data is converted from silver level. DARWIN EU® CC data quality dashboard SQL queries are implemented HUS ETL pipeline in addition of the standard quality control that is being done together with the DARWIN EU® CC upon onboarding.
10	Linkage	HUS is linked with Digital and Population Data Services Agency from where national death registry, social security number and address are imported. Data from many nation-wide registries can be combined at HUS, which is subject to obtaining special data permits.
11	Vital status	The source of vital status (date of death) is the Digital and Population Data Services Agency.
12	Limitations	No database-specific limitations documented. General limitations for the data type applicable.
13	Main references	No main reference provided.
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111134 Website: https://www.hus.fi/en

Finland: Tampere University Hospital patient cohort (FinOMOP-TaUH Pirha)

#	Section	Description
1	Data source identification and country	FinOMOP-TaUH Pirha (Tampere University Hospital patient cohort) Pirkanmaa, Finland
2	Data partner information section	FinOMOP Department of Research, Development and Education
3	Coverage and timespan	Data collection since: 2007 Extent: Regional. TaUH Research Database includes all specialities/all patient groups treated in the Tampere University Hospital,
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. Secondary, and tertiary care given in the region, including given clinical and pathology diagnoses,

#	Section	Description
		diagnostic and therapeutic procedures, laboratory findings, radiology and pathology reports, medication given in the hospital and electronic prescriptions, and continuous medical records (free text), including discharge letters since 2007.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records. The data is entered by the clinician at point of contact and processed in a data warehouse for secondary use.
6	General representativeness	The data covers patients visiting the hospital and therefore will not reflect the general population.
7	Data content /source coding	ICD10fi: Finnish modification of ICD10; SNOMED2-TMP: a local SNOMED2-based vocabulary of organ-diagnosis pairs in pathology; NCSPfi: Finnish national version of NCSP/Nomesco vocabulary for procedures; LABfi: Finnish national vocabulary (Kuntaliitto) for laboratory measurements; LABfi_TMP: Local additions to the national vocabulary for laboratory measurements; UNIT_FIN: measurement units; MICROBEfi_TMP: Microbe names in measurement data; ATC: all drugs except anticancer drugs; active ingredient (in Finnish): for anticancer drugs; VNR: Nordic medicine product identifiers (https://pharmaca.fi/en/health/pharma/vnr/); Hilmo_eala: Finnish national medical specialty vocabulary.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. In some cases, patients can be re-registered with a new ID. Patients are identified by their social security number, and each patient gets a patient ID in the EHR, associated with the social security number. However, there are cases when the social security number has changed, and the associated patient ID also has changed. Approximately 4% of all patient IDs in the database are such duplicates. Majority of such duplicates are newborns who have first been assigned a temporary social security number, which has later been changed to a permanent social security number. An unidentified person can have been given a temporary social security number, and sometimes a person can change their social security number, for example, because of threat (stalking) or gender reassignment. It is technically possible to link all the instances of a person in the database however this is not currently done.
9	Quality control (data source specific)	The source database tables are checked for completeness on a general level: whether there are null values in the fields, what the date ranges are for each table, are there gaps in the data, and how many unique patients can be found in each data table. Before mapping to the OMOP CDM, we have identified the data tables and the fields in the tables that are required to build the OMOP CDM. The White Rabbit and Rabbit-in-a-Hat tools were used in this process.
10	Linkage	Data from many nation-wide registries can be combined, which is subject to obtaining special data permits.
11	Vital status	The source of vital status (date of death) is the Digital and Population Data Services Agency.
12	Limitations	No database-specific limitations documented. General limitations for the data type applicable.
13	Main references	No main reference provided.
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111135 Website: https://www.pirha.fi/en/web/english

Finland: Finnish Care Register for Health Care (FinOMOP-THL)

#	Section	Description
1	Data source identification and country	FinOMOP-THL (Finnish Care Register for Health Care) Finland
2	Data partner information section	FinOMOP The Department of Data and Analytics
3	Coverage and timespan	Data collection since: 1998 Extent: Nation-wide. The current CDM population comprises all persons having been alive and residing in Finland since the beginning of 2011.
4	Healthcare setting / type of data	Primary care – General Practitioner, and primary care specialists (e.g. paediatricians), and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and other (specify). THL maintains health registers that cover both public and private, primary, and specialised inpatient, urgent and outpatient health care encounters in Finland, starting from 2011. The entire public sector and private inpatient encounters have been included since 2011, while private outpatient encounters, including occupational care, are included since 2020. Since 1998, the Care Register for Health Care (TerveysHilmo) has covered both public outpatient and inpatient specialized care and private inpatient care. Since 2009, the Finnish National Vaccination Register has covered all vaccinations from the public sector and from a large part of private vaccination providers, with the data coverage from both sections being very good to complete from 2020 onwards. Since 2011, the Register of Primary Health Care Visits (AvoHilmo) has covered public primary care. Since 2020, the register has also covered private outpatient care and occupational care. In addition, the CDM also contains positive COVID-19 test results from the Finnish National Infectious Diseases Register. The register itself covers all laboratory confirmations for around 70 specific microbes from 1995 onwards, but only COVID-19 has currently been mapped to the CDM. The CDM also includes prescription records from multiple different sources. Both Care Register for Health Care and Register of Primary Health Care Visits contain very basic prescription data recorded during health care encounters that include just the ATC-code and trade name of medication. More comprehensive prescription data from Kanta Prescription Centre, maintained by Social Insurance Institution of Finland (Kela), has been integrated into the CDM since its 2024 release. The Kanta Prescription records are based electronic prescriptions, which were adopted by most public health care providers in 2010 and by most private providers by 2017.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Registries. Data is entered by clinicians upon healthcare contact and processed by THL (Kela in the case of Kanta Prescription Centre).
6	General representativeness	The THL data has national coverage and is therefore well representative of the Finnish population. Using the complete population as a basis for the person table also serves to facilitate calculations on a population level, e.g. incidence rates.
7	Data content /source coding	The following coding systems have been OMOP-mapped, typically to a good level of completeness: ICD10fi Finnish Extension, ICPC-2, ATC, Toimenpideluokitus (procedure classification adapted from the Nordic Classification of Surgical Procedures (NCSP)), Terveydenhuollon erikoisalat (Hilmo specific provider speciality), Rokotustapa (AR/YDIN National classification for vaccine administration), Tupakointistatus (AR/YDIN National classification for smoking status). Vaccinations are identified on product level based on batch number, trade name, vaccine title, and ATC-code. This is mapped on brand and type in the OMOP CDM.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Each patient in THL has a unique identifier.

#	Section	Description
9	Quality control (data source specific)	The source data collection undergoes a structural and semantic validation before entry into the source database. Additionally, some coded variables undergo quality assessment against the respective code systems post entry into the database. The source registers are also assessed for completeness and coverage, with the aim of improving future collection in the areas where data is lacking.
10	Linkage	THL is already a linkage of multiple Finnish registries (see above).
11	Vital status	The National Population registry data forms the basis for forming the patient population. This ensures an up-to-date location (municipality of residence) of patients, as well as complete death occurrences (although not the cause of death).
12	Limitations	All drug records in the CDM are currently based on prescriptions. Kanta Prescription Centre also includes information on drug dispensing, but these have not currently been converted into OMOP CDM. Depending on the type of the medication, a single prescription can be for up to two years dosage of the drug. This means a patient can have up to two-year breaks between observations while actively using the drug. Observation of a prescription also does not mean that the patient necessarily bought or used the medication. The CDM does not currently have meaningful end dates or days of supply for drug exposure. This information is not available for Care Register for Health Care or Register of Primary Health Care Visits, and for Kanta Prescription Centre it is only available in unstructured, free-form format that has not been converted into OMOP CDM. The source system for prescription drug data is only available to the DP until Dec-2023, therefore drug_exposure records from this source are only available in that time range. Not all private health care records are covered for the CDM's entire follow-up time from 2011 onwards. For Register of Primary Health Care Visits and Finnish National Vaccination Register, records from private health care have been available from 2020 onwards. For Kanta Prescription Centre, the coverage of private health care records has been good from 2017 onwards. The inclusion of private health care mainly presents itself as an increase in the number of observations, meaning that it has to be accounted for when interpreting any time series data from the CDM.
13	Main references	Häkkinen, Pirjo; Mölläri, Kaisa; Saukkonen, Sanna-Mari; Väyrynen, Riikka; Mielikäinen, Lasse; Järvelin, Jutta "Hilmo - Sosiaali- ja terveydenhuollon hoitoilmoitus 2020 : Määrittelyt ja ohjeistus : Voimassa 1.1.2020 alkaen" Terveyden ja hyvinvoinnin laitos (2019):
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111187;https://catalogues.ema.europa.eu/data-source/1111191 Website: https://thl.fi/en/statistics-and-data/data-and-services/register-descriptions;https://www.kanta.fi/en/research-and-knowledge-management

France: Clinical Data Warehouse of Bordeaux University Hospital (CDW Bordeaux)

#	Section	Description
1	Data source identification and country	CDW Bordeaux (Clinical Data Warehouse of Bordeaux University Hospital) Nouvelle-Aquitaine, France
2	Data partner information section	Centre Hospitalier Universitaire de Bordeaux (CHUBX) Gironde / Nouvelle-Aquitaine
3	Coverage and timespan	Data collection since: 2005 Extent: Regional. It covers the population of Bordeaux metropolitan area, and possibly beyond, as the health care centre for referrals and expertise for the Nouvelle Aquitaine region. The database contains data from 2005 onwards.
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and claims data. The database currently captures information about patient demographics, visit details, conditions, procedures, drugs, measurements, and mortality.

