Study Protocol P3-C3-005

DARWIN EU® – Incidence rates of venous thromboembolic events in cancer patients

30/10/2024

Version 1.0

Table of Contents

Study Title	DARWIN EU® – Incidence rates of venous thromboembolic events in cancer patients
Protocol version	1.0
Date	30/10/2024
EU PAS number	Study not yet registered
Active substance	None
Medicinal product	None
Research question and objectives	This study aims to estimate incidence rates of venous thromboembolic events (deep vein thrombosis (DVT), pulmonary embolisms (PE), venous thromboembolism (VTE, composite of DVT and PE), pelvic venous thrombosis (PVT), splanchic vein thrombosis (SVT), including hepatic and extra-hepatic vein thrombosis, retinal vein thrombosis (RVT), including retinal central vein thrombosis, and disseminated intravascular coagulation (DIC) in adult patients (aged 18 and above) newly diagnosed with selected cancers in 2016-2023 (lung, breast, ovary, corpus uteri, prostate, pancreas, colorectal, stomach, oesophageal, liver, brain, bone, kidney, melanoma, lymphoma and leukaemia) and to describe their characteristics at the time of cancer diagnosis. The specific objectives of the study are: 1. To estimate the incidence rates of thromboembolic events in patients newly diagnosed with each type of selected cancers stratified by country/database, age group, sex, study sub-period, and cancer stage (when available) one and two years after cancer diagnosis. 2. To characterise cancer patients at the time of cancer diagnosis in terms of demographics, comorbidities, concomitant medications, as well as treatments received in the first 90 days after cancer diagnosis.
Country(ies) of study	Belgium, Denmark, Estonia, Finland, Germany, The Netherlands, Spain, United Kingdom
Author	Anton Barchuk (a.barchuk@darwin-eu.org) Talita Duarte-Salles (t.duarte@darwin-eu.org)

LIST OF ABBREVIATIONS

Abbreviation	Name
ATC	Anatomical Therapeutic Chemical
CDM	Common Data Model
CPRD	Clinical Practice Research Datalink
DARWIN EU®	Data Analysis and Real World Interrogation Network
DK-DHR	Danish Data Health Registries
DOI	Declaration Of Interests
DQD	Data Quality Dashboard
DRE	Digital Research Environment
DVT	Deep Venous Thrombosis
DIC	Disseminated Intravascular Coagulation
GP	General Practitioner
EHR	Electronic Health Record
EMA	European Medicines Agency
EBB	Estonian Biobank
ECOG	Eastern Cooperative Oncology Group
EGCUT	Estonian Genome Center at the University of Tartu
ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
ICD-O-3	International Classification of Diseases for Oncology, 3rd Edition
ICD-10	International Classification of Diseases, 10th revision
ICPC-1	International Classification of Primary Care
IPCI	Integrated Primary Care Information Project
LPD	Longitudinal Patient Database
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
PE	Pulmonary Embolism
PVT	Pelvic Venous Thrombosis
RVT	Retinal vein thrombosis
SNOMED	Systematized Nomenclature of Medicine
SVT	Splanchic Vein Thrombosis
UKBB	UK Biobank
VTE	Venous Thromboembolism

1. TITLE

DARWIN EU® – Incidence rates of venous thromboembolic events in cancer patients

2. RESPONSIBLE PARTIES – STUDY TEAM

Study team role	Names	Organisation
Study Project Manager / Principal Investigator	Talita Duarte-Salles Anton Barchuk	Erasmus MC
Data Scientist	Cesar Barboza Ger Inberg Maarten van Kessel Adam Black Ross Williams	Erasmus MC
Epidemiologist	Berta Raventós Julieta Politi	Erasmus MC
Clinical Domain Expert	Anton Barchuk	Erasmus MC

Data Partner*	Names	Organisation
CPRD GOLD and UKBB	Antonella Delmestri	University of Oxford
DK-DHR	Claus Møldrup Elvira Bräuner Susanne Bruun Tine Iskov Kopp Cæcilie Brinth Christiansen	Danish Medicines Agency
EBB	Marek Oja Raivo Kolde	Estonian Biobank, Estonia
FinOMOP-HILMO	Anna Hammais Gustav Klingstedt	Finnish Care Register for Health Care, Finland
FinOMOP-HUS	Eric Fey Kimmo Porkka Tiina Wahlfors	Hospital District of Helsinki and Uusimaa, Finland
IQVIA DA Germany and IQVIA LPD Belgium	Gargi Jadhav Isabella Kacmarczyl Akram Mendez Hanne van Ballegooijen Dina Vojinovic	IQVIA
IPCI	Katia Verhamme	Integrated Primary Care Information, Netherlands
SIDIAP	Anna Palomar-Cros Irene López-Sánchez Agustina Giuliodori	IDIAPJGol

*Data partners' role is only to execute code at their data source, review and approve their results. These people do not have an investigator role.

3. ABSTRACT

Title

DARWIN EU® – Incidence rates of venous thromboembolic events in cancer patients

Rationale

Cancer-associated venous thrombosis is relatively common: from 20% to 30% of all primary venous thromboembolic events are cancer-associated. While cancer patients have an increased risk of developing venous thromboembolism (VTE) compared to individuals without underlying malignancies, it is also recognised as one of the major causes of death in cancer patients. Still, the reported incidence varies across different populations and cancer types and can also be attributed to variations in patient characteristics, management options, and the cancer stage at diagnosis. The incidence of VTE was found to be higher in cases of renal cell, ovarian, pancreatic, stomach, and lung cancers, as well as acute myelogenous leukaemia and non-Hodgkin lymphoma during the four months immediately preceding the cancer diagnosis. When investigating a safety signal, reliable information on background risk is crucial to assess potential associations with oncological treatments.

Research Objectives

This study aims to estimate incidence rates of venous thromboembolic events (deep vein thrombosis (DVT), pulmonary embolisms (PE), venous thromboembolism (VTE, composite of DVT and PE), pelvic venous thrombosis (PVT), splanchic vein thrombosis (SVT), including hepatic and extra-hepatic vein thrombosis, retinal vein thrombosis (RVT), including retinal central vein thrombosis, and disseminated intravascular coagulation (DIC) in adult patients (aged 18 and above) newly diagnosed with selected cancers in 2016-2023 (lung, breast, ovary, corpus uteri, prostate, pancreas, colorectal, stomach, oesophageal, liver, brain, bone, kidney, melanoma, lymphoma and leukaemia) and to describe their characteristics at the time of cancer diagnosis.

The specific objectives of the study are:

1. To estimate the incidence rates of thromboembolic events in patients newly diagnosed with each type of selected cancers stratified by country/database, age group, sex, study sub-period, and cancer stage (when available) one and two years after cancer diagnosis.
2. To characterise cancer patients at the time of cancer diagnosis in terms of demographics, comorbidities, concomitant medications, as well as treatments received in the first 90 days after cancer diagnosis.

Research Methods

Study design

Population-based cohort study.

Population

The study population will include all individuals aged 18 years and above with a primary diagnosis of one of the selected cancers (lung, breast, ovary, corpus uteri, prostate, pancreas, colorectal, stomach, oesophageal, liver, brain, bone, kidney, melanoma, lymphoma and leukaemia) in the study period from 01/01/2016 to 31/12/2023. Only patients with the first and one cancer diagnosis (except non-melanoma skin cancer) will be included. Cancer cases and thromboembolic events will be identified based on appropriate computable phenotyping algorithms. Conditions in the OMOP CDM use the Systematised

Nomenclature of Medicine (SNOMED) as the standard vocabulary for diagnosis codes. The International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) will also be considered for cancer diagnoses. Additional eligibility of a minimum of 1 year of potential follow-up time will be imposed to ensure sufficient time to capture potential outcomes of interest.

Variables

Outcomes will include thromboembolic events, in particular, DVT, PE, VTE (composite of DVT and PE), PVT, SVT, RVT, DIC.

Data sources

1. Clinical Practice Research Datalink GOLD (CPRD GOLD), United Kingdom
2. Danish Data Health Registries (DK-DHR), Denmark
3. Estonian Biobank (EBB), Estonia
4. Finnish Care Register for Health Care (FinOMOP-HILMO) Finland
5. Hospital District of Helsinki and Uusimaa (FinOMOP-HUS), Finland
6. IQVIA Disease Analyzer Germany (IQVIA DA Germany), Germany
7. IQVIA Longitudinal Patient Database Belgium (IQVIA LPD Belgium), Belgium
8. Integrated Primary Care Information (IPCI), Netherlands
9. The Information System for Research in Primary Care (SIDIAP), Spain.
10. UK Biobank (UKBB), United Kingdom

Sample size

No sample size will be calculated as this is a descriptive study. However, analysis by strata will be limited to databases with enough cases (5 or more) to provide meaningful results. Summary measures of occurrence will not be calculated for strata with counts less than 5.

Data analyses

Analyses will be conducted separately for each database and carried out in a federated manner, allowing analyses to be run locally without sharing patient-level data.

The incidence of thromboembolic events (Objective 1) will be estimated over one and two years after the selected cancer diagnosis. Each cancer type and outcome will be assessed separately. Large-scale patient-level characterisation (Objective 2) will be conducted at the index date. Age and sex will be described at the time of diagnosis. The medical history and medication will be assessed at the index date.

For all analyses, absolute and relative frequencies will be reported. A minimum cell count of 5 will be used when reporting results, with any smaller counts reported as "<5" and zero counts as "0". Overall analyses will be done separately for each database. Further stratification by age category (18-29; 30-39; 40-49; 50-59; 60-69; 70-79; 80-89; 90 and over), sex, study sub-period, cancer stage will be conducted when possible (minimum cell count reached and data available). The following sub-periods will be also used: 2016-2019, 2020-2023.

4. AMENDMENTS AND UPDATES

None.

5. MILESTONES

Study deliverables	Timeline
Draft Study Protocol	October 2024
Final Study Protocol	October/November 2024 (or upon EMA acceptance)
Creation of Analytical code and Phenotyping	November/December 2024
Execution of Analytical Code on the data	January 2024 (depending on IRB approval dates)
Draft Study Report	February 2024
Final Study Report	February 2024

6. RATIONALE AND BACKGROUND

Cancer-associated venous thrombosis is relatively common: from 20% to 30% of all primary venous thromboembolic events are cancer-associated (Timp, 2013). While cancer patients have an increased risk of developing venous thromboembolism (VTE) compared to individuals without underlying malignancies (Blom, 2005), it is also recognised as one of the major causes of death in cancer patients (Wang, 2017). Still, the reported incidence varies across different populations and cancer types and can also be attributed to variations in patient characteristics, management options, and the cancer stage at diagnosis (Timp, 2013).

