Study Protocol

P5-C1-007

DARWIN EU® - Cancer occurrence during pregnancy

31/07/2026

Version 3.0

Authors: Berta Raventós, Rana Jajou, Talita Duarte-Salles, Cindy Porras, Natasha Yefimenko, Ger Inberg, Arianna Alamshahi, Ross Williams

Public

CONTENTS

Study title	DARWIN EU® - Cancer occurrence during pregnancy
Protocol version	V3.0
Date	31/07/2026
EUPAS number	EUPAS1000001059
Active Substance	Not applicable
Medicinal Product	Not applicable
Research question and objectives	The aim of this study is to describe the occurrence of cancer during pregnancy and to describe the types of cancer diagnosed, with and without inclusion of the 12 months following pregnancy end. The specific objectives of this study are: 1. To estimate the occurrence of cancer diagnosed during pregnancy, with and without inclusion of the 12 months following pregnancy end. 2. To characterise all cancers diagnosed during pregnancy, with and without inclusion of the 12 months following pregnancy end.
Countries of study	Norway, Spain, the United Kingdom.
Authors	Berta Raventós (b.raventos@darwin-eu.org) Rana Jajou (r.jajou@darwin-eu.org) Talita Duarte-Salles (t.duarte@darwin-eu.org) Cindy Porras (c.porras@darwin-eu.org) Natasha Yefimenko (n.yefimenkonosova@darwin-eu.org) Ger Inberg (g.inberg@darwin-eu.org) Arianna Alamshahi (a.alamshahi@darwin-eu.org) Ross Williams (r.williams@darwin-eu.org)

LIST OF ABBREVIATIONS

Acronyms/terms	Description
BRCA1	BReast CAncer gene 1
BRCA2	BReast CAncer gene 2
CC	Coordination Centre
CDM	Common Data Model
CI	Confidence Interval
CPRD GOLD	Clinical Practice Research Datalink GOLD
DARWIN EU®	Data Analysis and Real-World Interrogation Network
DRE	Digital Research Environment
DTZ	Data Transfer Zone
EHR	Electronic Health Records
EMA	European Medicines Agency
ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
EU	European Union
EUPAS	EU Post-Authorisation Studies Register
GDPR	General Data Protection Regulation
GP	General Practitioner
ICD	International Classification of Diseases
IDIAPJGol	Institut d'Investigació en Atenció Primària de Salut Jordi Gol
IQR	Interquartile Range
IRB	Institutional Review Board
MBRN	Medical Birth Registry of Norway
NLHR	Norwegian Linked Health Registry data
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
Oxford NDORMS	University of Oxford - Nuffield Department of Orthopeadics, Rheumatology, and Musculoskeletal Sciences (Oxford NDORMS)
PET	Perinatal Expansion Table
RWE	Real-World Evidence
SNOMED	Systematized Nomenclature of Medicine
SIDIAP	The Information System for Research in Primary Care
UiO	University of Oslo

1. TITLE

DARWIN EU® - Cancer occurrence during pregnancy

2. DESCRIPTION OF THE STUDY TEAM

Study team role	Names	Organisation
Principal Investigator	Berta Raventós	Erasmus MC
Co-Principal Investigators	Rana Jajou Talita Duarte-Salles	Erasmus MC
Epidemiologist	Cindy Porras[1]	Erasmus MC
Data Scientists	Ger Inberg Arianna Alamshahi Cesar Barboza Maarten van Kessel Sumiya Abdirashid Cristiana Di Tullio Ross Williams	Erasmus MC
Study Manager	Natasha Yefimenko	Erasmus MC
Data source	**Names**	**Data Partner Organisation***
Norwegian Linked Health Registry data (NLHR)	Saeed Hayati Petrica Ioan Olteanu Hedvig Nordeng	University of Oslo (UiO)
The Information System for Research in Primary Care (SIDIAP)	Agustina Giuliodori Elena Roel Laura Granés Irene López	Institut d'Investigació en Atenció Primària de Salut Jordi Gol (IDIAPJGol)
Clinical Practice Research Datalink GOLD (CPRD GOLD)	Antonella Delmestri	University of Oxford - Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences (Oxford NDORMS)

*Data partners do not have an investigator role. Data partners execute code at their data source, review, and approve their results.
[1]Added to the team on 8 July 2026.

3. ABSTRACT

DARWIN EU® - Cancer occurrence during pregnancy

Rationale and background

Few studies have examined the incidence of cancer during pregnancy, and existing evidence from studies conducted in Europe is limited. Real-World Evidence (RWE) regarding the occurrence of cancer during pregnancy can help address this gap by providing more recent estimates from different European countries, informing future research priorities, and contributing to evidence-based decision-making in this vulnerable population.

Objectives

The aim of this study is to describe the occurrence of cancer during pregnancy and to describe the types of cancer diagnosed, with and without inclusion of the 12 months following pregnancy end.

The specific objectives of this study are:

1. To estimate the occurrence of cancer diagnosed during pregnancy, with and without inclusion of the 12 months following pregnancy end.

2. To characterise all cancers diagnosed during pregnancy, with and without inclusion of the 12 months following pregnancy end.

Methods

Study design

- A disease epidemiology study (objective 1).
- A characterisation study (objective 2).

Individuals will be followed up from pregnancy start date (index date) to the earliest of pregnancy end, outcome occurrence, loss to follow-up, death, or end of the observation period (latest available data), whichever occurs first. The end of the post-partum period (i.e. 12 months following pregnancy end) will also be considered as an endpoint for incidence calculations (both analyses will be conducted separately). For objective 2, the index date will be defined as the date of cancer diagnosis.

Population

All pregnancies starting from 01/01/2010 (or start of data availability, if later) to one year before the end of data availability occurring in females aged 12 to 55 years at the index date. Pregnancies with implausible duration will be excluded. For incidence calculations, pregnancies with history of cancer (except for non-melanoma skin cancer) in the 5 years prior to the index date will also be excluded.

Data sources

Norway: Norwegian Linked Health Registry data (NLHR); Spain: The Information System for Research in Primary Care (SIDIAP); The United Kingdom: Clinical Practice Research Datalink GOLD (CPRD GOLD)

Statistical analysis

Two estimands (occurrence of new, relapsed, or recurrent cancer diagnoses; incidence of new cancers) will be calculated in all pregnancies recorded in the PET and will be expressed as number of cancers per 100,000 pregnancies over the entire study period. Cancer will be defined as any malignant cancer (excluding non-melanoma skin cancer). Both estimands will be calculated with and without the 12-month post-partum period following pregnancy end.