#	Section	Description
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Inpatient hospital billing systems, and Biobank. The integrated data is extracted from the hospital production information system via a real-time research database. The data is then processed and quality controlled by a team dedicated to maintaining the database. Internal evaluations were carried out to ensure consistency between the research database and the patient bedside software.
6	General representativeness	This is the 6th largest metropolitan area in France, and CHUBX is the largest hospital in the region. More than 75% of the patients administered to Bordeaux university hospital reside in the Gironde departments, with almost 50% coming directly from the Bordeaux metropolitan area. The hospital also captures additional cases from Nouvelle-Aquitaine region.
7	Data content /source coding	Diagnosis source data is coded in ICD-10 terminology. Procedures are coded in CCAM (French terminology). Laboratory measurements are coded in local terminology and partially mapped to LOINC. Drugs are coded through a local terminology and then mapped to UCD (French terminology), as well as ATC codes.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. We use the hospital's unique identifier to generate the patient identifier in OMOP. If two identities are merged at the hospital, the merge is taken into account in the CDW. An automatic (hourly) detection of suspected duplicated identities has been implemented at the hospital since 2020, with merging of duplicated identities by a specialized team. Identities since 2015 were processed retrospectively. Thus, the rate of identity duplication in the database is low, especially since 2015.
9	Quality control (data source specific)	- The integrated data comes from the hospital production information system through a real-time replicated database. Consistency evaluations between the replicated database and the production system are performed by the technical team in charge of maintaining the replicated database. - In the same way, consistency checks are performed between the replicated database and the data integrated into the i2b2 CDW. In addition, dashboards enable monitoring the data integrated into the i2B2 CDW, in particular by controlling the amount of data available over time, and its evolution, according to the various data sources. - An internal evaluation was carried out to ensure the consistency between the data integrated into i2b2 and the data available in the software used at the patient's bedside. In addition, many use cases were performed on the i2b2 CDW, with return to the patient chart and comparison of the data integrated into i2b2 and the data available in the care file.
10	Linkage	Death certificates (without the cause of death).
11	Vital status	The database is linked to the French death registry.
12	Limitations	CDW Bordeaux is limited to events captured in the hospital setting and thus does not include patient events not treated by the hospital (e.g. rare cancers). Patient events that are not included in CDW Bordeaux are rare disease treatments or specialist events that occur outside of CHUBX.
13	Main references	Cossin S, Diouf S,Griffier R,Le Barrois d'Orgeval P,Diallo G,Jouhet V "Linkage of Hospital Records and Death Certificates by a Search Engine and Machine Learning." JAMIA open (2021): 33709061
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111112 Website: https://www.chu-bordeaux.fr/

Germany: InGef Research Database (InGef RDB)

#	Section	Description
1	Data source identification and country	InGef RDB (InGef Research Database) Germany
2	Data partner information section	Institute for Applied Health Research Berlin GmbH (InGef) InGef is a research service provider within the German statutory health insurance system.
3	Coverage and timespan	Data collection since: 2014 Extent: Nation-wide. The Research Database contains approximately 10.5 million insured persons from 50 of the 94 statutory health insurances in Germany (as of Oct 2025). All individuals in the research database are included in the OMOP CDM. Due to data protection laws for personal data in Germany, only the last 10 years of data are available in the RDB and thus in the OMOP CDM.
4	Healthcare setting / type of data	Primary care – General Practitioner, and community pharmacists, and primary care specialists (e.g. paediatricians), and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and claims data. InGef is a research service provider within the German statutory health insurance system. In general, the data in the RDB reflects most of the health services paid for by the statutory health insurances. In the OMOP CDM, the health services are reduced to primary and secondary care. This means that treatments by GPs and specialists (e.g. paediatricians), prescription medicines dispensed by pharmacies, hospital inpatient and outpatient care, as well as information about all insured individuals are included. The following data elements are presented in the OMOP CDM: demographic information, diagnoses, procedures, dispensing drugs and advanced therapy medicinal products, vaccinations, pregnancy data (via diagnoses and procedures) and contraception.
5	Data collection process	Insurance/administrative claims. In Germany, data exchange between medical service providers and statutory health insurances is organized via central data collection centers. In addition to receiving and forwarding data, these centers also store data and provide access to third parties under strict regulations in data warehouses. The RDB contains selected, condensed and anonymized information from one of these data warehouses. Data sovereignty remains with the individual health insurance companies.
6	General representativeness	The RDB covers about 11% of the German population and is comparable to the German population in terms of the distribution of age and sex. Most health insurances that contribute to the RDB have nationwide coverage, meaning that the database covers all regions of Germany. Since almost all services covered by statutory health insurances are specified in national legislation, healthcare provision all over Germany is well represented in the RDB. Additionally, in Germany it is very common to stay with the same health insurance throughout life, which results in a good longitudinal coverage over the entire period of 10 years.
7	Data content / source coding	The coding in the research database complies with national classification and coding rules in Germany. Diagnoses are coded according to ICD-10-GM. Inpatient and outpatient surgeries or procedures are recorded as OPS codes (German classification of Operations and Procedures). The dispensing of drugs in pharmacies is recorded using the PZN (pharmaceutical registration number). For drugs that miss a PZN-to-RxNorm mapping, the ATC code is used instead. In some cases, dispensed drugs can be coded using OPS codes (e.g. in hospitals) or EBM codes (fee schedule for outpatient treatments).
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network.
9	Quality control (data source specific)	The data transmitted by healthcare providers complies with the standardized requirements and formats of the Association of Statutory Health Insurances (GKV-SV). Before being imported into the research database, the data elements are checked for data format, completeness and

#	Section	Description
		plausibility. After each update of the research database, various counts are compared with the previous update to verify completeness. Due to the anonymity of the database, direct validation of the data (e.g., using medical records as the gold standard) is not possible.
10	Linkage	Due to the anonymization of the source data, linkage is not possible.
11	Vital status	The date of death is recorded as the last day of the quarter in which the death occurred (i.e. 30/31st of Mar/Jun/Sept/Dec) as reported to the health insurance (no linkage to death registry). The cause of death is not available.
12	Limitations	Ambulatory diagnoses and procedures are summarised in the source on a quarterly basis. Both are mapped to the observation table with the date set to the last day of the respective quarter (i.e. 30/31st of Mar/Jun/Sept/Dec) and the concept "History of event within 3 months" (observation_concept_id 1340222), with the actual diagnosis or procedure concept_id recorded in the field "value_as_concept_id". There is no vocabulary for the German pharmaceutical product codes (PZN). A direct source-to-standard-mapping has been done manually by InGef but is incomplete. The drug exposure duration is unknown. Following OMOP conventions, the end date is always set to dispensing date + 29. Outpatient and inpatient procedures are recorded as OPS codes (German Procedure Classification), for which the vocabulary is incomplete. Pharmacy & Hospital data having a general lag of 2 calendar months, while outpatient data has a lag in delivery of up to 9 calandar months before it can be considered complete. Data deliveries occur separately per health insurance, and the information about which health insurance a person is insured with is removed during the anonymisation process. This means that, per patient, we cannot explicitly determine whether the lack of data means that their insurance ended/they were not diagnosed or treated/they did not die in that time period, or whether we don't see any records because their insurance company has not yet delivered its most recent data for that sector. This means that conclusions drawn using data for the last few quarters in the release could be inconsistent. Drugs mostly contain pharmancy-dispensed drugs. We generally don't have records of drugs administered by a healthcare professional as part of a treatment (inpatient as well as outpatient. An exception are drugs that hospitals bill separately using OPS procedures, for example expensive treatments.
13	Main references	Andersohn F, Walker J "Characteristics and external validity of the German Health Risk Institute (HRI) Database." Pharmacoepidemiology and drug safety (2016): 26530279
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111207 Website: https://www.ingef.de/en/

Greece: Papageorgiou General Hospital (PGH)

#	Section	Description
1	Data source identification and country	PGH (Papageorgiou General Hospital) Central Macedonia, Greece
2	Data partner information section	Papageorgiou General Hospital (PGH) Program Management Office
3	Coverage and timespan	Data collection since: 1999 Extent: Other. Patients from across the Macedonia region and from neighbouring countries as well.
4	Healthcare setting / type of data	Primary care – General Practitioner, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and other (specify). long-term/skilled nursing facility