The association between cancer and thromboembolic events was assessed in various studies. In a large cohort of cancer patients, the incidence of thromboembolic events was found to be higher in cases of renal cell, ovarian, pancreatic, stomach, and lung cancers, as well as acute myelogenous leukaemia and non-Hodgkin lymphoma during the four months immediately preceding the cancer diagnosis (White, 2005). The prevalence of thromboembolic events at the time of diagnosis was highest for pancreatic cancer and lowest for breast cancer in another registry-based study (Ohashi, 2020). All cancer sites showed an increased prevalence in the thromboembolic events incidence cohort (Petterson, 2015). The incidence of thromboembolic events was notably high during the first few months of chemotherapy in a cohort of cancer patients followed for up to 12 months, with higher odds observed in pancreatic, gastric, and lung cancers (Khorana, 2013). Certain cancer medications have also been associated with higher risks of thromboembolic events (Nalluri, 2008; Khorana, 2013). Additionally, major surgery is known to be associated with thromboembolic events, with an increased risk that persists for 90 to 120 days post-surgery (Björklund, 2024). A high rate of recurrent thromboembolic events is observed over time following the discontinuation of anticoagulant therapy in patients with cancer-associated thrombosis (van Hylckama Vlieg, 2023).

When a safety signal of this nature appears in cancer populations, it can be challenging to assess a potential association with new treatment options without reliable information on background risk. This study

addresses this knowledge gap by generating background incidence rates of thromboembolic events events among patients with selected cancer types.

7. RESEARCH QUESTION AND OBJECTIVES

This study aims to estimate incidence rates of venous thromboembolic events (deep vein thrombosis (DVT), pulmonary embolisms (PE), venous thromboembolism (VTE, composite of DVT and PE), pelvic venous thrombosis (PVT), splanchic vein thrombosis (SVT), including hepatic and extra-hepatic vein thrombosis, retinal vein thrombosis (RVT), including retinal central vein thrombosis, and disseminated intravascular coagulation (DIC) in adult patients (aged 18 and above) newly diagnosed with selected cancers in 2016-2023 (lung, breast, ovary, corpus uteri, prostate, pancreas, colorectal, stomach, oesophageal, liver, brain, bone, kidney, melanoma, lymphoma and leukaemia) and to describe their characteristics at the time of cancer diagnosis.

The specific objectives of the study are:

1. To estimate the incidence rates of thromboembolic events in patients newly diagnosed with each type of selected cancers stratified by country/database, age group, sex, study sub-period, and cancer stage (when available) one and two years after cancer diagnosis.

2. To characterise cancer patients at the time of cancer diagnosis in terms of demographics, comorbidities, concomitant medications, as well as treatments received in the first 90 days after cancer diagnosis.

A description of the proposed objectives to be achieved in the study is found in **Table 1**.

Table 1. Primary and secondary research questions and objectives.

A. Primary research question and objective.

Objective:	To estimate the incidence rates of thromboembolic events in patients newly diagnosed with each type of selected cancers stratified by country/database, age group, sex, study sub-period, and cancer stage (when available) one and two years after cancer diagnosis.
Hypothesis:	N/A
Population *(mention key inclusion-exclusion criteria):*	The study population will include all individuals aged 18 years and above with a primary diagnosis of selected cancers in the study period from 01/01/2016 to 31/12/2023. Selected cancer types include: • Malignant neoplasm of oesophagus (specified in the protocol as oesophageal cancer), • Malignant neoplasm of stomach (stomach cancer), • Malignant neoplasm of colon, rectosigmoid junction, rectum, anus and anal canal (colorectal cancer), • Malignant neoplasm of liver and intrahepatic bile ducts (liver cancer), • Malignant neoplasm of pancreas (pancreatic cancer),

	• Malignant neoplasm of trachea, bronchus and lung (lung cancer), • Malignant neoplasms of bone and articular cartilage • (bone cancer) • Malignant melanoma of skin (skin melanoma) • Malignant neoplasm of breast (breast cancer), • Malignant neoplasm of corpus uteri (corpus uteri cancer including endometrial cancer), • Malignant neoplasm of ovary (ovarian cancer), • Malignant neoplasm of prostate (prostate cancer), • Malignant neoplasm of kidney, except renal pelvis (kidney cancer), • Malignant neoplasm of meninges, brain and spinal cord, cranial nerves and other parts of central nervous system (brain cancer), • Malignant neoplasms, stated or presumed to be primary, of lymphoid, haematopoietic and related tissue (lymphoma and leukaemia) Only patients with the first and one cancer diagnosis (except non-melanoma skin cancer) will be included. Additional eligibility of a minimum of 1 year of potential follow-up time will be imposed to ensure sufficient time to capture potential outcomes of interest.
Exposure:	N/A
Comparator:	N/A
Outcome:	Venous thromboembolic events: • deep vein thrombosis (DVT), • pulmonary embolisms (PE), • venous thromboembolism (VTE, composite of DVT and PE), • pelvic venous thrombosis (PVT), • splanchic vein thrombosis (SVT), including hepatic and extra-hepatic vein thrombosis, • retinal vein thrombosis (RVT), including retinal central vein thrombosis • disseminated intravascular coagulation (DIC)
Time *(when follow up begins and ends):*	For each outcome, study participants will be followed up from the date of the selected cancer diagnosis (index date) until the earliest of the following events: occurrence of the outcome, end of follow up (1 year for the main analysis or 2 years), loss to follow-up, end of data availability, or date of death.
Setting:	Routinely collected data from 10 databases in 8 European countries.

Main measure of effect:	Proportions, incidence rates and incidence rate ratios if information on stage is available

B. Secondary research question and objective.

Objective:	To characterise cancer patients at the time of cancer diagnosis in terms of demographics, comorbidities, concomitant medications, as well as treatments received in the first 90 days after cancer diagnosis.
Hypothesis:	N/A
Population *(mention key inclusion-exclusion criteria):*	The study population will include all individuals aged 18 years and above with a primary diagnosis of selected cancers in the study period from 01/01/2016 to 31/12/2023. Selected cancer types include: • Malignant neoplasm of oesophagus (specified in the protocol as oesophageal cancer, corresponding), • Malignant neoplasm of stomach (stomach cancer), • Malignant neoplasm of colon, rectosigmoid junction, rectum, anus and anal canal (colorectal cancer), • Malignant neoplasm of liver and intrahepatic bile ducts (liver cancer), • Malignant neoplasm of pancreas (pancreatic cancer), • Malignant neoplasm of trachea, bronchus and lung (lung cancer), • Malignant neoplasms of bone and articular cartilage • (bone cancer) • Malignant melanoma of skin (skin melanoma) • Malignant neoplasm of breast (breast cancer), • Malignant neoplasm of corpus uteri (corpus uteri cancer including endometrial cancer), • Malignant neoplasm of ovary (ovarian cancer), • Malignant neoplasm of prostate (prostate cancer), • Malignant neoplasm of kidney, except renal pelvis (kidney cancer), • Malignant neoplasm of meninges, brain and spinal cord, cranial nerves and other parts of central nervous system (brain cancer), • Malignant neoplasms, stated or presumed to be primary, of lymphoid, haematopoietic and related tissue (Lymphoma and leukaemia) Only patients with the first and one cancer diagnosis (except non-melanoma skin cancer) will be included. Additional eligibility of a

	minimum of 1 year of potential follow-up time will be imposed to ensure sufficient time to capture potential outcomes of interest.
Exposure:	N/A
Comparator:	N/A
Outcome:	Medical History: asthma, COPD, chronic liver disease, Crohn's disease, Diabetes mellitus, gastro-oesophageal reflux disease (GERD), GI-Bleeding, Human Immunodeficiency Virus (HIV), Hyperlipidaemia, Hypertension, Obesity, Osteoarthritis, Pneumonia, Psoriasis, Renal impairment, Ulcerative Colitis, Viral Hepatitis, Visual system disorder [General] -- Schizophrenia, Dementia, Parkinson, Depressive disorder, Anxiety, Attention Deficit Hyperactivity Disorder (ADHS) [Neurology], thromboembolic events outcomes of interest mentioned above --- Any cancer except non-melanoma skin cancer (for quality assessment purposes, this should be 0 in our study population before index date). Medication use: Agents acting on the renin-angiotensin system, Antibacterials for systemic use, Antidepressants, Antiepileptics, Anti-inflammatory and antirheumatic products, Antineoplastic agents, Anti-psoriatic, Antithrombotic agents, Beta blocking agents, Calcium channel blockers, Diuretics, Drugs for acid related disorders, Drugs for obstructive airway diseases, Drugs used in diabetes, Immunosuppressants, Lipid modifying agents, Opioids, Psycholeptics, Psychostimulants, agents used for ADHD and nootropics [General] -- contraceptives [contraceptives].
Time *(when follow up begins and ends):*	Medical History at the index date and medication at the index date and from 1 to 90 days.
Setting:	Routinely collected data from 10 databases in 8 European countries.
Main measure of effect:	Proportions

8. RESEARCH METHODS

8.1 Study type and study design

This will be a population-level descriptive epidemiology and patient-level characterisation study. As described in the DARWIN EU® Complete Catalogue of Standard Data Analyses (**Table 2**). A retrospective cohort study of all newly diagnosed adult patients with selected cancers will be conducted.

Table 2. Description of potential study types and related study designs.

Study type	Study design	Study classification
Population-level descriptive epidemiology	Population-level cohort	Off the shelf
Patient-level characterisation	Cohort analysis	Off the shelf

8.2 Study setting and data sources

This study will use routinely collected health data from 10 databases from 8 European countries. All of them have been previously mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). The databases to be included in the study are:

1. Clinical Practice Research Datalink GOLD (CPRD GOLD), United Kingdom
2. Danish Data Health Registries (DK-DHR), Denmark
3. Estonian Biobank (EBB), Estonia
4. Finnish Care Register for Health Care (FinOMOP-HILMO) Finland
5. Hospital District of Helsinki and Uusimaa (FinOMOP-HUS), Finland
6. IQVIA Disease Analyzer Germany (IQVIA DA Germany), Germany
7. IQVIA Longitudinal Patient Database Belgium (IQVIA - LPD Belgium), Belgium
8. Integrated Primary Care Information (IPCI),Netherlands
9. The Information System for Research in Primary Care (SIDIAP), Spain.
10. UK Biobank (UKBB), United Kingdom

Information on data sources is available in **Table 3**. Databases were selected based on the person count of cancer diagnoses under interest and thromboembolic events, and European representativeness. We also considered the lack of major issues and errors during the most recent internal and Darwin EU onboarding quality checks for the selected databases. Preliminary stratification by tumour stage will only be feasible in DK-DHR, EBB and FinOMOP-HUS.