Individuals diagnosed with cancer during pregnancy (including new, relapsed, and recurrent disease) will be described at the date of cancer diagnosis, in terms of type of cancer diagnosed (i.e. based on a pre-specified list of cancer types), and pregnancy outcome (e.g. live birth, still birth, miscarriage), pregnancy trimester at time of cancer diagnosis, previous history of cancer, gravidity (i.e. first versus subsequent pregnancy), and birth term status (i.e. pre-term versus term pregnancy). A sensitivity analysis will be performed, restricting pregnancies to those resulting in live birth.

4. AMENDMENTS AND UPDATES

None.

5. MILESTONES

Study milestones and deliverables	Planned dates*
Final Study Protocol	31 July 2026
Creation of Analytical code	August 2026
Execution of Analytical Code on the data	August–September 2026
Draft Study Report	12 October 2026
Final Study Report	To be confirmed by EMA

*Planned dates are dependent on obtaining approvals from the internal review boards of the data sources and deliverables review rounds.

6. RATIONALE AND BACKGROUND

Pregnancy-associated cancer is usually defined as cancer diagnosed during pregnancy and up to one year after the end of pregnancy.[1-5] The rationale for including the post-partum period is the potential for delayed diagnosis, as cancer-related symptoms may be attributed to physiological changes of pregnancy, diagnostic evaluation may be deferred during pregnancy, and less aggressive tumours may remain undiagnosed until after delivery.[1, 3] Prior studies suggest that cancers diagnosed in the post-partum period account for a substantial proportion of cancers and have been reported to exceed 64% of cases.[1-3, 6]

The incidence of pregnancy-related cancer varies across studies, likely due to differences in the study period, settings, and pregnancies included (i.e. all pregnancies, only pregnancies resulting in birth, or only pregnancies resulting in live birth). Crude incidence of pregnancy-associated cancer was estimated at 89.6 per 100,000 pregnancies including live births and induced abortions in Denmark (1997–2006),[5] and 122.9 per 100,000 pregnancies resulting in birth or abortion in Lombardy, Italy (2001–2012).[6] For pregnancies resulting in birth, incidence of pregnancy-associated cancer in Sweden (1973–2017) was estimated at 28 per 100,000 during pregnancy and 73.6 per 100,000 in the year after delivery. In studies conducted outside Europe, pregnancy-associated cancer incidence was estimated at 94 per 100,000 deliveries in California (1991–1999),[1] 137.3 per 100,000 deliveries in Australia (1994–2008),[3] and 153.1 and 146.5 per 100,000 deliveries (2003–2015) in two Canadian provinces, Alberta and Ontario, respectively.[2]

A statistically significant increase in crude incidence of pregnancy-associated cancer over time has been reported in studies from Denmark, Australia, and Norway.[3-5] In Denmark and Australia, this increase remained statistically significant in age-standardised rates, suggesting that the increasing trend in incidence cannot be explained by delayed childbearing and increasing maternal age alone.[3, 5] However, temporal increases have not been reported consistently. Canadian findings varied by province, showing a significant upward trend in Ontario but no temporal trend in Alberta, and no significant temporal trend was observed in Italy.[2, 6]

The most frequently recorded cancer types differed by study, but usually included melanoma of the skin and breast cancer.[1, 3-7] Thyroid cancer,[1, 3, 6] cervical cancers,[1, 4, 5, 7] and lymphomas [1, 6] were also reported amongst the most frequent.[1, 3-6] In studies conducted in Denmark and Sweden, the most frequent pregnancy-associated cancers were melanoma, cervical cancer, and breast cancer, whereas in an Italian study, breast cancer, thyroid cancer, and lymphomas were the most common.[5-7]

Cancer occurring during pregnancy poses specific challenges, as cancer treatments can directly impact maternal safety, foetal development, and pregnancy outcomes. Previous studies have reported higher rates of preterm birth among patients with cancer during pregnancy, with 48% of these live birth pregnancies occurring preterm.[8] As managing an oncological condition alongside a pregnancy is highly complex, clinicians, patients, and regulators face significant uncertainty when trying to balance effective maternal therapy with foetal protection. From a clinician's point of view, not only is cancer incidence (first-ever diagnosis) important, but every case where an individual is told, during pregnancy, that they and their doctor must make a treatment decision, irrespective of whether this concerns a new (incident) cancer, a relapse, or a recurrence.

To the best of our knowledge, very few studies have examined the incidence of pregnancy-associated cancer in European countries.[5, 6] Real-World Evidence (RWE) regarding the precise incidence of cancer during pregnancy can help address this gap by providing more recent estimates from different European countries, informing future research priorities, and contributing to evidence-based decision-making in this vulnerable population.

7. RESEARCH QUESTION AND OBJECTIVES

Research questions

How many females are diagnosed with cancer during pregnancy, and what types of cancer are diagnosed?

Research objectives

The aim of this study is to describe the occurrence of cancer during pregnancy and to describe the types of cancer diagnosed, with and without inclusion of the 12 months following pregnancy end.

The specific objectives of this study are:

1. To estimate the occurrence of cancer diagnosed during pregnancy, with and without inclusion of the 12 months following pregnancy end.
2. To characterise all cancers diagnosed during pregnancy, with and without inclusion of the 12 months following pregnancy end.

8. RESEARCH METHODS

8.1. Study design

A cohort study will be conducted using routinely collected health data from 3 data sources from 3 European countries. The study will comprise:

- A disease epidemiology study to address objective 1, assessing occurrence of cancer during pregnancy;
- A characterisation study to address objective 2, describing cancers diagnosed during pregnancy.

Both objectives will be addressed with and without the 12 months following pregnancy end.

This study will use the Observational Medical Outcomes Partnership (OMOP) Perinatal Expansion Table (PET), which provides a standardised structure for perinatal data.[9] Generated during the mapping of source data to the standardised format required by the OMOP Common Data Model (CDM), the PET has been implemented by several data sources within the DARWIN EU® Data Network and has been used in previous DARWIN EU® studies, including a study on benzodiazepine use during pregnancy (EUPAS1000000536), and the DARWIN EU® PeriNet study (EUPAS1000000812; Objective 4).