#	Section	Description
5	Data collection process	Insurance/administrative claims, and Outpatient electronic health records, and Inpatient hospital electronic health records, and Inpatient hospital billing systems. The data is gathered live (as the patient is examined) in the production EHR instance.
6	General representativeness	The database represents best the population of Northern Greece requiring hospital care. It captures data on hospitalizations, ICU admissions, medical procedures, and measurements. With a capacity of 745 beds, PGH supports over 200,000 hospitalization days annually. The OMOP CDM has around 1.41M patients.
7	Data content /source coding	Medications are coded using ATC codes. KEN («Κλειστά Ενοποιημένα Νοσήλεια») Coding System and Internal Hospital Coding System: This coding system is applied to procedural data, and it was enforced by the Greek Public Health Ministry. It has been mapped to ICD-10. Internal Hospital Coding System: This system is employed for coding medical devices. ICD-10 (International Classification of Diseases, Tenth Revision): This system is utilized for coding both 'Diagnoses' and 'Deaths' from the source. Internal LIS Coding System: This coding system, associated with the internal laboratory information system, is used for laboratory tests and examinations.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. In Greece, each patient is identified by a unique universal health insurance number.
9	Quality control (data source specific)	There is no formal quality assurance plan in place for the raw data collected during daily clinical practice, therefore, gaps or errors in the data should be anticipated.
10	Linkage	No known linkages.
11	Vital status	There are only in-hospital deaths available in the data.
12	Limitations	Drugs recorded at the active ingredient level. Drug dosages and routes are not captured at the moment. Drug exposure presents the drug ordered for the patient and not the administration (in the majority of the cases the patient is given the ordered drug but this might not be true for 100% of the cases). Drug exposure dates also refer to the order date. Medical devices have not been mapped. Medical history is missing. Patient progress is not available after discharge. Vital measurements like body temperature, BMI, blood pressure are not captured and Observations in general are missing.
13	Main references	Papapostolou G, Chytas A,Rekkas A,Bigaki M,Zeimpekis D,Dermentzoglou L,Tortopidis G,Natsiavas P "Real-World Data in Greece: Mapping the Papageorgiou General Hospital Data to the OMOP Common Data Model." Studies in health technology and informatics (2024): 39176625
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/institution/1000000215 Website: https://www.papageorgiou-hospital.gr/?lang=en

Hungary: Semmelweis University Clinical Data (SUCD)

#	Section	Description
1	Data source identification and country	SUCD (Semmelweis University Clinical Data) Budapest, Hungary
2	Data partner information section	Semmelweis University (SU) -
3	Coverage and timespan	Data collection since: 2010 Extent: Regional. The general catchment area of SU is the central region of the country, Budapest city and Pest county, although patients can be referred from anywhere in Hungary. The total population of Budapest and Pest county is approximately 4,200,000 people. The total population of Hungary is around 9,500,000.

#	Section	Description
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and claims data, and other (specify). diagnostic data (laboratory tests, radiology, pathology)
5	Data collection process	Insurance/administrative claims, and Outpatient electronic health records, and Inpatient hospital electronic health records, and Inpatient hospital billing systems, and Registries. Data is extracted directly from the source database. From there, the data entry in the system is heavily controlled and validated on the user interface before being made available for further research.
6	General representativeness	SU captures information on patients who are covered by the public health insurance system. This covers all Hungarian citizens, and therefore the database should mirror the source population well. Although, besides Semmelweis University Clinics, there are multiple hospitals in the region, and data on visits in other hospitals in not represented in the database. Therefore, the patient population is not directly representative of the general population.
7	Data content /source coding	Regarding SU's source data, procedures and diagnoses are coded in SNOMED, measurements are coded in LOINC, and drugs are stored in RxNorm and ATC.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients have a unique identifier (SSN).
9	Quality control (data source specific)	The clinical database is the source database and therefore it has to be treated as a trusted database. Data entry in the systems is heavily controlled by validation on the user interface, and there are large number of rules that controls the data on the insurer's side that has to be corrected in the system by the users to be able to close the encounters. OMOP mapping is done in the framework by EHDEN recognized partners under quality check by the EHDEN society.
10	Linkage	No known linkages.
11	Vital status	Source for vital status unknown.
12	Limitations	Medication prescribed in secondary care is fully present in our database, but medication given in the hospital is rarely documented. General limitations for the data type applicable. General practitioner data is not present in our database; therefore, this part of the patient journey is not represented.
13	Main references	No main reference provided.
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1000000184 Website: https://www.semmelweis.hu

Italy: Research Repository @Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico (POLIMI)

#	Section	Description
1	Data source identification and country	POLIMI (Research Repository @Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico) Lombardia, Italy
2	Data partner information section	Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico (POLIMI) Scientific Direction
3	Coverage and timespan	Data collection since: 2006 Extent: Regional.

#	Section	Description
		The geographic coverage area includes the city of Milan and the region of Lombardy. The region of Lombardy, including Milan, has a total population of around 10 million people.
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. The following data elements are collected: diagnoses (including rare diseases and pregnancy data), hospital admission and discharge information (including ICU data), drug prescription, dispensing and vaccination information and clinical measurements.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records. The data is collected from EHRs.
6	General representativeness	The database has a population of all age groups and 27% number of hospitalized patients/number of residents in Milan.
7	Data content /source coding	The drug prescription and dispensing have ATC and RxNorm codes. The diagnoses and procedures are coded in ICD-9-CM and SNOMED.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Yes, and there is an internal process to merge the patient back to the original.
9	Quality control (data source specific)	The source databases are the ones provided by each software provider, so there is no quality control except for some field data types during the input if the form allows it.
10	Linkage	There is a linkage with the ER, for each hospitalization we search if that admission is subsequent of an ER access. In that case we link the ER access to the hospitalization visit.
11	Vital status	There is no linkage with a death registry, and the cause of death is not captured. Only in-hospital deaths are covered.
12	Limitations	The drug route and quantity are currently missing. There is a low mortality rate because only in-hospital deaths are covered. There is also no information around devices and contraception. Only hospital data is captured, so no GP visits or registries like the death registry.
13	Main references	Ageno W, Cogliati C,Perego M,Girelli D,Crisafulli E,Pizzolo F,Olivieri O,Cattaneo M,Benetti A,Corradini E,Bertù L,Pietrangelo A "Clinical risk scores for the early prediction of severe outcomes in patients hospitalized for COVID-19." Internal and emergency medicine (2021): 33620680
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111226 Website:

The Netherlands: Netherlands Cancer Registry (NCR)

#	Section	Description
1	Data source identification and country	NCR (Netherlands Cancer Registry) Netherlands
2	Data partner information section	Integraal Kankercentrum Nederland (IKNL) Department of Research and Development
3	Coverage and timespan	Data collection since: 1989 Extent: Nation-wide. The NCR compiles clinical data of all individuals newly diagnosed with cancer in the Netherlands. Data managers register newly diagnosed cancer patients, starting in 1989, on a national basis, with 3 million patients included.

#	Section	Description
4	Healthcare setting / type of data	Hospital inpatient care, and other (specify). The NCR is a disease-specific registry that contains characteristics of patients, details about the cancer diagnosis and the primary treatment, and information on the vital status which is updated once a year.
5	Data collection process	Registries. Data managers employed by the Netherlands Comprehensive Cancer Organisation (IKNL) enter and process the data.
6	General representativeness	The data has nationwide coverage in The Netherlands of people having a cancer diagnosis since 1989. Data since 1993 is OMOP standardised.
7	Data content /source coding	ATC codes, ICD-O-3. The diagnosis contains the type of cancer (ICD-O-3) and the stage (TNM). In addition, internally developed coding for diagnostic/therapeutic procedures.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. No, we use identifiable data for linkage and identification across hospitals.
9	Quality control (data source specific)	All data is collected and checked by data managers. These are some of the quality control procedures: - All data managers receive a training when they start, and they get yearly refresher courses. - The registration application limits what can be registered, depending on certain disease characteristics. This prevents a lot of errors. - A team of data managers is responsible for quality (werkgroep kwaliteit). - Random samples of registered data are checked. - Automatic checks are done. - Researchers, who use the data, can request quality checks if they suspect issues.
10	Linkage	Established linkages: - Pathology data from PALGA - Vital status by Personal Records Database
11	Vital status	Yealy updated linkage with Personal Records Database
12	Limitations	The database only has patients with a cancer diagnosis. The primary treatment is registered. There are some lab results, mainly around the diagnosis and relevant for primary treatment. There is no sociodemographic information. Only ATC codes are available for some drugs, and this can only be mapped to ingredient level in OMOP.
13	Main references	Gijs Geleijnse, RuRu Chun-Ju Chiang, Melle Sieswerda, Melinda Schuurman, K. C. Lee, Johan van Soest, Andre Dekker, Wen-Chung Lee & Xander A. A. M. Verbeek "Prognostic factors analysis for oral cavity cancer survival in the Netherlands and Taiwan using a privacy-preserving federated infrastructure" Scientific Reports (2020):
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111117 Website: https://iknl.nl/international

Norway: Cancer Registry Norway (CRN)

#	Section	Description
1	Data source identification and country	CRN (Cancer Registry Norway) Norway
2	Data partner information section	Norwegian Institute of Public Health (NIPH) Cancer Registry of Norway
3	Coverage and timespan	Data collection since: 1953 Extent: Nation-wide.