The selected databases also fulfil the criteria required to capture outcomes of interest and relevant data to conduct a patient-level characterisation of newly diagnosed cancer patients across different European settings and regions. Not all databases have all outcomes of interest. DVT is not present in IQVIA - LPD Belgium. SVT counts are limited in EBB, IQVIA - LPD Belgium. RVT is present only in CPRD GOLD, SIDIAP and UKBB.

Table 3. Description of the selected data sources.

Country	Name of Database	Justification for Inclusion	Health Care setting	Type of Data	Number of active subjects	Data lock for the last update
GB	CPRD GOLD	Covers primary care setting, data on cancer diagnoses, comorbidities, medications, and date of death.	Primary care, hospital care (OP)	EHRs	17.5 M	2024-04-15
DK	DK-DHR	National health data database which includes information on cancer diagnoses, staging and medical history	All settings	EHRs, registries, claims	8.5 M	2024-05-15
EE	EBB	Contains health insurance claims, digital prescriptions, discharge information and causes of death through linkage with the national death register. Data is linked to cancer registry.	Primary care, hospital care (IP and OP)	EHRs, claims, registries, biobank	0.2 M	2023-03-20
FI	FinOMOP - HILMO	Nation-wide hospital registry data with high-quality information on cancer diagnoses and mortality.	Hospital care (IP and OP)	EHRs, registries	7.1 M	2024-06-24
FI	FinOMOP - HUS	Hospital registry which includes information on cancer patients and medical history.	Hospital care (IP and OP)	EHRs	3.5 M	2024-05-03

Country	Name of Database	Justification for Inclusion	Health Care setting	Type of Data	Number of active subjects	Data lock for the last update
DE	IQVIA DA Germany	Covers primary care setting with information on cancer diagnoses and medical history.	Primary care	EHRs	43.1 M	2024-03-25
BE	IQVIA LPD Belgium	Covers primary care setting with information on cancer diagnoses and medical history.	Primary care	EHRs	1.1 M	2024-03-25
NL	IPCI	Covers primary care setting, data on cancer diagnoses previously validated, information available on comorbidities, medications, and date of death.	Primary care	EHRs	2.9 M	2024-08-29
ES	SIDIAP	Covers primary care setting with information on cancer diagnoses and medical history.	Primary care	EHRs	8.6 M	2023-03-20
GB	UKBB	Genetic data on biobank participants linked to EHRs from primary care, hospitalisations, cancer registrations and mortality.	Primary care, hospital care (IP and OP)	EHRs, registries, biobank	0.5 M	2024-02-16

BE = Belgium, DE = Germany, DK=Denmark, EE = Estonia, ES=Spain, EHR=Electronic Health Record, FR = France, GB = United Kingdom of Great Britain and Northern Ireland, FI = Finland, IP = Inpatient, NL = the Netherlands, OP = outpatient

Clinical Practice Research Datalink GOLD (Oxford) (CPRD GOLD)

The CPRD GOLD database collects data in the United Kingdom and is maintained by Clinical Practice Research Datalink. CPRD GOLD contains data from all four UK constituent countries (England, Scotland, and Northern Ireland). A database contains data from 1987 to present. Data are collected from general practitioner (GP) clinics that use the Vision® software system. GP clinics are responsible for non-emergency care and referrals. Data is collected from patient records stored at GP clinics. Over 98% of the UK population is registered with a GP. Covering 4.6% of the current UK population, CPRD GOLD includes 4.9% of contributing GP practices. Compared with the UK Census 2011, CPRD patients are broadly representative of the UK population in terms of age, sex, and ethnicity and comparable to the Health Survey for England for body mass index distribution. The CPRD GOLD may not be representative of all practices in the UK based on geography and size. It obtains data from electronic health records.

The database currently holds information about the person (demographics), visits (inpatient and outpatient), conditions and procedures (billing codes), drugs (outpatient prescriptions and inpatient orders and administrations), measurements (laboratory tests and vital signs), and dates of death (in or out-hospital death). CPRD GOLD uses SNOMED medical terms. Transformation-level validation checks for referential integrity between records ensure that no orphan records are included in the database (for example, all event records link to a patient). Duplicate records are identified and removed. In contrast, research-quality-level validation covers the actual content of the data. CPRD provides a patient-level data quality metric as a binary 'acceptability' flag. This is based on recording and internal consistency of key variables, including date of birth, practice registration date, and transfer out date. Validation of the CPRD has shown a high positive predictive value for some diagnoses, and where evaluated, comparisons of incidence with other UK data sources are also broadly similar. Research into data quality has shown large variations in the inter-practice recording of data. Established linkages include, but are not limited to, Hospital Episode Statistics (hospitalisation data), Office for National Statistics (mortality data including causes of death), Index of Multiple Deprivation and Townsend scores (deprivation data), and disease registries. Vital status (death date and causes) is obtained from the Office for National Statistics.

CPRD GOLD is limited to GP records. General practices receive information about patient contacts with secondary care, which must be manually entered into the patient record. However, the database also combines data from various sources through effective linkages. Primary care data quality is variable because GPs enter data during routine consultations, not for research. The database was described by Herrett et al., 2015. The database description is also available at cprd.com.

Danish Data Health Registries (DK-DHR)

Danish health data is collected, stored, and managed in national health registers at the Danish Health Data Authority. It covers the entire population, which makes it possible to study the development of diseases and their treatment over time. There are no gaps in terms of gender, age and geography in Danish health data due to mandatory reporting on all patients from cradle to grave in all hospitals and medical clinics. Personal identification numbers link data across registers, so we have data on all Danes throughout their lives, regardless of whether they have moved around the country. High data quality due to standardisation, digitisation and documentation means Danish health data is not based on interpretation. The Danish Health Data Authority is responsible for the national health registers and maintaining and developing standards and classifications in the Danish healthcare system. The legislation ensures a balance between personal data protection and use. In the present database, we have access to the following registries for the entire Danish population of 5.9 million persons from 1.1.1995: The Central Person Registry (CPR), The National Patient Registry (LPR), The Register of Pharmaceutical Sales (LSR), The National Cancer Register (CAR), The Cause of Death registry (DAR), The Clinical Laboratory Information Register (LAB), COVID-19 test and

vaccination Registries (SSI-OVD, SSI-DDV), The complete Vaccination registry (DDV_all). All data registered from 1.1.1995 will be included.

Estonian Biobank (EBB)

The EBB collects data in Estonia and is maintained at the Estonian Genome Center, University of Tartu. It is a nationwide database, and the network of recruitment offices for EBB covers all 15 counties of the country. The Estonian Genome Project Foundation initiated the Estonian Biobank Project in 1999, which was transformed into the Estonian Genome Center of the University of Tartu (EGCUT) in 2007. The data is available from 2004 onwards. The Estonian Biobank cohort is a volunteer-based sample of the Estonian resident adult population. EBB represents an Estonian population-based cohort size of 52,000 participants aged 18 years and older recruited at GP offices, private practices, and hospitals or in the recruitment offices of the Estonian Genome Center. The age, sex, and geographical distribution closely reflect those of the Estonian adult population and encompass nearly 5% of the population. Overall, the representation of men in the biobank is 3.4%, and women's is 5.5%. Older people tend to participate less frequently; however, all age groups are well represented. The database obtains data from the biobank records. All participants have undergone a standardised health assessment, including providing 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.

For all starting data collectors, the first ten questionnaires were monitored for completeness and illogical answers, and after that, 10% were selected randomly for monitoring; 21% were inspected and corrected when necessary. From the monitored questionnaires, 99% were classified as high quality, meaning all the fields were filled in, and the answers appeared logical. Follow-up data are available via linkage with national health-related registries and re-examination of participants. Furthermore, electronic health records are updated every half year for phenotypic outcome information. The EBB database is regularly linked with national registries, hospital databases, and the national health insurance fund database, which holds treatment and service bills. Vital status (death date and causes) is obtained from the Causes of Death Registry. Participation in the EBB cohort is voluntary; therefore, the biobank does not represent a classical random sample and could be subject to recruitment bias. Although recruitment was open to everyone, there is a disproportion of ethnic Estonians and ethnic Russians in the biobank, with Estonians being overrepresented and Russians underrepresented. Also, the limited depth of collected data can sometimes limit the number of projects in which the data can be used. The database was described by Leitsalu et al., 2015, 10.1093/ije/dyt268. The database description is also available at genomics.ut.ee/en/content/estonian-biobank.

Finnish Care Register for Health Care (FinOMOP - HILMO)

The Finnish Care Register for Health Care (fi: Hoitoilmoitusrekisteri) continues the former Hospital Discharge Register, which originally gathered data on hospital discharge (Sund et al., 2012). The Care Register has comprehensive data on services and service users nationwide, including Finnish public inpatient and outpatient primary and specialised care. Since 1998, the register has covered public outpatient and inpatient specialised care and private inpatient care (TerveysHilmo). Since 2011, the register has covered public primary care (AvoHilmo). Since 2020, the register has covered private outpatient care and occupational care. The CDM is currently produced from the data collection on inpatient and outpatient specialised care (TerveysHilmo) and is limited to observation periods commencing after 01/01/2015. The Register of Primary Health Care Visits (AvoHilmo) is currently outside the scope of the CDM and will be added to CDM during the remainder of 2023. The inclusion of data collected before 2015 is also being planned. The National Population Registry is also used as a source for the CDM database. The National

Population Registry data forms the basis for forming the patient population. This ensures up-to-date location (municipality of residence) of patients and complete death occurrences (although not the cause of death). Using the complete population as a basis for the person table also facilitates calculations on a population level, e.g. incidence rates. The HILMO database has been used to assess the quality of cancer registry data in Finland (Leinonen et al., 2017).