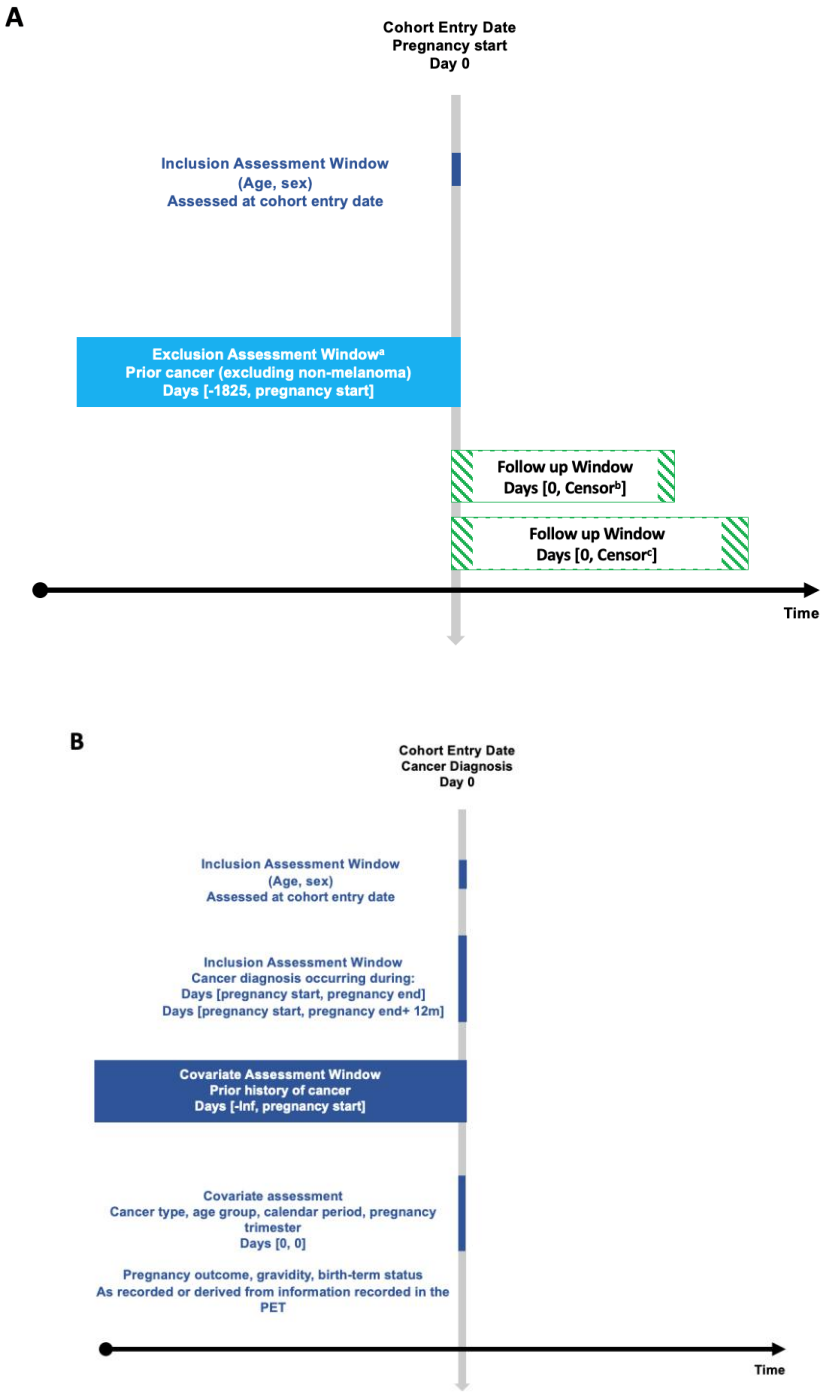

Figure 1. Graphical depiction of the study design for Objective 1 (A) and 2 (B).

a. The exclusion of pregnancies with a prior history of cancer (except for non-melanoma skin cancer), 5 years prior to the index date, will only be applied for incidence calculations.
b. End of follow-up will be defined as the earliest of: end of pregnancy, outcome occurrence, loss to follow-up, death, or end of the observation period (latest available data), whichever occurs first.
End of follow-up will be defined as the earliest of: end of the 12-month period following pregnancy end, outcome occurrence, loss to follow-up, death, or end of the observation period (latest available data), whichever occurs first.

Inf=All prior history; PET=Perinatal Expansion Table.

8.2. Follow-up

Analyses will be conducted at the pregnancy episode level. Individuals will be allowed to contribute with multiple entries, corresponding to each pregnancy recorded during the study period, provided they meet the criteria specified in **Section 8.3**.

For Objective 1, follow-up will start on the pregnancy start date (**Figure 1A**). According to the PET specifications, this date is based on ultrasound estimates or, if ultrasound information is unavailable, calculated from the last menstrual period (see **Section 8.4** for additional information).[8] End of follow-up will be defined as the earliest of pregnancy end, outcome occurrence, loss to follow-up, death, or end of the observation period (latest available data), whichever occurs first. The end of the post-partum period (i.e. 12 months following pregnancy end) will also be considered as an endpoint for Objective 1 (both analyses will be conducted separately). If an individual has two pregnancy episodes separated by less than 12 months, follow-up for the first pregnancy will be censored at the start of the second pregnancy, and follow-up will then begin for the second pregnancy.

For Objective 2, the index date will be defined at the date of cancer diagnosis, provided that this occurs during pregnancy or during the 12 months following pregnancy end (**Figure 1B**). No follow-up will be performed.

8.3. Study population with inclusion and exclusion criteria

All recorded pregnancy episodes of females aged 12 to 55 years at pregnancy start (i.e. index date) will be considered for inclusion. As described in **Section 8.2**, individuals will be allowed to contribute with multiple entries, corresponding to each pregnancy recorded during the study period, provided that the pregnancies meet the criteria specified below:

Inclusion criteria

- Pregnancies starting in the period of 01/01/2010 (or start of data availability, if later) until one year before the end of data availability.
- Females aged 12 to 55 years, at the index date.

Exclusion criteria:

- For incidence calculations only: Pregnancies with a prior history of cancer (except for non-melanoma skin cancer), 5 years prior to the index date.

Some prior quality checks will be performed to ensure the plausibility and observability of the pregnancies recorded. Pregnancies with the following characteristics will be excluded:

- Pregnancies among individuals not observed in the data source at the pregnancy start or end date.
- Pregnancies with implausible duration (i.e. pregnancy end date < pregnancy start date, <1 day, or > 44 weeks of gestation)
- Pregnancies with conflicting dates (i.e. distinct pregnancies with overlapping dates)

8.4. Study setting and data sources

This study will be conducted using routinely collected data from 3 data sources within the DARWIN EU® Data Network from 3 European countries. All data were *a priori* mapped to the OMOP CDM,[10-12] and have already implemented the PET.

Table 1. Data sources.

Country	Name of Data source	Health Care setting	Type of Data	Number of active individuals[1]	Calendar period covered by each data source[2,3]	Contributing to
Norway	Norwegian Linked Health Registry data (NLHR)	Primary care, Secondary care (inpatient and outpatient)	Registries	5.55M	2008–2023	All objectives
Spain	The Information System for Research in Primary Care (SIDIAP)	Primary care, inpatient care	EHR	6.07M	2006–2024	All objectives
The United Kingdom	Clinical Practice Research Datalink GOLD (CPRD GOLD)	Primary care	EHR	2.37M	1987–2025	All objectives

EHR=Electronic health records.
[1]Defined as the maximum number of individuals in observation in the last 6 months of data.
[2]For NLHR, different registries that comprise each data source have different start dates. Primary care data and birth registry data start in 2008, while secondary care data, drug, vaccination, and infectious disease registry start in 2018.
[3]Current data releases include data up to: December 2023 (NLHR); December 2024 (SIDIAP); and November 2025 (CPRD GOLD).