#	Section	Description
		The catchment area is all of Norway. The approximate population is around 5.6 Million. The cancer-specific clinical registries have collected nationwide data since the mid-2010s and are continuously updated.
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. The following data elements are collected: diagnoses, pathology, lab reports, clinical notifications, deaths, procedures, measurements, drug prescription, and dispensing. The CRN also includes data from various clinical cancer registries for specific cancer types, providing extended information on diagnosis, treatment and follow-up. For example, there are clinical registries for lung, breast, gynecological (ovarian, cervical), prostate, melanoma, sarcoma, and other cancers, which supplement the main incidence registry.
5	Data collection process	Registries. Sources of information for the cancer registry are: clinical notifications from hospitals and specialists (public and private), pathology reports from pathology departments (public and private), radiation therapy and SACT-information from hospitals, data from the Norwegian Patient Registry, the Population Registry and the Cause of Death Registry, and PREM and PROM from patients. The data sources send data to the Cancer Registry, mainly in electronic form. Data is received and pre-processed before they are further enriched and processed by the data controllers, following a standard SOP for data quality checks. In addition, strict quality assessment procedures are carried out to manually evaluate the quality of the information regarding general statistics and cancer-specific domain data.
6	General representativeness	The Cancer Registry of Norway is a population-based cancer registry covering the entire Norwegian population. Registration of cancer patients in the CRN is mandatory in Norway, with no possibility for opt-out. The registry therefore includes all cancer patients (including premalignant cases and some benign tumours of the CNS) and provides a comprehensive representation of the national cancer population.
7	Data content /source coding	ICD03 and ICD10 are used for topography/morphology/disease of the cancer diagnosis. Some examples of other used classifications are ISUP, ECOG, TNM and SNOMED. The cancer registry also has extensive in-house classifications developed through over 70 years of cancer registration. Drugs prescriptions and dispensing is recorded with ATC codes, indications for the prescriptions are recorded with ICD10 codes. All information on the content of the CRN can be found in our online metadata-catalogue at https://metadata.kreftregisteret.no
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. All patients have 11-digit unique national identifiers for identification and linkage.
9	Quality control (data source specific)	The CRN follows a strict quality assurance regime, including validation of all incoming data and a rule engine that enforces international, national, and internal standards uniformly. Manual checks are performed for specific cancer types to identify unlikely data, such as melanoma cases with tumour diameters exceeding 2 cm. Data quality is documented in scientific articles and annual statistics, including metrics like morphological verification rate, estimated completeness, proportion of cases identified only via death certificates, and male-to-female ratios.
10	Linkage	Every citizen has a unique personal identifier, and via this identifier, patients can be linked to other national data sources.
11	Vital status	Vital status is regularly updated in the Cancer Registry of Norway through linkage with the National Population Register. Information on cause for death is obtained from the Cause of death Registry through linkage.
12	Limitations	No database-specific limitations documented. General limitations for the data type applicable.

#	Section	Description
13	Main references	Larsen IK, Småstuen M,Johannesen TB,Langmark F,Parkin DM,Bray F,Møller B "Data quality at the Cancer Registry of Norway: an overview of comparability, completeness, validity and timeliness." European journal of cancer (Oxford, England : 1990) (2009): 19091545
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111128 Website: https://www.kreftregisteret.no/en/

Spain: Hospital Universitario 12 de Octubre (H12O)

#	Section	Description
1	Data source identification and country	H12O (Hospital Universitario 12 de Octubre) Comunidad de Madrid, Spain
2	Data partner information section	Hospital Unibrtdizstio 12 de Octubre (H12O) Instituto de Investigación (i+12)
3	Coverage and timespan	Data collection since: 2015 Extent: Regional. H12O is a national reference for the treatment of certain pathologies covering patients in the southern area of Madrid region.
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and other (specify). The H12O database, INFOBANCO, contains information from the different health domains (laboratory, prescriptions, treatments, administrative, diagnoses, etc.). In addition, information is also obtained from other data sources, such as the pathological anatomy system, which provides information about sample analysis, and the cost system, containing information on the cost associated with a contact with the hospital. Work for the inclusion of further data is ongoing, among others, radiological information, or PROMs.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Inpatient hospital billing systems, and Registries, and Other. Data is entered by clinicians and processed in the regional 'INFOBANCO' platform. This platform allows combining data from multiple heterogeneous sources, and provides mechanisms for data governance, adequacy, interrogation, visualization and analysis for real-world evidence generation and decision support.
6	General representativeness	H12O only covers patients visiting the hospital due to a medical condition and is therefore not necessarily representative of the general population.
7	Data content /source coding	Diagnosis: ICD-10-CM, IDC-9, SNOMED CT, ORPHA; Procedures: ICD-10-PCS, ICD-9, SNOMED CT; Medication: ATC and SNOMED CT; Laboratory tests: LOINC; Histopathology: ICD-O-3.1 and SNOMED CT; Clinical observations: SNOMED CT
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients cannot be registered with different IDs. Each patient has a unique identifier in the hospital's EHR (NHC), a regional unique identifier (CIPA), and a national unique identifier (CIP-SNS).
9	Quality control (data source specific)	The EHR contains certain rules for recording information, e.g. it does not allow closing reports that do not have a coded diagnosis. In addition, it allows the use of terminologies, such as SNOMED or LOINC, for recording information, which helps to reduce unstructured data.
10	Linkage	The hospital data can be linked with pharmacy and the Madrid public health service.
11	Vital status	The hospital is connected to the CIBELES system of the Madrid public health service, which allows the unique identification of users and contains the date of death of the patient. In this

#	Section	Description
		way, the hospital has the information of the death, regardless of whether it occurred in the hospital or not.
12	Limitations	No database-specific limitations documented. General limitations for the data type applicable.
13	Main references	Miguel Pedrera-Jiménez, Noelia Garcia Barrio, Antonio J Díaz Holgado, Pablo Serrano-Balazote, et al. "INFOBANCO: Standardized platform for management, semantic interoperability, and transparent reuse of EHRs" Conference: EIT Health German-Spanish Symposium on Health DataAt: Mannheim, Germany (2022):
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111145 Website: https://cpisanidadcm.org/infobanco

Spain: Institut Municipal Assistència Sanitària Information System (IMASIS)

#	Section	Description
1	Data source identification and country	IMASIS (Institut Municipal Assistència Sanitària Information System) Catalunya, Spain
2	Data partner information section	Consorci Mar Parc de Salut Barcelona (PSMAR) Direcció de Control de Gestió (Management and Control Department)
3	Coverage and timespan	Data collection since: 1990 Extent: Regional. The primary catchment area covers the Barcelona city area, in Catalonia (Spain). However, all recorded visits are included, which may also include patients residing outside this area.
4	Healthcare setting / type of data	Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. The available information includes diagnoses, procedures, drug administrations, dispensations and prescriptions, laboratory tests, visit types, cancer data, and vaccinations. These data are linked to primary care centers, which currently contribute information on diagnoses, and treatments, as well as to mortality data.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Other. The data collection process begins at the point of care, when a patient is attended in any facility within our hospital network or associated primary care centers. Clinical information is recorded directly into the electronic health records (EHRs) and subsequently complemented with data from external sources, including the regional vaccine registry, the cancer registry, the national mortality registry, and outpatient treatment records. All information is consolidated into a relational database using standardized formats and subjected to multiple quality control procedures to ensure completeness and consistency. The resulting datasets are then extracted, transformed, and loaded (ETL) into an OMOP Common Data Model (CDM) database on a quarterly to semi-annual basis.
6	General representativeness	The data source includes patients across all socioeconomic levels, ages, and genders, suggesting that the dataset is representative of the population of Barcelona requiring secondary (hospital-based) healthcare at any given time.
7	Data content /source coding	Health outcomes, diagnoses, and procedures are recorded using ICD-9-CM and ICD-10-CM classification. Laboratory tests are recorded in the source data using LOINC terminology and drugs using the national codes classification. Cancer registry is based on ICD-O-3 classification. Additionally, a set of demographics data are available such as age and sex.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. A patient cannot be registered under different ID numbers. They are registered under the same ID, even when changing practice of re-registration or admission.

#	Section	Description
9	Quality control (data source specific)	Several quality control procedures have been implemented during the creation and update of IMASIS. The process includes a series of filters that exclude any data whose format or content do not match the specifications of the target variables. Some examples of these actions include the following. - Records containing codes that do not belong to the standardized vocabularies used for encoding are reviewed within their respective domains and, if necessary, discarded. - Records with implausible values, such as dates preceding a patient's birthdate, are identified and corrected or removed. - Before each IMASIS update, the number of records in the source files is verified. Record counts from the new IMASIS instance are compared with both the source files and previous IMASIS versions to detect potential inconsistencies or anomalies in the source data. - In addition, if the Data Quality Dashboard (DQD) package identifies quality issues in the OMOP instance, for example diagnoses recorded for patients with an implausible gender, these discrepancies are reviewed and corrected in the source data.
10	Linkage	Information of diagnosis and treatments received in outpatient settings and in primary care as well as national mortality registry, cancer registry and regional vaccination registry.
11	Vital status	The hospital's death registry records the age at death for each patient when the event occurs within the hospital. In addition, deaths occurring outside the hospital are captured through linkage with the national mortality registry; however, in these cases, the specific cause of death is not currently available.
12	Limitations	No database-specific limitations documented. General limitations for the data type applicable.
13	Main references	Mayer MA, Furlong LI,Torre P,Planas I,Cots F,Izquierdo E,Portabella J,Rovira J,Gutierrez-Sacristan A,Sanz F "Reuse of EHRs to Support Clinical Research in a Hospital of Reference." Studies in health technology and informatics (2015): 25991136
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/104316 Website: https://www.parcdesalutmar.cat/en/

Sweden: Health Impact - Swedish Population Evidence Enabling Data-linkage (HI-SPEED)

#	Section	Description
1	Data source identification and country	HI-SPEED (Health Impact - Swedish Population Evidence Enabling Data-linkage) Sweden
2	Data partner information section	SMPA-GU, Läkemedelsverket, Box 26, 751 03 Uppsala, Sweden - Box 469, 405 30 Gothenburg, Sweden Pharmacoepidemiology and Analysis Department (FeA) - School of Public Health and Community Medicine, Institute of Medicine
3	Coverage and timespan	Data collection since: 2015 Extent: Nation-wide. The catchment area includes the whole of Sweden, covering the full population of approximately 11.7 million.
4	Healthcare setting / type of data	Primary care – General Practitioner, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. Primary care (GPs) is available only for the 2 largest regions (~40% of national population) The following data elements are collected: Socio-demographics, dispensed drug prescriptions, cause of death, diagnoses and procedures from secondary (specialist) care and inpatient visits or clinical events, as well as from primary care visits (40%pop only).
5	Data collection process	Registries. The data is acquired from the relevant Swedish national and regional registries, only once all legislative, GDPR and ethical approvals have been granted. Therefore only relevant data is