Hospital District of Helsinki and Uusimaa (FinOMOP - HUS)

The HUS data lake is a comprehensive, integrated data source derived in real-time from all patients who visit the HUS hospitals and receive treatment (Vikkula et al., 2023). HUS is responsible for specialised healthcare in Finlands' Uusimaa region and the treatment of many rare and severe diseases nationally centralised to HUS. HUS's catchment area covers about 2.2 million people. In 2023, there were 2.43 million booked appointments and 255,896 emergency department visits for specialist medical care. 691,702 patients received any treatment in HUS specialist medical care and at emergency departments, and 86,849 surgical procedures were performed. All visits, examinations, laboratory tests, procedures, and treatments are recorded 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 past and current source systems are available.

IQVIA Disease Analyzer Germany (IQVIA DA Germany)

Germany DA is collected from extracts of patient management software used by GPs and specialists from ambulatory care settings. Patients visiting multiple providers are not cross-identified for data protection reasons and, therefore, recorded as separate in the system. Dates of service include from 1992 through the present. The first and last consultation dates define observation time. Germany has no mandatory GP system, and patients can choose specialists. Drugs are recorded as prescriptions of marketed products. No explicit registration or approval is needed for drug utilisation studies.

IQVIA Longitudinal Patient Database Belgium (IQVIA LPD Belgium)

Belgium Longitudinal patient data (LPD) is collected from GP prescribing systems and contains patient records on all signs and symptoms, diagnoses and prescribed medications. The information recorded allows patients and doctors to be monitored longitudinally. Data are recorded directly in real-time during patient consultations via a practice management software system. It is used in studies to provide various market insights such as treatment trends, patient pathway analysis and treatment compliance. The panel of contributing physicians (a stable 300 GPs) is maintained as a representative sample of Belgium's primary care physician population according to three criteria known to influence prescribing: age, sex and geographical distribution. Currently, the database covers 1.1 M cumulative patients from 2012 through to the present. The panel consists of a stable 300 GPs that are geographically well-spread. The total number of active GPs in Belgium is 15.602. The regional geographical spread of physicians in the LPD data is also representative of the distribution across the country: 57% GPs in the North (compared to 54% nationally), 31% in the South (33% nationally) and 12% in Brussels (13%). The data provider has more than 2.250 GPs under contract, so a replacement is easily found in case of a dropout. Drugs obtained over the counter by the patient outside the prescription system are not reported. No explicit registration or approval is needed for drug utilisation studies.

Integrated Primary Care Information (IPCI)

The database collects data in the Netherlands. It was started in 1992 by the Department of Medical Informatics of the Erasmus University Medical Center in Rotterdam, the Netherlands. The current database contains patient records from 2006 onwards when the size of the database started to increase significantly. IPCI is a nationwide Dutch database. However, it mainly covers the central part of the country, including the most densely populated and non-urban areas. The IPCI database contains data from records of general

practitioners' (GPs) practices. It contains information on all patients registered with GPs responsible for non-emergency care and referrals. More than 99% of the Dutch population has health insurance, and almost all citizens are registered with a general practitioner. Over 12 months, around 78% of the population has at least one contact with their GP. IPCI included around 350 GP practices out of around 5000 in the country (~ 7%). The demographic composition of the IPCI population mirrors that of the general Dutch population in terms of age and sex.

IPCI obtains data from computer-based patient records. Patient-level data includes demographic information, complaints and symptoms, diagnoses, laboratory test results, lifestyle factors, and correspondence with secondary care, such as referral and discharge letters. Dutch GPs use the International Classification of Primary Care (ICPC-1) coding for complaints, symptoms, and diagnoses, an international standard developed and updated by the World Organization of Family Doctors (WONCA) International Classification Committee.

Extensive quality control steps are performed before each data release. These include comparing patient characteristics between practices and checks to identify abnormal temporal data patterns in practices. Additional checks include over 200 indicators related to population characteristics (e.g., reliability of birth and mortality rates) and medical data (e.g., availability of durations of prescriptions, completeness of laboratory results, availability of hospital letters and prescriptions, the proportion of patients with blood pressure measurement, etc.).

IPCI is not linked with other databases. Vital status (death date and cause) is collected based on GP records. The main limitation is that IPCI is limited to GP records, and although it contains information on referrals and discharge letters, it may not capture specific hospital information. The database profile was described by de Ritter et al., 2022. The database description is also available at ipci.nl.

The Information System for Research on Primary Care (SIDIAP)

The database collects data in Spain and is maintained by the SIDIAP team, supported by the Catalan Health Institute and the Institute for Primary Health Care Research Jordi Gol i Gurina. It contains patient records from 2005 onwards and is updated every six months. It is a regional database covering the region of Catalonia. SIDIAP collects data in the primary care setting. It contains data from the population registered in over 280 primary care practices throughout Catalonia. Approximately 80% of the population registered with primary care is covered by SIDIAP. The demographic composition within SIDIAP closely mirrors that of the broader Catalan population, encompassing a representative spectrum of geographic distribution, age, and sex proportions. SIDIAP obtains data from electronic health records. The dataset covers demographics, all-cause mortality, disease diagnoses, prescription and dispensation records of drugs, results of laboratory tests, socio-economic indicators, vaccination records, lifestyle information, parent-child linkage, and various clinical parameters. Diseases are classified under the International Classification of Diseases 10th revision (ICD-10)

Quality checks have been implemented, including central identification of duplicate patient IDs and visual inspection for temporal patterns in the registry of a certain variable. Furthermore, the data undergoes assessment for availability (longitudinally and reliability), plausibility (range checks and unusual values), and visualisation tools. Specifically, for biochemistry data, consistency for measurements taken in different laboratories is assessed, and unit conversion is undertaken when needed. SIDIAP is linked with numerous other databases. It integrates data from external sources, including laboratory biomarker data, drug prescription and dispensation records, hospital discharge records, mental health centres, and other specific disease registries. Vital status (death date and cause) is collected through linkage with the civil registry. The main limitation is that SIDIAP covers only primary health care records. However, it is combined with data

from various other sources through effective linkages. The database profile was described by de Recalde et al., 2022. The database description is also available at sidiap.org.

UK BioBank (UKBB)

UK Biobank is a powerful biomedical database that can be accessed globally to enable discoveries to improve public health. UK Biobank contains in-depth genetic, biomarker, imaging and health information from over half a million volunteers living in the UK aged 40–69 years at the time of recruitment (2006–2010). UK Biobank has collected unprecedented biological and medical data for a large-scale, long-term prospective study. With their consent, they regularly provide blood, urine and saliva samples and detailed information about their lifestyle, which is then linked to their health-related records (e.g. primary care data, hospital data, cancer registry) to provide a deeper understanding of how individuals experience diseases. Since 2012, the UK Biobank database, the largest and richest of its kind, has been open to applications from researchers. The resource is available in anonymised format to scientists from the UK and worldwide, subject to verification that the research is health-related and in the public interest. Researchers are required to publish their results in an open-source publication site or an academic journal and return their findings to the UK Biobank. At the time of writing, nearly 3,600 research applications have been approved for using UK Biobank data, and 3,239 peer-reviewed articles based on them have been published.

8.3 Study period

The study period will be from 01/01/2016 to 31/12/2023 or the end of available data in each source if it comes earlier (see **Table 3** Data lock for the last update column for more details).

8.4 Follow-up

Study participants will be followed up from the date of cancer diagnosis (index date, **Table 4**) until the first occurrence of any of the following events: occurrence of the outcome, loss to follow-up, end of follow-up (one or two years), end of data availability, or date of death.

Table 4. Operational definition of time 0 (index date) and other primary time anchors.

Study population name(s)	Time Anchor Description (e.g. time 0)	Number of entries	Type of entry	Washout window	Care Setting[1]	Code Type[2]	Diagnosis position	Incident with respect to…	Measurement characteristics/ validation	Source of algorithm
Cancer patients	Date of cancer diagnosis	Single entry	Incident	[any time, -1]	IP, OP, OT	SNOMED, ICD-O-3	Any	Any cancer diagnosis except non-melanoma skin cancer	N/A	N/A

[1] IP = inpatient, OP = outpatient, OT = other, n/a = not applicable

[2] SNOMED = Systematized Nomenclature of Medicine, ICD-O-3: International Classification of Diseases for Oncology, 3rd Edition

8.5 Study population with inclusion and exclusion criteria

The study population will include all individuals aged 18 years and above with a primary diagnosis of selected cancers in the study period. Cancer types will include lung, breast, ovary, endometrium, prostate, pancreas, colorectal, stomach, oesophageal, liver, brain, bone, kidney, melanoma, lymphoma and leukaemia. Only patients with the first and one cancer diagnosis (except non-melanoma skin cancer) will be included. Cancer cases and thromboembolic events will be identified based an appropriate computable phenotyping algorithms. Conditions in the OMOP CDM use the Systematised Nomenclature of Medicine (SNOMED) as the standard vocabulary for diagnosis codes. For cancer diagnoses, the International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) will also be considered. Algorithms to reproduce cancer phenotypes will be shown along with the study results.

Additional eligibility of a minimum of 1 year of potential follow-up time will be imposed to ensure sufficient time to capture potential outcomes of interest. For instance, in a database with data up to 31/12/2023, cancer cases diagnosed from 01/01/2016 up to 01/01/2023 will be included. A prior history requirement of one year of observation prior to the index date will be imposed. This is especially relevant for primary care databases, where a minimum observation period of one year before a cancer diagnosis is required to detect prevalent cases.

Appendix II provides a preliminary code list for each cancer. The code list might be modified after cohort diagnostics

The operational definitions of the inclusion and exclusion criteria are presented in Error! Reference source not found. and **Table 6**, respectively.

Table 5. Operational definitions of inclusion criteria.

Criterion	Details	Order of application	Assessment window	Care Settings[1]	Code Type	Diagnosis position	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Patients newly diagnosed with selected cancers	Primary selected cancer	After	-	IP, OP, OT	SNOMED, ICD-O-3	N/A	Cancer patients	N/A	N/A
Age	Participants aged 18 or above	After	At index date	IP, OP, OT	N/A	N/A	Cancer patients	N/A	N/A
Minimum prior observation period of 365 days	Only participants with a minimum observation period of 365 days prior to diagnosis of chondrosarcoma (index date) s	Before	365 days	OP, OT	N/A	N/A	Cancer patients	N/A	N/A
Minimum potential follow-up time	Only participants with a cancer diagnosis (index date) occurring one year prior to end of data availability in the database will be included	After	[0, 365]	IP, OP, OT	N/A	N/A	Cancer patients	N/A	N/A

[1] IP = inpatient, OP = outpatient, OT = other, n/a = not applicable.