Data sources selection

These data sources fulfil the criteria required in terms of data quality, completeness, timeliness, and representativeness while covering different regions of Europe (**Annex II**).

All the included data sources have pregnancy-related data mapped to the PET, which will be used in the context of this study. All data sources included have previously participated, or are currently participating, in multi-data source studies on pregnancy using the PET,[13-15] including the DARWIN EU® PeriNet study.

Importantly, NLHR populated the PET using data from the Medical Birth Registry of Norway (MBRN) with mandatory reporting of pregnancies starting from gestational week 12. Pregnancies ending before week 12 are identified based on a pregnancy algorithm developed by the University of Oslo (UiO), which uses data from primary and secondary care to complement information from the birth registry.[16] The UiO algorithm is applied to the source data prior to mapping to the OMOP CDM. Therefore, pregnancies included in the PET in NLHR will comprise those identified from the birth registry and those identified through the UiO algorithm. In SIDIAP and CPRD GOLD, the PET is derived from primary care data, including reproductive health clinics in primary care in SIDIAP. In CPRD GOLD, pregnancy episodes are inferred based on the algorithm developed by Matcho et al.[17]

For SIDIAP, linkage to hospital discharge data will be requested based on findings from a previous validation study of cancer diagnoses conducted within this data source, which showed that incorporating these data increased sensitivity for the identification of all cancers.[18]

Pregnancy start data

In NLHR, pregnancy start data differs depending on the length of the pregnancy. If the pregnancy length is over 12 weeks and the pregnancy is recorded in the MBRN, then the gestational length in days is primarily calculated based on the ultrasound estimated due date. If information from an ultrasound is not available, the gestational length is calculated based on the first day of the last menstrual period. For induced abortions, a gestational length based on the size of the uterus is also used. If pregnancy length is under 12 weeks and the pregnancy is derived from the UiO algorithm, then the pregnancy start is based on 45 International Classification of Diseases (ICD) 10th revision codes, which are used as markers. For individuals for whom no pregnancy start markers were available from the primary and secondary care registries, outcome-specific estimates of gestational age are used: 8 weeks (i.e. 56 days) for elective terminations and ectopic pregnancies, 10 weeks (i.e. 70 days) for miscarriages, and 11 weeks for molar pregnancies (i.e. 77 days). Gestational age for 66.3% of pregnancies ending in miscarriage and 47.2% of pregnancies ending in elective termination is inferred based on pregnancy markers. For the remaining pregnancies ending before 12 weeks of gestation (around 34% of miscarriages and 53% of elective terminations), gestational age is assigned using pre-specified durations of 10 and 8 weeks, respectively.[16]

In SIDIAP, pregnancy start corresponds to the date of the last menstrual period which is, in most cases, corrected based on ultrasound and recorded in the electronic health record.

In CPRD GOLD, pregnancy start corresponds mainly with the date of the last menstrual period or with the date of an antenatal visit in primary care.

8.5. Study period

The study period will be from 01/01/2010 (or data availability start, if later) to 31/12/2025 (or data availability end, if earlier).

8.6. Variables

8.6.1. Exposure

Not applicable.

8.6.2. Outcome

Objective 1:

Two outcomes will be used for this objective.

- Any cancer diagnosis. This will be any malignant cancer (excluding non-melanoma skin cancer), including new, relapsed, and recurrent disease.
- Diagnosis of a new cancer (excluding non-melanoma skin cancer). This will be a subset of the first, for which a washout period of 5 years will be used to exclude previous cancer cases (see **Section 8.3**).

Objective 2:

No outcome will be assessed as part of Objective 2.

Concept sets used for the identification of outcomes will be developed during the study execution following the standard DARWIN EU® phenotyping processes, which involve the review of code lists by clinical experts and the review of phenotypes after their execution in the participating data sources. Outcomes will be identified as the presence of records indicating the disease. Condition records will be identified using Systematized Nomenclature of Medicine (SNOMED) codes. For the operational definition, the outcome will be assigned a duration of up to 1 year, unless the patient's observation period ends earlier, in which case the cohort end date will be defined as the end of observation.

8.6.3. Covariates and other variables

Objective 1:

Covariates for this objective will only include age groups (12–34; 35–55 years), which will be defined at the index date (i.e. date of pregnancy start) and will be used for stratification.

Objective 2:

Covariates for Objective 2 will include:

- Cancer types, described at the index date (i.e. date of cancer diagnosis) based on the list of cancer types provided below:
    - Breast cancer
    - Melanoma skin cancer
    - Thyroid cancer
    - Cervical cancer
    - Ovarian cancer
    - Lymphoma
    - Other malignancies (i.e. all malignancies excluding those listed above, and non-melanoma skin cancer)
- Prior history of cancer (i.e. any malignant cancer, excluding non-melanoma skin cancer), described at index date (i.e. date of cancer diagnosis)
- Pregnancy trimester, described at index date (i.e. date of cancer diagnosis):
    - First trimester, defined in days, as: [pregnancy start, pregnancy start +90]
    - Second trimester, defined in days, as: [pregnancy start+91, pregnancy start +180]
    - First and second trimester (combined), defined in days, as: [pregnancy start, pregnancy start +180]
    - Third trimester, defined in days, as: [pregnancy start+181, pregnancy end]
    - Post-partum period (for analyses including the 12-month period after pregnancy end), defined in days, as: [pregnancy end, pregnancy end +365]
- Gravidity (i.e. number of previous pregnancies regardless of gestational length), described at pregnancy start date. This variable will be described based on information available in the PET, and will be assessed as a categorical variable with the following groups:
    - First pregnancy (gravidity of 0)
    - Subsequent pregnancies (gravidity ≥ 1)
- Pregnancy outcome, described at pregnancy end. This variable will be described based on information available in the PET, and will include:
    - Live birth
    - Miscarriage (<20 weeks)
    - Stillbirth (≥ 20 weeks)
    - Elective termination of pregnancy

  - Discordant (different outcomes in multiple pregnancies)
  - Unknown
- Birth term status, described at pregnancy end. This variable will be described based on pregnancy start and end dates recorded in the PET, and will comprise of:
  - Pre-term birth: Live births ending between 20 and 36 weeks of gestation (equivalent to 140–258 days after pregnancy start)

    Term birth: Live births ending at ≥37 weeks of gestation (equivalent to ≥259 days after pregnancy start)

Covariates for stratification for this objective are as follows:

- Age groups (12–34; 35–55 years), defined at the index date.
- Calendar period (2010–2014; 2015–2019; 2020–end of data), defined at the index date.