#	Section	Description
		passed on, which will then be entered and processed by the study team. The data are updated several times annually.
6	General representativeness	The coverage includes all patients of all sociodemographic characteristics. Therefore it should mirror the source population to a very good extent.
7	Data content /source coding	Medicines are coded with ATC and NPLID (National Product ID), ICD10-SE is used for diagnoses and the Swedish procedure coding system (KVA) is used for clinical procedures.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients have a uniquely id across datasets.
9	Quality control (data source specific)	The source data are obtained from the relevant Swedish National and Regional Registers. The registers perform some regular quality controls on their data. After receiving the data, we perform additional checks and cleaning. We also run regular quality checks on the data we manage.
10	Linkage	Data on specialist care is acquired from the National Patient Register, mortality information is provided by the Cause-Of-Death Registry. Drug data is provided by the Patient Drug Register. And similarly for other registers. Data are linked very accurately using the national personal ID number and pseudonymised before delivery to HI-SPEED. All data are updated 2-4 times per year.
11	Vital status	Data on death and underlying + contributing causes-of-death are extracted from the Cause-of-Death registry (i.e. based on death certificates).
12	Limitations	General limitations for the data type applicable. This is a research project where all studies require ethics approval. Data collection since: 2015 for most data, except prescribed drug register (from 2018), and some COVID-related data (tests, vaccination) from 2020 Primary care is only available for a subset.
13	Main references	Nyberg F, Franzén S,Lindh M,Vanfleteren L,Hammar N,Wettermark B,Sundström J,Santosa A,Björck S,Gisslén M "Swedish Covid-19 Investigation for Future Insights - A Population Epidemiology Approach Using Register Linkage (SCIFI-PEARL)." Clinical epidemiology (2021): 34354377
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/node/4463/ Website: https://www.gu.se/en/research/scifi-pearl

The United Kingdom: UK BioBank (UKBB)

#	Section	Description
1	Data source identification and country	UKBB (UK BioBank) United Kingdom
2	Data partner information section	University of Oxford - Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences (Oxford NDORMS)
3	Coverage and timespan	Data collection since: 2006 Extent: Nation-wide. People recruited from whole of the UK.
4	Healthcare setting / type of data	Primary care – General Practitioner, and primary care specialists (e.g. paediatricians), and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and other (specify). UK Biobank is made by a rich variety of data sources, which include genetic data, primary care data, hospital inpatient data, death data, and cancer registry.

#	Section	Description
5	Data collection process	Inpatient hospital electronic health records, and Registries, and Biobank. The baseline assessment is consisting of both patient reported data (questionnaire) and physical measurements. GP and hospital data, as well as death and cancer registry records, are linked afterwards and are subject to data validation: https://biobank.ctsu.ox.ac.uk/~bbdatan/Data_cleaning_overall_doc_showcase_v1.pdf
6	General representativeness	The database population consists of volunteers aged 40-69years. We can expect that volunteers might have been more willing to participate if living closer to one of the 22 recruitment centres or if more interested in health issues compared to the general population. These aspects might have introduced an unavoidable bias in the cohort.
7	Data content /source coding	READ2, READ3, DM+D, ICD9, ICD10, OPCS3, OPCS4, ICD-O-3 are used.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5.4 and the OMOP standard vocabularies (SNOMED, RxNorm, LOINC). The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. No.
9	Quality control (data source specific)	All UK BioBank data are provided already curated and each of the many datasets have specific curation algorithms and procedures. As always, primary care and hospital data, which come from real-world setting, need special attention regarding data quality. Please refer to the link below for specific details https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/primary_care_data.pdf https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/HospitalEpisodeStatistics.pdf
10	Linkage	The database contains liked data from death and cancer registry, GP and hospital for the participants.
11	Vital status	Linked to the national death registry.
12	Limitations	The UKBB source data will not be updated anymore and have no new records after December 2022. There is no day's supply information captured in the source. GP prescription data are available for 45% of the cohort. There is no information on dispensed medicines. GP laboratory tests and results are not available.
13	Main references	Hewitt J, Walters M,Padmanabhan S,Dawson J "Cohort profile of the UK Biobank: diagnosis and characteristics of cerebrovascular disease." BMJ open (2016): 27006341
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111233 Website: https://www.ukbiobank.ac.uk/

ANNEX II. Fitness for use assessment

Theme/Issue	Croatia (NAJS)	Denmark (DK-DHR)	Estonia (EBB)	Finland (FinOMOP-HUS)	Finland (FinOMOP-TaUH Pirha)	Finland (FinOMOP-ACI Varha)	Finland (FinOMOP-THL)	France (CDW Bordeaux)	Germany (InGef RDB)
Care settings covered	Primary, secondary, inpatient	Pharmacy, secondary, inpatient	Broad incl. pharmacy	Secondary, inpatient	Secondary, inpatient	Secondary, inpatient	Primary, secondary, inpatient	Secondary, inpatient	Broad incl. pharmacy
Validated cancer registry linkage	Available	Available (registry only)	Partially available	Not available	Not available	Not available	Not available	Not available	Not available
Cancer stage & grade information	Partially available	Available (registry only)	Available	Available	Available	Available	Not available	Available	Not available
Recurrence/ progression data	Available	Available	Not available	Available	Not available	Not available	Not available	Available	Available
Performance status (ECOG etc.)	Not available	Not available	Available	Available	Available	Available	Available	Available	Partially available
Biomarker information	Not available	Not available (future update)	Available	Available	Available	Available	Not available	Available	Not available
Treatment detail completeness	Not available	Partially available	Available	Available	Available	Available	Not available	Available	Partially available
IRB timelines	Feasible (~1 month)	Feasible (umbrella approval)	Feasible (~1 month)	Feasible (~2 months)	Feasible (~2 months)	Feasible (~2 months)	Feasible (~2 months)	Feasible (~2 weeks)	Feasible (umbrella approval)

Theme/Issue	Greece (PGH)	Hungary (SUCD)	Italy (POLIMI)	Netherlands (NCR)	Norway (CRN)	Spain (H12O)	Spain (IMASIS)	Sweden (HI-SPEED)	UK (UKBB)
Care settings covered	Primary, secondary, inpatient	Secondary, inpatient	Secondary, inpatient	Inpatient only	Secondary, inpatient	Secondary, inpatient	Secondary, inpatient	Primary*, secondary, inpatient	Primary, secondary, inpatient
Validated cancer registry linkage	Not available	Not available	Not available	Available	Available	Not available	Not available	Not available	Partially available
Cancer stage & grade information	Not available	Available	Not available	Available	Available	Available	Available	Available	Not available
Recurrence/ progression data	Not available	Available	Not available	Available	Available	Not available	Available	Available	Not available
Performance status (ECOG etc.)	Not available	Available	Available	Available	Available	Available	Available	Not available	Not available
Biomarker information	Not available	Available	Not available	Available	Available	Available	Available	Not available	Not available
Treatment detail completeness	Partly available	Available	Not available	Partially available	Available	Available	Available	Not available	Not available
IRB timelines	Feasible (~2 months)	Feasible (~2 months)	Feasible (~1 month)	Feasible (~3 months)	Feasible (~1 month)	Feasible (~3 months)	Feasible (~3 months)	Feasible (~1 month)	Feasible (~3 months)

ANNEX III. Operational and reporting considerations

DATA MANAGEMENT

Data management

All data sources have previously mapped their data to the OMOP common data model. This enables the use of standardised analytics and using DARWIN EU® tools across the network, since the structure of the data and the terminology system is harmonised. The OMOP CDM was developed and maintained by the Observational Health Data Sciences and Informatics (OHDSI) initiative and is described in detail on the wiki page of the CDM: https://ohdsi.github.io/CommonDataModel and in The Book of OHDSI: http://book.ohdsi.org.

The analytic code for this study will be written in R and will use standardised analytics wherever possible. Each data partner will execute the study code against their data source containing patient-level data and then return the results (csv files), which will only contain aggregated data. The results from each of the contributing data sources will then be combined in tables and figures for the study report.

Data storage and protection

For this study, participants from various EU member states will process personal data from individuals that is collected in national/regional electronic health record data sources. Due to the sensitive nature of this personal medical data, it is important to be fully aware of ethical and regulatory aspects and to strive to take all reasonable measures to ensure compliance with ethical and regulatory issues on privacy.

All data sources used in this study are already used for pharmaco-epidemiological research and have a well-developed mechanism to ensure that European and local regulations dealing with ethical use of the data and adequate privacy control are adhered to. In agreement with these regulations, rather than combining person level data and performing only a central analysis, local analyses will be run, which generate non-identifiable aggregate summary results.

The output files are stored in the DARWIN EU® Digital Research Environment (DRE). These output files do not contain any data that allow identification of subjects included in the study. The DRE implements further security measures to ensure a high level of stored data protection to comply with the local implementation of the General Data Protection Regulation (GDPR) (EU) 679/20161 in the various member states.

QUALITY CONTROL

Data source quality control

When defining drug cohorts, non-systemic products will be excluded from the list of included codes summarised on the ingredient level.