Table 6. Operational definitions of exclusion criteria.

Criterion	Details	Order of application	Assessment window	Care Settings[1]	Code Type	Diagnosis position	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
History of cancer diagnosis	Participants with a diagnosis of cancer (any, excluding non-melanoma skin cancer) any time prior index date	After	Any time prior to cancer diagnosis	OP, IP, OT	SNOMED, ICD-O-3	Any	Cancer patients	N/A	N/A
Multiple primary tumors	Participants with a two diagnoses of cancer (any, excluding non-melanoma skin cancer) at index date	After	At index date	OP, IP, OT	SNOMED, ICD-O-3	Any	Cancer patients	N/A	N/A

[1] IP = inpatient, OP = outpatient, OT = other, n/a = not applicable

8.6 Variables

8.6.1. Exposures

None.

8.6.2. Outcomes

Conditions of interest will include thromboembolic events, in particular, (deep vein thrombosis (DVT), pulmonary embolisms (PE), venous thromboembolism (VTE, composite of DVT and PE), pelvic venous thrombosis (PVT), splanchic vein thrombosis (SVT), including hepatic and extra-hepatic vein thrombosis, retinal vein thrombosis (RVT), including retinal central vein thrombosis, and disseminated intravascular coagulation (DIC).

The operational definition of the outcomes is presented in the **Table** .

A preliminary list of codes is provided in the **Appendix I**.

Table 7. Operational definitions of outcome.

Outcome name	Details	Primary outcome	Type of outcome	Washout window	Care Settings[1]	Code Type	Diagnosis Position	Applied to study populations	Measurement characteristics/ validation	Source of algorithm
DVT	See Appendix I	Yes	Binary	[-365,-1]	IP, OP, OT	SNOMED	Any	Cancer patients	N/A	N/A
PE	See Appendix I	Yes	Binary	[-365,-1]	IP, OP, OT	SNOMED	Any	Cancer patients	N/A	N/A
VTE	See Appendix I	Yes	Binary	[-365,-1]	IP, OP, OT	SNOMED	Any	Cancer patients	N/A	N/A
PVT	See Appendix I	Yes	Binary	[-365,-1]	IP, OP, OT	SNOMED	Any	Cancer patients	N/A	N/A
SVT	See Appendix I	Yes	Binary	[-365,-1]	IP, OP, OT	SNOMED	Any	Cancer patients	N/A	N/A
RTV	See Appendix I	Yes	Binary	[-365,-1]	IP, OP, OT	SNOMED	Any	Cancer patients	N/A	N/A
DIC	See Appendix I	Yes	Binary	[-365,-1]	IP, OP, OT	SNOMED	Any	Cancer patients	N/A	N/A

[1] IP = inpatient, OP = outpatient, OT = other, n/a = not applicable, DVT = deep vein thrombosis, PE = pulmonary embolisms, VTE = venous thromboembolism, PVT = pelvic venous thrombosis, SVT = splanchic vein thrombosis, RVT = retinal vein thrombosis, DIC = DIC.

8.6.3. Other covariates, including confounders, effect modifiers and other variables

The age at the index date (primary cancer diagnosis) will be described. The following age grouping will be used: 18-29; 30-39; 40-49; 50-59; 60-69; 70-79; 80-89; 90 and over. The sex (male/female) of study participants will also be identified.

Large-scale patient-level characterisation will be conducted at the time of diagnosis. Age and sex at the time of cancer diagnosis will be described for each generated study cohort. Medical history will be assessed at the time of diagnosis. Medication use history will be reported at the time of diagnosis. We will also report medication use for 1 to 90 days post-index date.

A list of pre-specified co-morbidities and co-medications will be described. Co-morbidities and co-medications were selected based on definitions that were previously used in other DARWIN EU studies. These will include:

- Medical History: asthma, COPD, chronic liver disease, Crohn's disease, Diabetes mellitus, gastro-oesophageal reflux disease (GERD), GI-Bleeding, Human Immunodeficiency Virus (HIV), Hyperlipidaemia, Hypertension, Obesity, Osteoarthritis, Pneumonia, Psoriasis, Renal impairment, Ulcerative Colitis, Viral Hepatitis, Visual system disorder [General] -- Schizophrenia, Dementia, Parkinson, Depressive disorder, Anxiety, Attention Deficit Hyperactivity Disorder (ADHS) [Neurology], thromboembolic events outcomes of interest mentioned above --- Any cancer except non-melanoma skin cancer (for quality assessment purposes, this should be 0 in our study population before index date).
- Medication use: Agents acting on the renin-angiotensin system, Antibacterials for systemic use, Antidepressants, Antiepileptics, Anti-inflammatory and antirheumatic products, Antineoplastic agents, Anti-psoriatic, Antithrombotic agents, Beta blocking agents, Calcium channel blockers, Diuretics, Drugs for acid related disorders, Drugs for obstructive airway diseases, Drugs used in diabetes, Immunosuppressants, Lipid modifying agents, Opioids, Psycholeptics, Psychostimulants, agents used for ADHD and nootropics [General] -- contraceptives [contraceptives].

If available, AJCC\UICC TNM stage groups, and WHO/ECOG performance status will also be described and used for stratification in objective 1.

The operational definition of the covariates is described in the

Table 8. Operational definitions of covariates.

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type	Diagnosis Position	Applied to study populations	Measurement characteristics/ validation	Source for algorithm
Co-morbidities	Large-scale patient-level characterisation with regard to underlying comorbidities	Binary	[0,0]	IP, OP, OT	SNOMED	N/A	Cancer patients	N/A	N/A
Medication use	Large-scale patient-level characterisation with regard to use of concomitant drugs	Binary	[0,0], [1,90]	IP, OP, OT	RxNorm	N/A	Cancer patients	N/A	N/A

[1] IP = inpatient, OP = outpatient, OT = other, n/a = not applicable

8.

Table 8. Operational definitions of covariates.

Characteristic	Details	Type of variable	Assessment window	Care Settings [1]	Code Type	Diagnosis Position	Applied to study populations	Measurement characteristics/ validation	Source for algorithm
Co-morbidities	Large-scale patient-level characterisation with regard to underlying comorbidities	Binary	[0,0]	IP, OP, OT	SNOMED	N/A	Cancer patients	N/A	N/A
Medication use	Large-scale patient-level characterisation with regard to use of concomitant drugs	Binary	[0,0], [1,90]	IP, OP, OT	RxNorm	N/A	Cancer patients	N/A	N/A

[1] IP = inpatient, OP = outpatient, OT = other, n/a = not applicable

8.7 Study size

No sample size will be calculated as this is a descriptive study. However, analysis by strata will be limited to databases with enough cases (5 or more) to provide meaningful results. Summary measures of occurrence will not be calculated for strata with counts less than 5.

8.8 Analysis

8.1.1. Federated network analyses

Analyses will be conducted separately for each database and will be carried out in a federated manner, allowing analyses to be run locally without sharing patient-level data. Before study initiation, the analytics are tested on a subset of the data sources or on a simulated set of patients, and quality control checks are performed. Once all the tests are passed, the final study code is released in the version-controlled Study Repository for execution against all the participating data sources.

The data partners locally execute the analytics against the OMOP CDM in R Studio and review and approve the default aggregated results before returning them to the Coordination Centre. Sometimes, multiple execution iterations are performed, and additional fine-tuning of the code base is needed. A service desk will be available for support during the study execution.

The study results of all data sources are checked, after which they are made available to the team in the Digital Research Environment (DRE), and the Study Dissemination Phase can start. All results are locked and timestamped for reproducibility and transparency.

8.8.2 Patient privacy protection

Cell suppression will be applied as required by databases to protect people's privacy. Cell counts < 5 will be masked.

8.8.3 Statistical model specification and assumptions of the analytical approach considered

Data analyses

This study will involve the analysis described in **Table 9**.

Table 9. Description of study types and type of analysis.

Study type	Study classification	Type of analyses
Patient-level characterisation	Off-the-shelf	- Large-scale characterisation - Patient-level characteristics - Prognosis / progression to a pre-specified outcome - Standard care description
Population-level descriptive epidemiology	Off-the-shelf	- Incidence rates of the condition of interest - Prevalence rates of the condition of interest

The incidence rates of thromboembolic events (Objective 1) will be estimated over one and two years after the selected cancer diagnosis. The denominator will include the person-time contributed by the cancer patients from the date of diagnosis till the event or end of the respective period (one or two years). The numerator will include the outcomes of interest. Incidence rates will be presented per person-time with 95% confidence intervals derived using the exact method. Additionally, incidence rate ratios will be calculated when information on cancer stage is available with group AJCC/UICC stage as a reference. Each cancer type and outcome will be assessed separately.

Large-scale patient-level characterisation (Objective 2) will be conducted at the index date. Age and sex will be described at the time of diagnosis as well as follow-up time after the index date. The medical history and medication will be assessed at the index date. Medication use will be at the index date and from 1 to 90 days after diagnosis.

For all analyses, absolute counts and proportions will be reported. A minimum cell count of 5 will be used when reporting results, with any smaller counts reported as "<5" and zero counts as "0". Overall analyses will be done separately for each database.

Further stratification by age category (18-29; 30-39; 40-49; 50-59; 60-69; 70-79; 80-89; 90 and over), sex, study sub-period, and cancer stage will be conducted when possible (minimum cell count reached and data available). The following sub-periods will also be used: 2016-2019 and 2020-2023.

R-packages

The following R packages will be used: "CohortDiagnostics" (https://github.com/OHDSI/CohortDiagnostics/) to evaluate the phenotype algorithms developed for the study.

"IncidencePrevalence" (https://github.com/darwin-eu/IncidencePrevalence) to estimate incidence rates by dividing the number of events by the person-time. Incidence rates will be calculated using function *estimateIncidence()* with denominator person-time specified by *generateDenominatorCohortSet()*.

"PatientProfiles" (https://github.com/darwin-eu-dev/PatientProfiles) and "CohortCharacteristics" (https://github.com/darwin-eu-dev/CohortCharacteristics) will be used for summarising characteristics of cohorts of patients with cancer.

8.9 Evidence synthesis

Results from analyses described in section 8.8 will be presented separately for each database and no meta-analysis of results will be conducted.