As described in **Section 8.2**, concept sets used for the identification of outcomes will be developed during the study execution following the standard DARWIN EU® phenotyping process.

8.7. Study size

No sample size has been calculated, as this is a descriptive disease epidemiology and characterisation study which will not test a specific hypothesis. The sample size will therefore be determined by the availability of data on recorded pregnancy episodes and conditions of interest.

Based on a preliminary feasibility assessment using characterisation results from the DARWIN EU® PeriNet study, the expected number of pregnancies in each participating data source will range from 360,000 to 460,000 (**Table S1**). The number of pregnancies captured in NLHR is expected to be higher, as the previous study restricted the study period for this data source from 2018 onwards. The proportion of pregnancies resulting in live birth was estimated at approximately 70% in NLHR and CPRD GOLD, and 90.4% in SIDIAP (see **Table S1**, and sensitivity analyses described in **Section 8.8**). Pregnancy episodes with a diagnosis of cancer recorded during pregnancy were rare, with >5 counts observed for only a few cancer types. In NLHR, approximately 200 cases were observed for breast cancer; 100–200 for gynaecological cancers, melanoma, and thyroid cancer; 50–100 for bone and central nervous system cancers; and <20 for digestive, genitourinary, head and neck, haematological, and respiratory cancers. In SIDIAP and CPRD GOLD, the cancer types with >5 counts during pregnancy were breast cancer (<50), thyroid cancer (<50), melanoma (<30), and gynaecological cancers (in CPRD GOLD only; <30).

8.8. Analysis

8.8.1. Federated network analyses

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

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

8.8.2. Data privacy protection

The data partners will locally execute the analytics against the OMOP CDM in R Studio and review and approve the default aggregated results. They will then be made available to the Principal Investigators and study team in secure online repository of DTZ (Data Transfer Zone). All results will be locked and

timestamped for reproducibility and transparency. The study results of all data sources will be checked, after which they are made available to the team, and the Study Dissemination Phase can start. All analyses will be conducted separately for each data source, and will be carried out in a federated manner, allowing analyses to be run locally without sharing patient-level data. Cell counts <5 will be suppressed when reporting results to comply with the data source's privacy protection regulations.

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

All analyses will be calculated based on OMOP CDM mapped data and will use the R package *CohortCharacteristics,* developed by DARWIN EU®.[19]

Objective 1

Two estimands (occurrence of new, relapsed, or recurrent cancer diagnoses; incidence of new cancers) will be calculated in all pregnancies recorded in the PET and will be calculated as number of cancers per 100,000 pregnancies, with and without the 12-month period following pregnancy end (both analyses will be conducted separately). Both estimands will be calculated over the entire study period and will be stratified by age group (see **Section 8.6.3**).

The 95% Confidence Intervals (CI) will be calculated assuming that the number of cancers follow a Poisson distribution.

Objective 2

Individuals diagnosed with cancer during pregnancy (including new, relapsed, or recurrent disease), with and without inclusion of the 12 months following pregnancy end, will be described using counts and percentages for categorical variables, and medians with interquartile ranges (IQR) for continuous variables.

The list of covariates that will be described is specified in **Section 8.6.3** and consists of: the type of cancer diagnosed, the pregnancy trimester in which it was diagnosed, and prior history of cancer. Gravidity, birth term status, and the outcome of the pregnancy in which the cancer was diagnosed will also be described. In addition, a basic characterisation in terms of demographic characteristics will be performed, including age at the index date (as a continuous variable) and time in observation before and after the index date (see **Section 8.8.4**).

Results will be stratified by age group and calendar period, as described in **Section 8.6.3**.

Sensitivity analysis

We will conduct a sensitivity analysis by restricting pregnancies included to those resulting in live birth (**Table 2**). Based on the DARWIN EU® PeriNet study (Objective 4), the proportion of pregnancies resulting in live birth was estimated at approximately 70% in NLHR and CPRD GOLD, and 90.4% in SIDIAP, while the proportion of pregnancy losses (including both miscarriage and elective termination of pregnancy) was approximately 29% in NLHR and CPRD GOLD, and 8.8% in SIDIAP. The proportion of pregnancies ending in stillbirth in participating data sources as estimated in DARWIN EU® PeriNet is ≤0.6% (**Table S1**).

Given the differential capture of pregnancies resulting in early gestational outcomes across data sources, this sensitivity analysis will help assess incidence of cancer in a more clearly defined denominator, which will support the interpretation of findings and facilitates comparison with the literature.

Table 2. Sensitivity analyses – rationale, strengths, and limitations.

	What is being varied? How?	Why?	Strengths of the sensitivity analysis compared to the primary	Limitations of the sensitivity analysis compared to the primary
Add live birth restriction	Pregnancies will be required to result in live birth to be included in the study.	Early pregnancy loss is not consistently captured across data sources. This analysis will help assess incidence of cancer in a more clearly defined denominator, which will be consistent across data sources.	To support the interpretation of findings and facilitate comparison with the literature.	This analysis will restrict the number of pregnancies included, particularly in NLHR and CPRD GOLD. This will likely reduce the number of cases captured, which may limit the feasibility of producing stratified results.

NLHR=Norwegian Linked Health Registry data; CPRD GOLD=Clinical Practice Research Datalink GOLD.

All analyses

Cell counts <5 will be suppressed when reporting results to comply with the data source's privacy protection regulations.

Regarding missing data, analyses for all objectives will be conducted using available data without imputation. As such, no specific methods will be applied to address missingness, and the absence of a recorded entry (e.g. condition) will be interpreted as the individual not having the event of interest (e.g. outcome).

8.8.4. Output

Output will include a PDF report, including an executive summary and the following tables and figures.

- Table 1. Number of pregnancies included in the study, by data source.
- Table 2. Occurrence of cancer (including new, relapsed, and recurrent) during pregnancy, with and without the 12-month period following pregnancy end, by data source and age group.
- Table 3. Incidence of new cancer during pregnancy with and without the 12-month period following pregnancy end, by data source and age group.
- Table 4. Characterisation of cancers (new, relapsed, or recurrent) diagnosed during pregnancy (excluding the 12-month period following pregnancy end), by data source.
- Table 5. Characteristics of pregnancies with cancer (new, relapsed, or recurrent) diagnosed during pregnancy (excluding the 12-month period following pregnancy end), by data source.
- Table 6. Characterisation of cancers (new, relapsed, or recurrent) diagnosed during pregnancy (excluding the 12-month period following pregnancy end), by data source and age group.
- Table 7. Characterisation of cancers (new, relapsed, or recurrent) diagnosed during pregnancy (excluding the 12-month period following pregnancy end), by data source and calendar period.
- Table 8. Occurrence of cancer (including new, relapsed, and recurrent) during pregnancy resulting in live birth, with and without the 12-month period following pregnancy end, by data source (sensitivity analysis).
- Table 9. Characterisation of cancers (new, relapsed, or recurrent) diagnosed during pregnancy (excluding the 12-month period following pregnancy end) resulting in live birth, by data source (sensitivity analysis)

An interactive dashboard (i.e. Shiny app) will be generated by incorporating all the results (tables and figures). Main results will be presented in the PDF report, while all other outputs will be made available through the Shiny app.