When defining cohorts for indications, a systematic search of possible codes for inclusion will be identified using the *CodelistGenerator* R package (https://github.com/darwin-eu/CodelistGenerator). This package allows the user to define a search strategy and will use this to query the vocabulary tables of the OMOP common data model so as to find potentially relevant codes. In addition, the *CohortDiagnostics* (https://github.com/OHDSI/CohortDiagnostics) and *DrugExposureDiagnostics* (https://cran.r-project.org/web/packages/DrugExposureDiagnostics/index.html) R packages will be run, if needed, to assess the use of different codes across the data sources contributing to the study and identify any codes potentially omitted in error. The *DrugExposureDiagnostics* package evaluates ingredient-specific attributes and patterns in drug exposure records.

The study code will be based on DARWIN EU® R packages: *PatientProfiles* to identify person characteristics (https://darwin-eu.github.io/PatientProfiles), *DrugUtilisation* to characterise the drug use (https://doi.org/10.1002/pds.5809), *CohortCharacteristics* to characterise the cohort by indication (https://darwin-eu.github.io/CohortCharacteristics/), and *TreatmentPatterns* to facilitate the standardised

development and analysis of treatment patterns (https://doi.org/10.1016/j.cmpb.2022.107081. These packages will include numerous automated unit tests to ensure the validity of the codes, alongside software peer review and user testing. The R package will be made publicly available via GitHub.

PLANS FOR DISSEMINATING AND COMMUNICATING STUDY RESULTS

A PDF report, including an executive summary and the specified tables and/or figures, will be submitted to EMA by the DARWIN EU® CC upon completion of the study.

An interactive dashboard incorporating all the results (tables and figures) will be provided alongside the PDF report. The full set of underlying aggregated data used in the dashboard will also be made available, if requested.

ANNEX IV. List of stand-alone documents

Table S1. Preliminary list of concepts used to describe selected cancers (concept sets include all descendants of listed concepts).

Cancer type	Concept name	Concept ID	ICD-10 equivalent group codes	SNOMED equivalent code
Breast Cancer	Malignant tumour of breast	4112853	C50	254837009
Lung Cancer	Malignant neoplasm of lower respiratory tract	4334322	C33, 34	430621000
Prostate Cancer	Malignant neoplasm of prostate	4163261	C61	399068003
Colorectal Cancer	Malignant tumour of large intestine	443396	C18–21	363510005
Skin Melanoma	Malignant melanoma of skin	141232	C43	93655004
Bladder Cancer	Malignant neoplasm of urinary bladder	197508	C67	399326009
Oesophageal cancer	Malignant tumour of oesophagus	4181343	C15	363402007

ICD-10 = International Classification of Diseases, 10th Revision.

Table S2. List of concepts used to describe pharmacological treatments for cancer (concept sets include all descendants of listed concepts).

Concept name	Concept ID
abarelix	19010868
abemaciclib	792649
abiraterone	40239056
adagrasib	1301337
afatinib	43533090
aflibercept	40244266
aldesleukin	1309770
alectinib	35604657
alpelisib	1396423
aminoglutethimide	1503057
amivantamab	1536935
amrubicin	36854863
amrubicin hydrochloride	35198076
anastrozole	1348265
apalutamide	963987
atezolizumab	42629079
aumolertinib	36861628
avelumab	1593273
bcg, live, connaught strain	19023835

Concept name	Concept ID
bcg, live, montreal strain	19013730
bcg, live, tice strain	19086176
belotecan	36850058
bevacizumab	1397141
bicalutamide	1344381
binimetinib	1559930
brigatinib	1594187
buserelin	19033280
cabazitaxel	40222431
cabozantinib	43012292
camrelizumab	36853469
capecitabine	1337620
capivasertib	747200
capmatinib	1145637
carboplatin	1344905
carmustine	1350066
cemiplimab	35200783
ceritinib	44818466
cetuximab	1315411
chidamide	35884379
chlorambucil	1390051
cisplatin	1397599
cobimetinib	35605804
crizotinib	40242675
cyclophosphamide	1310317
cyproterone	19010792
dabrafenib	43532299
dacarbazine	1311409
dacomitinib	35200803
darolutamide	1361291
dasatinib	1358436
datopotamab	1254111
datopotamab deruxtecan	1465264
degarelix	19058410
dexamethasone	1518254
diethylstilbestrol	1525866
docetaxel	1315942
dostarlimab	1536789

Concept name	Concept ID
doxorubicin	1338512
durvalumab	1594034
elacestrant	1301646
encorafenib	1559914
enfortumab	37498261
ensartinib	1466160
entrectinib	37496447
enzalutamide	42900250
epirubicin	1344354
erdafitinib	1511250
eribulin	40230712
erlotinib	1325363
estradiol	1548195
estramustine	1349025
ethinyl estradiol	1549786
etoposide	1350504
everolimus	19011440
exemestane	1398399
fluorouracil	955632
flutamide	1356461
formestane	19020079
freeze-dried bcg(japanese strain)	35834910
fruquintinib	747237
fulvestrant	1304044
furmonertinib	36855491
gefitinib	1319193
gemcitabine	1314924
gimeracil	36878778
goserelin	1366310
histrelin	1366773
hyaluronidase	19073699
hydrocortisone	975125
hydroxyurea	1377141
icotinib	35884401
ifosfamide	19078187
imatinib	1304107
imlunestrant	36855119
inavolisib	1735491

Concept name	Concept ID
interferon alfa-2b	1380068
ipilimumab	40238188
irinotecan	1367268
ixabepilone	19025348
ketoconazole	985708
lapatinib	1359548
larotrectinib	1355705
lazertinib	1735080
lenvatinib	46221433
letrozole	1315946
leucovorin	1388796
leuprolide	1351541
levamisole	1389464
levoleucovorin	40168303
lifileucel	42609269
lomustine	1391846
lorlatinib	35201733
margetuximab	739518
megestrol	1300978
melphalan	1301267
methotrexate	1305058
mitomycin	1389036
mitoxantrone	1309188
mobocertinib	701915
nadofaragene firadenovec	746328
necitumumab	35606215
nedaplatin	35198201
neratinib	793846
nilotinib	1394023
nilutamide	1315286
nintedanib	45775396
niraparib	1593861
nivolumab	45892628
nogapendekin alfa	36853973
nogapendekin alfa inbakicept	1734398
olaparib	45892579
osimertinib	35605522
oteracil	36878777

Concept name	Concept ID
oxaliplatin	1318011
paclitaxel	1378382
palbociclib	45892075
panitumumab	19100985
peginterferon alfa-2a	1714165
pembrolizumab	45775965
pemetrexed	1304919
penpulimab	1466158
pertuzumab	42801287
pralsetinib	37002765
prednisone	1551099
procarbazine	1351779
pyrotinib	36856361
radium-223	902727
raltitrexed	19038536
ramucirumab	44818489
regorafenib	42903460
relatlimab	779239
relugolix	739471
repotrectinib	747362
retifanlimab	1302024
ribociclib	1592911
ripretinib	1145752
rucaparib	1718850
sacituzumab	1145484
sargramostim	1308432
selpercatinib	1145791
semustine	36853363
serplulimab	36861273
sintilimab	36854475
sipuleucel-t	40224095
sotorasib	1536963
sugemalimab	36854773
sunvozertinib	1253903
talazoparib	35201068
taletrectinib	36862664
tamoxifen	1436678
tarlatamab	1734429

Concept name	Concept ID
tegafur	19056756
telisotuzumab	1466283
temozolomide	1341149
teniposide	19136750
tepotinib	739739
thiotepa	19137385
tipiracil	35602757
tislelizumab	42609339
topotecan	1378509
toremifene	1342346
toripalimab	747052
trametinib	43532497
trastuzumab	1387104
tremelimumab	741851
trifluridine	905078
triptorelin	1343039
tucatinib	1145571
valrubicin	19012543
vandetanib	40238052
vemurafenib	40241937
vinblastine	19008264
vincristine	1308290
vindesine	19008336
vinflunine	36878852
vinorelbine	1343346
zenocutuzumab	1464913
zolbetuximab	1735539
zongertinib	1253620

Table S3. Preliminary list of concepts to identify surgical procedures and radiotherapy.

Concept name	Concept ID (including descendants)	Vocabulary
Surgical procedure	4301351	SNOMED
Brachytherapy	40317890	SNOMED
Radiotherapy	1242725	SNOMED

SNOMED = Systemised Nomenclature of Medicine.

Table S4. Preliminary list of concepts to identify metastatic disease and carcinomatosis.