9. DATA MANAGEMENT

9.1 Data management

All databases are mapped to the OMOP common data model. This enables the use of standardised analytics and tools across the network since the structure of the data and the terminology system is harmonised. The OMOP CDM is 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. Each data partner will execute the study code against their database containing patient-level data and will then return the results set which will only contain aggregated data. The results from each of the contributing data sites will then be combined in tables and figures for the study report.

9.2 Data storage and protection

For this study, participants from various EU member states will process personal data from individuals which is collected in national/regional electronic health record databases. 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 databases 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 Digital Research Environment. These output files do not contain any data that allows the 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.

10. QUALITY CONTROL

General database quality control

Several open-source quality control mechanisms for the OMOP CDM have been developed (see Chapter 15 of The Book of OHDSI http://book.ohdsi.org/DataQuality.html). In particular, data partners are expected to run the OHDSI Data Quality Dashboard tool (https://github.com/OHDSI/DataQualityDashboard). This tool provides numerous checks relating to the conformance, completeness and plausibility of the mapped data. Conformance focuses on checks that describe the compliance of the representation of data against internal or external formatting, relational, or computational definitions; completeness in the sense of data quality is solely focused on quantifying missingness or the absence of data, while plausibility seeks to determine the believability or truthfulness of data values. Each of these categories has one or more subcategories and are evaluated in two contexts: validation and verification. Validation relates to how well data aligns with external benchmarks with expectations derived from known true standards. In contrast, verification relates to how well data conforms to local knowledge, metadata descriptions, and system assumptions.

Study-specific quality control

When defining selected cancers, outcomes and co-morbidities, a systematic search of possible codes for inclusion was identified using CodelistGenerator R package (https://github.com/darwin-eu/CodelistGenerator). This software allows the user to define a search strategy and using this will then query the vocabulary tables of the OMOP CDM to find potentially relevant codes. The codes returned will be then reviewed by two clinical epidemiologists to consider their relevance. In addition, the CohortDiagnostics R package (https://github.com/OHDSI/CohortDiagnostics) will be run to assess the use of different codes across the databases contributing to the study and identify any codes potentially omitted in error. This will allow for a consideration of the validity of the study cohort of patients with the selected

cancers and co-morbidities in each of the databases and inform decisions around whether multiple definitions are required.

The study code will be based on three R packages currently being developed to (1) estimate incidence rates of thromboembolic events ("IncidencePrevalence"), (2) characterise demographic and clinical characteristics ("PatientProfiles" and "CohortCharacteristics"). These packages will include numerous automated unit tests to ensure the validity of the codes, alongside software peer review and user testing. The study code will be made publicly available via GitHub.

11. LIMITATIONS OF THE RESEARCH METHODS

The study will be informed by routinely collected health care data, and so data quality issues must be considered. In particular, the identification of cancer patients and thromboembolic events may vary across databases. While relatively few false positives would be expected, false negatives may be more likely, especially for databases without patient-level linkage to secondary care data. Underestimation of thromboembolic events is also possible, particularly for rare events with complex diagnoses.

Given the large number and diverse nature of participating data sources, it is important to note that differences in patient representations might arise from disparate coding practices and specifics of data capture. The granularity or detail of concepts representing clinical facts can vary across source terminologies (e.g., ICD-10, Read codes), influencing how information is later transformed into standardised vocabularies (Ostropolets et al., 2021). The preliminary code lists created to identify cancer patients include codes from standard vocabularies used in tumour registries, such as ICD-O-3 codes. However, most databases will only have information on cancer diagnoses using SNOMED codes, which may not be granular enough to cover all the topology and histology details of cancer (Campbell et al., 2014). ICD-O-3 codes are only available at DK-DHR, EBB and UKBB.

The large-scale characterisation will provide an overview of the characteristics, comorbidities, and medication use, including anticancer treatment is available, of cancer patients. However, our study will not differentiate the time before and after cancer treatment initiation, and therefore it will not be able to disentangle the risk of thromboembolic events posed by cancer treatments from the risk posed by cancer itself.

In addition, for the calculation of incidence rates (Objective 1), we will apply a one-year washout to exclude patients who experienced the thromboembolic events under study prior to index date. Given that acute thromboembolic events can be the first manifestation of an occult malignancy, we will miss patients with thromboembolic events who are subsequently diagnosed with cancer.

12. MANAGEMENT AND REPORTING OF ADVERSE EVENTS/ADVERSE REACTIONS

Adverse events/adverse reactions will not be collected or analysed as part of this evaluation. The nature of this non-interventional evaluation, through the use of secondary data, does not fulfil the criteria for reporting adverse events, according to module VI, VI.C.1.2.1.2 of the Good Pharmacovigilance Practices (https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guideline-good-pharmacovigilance-practices-gvp-module-vi-collection-management-submission-reports_en.pdf).

13. GOVERNANCE BOARD ASPECTS

All data sources require approval from their respective IRB boards, with the exception of IQVIA DA Germany and IQVIA LPD Belgium which will not require any further specific approvals to undertake this study.

14. 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. If requested, the full set of underlying aggregated data used in the dashboard will also be made available.

15. OTHER ASPECTS

None.

16. REFERENCES

Timp, J. F., Braekkan, S. K., Versteeg, H. H., & Cannegieter, S. C. (2013). Epidemiology of cancer-associated venous thrombosis. Blood, The Journal of the American Society of Hematology, 122(10), 1712-1723.

Blom, J. W., Doggen, C. J., Osanto, S., & Rosendaal, F. R. (2005). Malignancies, prothrombotic mutations, and the risk of venous thrombosis. Jama, 293(6), 715-722.

Björklund, J., Rautiola, J., Zelic, R., Edgren, G., Bottai, M., Nilsson, M., ... & Akre, O. (2024). Risk of Venous Thromboembolic Events After Surgery for Cancer. JAMA Network Open, 7(2), e2354352-e2354352.

Khorana, A. A., Dalal, M., Lin, J., & Connolly, G. C. (2013). Incidence and predictors of venous thromboembolism (VTE) among ambulatory high-risk cancer patients undergoing chemotherapy in the United States. Cancer, 119(3), 648-655.

Nalluri, S. R., Chu, D., Keresztes, R., Zhu, X., & Wu, S. (2008). Risk of venous thromboembolism with the angiogenesis inhibitor bevacizumab in cancer patients: a meta-analysis. Jama, 300(19), 2277-2285.

Ohashi, Y., Ikeda, M., Kunitoh, H., Sasako, M., Okusaka, T., Mukai, H., ... & Sakon, M. (2020). Venous thromboembolism in cancer patients: report of baseline data from the multicentre, prospective Cancer-VTE Registry. Japanese journal of clinical oncology, 50(11), 1246-1253.

Petterson, T. M., Marks, R. S., Ashrani, A. A., Bailey, K. R., & Heit, J. A. (2015). Risk of site-specific cancer in incident venous thromboembolism: a population-based study. Thrombosis research, 135(3), 472-478.

van Hylckama Vlieg, M. A., Nasserinejad, K., Visser, C., Bramer, W. M., Ashrani, A. A., Bosson, J. L., ... & Geijteman, E. C. (2023). The risk of recurrent venous thromboembolism after discontinuation of anticoagulant therapy in patients with cancer-associated thrombosis: a systematic review and meta-analysis. EClinicalMedicine, 64.

White, R. H., Chew, H. K., Zhou, H., Parikh-Patel, A., Harris, D., Harvey, D., & Wun, T. (2005). Incidence of venous thromboembolism in the year before the diagnosis of cancer in 528 693 adults. Archives of internal medicine, 165(15), 1782-1787.

Wang, T. F., Li, A., & Garcia, D. (2018). Managing thrombosis in cancer patients. Research and Practice in Thrombosis and Haemostasis, 2(3), 429-438.

Herrett, E., Gallagher, A. M., Bhaskaran, K., Forbes, H., Mathur, R., van Staa, T., & Smeeth, L. (2015). Data Resource Profile: Clinical Practice Research Datalink (CPRD). International journal of epidemiology, 44(3), 827–836.

de Ridder, M. A. J., de Wilde, M., de Ben, C., Leyba, A. R., Mosseveld, B. M. T., Verhamme, K. M. C., van der Lei, J., & Rijnbeek, P. R. (2022). Data Resource Profile: The Integrated Primary Care Information (IPCI) database, The Netherlands. International journal of epidemiology, 51(6), e314–e323.

Leinonen, M. K., Miettinen, J., Heikkinen, S., Pitkäniemi, J., & Malila, N. (2017). Quality measures of the population-based Finnish Cancer Registry indicate sound data quality for solid malignant tumours. European journal of cancer (Oxford, England: 1990), 77, 31–39.

Vikkula, J., Uusi-Rauva, K., Ranki, T., Toppila, I., Aalto-Setälä, M., Pousar, K., Vassilev, L., Porkka, K., Silvennoinen, R., & Brück, O. (2023). Real-world evidence of multiple myeloma treated from 2013 to 2019 in the Hospital District of Helsinki and Uusimaa, Finland. Future oncology (London, England), 19(30), 2029–2043.

Sund R. (2012). Quality of the Finnish Hospital Discharge Register: a systematic review. Scandinavian journal of public health, 40(6), 505–515.

Recalde, M., Davila-Batista, V., Díaz, Y., Leitzmann, M., Romieu, I., Freisling, H., & Duarte-Salles, T. (2021). Body mass index and waist circumference concerning the risk of 26 types of cancer: a prospective cohort study of 3.5 million adults in Spain. BMC medicine, 19(1), 10.

Ostropolets A, Reich C, Ryan P, et al. Characterizing database granularity using SNOMED-CT hierarchy. AMIA Annu Symp Proc. 2021;2020:983-992. Published 2021 Jan 25.

Campbell, W. S., Campbell, J. R., West, W. W., McClay, J. C., & Hinrichs, S. H. (2014). Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings. Journal of the American Medical Informatics Association : JAMIA, 21(5), 885–892.