Mock tables are included below. Tables 6 and 7 will follow the same structure of **Mock Table 4**, presented stratified by age group and by calendar period, respectively. Tables 4 to 7 include characterisation results excluding the 12-month period following pregnancy end. Results obtained including the 12-month period after pregnancy end will be made available in the Annex or through the Shiny app, depending on relevance.

The current list of tables and figures may be adapted during the report writing to include additional elements that support the interpretation of findings. Depending on relevance, some of the tables of the main analyses (i.e. stratified results, depending on counts) will be placed in the Annex. Results from the sensitivity analyses will be presented as overall estimates in the report, with stratified results available in the Shiny app. Tables 8 and 9 will follow the same structure as **Mock Table 2** (reporting results per 100,000 live births instead of pregnancies) and **Mock Table 4**, respectively.

Mock Table 1. Number of pregnancies included in the study, by data source.

Data source[1]	Records (N)	Subjects (N)	Excluded records (N)	Excluded subjects (N)
Qualifying initial records[2]				
In observation at pregnancy start and end				
Pregnancies with plausible durations				
Pregnancies with conflicting dates				
Pregnancy start between 01/01/2010 and 1 year before end of data availability				
Age 12 to 55 at index date				
Female				

N=Number.
[1]Attrition will be reported for all participating data sources.
[2]Pregnancy episodes recorded in the Perinatal Expansion Table (PET).

Mock Table 2. Occurrence of cancer (including new, relapsed, and recurrent disease) during pregnancy with and without the 12-month period following pregnancy end, by data source and age group.

Data source	Age group (in years)	Number of pregnancies	Number of cancer diagnoses	Cases per 100,000 pregnancies
Time period at risk: Pregnancy				
Data source 1	Overall			
	12–34			
	35–55			
Data source 2	Overall			

Data source	Age group (in years)	Number of pregnancies	Number of cancer diagnoses	Cases per 100,000 pregnancies
	12–34			
	35–55			
Data source 3	Overall			
	12–34			
	35–55			
Time period at risk: Pregnancy +12m after				
Data source 1	Overall			
	12–34			
	35–55			
Data source 2	Overall			
	12–34			
	35–55			
Data source 3	Overall			
	12–34			
	35–55			

Mock Table 3. Incidence of new cancer during pregnancy, with and without the 12-month period following pregnancy end, by data source and by age group.

Data source	Age group (in years)	Number of pregnancies	Number of cancer diagnoses	Incidence per 100,000 pregnancies
Time period at risk: Pregnancy				
Data source 1	Overall			
	12–34			
	35–55			
Data source 2	Overall			
	12–34			
	35–55			
Data source 3	Overall			
	12–34			
	35–55			
Time period at risk: Pregnancy +12m after				
Data source 1	Overall			
	12–34			
	35–55			
Data source 2	Overall			
	12–34			
	35–55			

Data source	Age group (in years)	Number of pregnancies	Number of cancer diagnoses	Incidence per 100,000 pregnancies
Data source 3	Overall			
	12–34			
	35–55			

Mock Table 4. Characterisation of cancers (new, relapsed, or recurrent) diagnosed during pregnancy (excluding the 12-month period following pregnancy end), by data source.

	Data source 1	Data source 2	Data source 3
Number of records[1], N			
Number of subjects[2], N			
Cohort start date (range)			
Cohort end date (range)			
Age (in years) at index date, median [IQR]			
Age group (in years) at index date, N (%)			
- 12–34			
- 35–55			
Prior observation, median [IQR]			
Future observation, median [IQR]			
Days in cohort, median [IQR]			
Calendar period at index date, N (%)			
- 2010–2014			
- 2015–2019			
- 2020–end of data			
Prior history of cancer[3], N (%)			
By cancer type, N (%)			
- Breast cancer			
- Melanoma skin cancers			
- Thyroid cancer			
- Cervical cancer			
- Ovarian cancer			
- Lymphoma			
- Other malignancies			

N=Number; IQR=Interquartile range.
[1]Number of cancer diagnoses.
[2]Number of pregnant females diagnosed with cancer.
[3]Defined as cancer cases occurring at any time before the pregnancy start.

Mock Table 5. Characteristics of pregnancies with cancer (new, relapsed, or recurrent) diagnosed during pregnancy (excluding the 12-month period following pregnancy end), by data source.

	Data source 1	Data source 2	Data source 3
Number of records[1], N			
Number of subjects[2], N			
Pregnancy trimester, at cancer diagnosis, N (%)			
- First			
- Second			
- First and second trimester (combined)			
- Third			
Gravidity, N (%)			
- First pregnancy			
- Subsequent pregnancies			
Birth term status[3], N (%)			
- Pre-term birth			
- Term birth			
Pregnancy outcome[4], N (%)			
- Live birth			
- Miscarriage			
- Stillbirth			
- Elective termination			
- Discordant			
- Unknown			

N=Number.
[1]Number of incident cancer diagnoses.
[2]Number of pregnant females with incident cancer.
[3]Pre-term birth is defined as live births ending between 20 and 36 weeks of gestation. Term birth is defined as a live birth ending at ≥37 weeks of gestation.
[4]Outcomes will be based on those recorded in the Pregnancy Expansion Table (PET). According to the PET implementation these include: live birth, miscarriage, stillbirth (≥20 weeks), elective termination of pregnancy, discordant (different outcomes in multiple pregnancies), unknown. Additional categories may be available depending on data source granularity on pregnancy outcome recording.

8.9. Evidence synthesis

Results from analyses described in **Section 8.8** will be presented separately for each data source. No meta-analysis of results will be conducted.