Name	Vocabulary	Code	ID
Carcinomatosis of liver	ICDO3	8010/9-C22.0	44503831
Carcinomatosis of lung, NOS	ICDO3	8010/9-C34.9	44505002
Carcinomatosis of peritoneum, NOS	ICDO3	8010/9-C48.2	44505266
Carcinomatosis of skin, NOS	ICDO3	8010/9-C44.9	44505049
Carcinomatosis of specified parts of peritoneum	ICDO3	8010/9-C48.1	44505257
Metastatic acantholytic squamous cell carcinoma	SNOMED	1254749009	37165395
Metastatic adenocarcinoma	SNOMED	354081000119101	604497
Metastatic adenocarcinoma of unknown origin	SNOMED	307226002	4145870
Metastatic adenosquamous carcinoma	SNOMED	1264116009	37167301
Metastatic amelanotic malignant melanoma to skin	SNOMED	1332486000	1075919
Metastatic anaplastic thyroid carcinoma	SNOMED	1332484002	1075917
Metastatic apocrine adenocarcinoma	SNOMED	1290695006	1244718
Metastatic balloon cell malignant melanoma	SNOMED	1332483008	1075916
Metastatic cholangiocarcinoma	SNOMED	1290690001	1244716
Metastatic clear cell renal cell carcinoma	SNOMED	1187285005	37162300
Metastatic collecting duct renal cell carcinoma	SNOMED	1286896000	37168858
Metastatic cystic renal cell carcinoma	SNOMED	1286898004	37168860
Metastatic desmoplastic malignant melanoma	SNOMED	1335987009	1076162
Metastatic eccrine porocarcinoma	SNOMED	1290697003	1244720
Metastatic embryonal rhabdomyosarcoma	SNOMED	1279882001	37168383
Metastatic endometrial stromal sarcoma	SNOMED	1332485001	1075918
Metastatic extraskeletal osteosarcoma	SNOMED	1251465007	37165253
Metastatic fibrosarcoma	SNOMED	1336176005	1076220
Metastatic ganglioneuroblastoma	SNOMED	1336172007	1076218
Metastatic germinoma	SNOMED	1339051005	1076302
Metastatic glassy cell carcinoma	SNOMED	1336188007	1076225

Name	Vocabulary	Code	ID
Metastatic hepatocellular carcinoma	SNOMED	1186612004	37162050
Metastatic infantile fibrosarcoma	SNOMED	1339049006	1076300
Metastatic inflammatory carcinoma	SNOMED	1336190008	1076227
Metastatic lentigo maligna melanoma	SNOMED	1336186006	1076224
Metastatic lobular carcinoma	SNOMED	1290700004	1244722
Metastatic malignant glioma	SNOMED	1339050006	1076301
Metastatic malignant melanoma	SNOMED	443493003	40481908
Metastatic malignant melanoma of unknown primary	SNOMED	1296851009	1245145
Metastatic malignant neoplasm	SNOMED	128462008	432851
Metastatic malignant neoplasm to areola	SNOMED	1306323006	1075532
Metastatic malignant neoplasm to ectopic site of breast	SNOMED	1306322001	1075531
Metastatic malignant neoplasm to nipple	SNOMED	1306321008	1075530
Metastatic malignant neoplasm to respiratory and digestive systems	SNOMED	269473008	4147162
Metastatic myxofibrosarcoma	SNOMED	1336182008	1076222
Metastatic neuroblastoma	SNOMED	704152002	45772700
Metastatic neuroblastoma to paraspinal area	SNOMED	1251459007	37165252
Metastatic non-small cell lung cancer	SNOMED	457721000124104	36684857
Metastatic nuclear protein in testis associated carcinoma to bone	SNOMED	932947661000119102	1450873
Metastatic nuclear protein in testis associated carcinoma to liver	SNOMED	417437921000119101	1450482
Metastatic oxyphilic adenocarcinoma	SNOMED	1264118005	37167302
Metastatic papillary squamous cell carcinoma	SNOMED	1264258003	37167409
Metastatic pilomatrix carcinoma	SNOMED	1290698008	1244721
Metastatic pleomorphic rhabdomyosarcoma	SNOMED	1336185005	1076223
Metastatic renal cell carcinoma	SNOMED	702392008	45773365
Metastatic rhabdomyosarcoma	SNOMED	1269121005	37167895
Metastatic sarcoma	SNOMED	443144000	40480137
Metastatic sebaceous adenocarcinoma	SNOMED	1264367001	37167429

Name	Vocabulary	Code	ID
Metastatic small cell neuroendocrine carcinoma	SNOMED	1240362006	37165127
Metastatic spindle cell melanoma	SNOMED	1336177001	1076221
Metastatic spindle cell rhabdomyosarcoma	SNOMED	1336170004	1076216
Metastatic squamous cell carcinoma	SNOMED	403906006	4301510
Metastatic superficial spreading melanoma	SNOMED	1336173002	1076219
Metastatic synovial sarcoma	SNOMED	1336171000	1076217
Metastatic well-differentiated neuroendocrine neoplasm to bone	SNOMED	16067571000119106	1450198
Metastatic well-differentiated neuroendocrine neoplasm to peritoneum	SNOMED	16067651000119101	1450199
Metastatic well-differentiated neuroendocrine tumour	SNOMED	1290060007	1244565
Oligometastatic malignant neoplasm	SNOMED	1255358009	37165664

SNOMED = Systemised Nomenclature of Medicine.

ANNEX V. ENCePP checklist for study protocols

ENCePP Checklist for Study Protocols (Revision 4)

Study title: DARWIN EU® - Enhanced Preparedness of the DARWIN EU® Network for regulatory questions in oncology – DARWIN EU® OncoNet

EU PAS Register® number: EUPAS1000000824
Study reference number: P4-C4-002

Section 1: Milestones		**Yes**	**No**	**N/A**	**Section Number**
1.1	Does the protocol specify timelines for				
	1.1.1 Start of data collection[1]	☒	☐	☐	8.5
	1.1.2 End of data collection[2]	☒	☐	☐	8.5
	1.1.3 Progress report(s)	☐	☐	☒	
	1.1.4 Interim report(s)	☐	☐	☒	
	1.1.5 Registration in the EU PAS Register®	☐	☒	☐	
	1.1.6 Final report of study results.	☒	☐	☐	5

Comments:

Section 2: Research question		**Yes**	**No**	**N/A**	**Section Number**
2.1	Does the formulation of the research question and objectives clearly explain:	☐	☐	☐	
	2.1.1 Why the study is conducted? (e.g. to address an important public health concern, a risk identified in the risk management plan, an emerging safety issue)	☒	☐	☐	6
	2.1.2 The objective(s) of the study?	☒	☐	☐	7
	2.1.3 The target population? (i.e. population or subgroup to whom the study results are intended to be generalised)	☒	☐	☐	8.3
	2.1.4 Which hypothesis(-es) is (are) to be tested?	☐	☐	☒	
	2.1.5 If applicable, that there is no *a priori* hypothesis?	☐	☐	☒	

Comments:

[1] Date from which information on the first study is first recorded in the study dataset or, in the case of secondary use of data, the date from which data extraction starts.
[2] Date from which the analytical dataset is completely available.

Section 3: Study design		**Yes**	**No**	**N/A**	**Section Number**
3.1	Is the study design described? (e.g. cohort, case-control, cross-sectional, other design)	☒	☐	☐	8.1
3.2	Does the protocol specify whether the study is based on primary, secondary or combined data collection?	☒	☐	☐	8.4
3.3	Does the protocol specify measures of occurrence? (e.g., rate, risk, prevalence)	☒	☐	☐	8.8.3
3.4	Does the protocol specify measure(s) of association? (e.g. risk, odds ratio, excess risk, rate ratio, hazard ratio, risk/rate difference, number needed to harm (NNH))	☐	☐	☒	
3.5	Does the protocol describe the approach for the collection and reporting of adverse events/adverse reactions? (e.g. adverse events that will not be collected in case of primary data collection)	☐	☐	☒	

Comments:

Section 4: Source and study populations		**Yes**	**No**	**N/A**	**Section Number**
4.1	Is the source population described?	☒	☐	☐	8.3
4.2	Is the planned study population defined in terms of:				
	4.2.1 Study time period	☒	☐	☐	8.5
	4.2.2 Age and sex	☒	☐	☐	8.3
	4.2.3 Country of origin	☒	☐	☐	8.4
	4.2.4 Disease/indication	☒	☐	☐	8.3
	4.2.5 Duration of follow-up	☒	☐	☐	8.2
4.3	Does the protocol define how the study population will be sampled from the source population? (e.g. event or inclusion/exclusion criteria)	☒	☐	☐	8.3

Comments:

Section 5: Exposure definition and measurement		**Yes**	**No**	**N/A**	**Section Number**
5.1	Does the protocol describe how the study exposure is defined and measured? (e.g. operational details for defining and categorising exposure, measurement of dose and duration of drug exposure)	☐	☐	☒	
5.2	Does the protocol address the validity of the exposure measurement? (e.g. precision, accuracy, use of validation sub-study)	☐	☒	☐	
5.3	Is exposure categorised according to time windows?	☒	☐	☐	8.8.3
5.4	Is intensity of exposure addressed? (e.g. dose, duration)	☒	☐	☐	8.8.3

Section 5: Exposure definition and measurement		**Yes**	**No**	**N/A**	**Section Number**
5.5	Is exposure categorised based on biological mechanism of action and taking into account the pharmacokinetics and pharmacodynamics of the drug?	☐	☒	☐	
5.6	Is (are) (an) appropriate comparator(s) identified?	☐	☐	☒	

Comments:

Section 6: Outcome definition and measurement		**Yes**	**No**	**N/A**	**Section Number**
6.1	Does the protocol specify the primary and secondary (if applicable) outcome(s) to be investigated?	☐	☐	☒	
6.2	Does the protocol describe how the outcomes are defined and measured?	☒	☐	☐	8.6.1
6.3	Does the protocol address the validity of outcome measurement? (e.g. precision, accuracy, sensitivity, specificity, positive predictive value, use of validation sub-study)	☐	☒	☐	
6.4	Does the protocol describe specific outcomes relevant for Health Technology Assessment? (e.g. HRQoL, QALYs, DALYS, health care services utilisation, burden of disease or treatment, compliance, disease management)	☐	☐	☒	

Comments:

Section 7: Bias		**Yes**	**No**	**N/A**	**Section Number**
7.1	Does the protocol address ways to measure confounding? (e.g. confounding by indication)	☐	☐	☒	
7.2	Does the protocol address selection bias? (e.g. healthy user/adherer bias)	☒	☐	☐	9
7.3	Does the protocol address information bias? (e.g. misclassification of exposure and outcomes, time-related bias)	☒	☐	☐	9

Comments:

Section 8: Effect measure modification		**Yes**	**No**	**N/A**	**Section Number**
8.1	Does the protocol address effect modifiers? (e.g. collection of data on known effect modifiers, sub-group analyses, anticipated direction of effect)	☐	☐	☒	

Comments:

Section 9: Data sources		**Yes**	**No**	**N/A**	**Section Number**
9.1	Does the protocol describe the data source(s) used in the study for the ascertainment of:				
	9.1.1 Exposure? (e.g. pharmacy dispensing, general practice prescribing, claims data, self-report, face-to-face interview)	☒	☐	☐	Annex I
	9.1.2 Outcomes? (e.g. clinical records, laboratory markers or values, claims data, self-report, patient interview including scales and questionnaires, vital statistics)	☒	☐	☐	Annex I
	9.1.3 Covariates and other characteristics?	☒	☐	☐	Annex I
9.2	Does the protocol describe the information available from the data source(s) on:				
	9.2.1 Exposure? (e.g. date of dispensing, drug quantity, dose, number of days of supply prescription, daily dosage, prescriber)	☒	☐	☐	Annex I
	9.2.2 Outcomes? (e.g. date of occurrence, multiple event, severity measures related to event)	☒	☐	☐	Anex I
	9.2.3 Covariates and other characteristics? (e.g. age, sex, clinical and drug use history, co-morbidity, co-medications, lifestyle)	☒	☐	☐	Annex I
9.3	Is a coding system described for:				
	9.3.1 Exposure? (e.g. WHO Drug Dictionary, Anatomical Therapeutic Chemical (ATC) Classification System)	☒	☐	☐	Annex I
	9.3.2 Outcomes? (e.g. International Classification of Diseases (ICD), Medical Dictionary for Regulatory Activities (MedDRA))	☒	☐	☐	Annex I
	9.3.3 Covariates and other characteristics?	☒	☐	☐	Annex I
9.4	Is a linkage method between data sources described? (e.g. based on a unique identifier or other)	☒	☐	☐	Annex I

Comments:

Section 10: Analysis plan		**Yes**	**No**	**N/A**	**Section Number**
10.1	Are the statistical methods and the reason for their choice described?	☒	☐	☐	8.8
10.2	Is study size and/or statistical precision estimated?	☒	☐	☐	8.7
10.3	Are descriptive analyses included?	☒	☐	☐	8.8.3
10.4	Are stratified analyses included?	☒	☐	☐	8.8
10.5	Does the plan describe methods for analytic control of confounding?	☐	☐	☒	
10.6	Does the plan describe methods for analytic control of outcome misclassification?	☐	☒	☐	
10.7	Does the plan describe methods for handling missing data?	☒	☐	☐	8.8.3

Section 10: Analysis plan	**Yes**	**No**	**N/A**	**Section Number**
10.8 Are relevant sensitivity analyses described?	☒	☐	☐	8.8.3

Comments:

Section 11: Data management and quality control	**Yes**	**No**	**N/A**	**Section Number**
11.1 Does the protocol provide information on data storage? (e.g. software and IT environment, database maintenance and anti-fraud protection, archiving)	☒	☐	☐	Annex III
11.2 Are methods of quality assurance described?	☒	☐	☐	Annex III
11.3 Is there a system in place for independent review of study results?	☐	☐	☒	

Comments:

Section 12: Limitations	**Yes**	**No**	**N/A**	**Section Number**
12.1 Does the protocol discuss the impact on the study results of:				
12.1.1 Selection bias?	☒	☐	☐	9
12.1.2 Information bias?	☒	☐	☐	9
12.1.3 Residual/unmeasured confounding? (e.g. anticipated direction and magnitude of such biases, validation sub-study, use of validation and external data, analytical methods).	☐	☐	☒	
12.2 Does the protocol discuss study feasibility? (e.g. study size, anticipated exposure uptake, duration of follow-up in a cohort study, patient recruitment, precision of the estimates)	☒	☐	☐	8.7

Comments:

Section 13: Ethical/data protection issues	**Yes**	**No**	**N/A**	**Section Number**
13.1 Have requirements of Ethics Committee/ Institutional Review Board been described?	☒	☐	☐	Annex I
13.2 Has any outcome of an ethical review procedure been addressed?	☐	☐	☒	
13.3 Have data protection requirements been described?	☒	☐	☐	Annex III

Comments:

Section 14: Amendments and deviations	**Yes**	**No**	**N/A**	**Section Number**
14.1 Does the protocol include a section to document amendments and deviations?	☒	☐	☐	4

Comments:

Section 15: Plans for communication of study results	**Yes**	**No**	**N/A**	**Section Number**
15.1 Are plans described for communicating study results (e.g. to regulatory authorities)?	☒	☐	☐	Annex III
15.2 Are plans described for disseminating study results externally, including publication?	☐	☒	☐	

Comments:

Name of the main author of the protocol: Talita Duarte-Salles

Date: 15/04/2026

Signature: TDS

ANNEX VI. Glossary

Additional definitions are available in the EMA Glossary of terms https://www.ema.europa.eu/en/about-us/glossaries.

Aggregated Data

Data collected and combined from multiple sources to generate summary information, typically anonymised.

Benefit-Risk Assessment

Evaluation of the positive therapeutic effects of a medicine compared to its risks (e.g. side effects).

Common Data Model (CDM)

A standardised data structure that enables data from multiple sources to be harmonised, making analysis consistent and reproducible. DARWIN EU® utilises the OMOP CDM maintained by the OHDSI community.

Complex Studies (C3)

Studies requiring the development or customisation of specific study designs, protocols, and Statistical Analysis Plans (SAPs), with extensive collection or extraction of data. Examples include etiological studies measuring the strength and determinants of an association between an exposure and the occurrence of a health outcome in a defined population considering sources of bias, potential confounding factors, and effect modifiers.

Coordination Centre (CC)

The central hub responsible for managing and overseeing the activities within DARWIN EU®. It is based at Erasmus University Medical Center in Rotterdam, the Netherlands.

Data Access

The process of obtaining permission to use specific datasets for regulatory or scientific studies.

Data Quality Framework

A set of standards and procedures to ensure accuracy, completeness, timeliness, and consistency of data used in DARWIN EU®.

Data Source

A data source or repository of structured health-related data, such as electronic health records (EHRs), insurance claims, or registries.

DARWIN EU®

The European Medicines Agency's (EMA) federated network of real-world data sources designed to generate evidence to support regulatory decision-making.

EMA (European Medicines Agency)

The regulatory body responsible for the evaluation and supervision of medicinal products in the EU, overseeing DARWIN EU®.

Evidence Generation

The process of analysing real-world data to produce scientific information that can inform healthcare or regulatory decisions.

Federated Network

A data infrastructure where data remain at their original location but can be analysed in a harmonised way across multiple partners using a common model and tools.

GDPR (General Data Protection Regulation)

The EU regulation governing the protection of personal data and privacy, crucial to how DARWIN EU® handles health data.

Health Technology Assessment (HTA)

A systematic evaluation of properties and impacts of health technology, often using DARWIN EU® data to support assessments.

Metadata

Descriptive information about a data source (e.g. its content, quality, and structure), essential for identifying relevant data sources in DARWIN EU® studies.

Off-the-Shelf Studies (OTS)

Studies for which a standard protocol per study/analysis type and standardised analytics may be developed and applied or adapted, typically relating to a descriptive research question. This includes studies on disease epidemiology, for example, the estimation of the prevalence or incidence of health outcomes in defined time periods and population groups, or drug utilisation studies at the population or patient level.

OHDSI (Observational Health Data Sciences and Informatics)

An open-science collaborative community that develops tools and standards (including the OMOP CDM) to enable large-scale analytics of observational health data. OHDSI provides the technical and scientific foundation for DARWIN EU®'s analytical ecosystem.

Patient-Level Data

Data related to individuals, de-identified, used for longitudinal or detailed analyses.

OMOP (Observational Medical Outcomes Partnership)

A common data model (CDM) that standardises the structure and content of observational healthcare data, enabling systematic analysis across disparate datasets. DARWIN EU® uses the OMOP CDM to ensure interoperability and consistency in real-world evidence generation.

Real-World Data (RWD)

Data relating to individual health status or healthcare delivery that is collected from routine clinical practice rather than from randomised controlled trials.

Real-World Evidence (RWE)

Clinical evidence derived from the analysis of RWD, used to inform decisions by regulators, payers, or clinicians.

Regulatory Decision-Making

The process by which authorities like EMA assess data to authorise, monitor, or modify the use of medicines in the EU.

Routine Repeated Studies (RR)

Studies that are either Off-the-Shelf or Complex studies repeated on a regular basis, following the same protocol and study code, but with updated data and/or different data partners.

Study Protocol

A detailed plan describing how a specific real-world study will be conducted, including objectives, design, data sources, and analyses.

Very Complex Studies (C4)

Studies which cannot rely only on electronic health care data sources, or which would require complex methodological work, for example, due to the occurrence of events that cannot be defined by existing diagnosis codes, including events that do not yet have a diagnosis code, where it may be necessary to combine a diagnosis code with other data such as results of laboratory investigations. These studies might require the collection of data prospectively, or the inclusion of new (not previously onboarded) data sources.