17. ANNEXES

Appendix I

Table A1 Preliminary code list for deep vein thrombosis (DVT)

Concept Id	SNOMED Code	Concept Name
762047	286411000119108	Acute bilateral thrombosis of subclavian veins
762148	293471000119106	Acute deep vein thrombosis of bilateral iliac veins
761444	15711361000119101	Acute deep vein thrombosis of bilateral lower limbs following coronary artery bypass graft
35616028	293491000119107	Acute deep vein thrombosis of left iliac vein
35615035	15711401000119105	Acute deep vein thrombosis of left lower limb following procedure
761416	15708441000119103	Acute deep vein thrombosis of left upper limb following coronary artery bypass graft
35615031	15708401000119100	Acute deep vein thrombosis of left upper limb following procedure
43531681	651000119108	Acute deep vein thrombosis of lower limb
35616027	293481000119109	Acute deep vein thrombosis of right iliac vein
35615034	15711241000119106	Acute deep vein thrombosis of right lower limb following procedure
761415	15708281000119109	Acute deep vein thrombosis of right upper limb following coronary artery bypass graft
35615030	15708201000119101	Acute deep vein thrombosis of right upper limb following procedure
44782746	132281000119108	Acute deep venous thrombosis
44782751	134961000119104	Acute deep venous thrombosis of axillary vein
762008	285321000119103	Acute deep venous thrombosis of bilateral axillary veins
760875	12237551000119104	Acute deep venous thrombosis of bilateral calves
765155	285441000119102	Acute deep venous thrombosis of bilateral ileofemoral veins
762017	285501000119103	Acute deep venous thrombosis of bilateral internal jugular veins
762417	350291000119100	Acute deep venous thrombosis of bilateral legs
762020	285561000119102	Acute deep venous thrombosis of bilateral popliteal veins
765546	285621000119106	Acute deep venous thrombosis of bilateral tibial veins
762004	285261000119104	Acute deep venous thrombosis of both upper extremities
44782742	132241000119103	Acute deep venous thrombosis of calf
44782747	132291000119106	Acute deep venous thrombosis of femoral vein
762015	285451000119100	Acute deep venous thrombosis of ileofemoral vein of left leg
765541	285461000119103	Acute deep venous thrombosis of ileofemoral vein of right lower extremity
44782748	132301000119107	Acute deep venous thrombosis of iliofemoral vein
44782752	135001000119100	Acute deep venous thrombosis of internal jugular vein
762009	285331000119100	Acute deep venous thrombosis of left axillary vein
760876	12237631000119109	Acute deep venous thrombosis of left calf
765540	285391000119101	Acute deep venous thrombosis of left femoral vein
765922	285511000119100	Acute deep venous thrombosis of left internal jugular vein

762418	350301000119104	Acute deep venous thrombosis of left lower extremity
765537	285271000119105	Acute deep venous thrombosis of left upper extremity
44782767	136781000119101	Acute deep venous thrombosis of lower extremity as complication of procedure
46270071	132111000119107	Acute deep venous thrombosis of lower limb due to and following coronary artery bypass grafting
762022	285581000119106	Acute deep venous thrombosis of politeal vein of right leg
44782743	132251000119101	Acute deep venous thrombosis of popliteal vein
762021	285571000119108	Acute deep venous thrombosis of popliteal vein of left leg
762010	285341000119109	Acute deep venous thrombosis of right axillary vein
760877	12237711000119106	Acute deep venous thrombosis of right calf
762013	285401000119104	Acute deep venous thrombosis of right femoral vein
762018	285521000119107	Acute deep venous thrombosis of right internal jugular vein
762419	350311000119101	Acute deep venous thrombosis of right lower extremity
762005	285281000119108	Acute deep venous thrombosis of right upper extremity
44782745	132271000119105	Acute deep venous thrombosis of thigh
44782744	132261000119104	Acute deep venous thrombosis of tibial vein
762026	285631000119109	Acute deep venous thrombosis of tibial vein of left leg
765156	285641000119100	Acute deep venous thrombosis of tibial vein of right leg
44782421	132321000119103	Acute deep venous thrombosis of upper extremity
764016	449691000124103	Acute deep venous thrombosis of upper extremity after coronary artery bypass graft
44782766	136771000119104	Acute deep venous thrombosis of upper extremity as complication of procedure
762048	286421000119101	Acute thrombosis of left subclavian vein
45757410	133421000119101	Acute thrombosis of mesenteric vein
762049	286431000119103	Acute thrombosis of right subclavian vein
36712892	143561000119108	Acute thrombosis of splenic vein
44782762	132611000119104	Acute thrombosis of subclavian vein
435887	49956009	Antepartum deep vein thrombosis
4179911	297156001	Axillary vein thrombosis
37109253	285381000119104	Bilateral acute deep vein thrombosis of femoral veins
40478951	444325005	Bilateral deep vein thrombosis of lower extremities
4042396	16750002	Deep thrombophlebitis
4046884	134399007	Deep vein thrombosis of leg related to air travel
3655221	860699005	Deep vein thrombosis of lower extremity due to intravenous drug use
4133004	128053003	Deep venous thrombosis
4181315	428781001	Deep venous thrombosis associated with coronary artery bypass graft

438820	56272000	Deep venous thrombosis in puerperium
45773536	703277001	Deep venous thrombosis of femoropopliteal vein
763942	448841000124100	Deep venous thrombosis of left lower extremity
761980	25820001000004100	Deep venous thrombosis of left upper extremity
443537	404223003	Deep venous thrombosis of lower extremity
4133975	128055005	Deep venous thrombosis of pelvic vein
40480555	443210003	Deep venous thrombosis of peroneal vein
4322565	427775006	Deep venous thrombosis of profunda femoris vein
763941	448831000124105	Deep venous thrombosis of right lower extremity
761928	20850001000004108	Deep venous thrombosis of right upper extremity
4207899	438785004	Deep venous thrombosis of tibial vein
4028057	128054009	Deep venous thrombosis of upper extremity
435565	195437003	Embolism and thrombosis of the vena cava
40481089	444816006	Embolism from thrombosis of vein of lower extremity
4119760	234044007	Iliofemoral deep vein thrombosis
4124856	234041004	Inferior mesenteric vein thrombosis
4096099	25114006	Phlebitis of deep veins of lower extremity
4281689	66923004	Phlegmasia alba dolens
4284538	66877004	Phlegmasia cerulea dolens
4309333	213220000	Postoperative deep vein thrombosis
46285905	978441000000108	Provoked deep vein thrombosis
46271900	710167004	Recurrent deep vein thrombosis
4033521	14534009	Splenic vein thrombosis
4055089	197001004	Superior mesenteric vein thrombosis
4230403	438646004	Thrombophlebitis of axillary vein
4069561	1748006	Thrombophlebitis of deep femoral vein
761831	16014391000119106	Thrombophlebitis of deep vein of bilateral lower limbs
761830	16014351000119101	Thrombophlebitis of deep vein of left lower limb
761808	16006271000119105	Thrombophlebitis of deep vein of left upper limb
761832	16014431000119101	Thrombophlebitis of deep vein of right lower limb
761809	16006311000119105	Thrombophlebitis of deep vein of right upper limb
4221821	40198004	Thrombophlebitis of deep veins of lower extremity
440750	95452006	Thrombophlebitis of deep veins of upper extremities
4176614	42861008	Thrombophlebitis of iliac vein

761821	16012151000119109	Thrombophlebitis of left deep femoral vein
761819	16012071000119101	Thrombophlebitis of left femoral vein
761820	16012111000119108	Thrombophlebitis of right deep femoral vein
761818	16011991000119109	Thrombophlebitis of right femoral vein
4110339	195412008	Thrombophlebitis of the anterior tibial vein
4111868	195425000	Thrombophlebitis of the common iliac vein
4110343	195427008	Thrombophlebitis of the external iliac vein
439314	195410000	Thrombophlebitis of the femoral vein
4109877	195426004	Thrombophlebitis of the internal iliac vein
4112171	195411001	Thrombophlebitis of the popliteal vein
4112172	195414009	Thrombophlebitis of the posterior tibial vein
4250765	7387004	Thrombophlebitis of tibial vein
42538533	762256003	Thrombosis of iliac vein
44811347	864191000000104	Thrombosis of internal jugular vein
765049	16730001000004104	Thrombosis of left peroneal vein
4317289	95446005	Thrombosis of mesenteric vein
4203836	438647008	Thrombosis of subclavian vein
4175649	427776007	Thrombosis of the popliteal vein
4149782	309735004	Thrombosis of vein of lower limb
4153353	371051005	Traumatic thrombosis of axillary vein
46285904	978421000000101	Unprovoked deep vein thrombosis
77310	266267005	Deep vein phlebitis and thrombophlebitis of the leg
4189004	413956008	Deep vein thrombosis of leg related to intravenous drug use

Table A2 Preliminary code list for pulmonary embolism (PE)

Concept Id	SNOMED code	Concept name
608954	15964661000119102	Acute cor pulmonale due to septic pulmonary embolism
4120091	233936003	Acute massive pulmonary embolism
45768439	706870000	Acute pulmonary embolism
45768888	707414004	Acute pulmonary thromboembolism
44782732	133971000119108	Chronic pulmonary embolism
45768887	707412000	Chronic pulmonary thromboembolism
45771016	707413005	Chronic pulmonary thromboembolism without pulmonary hypertension
4219469	82153002	Miscarriage with pulmonary embolism
4108681	194883006	Postoperative pulmonary embolus

4091708	280964006	Pulmonary air embolism
440417	59282003	Pulmonary embolism
37109911	723859005	Pulmonary embolism due to and following acute myocardial infarction
45757145	10759311000119104	Pulmonary embolism in childbirth
37016922	713078005	Pulmonary embolism on long-term anticoagulation therapy
43530605	1001000119102	Pulmonary embolism with pulmonary infarction
4119608	233938002	Pulmonary fat embolism
4253796	74315008	Pulmonary microemboli
45766471	703636009	Pulmonary oil microembolism
4121618	233935004	Pulmonary thromboembolism
4119610	233940007	Pulmonary tumor embolism
4236271	438773007	Recurrent pulmonary embolism
36713113	328511000119109	Saddle embolus of pulmonary artery
35615055	15964701000119109	Saddle embolus of pulmonary artery with acute cor pulmonale
40479606	441557008	Septic pulmonary embolism
4119607	233937007	Subacute massive pulmonary embolism
4119609	233939005	Subacute pulmonary fat embolism

Table A3 Preliminary code list for pelvic venous thrombosis (PVT)