9. STRENGTHS AND LIMITATIONS

The study will be performed by using routinely collected health care data and so data quality issues must be considered. In particular, the capture of events of interest may vary across data sources and will be limited to events recorded in the healthcare systems covered by each data source. NLHR provides nationwide data from primary and secondary care settings, with the latter available from 2018 onwards. SIDIAP and CPRD

GOLD are broadly representative of in terms of age and sex distribution to the Catalan and UK populations, respectively.[14, 15] Both provide data from primary care settings, with SIDIAP including hospital discharge data. The coverage of CPRD GOLD has changed considerably, with a substantial shift in the regional distribution of contributing general practitioner practices.[14] This has affected its regional coverage and population size, which has decreased over the years and should be considered when interpreting the number of cancer cases by calendar period.[14]

Pregnancies captured will originate from those recorded in the PET, and therefore, the study population will be limited to pregnancies recorded in this table of the OMOP CDM. Based on Objective 4 of DARWIN EU® PeriNet, pregnancy outcomes occurring in early gestation are underreported and may not be well captured in the included data sources, except for UiO, which uses an algorithm to identify pregnancies ending before gestational week 12.[11] Given the differential capture of outcomes across data sources, the denominator and the pregnancies covered may differ across data sources (**Table S1**). However, the main analysis will include all pregnancies recorded in the data to take early non-live pregnancy outcomes into consideration, as a substantial proportion (15–25%) of clinically recognised pregnancies are estimated to end in miscarriage.[15] Pregnancy outcomes will be described as part of the characterisation of cancer cases. In CPRD GOLD, the algorithm used to construct pregnancy episodes in CPRD GOLD does not adequately distinguish between induced and spontaneous abortions, and these will therefore likely be affected by misclassification.[17] The sensitivity analysis restricted to live births will estimate cancer incidence using a more clearly defined denominator across data sources, supporting interpretation of the findings and comparison with the literature.

The pregnancy start date will not be inferred as part of this study and will be used as recorded in the PET. Its ascertainment will therefore depend on how pregnancy start is recorded or derived in each data source (see **Section 8.4**)**.** According to the PET specifications, the pregnancy start date is based on ultrasound estimates or, if ultrasound information is unavailable, calculated from the last menstrual period.[8] Given the nature of routinely collected data, some uncertainty around the exact pregnancy start date cannot be ruled out, and further evaluation of the pregnancy start date recording across participating data sources is outside the scope of this study. In NLHR, gestational age for 66.3% of pregnancies ending in miscarriage and 47.2% of pregnancies ending in elective termination is inferred based on pregnancy markers. For the remaining pregnancies, gestational age is assigned using pre-specified durations of 10 and 8 weeks, respectively, meaning that the estimated pregnancy start date may be less precise for these cases.[11]

Some uncertainty regarding the exact date of cancer diagnosis may also exist. SIDIAP has previously conducted a validation study comparing cancer diagnoses recorded in this data source (with linkage to hospital discharge data) with those recorded in cancer registries.[18] The study found that SIDIAP captured approximately 76% of the cancer diagnoses identified in the registries, while also including a substantial number of additional cases not recorded in the registries. Furthermore, most cancer diagnoses in SIDIAP were recorded within three months of the corresponding registry diagnosis date. Although a three-month difference is relatively short, it may be particularly relevant when assessing cancers diagnosed during pregnancy, given the limited duration of the exposure period. Nevertheless, extending case ascertainment to include the 12-month post-partum period is expected to improve case identification and mitigate the impact of minor discrepancies in diagnosis dates.

Importantly, results from DARWIN EU® PeriNet, have shown very limited counts of cancer diagnoses among this population (see **Section 8.7**). At pregnancy start, between 0.3% and 1.2% of episodes had a prior history of cancer. During pregnancy, diagnoses were very limited and only surpassed 100 cases for NLHR (2018 onwards) for breast cancer, gynaecological cancers, melanoma, and thyroid cancer. In SIDIAP and CPRD GOLD (2010 onwards), the cancer types with >5 counts during pregnancy were breast cancer (<50), thyroid cancer (<50), melanoma (<30), and gynaecological cancers (in CPRD GOLD only; <30). Notably, the number of cancer cases observed during pregnancy in DARWIN EU® PeriNet varied across data sources, with NLHR identifying the highest number of cases. Although CPRD GOLD captured the largest number of

pregnancy episodes (**Table S1**), it identified fewer cancer diagnoses during pregnancy, suggesting differences in case ascertainment across data sources. While this may reflect variations in data capture, such as the reliance of CPRD GOLD on primary care records only, this was not explored in detail in DARWIN EU® PeriNet and warrants further investigation. Given that a substantial proportion of pregnancy-associated cancers are diagnosed in the post-partum period, the inclusion of diagnoses occurring within 12 months after pregnancy end is expected to increase the number of captured cases. Nevertheless, stratification by age group and calendar period will likely remain challenging.

Lastly, data on some covariates of interest will be unavailable, including BReast CAncer gene 1 (BRCA1) and BReast CAncer gene 2 (BRCA2), which is not available across participating data sources, and gravidity, which is not available in CPRD GOLD.

10. REFERENCES

1. Smith LH, Danielsen B, Allen ME, Cress R. Cancer associated with obstetric delivery: results of linkage with the California cancer registry. American Journal of Obstetrics and Gynecology. 2003;189(4):1128–35.

2. Metcalfe A, Cairncross ZF, Friedenreich CM, Ray JG, Nelson G, Fell DB, et al. Incidence of Pregnancy-Associated Cancer in Two Canadian Provinces: A Population-Based Study. International Journal of Environmental Research and Public Health. 2021;18(6):3100.

3. Lee Y, Roberts C, Dobbins T, Stavrou E, Black K, Morris J, et al. Incidence and outcomes of pregnancy-associated cancer in Australia, 1994–2008: a population-based linkage study. BJOG: An International Journal of Obstetrics & Gynaecology. 2012;119(13):1572–82.

4. Stensheim H, Møller B, Van Dijk T, Fosså SD. Cause-Specific Survival for Women Diagnosed With Cancer During Pregnancy or Lactation: A Registry-Based Cohort Study. Journal of Clinical Oncology. 2009;27(1):45–51.

5. Eibye S, Kjær SK, Mellemkjær L. Incidence of Pregnancy-Associated Cancer in Denmark, 1977–2006. Obstetrics & Gynecology. 2013;122(3):608–17.

6. Parazzini F, Franchi M, Tavani A, Negri E, Peccatori FA. Frequency of Pregnancy Related Cancer. International Journal of Gynecological Cancer. 2017;27(3):613–9.

7. Lundberg FE, Stensheim H, Ullenhag GJ, Hanna, Lindemann K, Fredriksson I, et al. Risk factors for the increasing incidence of pregnancy-associated cancer in Sweden – a population-based study. Acta Obstetricia et Gynecologica Scandinavica. 2024;103(4):669–83.

8. De Haan J, Verheecke M, Van Calsteren K, Van Calster B, Shmakov RG, Mhallem Gziri M, et al. Oncological management and obstetric and neonatal outcomes for women diagnosed with cancer during pregnancy: a 20-year international cohort study of 1170 patients. The Lancet Oncology. 2018;19(3):337–46.