Concept Id	SNOMED Code	Concept Name
762148	293471000119106	Acute deep vein thrombosis of bilateral iliac veins
35616028	293491000119107	Acute deep vein thrombosis of left iliac vein
35616027	293481000119109	Acute deep vein thrombosis of right iliac vein
765155	285441000119102	Acute deep venous thrombosis of bilateral ileofemoral veins
761461	15712201000119101	Acute deep venous thrombosis of bilateral pelvic veins
762015	285451000119100	Acute deep venous thrombosis of ileofemoral vein of left leg
765541	285461000119103	Acute deep venous thrombosis of ileofemoral vein of right lower extremity
44782748	132301000119107	Acute deep venous thrombosis of iliofemoral vein
761462	15712241000119104	Acute deep venous thrombosis of left pelvic vein
44782761	132601000119102	Acute deep venous thrombosis of pelvic vein
765229	15712281000119109	Acute deep venous thrombosis of right pelvic vein
608965	15968901000119104	Bilateral iliac vein thrombophlebitis
765152	293441000119104	Chronic deep vein thrombosis of bilateral iliac veins
35616026	293461000119100	Chronic deep vein thrombosis of left iliac vein

761439	15711001000119104	Chronic deep vein thrombosis of left pelvic vein
46271548	709687000	Chronic deep vein thrombosis of pelvic vein
35616025	293451000119102	Chronic deep vein thrombosis of right iliac vein
761441	15711081000119107	Chronic deep vein thrombosis of right pelvic vein
765542	285471000119109	Chronic deep venous thrombosis of bilateral ileofemoral veins
761440	15711041000119102	Chronic deep venous thrombosis of bilateral pelvic veins
44782740	132201000119100	Chronic deep venous thrombosis of iliofemoral vein
765543	285481000119107	Chronic deep venous thrombosis of left ileofemoral vein
762016	285491000119105	Chronic deep venous thrombosis of right ileofemoral vein
761013	132541000119101	Deep venous thrombosis of bilateral pelvic veins
4133975	128055005	Deep venous thrombosis of pelvic vein
4119760	234044007	Iliofemoral deep vein thrombosis
608964	15968861000119105	Left iliac vein thrombophlebitis
4158798	361278002	Mondor's phlebitis of the penis
4285751	67486009	Pelvic thrombophlebitis in puerperium
608963	15968821000119100	Right iliac vein thrombophlebitis
4176614	42861008	Thrombophlebitis of iliac vein
4317290	95449003	Thrombophlebitis of pelvic vein
4111868	195425000	Thrombophlebitis of the common iliac vein
4110343	195427008	Thrombophlebitis of the external iliac vein
4109877	195426004	Thrombophlebitis of the internal iliac vein
201045	26373009	Thrombosed external hemorrhoids
195294	75955007	Thrombosed hemorrhoids
608966	15969021000119101	Thrombosed internal hemorrhoid grade IV
201595	52931009	Thrombosed internal hemorrhoids
42538533	762256003	Thrombosis of iliac vein
606527	1145183007	Thrombosis of pampiniform plexus
4319327	95448006	Thrombosis of pelvic vein
4295878	76598006	Thrombosis of penile vein
762443	368351000119106	Thrombosis of superficial vein of penis

Table A4 Preliminary code list for splanchic vein thrombosis, including hepatic and extrahepatic (STV)

Concept Id	SNOMED Code	Concept Name
45757410	133421000119101	Acute thrombosis of mesenteric vein

36712892	143561000119108	Acute thrombosis of splenic vein
196715	82385007	Budd-Chiari syndrome
45757409	133411000119108	Chronic thrombosis of mesenteric vein
36712891	143551000119106	Chronic thrombosis of splenic vein
4301208	38739001	Hepatic vein thrombosis
4124856	234041004	Inferior mesenteric vein thrombosis
199837	17920008	Portal vein thrombosis
4033521	14534009	Splenic vein thrombosis
4055089	197001004	Superior mesenteric vein thrombosis
4318407	95447001	Thrombophlebitis of mesenteric vein
4317289	95446005	Thrombosis of mesenteric vein

Table A5 Preliminary code list for retinal vein thrombosis, including retinal central vein thrombosis (RVT)

Concept Id	Concept Code	Concept Name
4339013	232048009	Branch retinal vein occlusion with macular edema
4334248	232046008	Branch retinal vein occlusion with neovascularization
4199035	314000002	Branch retinal vein occlusion with no neovascularization
313761	68478007	Central retinal vein occlusion
4208221	312997008	Central retinal vein occlusion - ischemic
4339011	232040002	Central retinal vein occlusion - juvenile
4335591	232042005	Central retinal vein occlusion - juvenile with macular edema
4334888	232041003	Central retinal vein occlusion - juvenile with neovascularization
4208222	312998003	Central retinal vein occlusion - non-ischemic
4339010	232039004	Central retinal vein occlusion with macular edema
4334246	232038007	Central retinal vein occlusion with neovascularization
42535735	733325006	Combined occlusion by thrombus of retinal artery and retinal vein
4334247	232043000	Hemispheric retinal vein occlusion
4335592	232045007	Hemispheric retinal vein occlusion with macular edema
4336005	232044006	Hemispheric retinal vein occlusion with neovascularization
4216561	71719003	Thrombophlebitis of retinal vein
4187790	46085004	Thrombosis of retinal vein
312622	24596005	Venous retinal branch occlusion

Table A6 Preliminary code list for disseminated intravascular coagulation (DIC)

Concept Id	Concept Code	Concept Name
436093	67406007	Disseminated intravascular coagulation
4028488	13507004	Purpura fulminans

Appendix II

Preliminary code list for selected cancer types is attached as a standalone document

Appendix III: ENCePP checklist for study protocols

ENCePP Checklist for Study Protocols (Revision 4)

Study title: DARWIN EU® – Incidence rates of venous thromboembolic events in cancer patients

EU PAS Register® number: N/A
Study reference number (if applicable): P3-C3-005

Section 1: Milestones	**Yes**	**No**	**N/A**	**Section Number**
1. Does the protocol specify timelines for 1.1.1 Start of data collection 1.1.2 End of data collection 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.	X			5. Milestones, 8.2 Data Sources

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) 2.1.2 The objective(s) of the study? 2.1.3 The target population? (i.e. population or subgroup to whom the study results are intended to be generalised) 2.1.4 Which hypothesis(-es) is (are) to be tested? 2.1.5 If applicable, that there is no *a priori* hypothesis?	X			7. Research question and objectives 8. Research methods

Comments:

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)	X			8.1 Study type and Study Design

3.2	Does the protocol specify whether the study is based on primary, secondary or combined data collection?	X			8.2 Study Setting and Data Sources
3.3	Does the protocol specify measures of occurrence? (e.g., rate, risk, prevalence)	X			8.8 Analysis
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))	X			8.8 Analysis
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)			X	

Comments:

Section 4: Source and study populations	**Yes**	**No**	**N/A**	**Section Number**
4.1 Is the source population described?	X			8.5 Study Population
4.2 Is the planned study population defined in terms of:	X			
4.2.1 Study time period				8.3 Study Period
4.2.2 Age and sex				8.6.3. Other covariates
4.2.3 Country of origin				8.2 Study Setting and Data Sources
4.2.4 Disease/indication				8.6.1. Exposures
4.2.5 Duration of follow-up				8.4 Follow-up
4.3 Does the protocol define how the study population will be sampled from the source population? (e.g. event or inclusion/exclusion criteria)	X			8.5 Study Population with inclusion and exclusion criteria

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)			X	
5.2 Does the protocol address the validity of the exposure measurement? (e.g. precision, accuracy, use of validation sub-study)			X	
5.3 Is exposure categorised according to time windows?			X	
5.4 Is intensity of exposure addressed? (e.g. dose, duration)			X	

5.5 Is exposure categorised based on biological mechanism of action and taking into account the pharmacokinetics and pharmacodynamics of the drug?			X	
5.6 Is (are) (an) appropriate comparator(s) identified?			X	

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?	X			8.6.2. Outcomes
6.2 Does the protocol describe how the outcomes are defined and measured?	X			8.6.2. Outcomes
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)		X		
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)		X		

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)			X	
7.2 Does the protocol address selection bias? (e.g. healthy user/adherer bias)			X	
7.3 Does the protocol address information bias? (e.g. misclassification of exposure and outcomes, time-related bias)			X	

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)			X	

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)			X	
9.1.2 Outcomes? (e.g. clinical records, laboratory markers or values, claims data, self-report, patient interview including scales and questionnaires, vital statistics)	X			8.2 Study Setting and Data Sources
9.1.3 Covariates and other characteristics?	X			8.6.3. Other covariates
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)			X	
9.2.2 Outcomes? (e.g. date of occurrence, multiple event, severity measures related to event)			X	
9.2.3 Covariates and other characteristics? (e.g. age, sex, clinical and drug use history, co-morbidity, co-medications, lifestyle)	X			8.2 Study Setting and Data Sources
9.3 Is a coding system described for:				
9.3.1 Exposure? (e.g. WHO Drug Dictionary, Anatomical Therapeutic Chemical (ATC) Classification System)			X	
9.3.2 Outcomes? (e.g. International Classification of Diseases (ICD), Medical Dictionary for Regulatory Activities (MedDRA))	X			8.6.2. Outcomes
9.3.3 Covariates and other characteristics?	X			8.6.3. Other covariates
9.4 Is a linkage method between data sources described? (e.g. based on a unique identifier or other)			X	

Comments:

Section 10: Analysis plan	**Yes**	**No**	**N/A**	**Section Number**
10.1 Are the statistical methods and the reason for their choice described?	X			8.8 Analysis
10.2 Is study size and/or statistical precision estimated?			X	
10.3 Are descriptive analyses included?	X			8.8.2 Descriptive statistics
10.4 Are stratified analyses included?	X			8.8 Analysis
10.5 Does the plan describe methods for analytic control of confounding?			X	
10.6 Does the plan describe methods for analytic control of outcome misclassification?			X	
10.7 Does the plan describe methods for handling missing data?			X	
10.8 Are relevant sensitivity analyses described?			X	

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)	X			9. Data management
11.2 Are methods of quality assurance described?	X			10. Quality Control
11.3 Is there a system in place for independent review of study results?			X	

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? 12.1.2 Information bias? 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).	X			11. Limitations of the research 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)	X			Table 8.2. Description of the selected Data Sources.

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?	X			13. Governance board aspects
13.2 Has any outcome of an ethical review procedure been addressed?			X	
13.3 Have data protection requirements been described?	X			9.2 Data storage and protection

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?	X			4. Amendments and updates

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)?	X			14. Plans for disseminating and communicating study results
15.2 Are plans described for disseminating study results externally, including publication?	X			14. Plans for disseminating and communicating study results

Comments:

Name of the main author of the protocol: Talita Duarte-Salles & Anton Barchuk

Date: 30/10/2024

Signature: _____