9. Abellan A, Burn E, Trinh NTH, Burkard T, Callahan A, Fernández-Bertolín S, et al. Expanding the OMOP Common Data Model to Support Perinatal Research in Network Studies. Pharmacoepidemiology and Drug Safety. 2025;34(2).

10. Raventós B, Fernández-Bertolín S, Aragón M, Voss EA, Blacketer C, Méndez-Boo L, et al. Transforming the Information System for Research in Primary Care (SIDIAP) in Catalonia to the OMOP Common Data Model and Its Use for COVID-19 Research. Clinical Epidemiology. 2023;Volume 15:969–86.

11. Trinh NT, Houghtaling J, Bernal FL, Hayati S, Maglanoc LA, Lupattelli A, et al. Harmonizing Norwegian registries onto OMOP common data model: Mapping challenges and opportunities for pregnancy and COVID-19 research. International Journal of Medical Informatics. 2024;191:105602.

12. Kafatos G, Maskell J, Archangelidi O, Neasham D. Validation of the transformed clinical practice research datalink (CPRD) GOLD and aurum data into the OMOP common data model. Health Informatics Journal. 2025;31(3).

13. Mercade-Besora N, Raventós B, Trinh NT, Li H, Hayati S, Giuliodori A, et al. Effectiveness of COVID-19 mRNA primary and booster vaccination against infection and hospitalisation during pregnancy: a target trial emulation and meta-analysis of data from 4 European countries [pre-print]. 2025.

14. Prieto-Alhambra D, Mercadé-Besora N. Safety of mRNA COVID-19 vaccines during pregnancy (EUPAS1000000712) HMA-EMA Catalogues of real-world data sources and studies2025 [updated 18/08/2025. Available from: https://catalogues.ema.europa.eu/node/4615/administrative-details.

15. Politi J, Duarte-Salles T. DARWIN EU® – Utilisation of commonly used benzodiazepines during pregnancy and the incidence of pregnancy losses (EUPAS1000000536): HMA-EMA Catalogues of real-world data

sources and studies; 2025 [updated 10–04–2025. Available from: https://catalogues.ema.europa.eu/node/4437/administrative-details.

16. Nordeng H, Lupattelli A, Hilde, Marleen. Detecting and Dating Early Non-live Pregnancy Outcomes: Generation of a Novel Pregnancy Algorithm From Norwegian Linked Health Registries. Pharmacoepidemiology and Drug Safety. 2024;33(9).

17. Matcho A, Ryan P, Fife D, Gifkins D, Knoll C, Friedman A. Inferring pregnancy episodes and outcomes within a network of observational databases. PLOS ONE. 2018;13(2):e0192033.

18. Recalde M, Manzano-Salgado C, Díaz Y, Puente D, Garcia-Gil MDM, Marcos-Gragera R, et al. Validation Of Cancer Diagnoses In Electronic Health Records: Results From The Information System For Research In Primary Care (SIDIAP) In Northeast Spain. Clinical Epidemiology. 2019;Volume 11:1015–24.

19. Du M, Prats-Uribe A, Mercadé-Besora N, Lopez-Guell K, Guo Y, Alcalde-Herraiz M, et al. CohortCharacteristics: an R package for population characterisation in observational studies using the OMOP common data model. European Journal of Epidemiology. 2026.

11. ANNEXES

ANNEX I. Description of data sources

Norway: Norwegian Linked Health Registry data (NLHR)

#	Section	Description
1	Data source identification and country	NLHR (Norwegian Linked Health Registry data) Norway
2	Data partner information section	University of Oslo (UiO) Faculty of Mathematics and Natural Science – Department of Pharmacy
3	Coverage and timespan	Data collection since: 2008 Extent: Nation-wide. Norway has a universal public health care system, consisting of primary and specialist health care services covering a population of approximately 5.4 million inhabitants.
4	Healthcare setting / type of data	Primary care – General Practitioner, and primary care specialists (e.g. paediatricians), and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. The following registries are included: the Medical Birth Registry of Norway (MBRN), the Norwegian Prescription Registry (NorPD), the Norwegian Patient Registry (NPR), Norway Control and Payment of Health Reimbursement (KUHR), the Norwegian Surveillance System for Communicable Diseases (MSIS), the Norwegian Immunisation Registry (SYSVAK), the National Death Registry, and the National Registry (NR).
5	Data collection process	Registries. Many population-based health registries were established in the 1960s, with use of unique personal identifiers facilitating linkage between registries. Data in these health registries are used for health analysis, health statistics, improving the quality of healthcare, research, administration, and emergency preparedness.
6	General representativeness	The NLHR data covers the full Norwegian population.
7	Data content / source coding	NPR: ICD-10 for diagnosis, ATC and some special codes for drug use, Norwegian codes for clinical procedures (surgery (NCSP), medicine (NCMP) and diagnostic imaging, image-guided intervention, and nuclear medicine (NCRP)). KUHR: ICD-10 and ICPC-2 and ICPC-2B for diagnosis/procedure. NorPD: ATC. SYSVAK and MSIS: national classifications. MBRN: custom classifications by questionnaires (incl. check box variables in Maternity health care card)
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Linkage between the registries was facilitated using project-specific person IDs generated from unique personal identification assigned at birth or immigration for all legal residents in Norway.
9	Quality control (data source specific)	In-house data quality checks of rates of common conditions, drug exposures, and outcomes. We compare obtained rates with official national statistics (e.g., birth statistics, yearly rates of drug dispensing, and diagnosis by age and gender). We also review missing data and outliers and inform registry holders of any unusual patterns.
10	Linkage	The NLHR is, by definition, a linkage of datasets. Helsedata.no is one central portal to apply for 11 national health registries, including all the registries that have been mapped to the OMOP CDM.
11	Vital status	The national death registry is linked.
12	Limitations	(i) Diagnostic codes in the secondary care are limited to the comprehensive list requested from the registries. The list of codes will be revised annually with each data delivery. (ii) ICD-10 codes vary in granularity, ranging from 2-character to full-length codes. The granularity of data will be revised annually with each data delivery (iii) Drug dispensations to outpatients are included; however, over-the-counter (OTC) medications are not captured in the data. Only a few selected expensive drugs given in hospital, are included. (iv) The NLHR comprises the following core

#	Section	Description
		datasets: • Primary care data: covers the period from 2008 to 2023 • Hospital (secondary care) data, and drug prescription and vaccination data: cover the period from 2018 to 2023 The observation period is defined as follows: • Observation_period_start_date: the later of 2008 or the individual's birth date • Observation_period_end_date: the earlier of the individual's death date or the end of 2023 Immigration and emigration data are also considered when determining the observation period. While most individuals have a single continuous observation period, a small number have two non-overlapping periods. These differing coverage periods across datasets can affect study outcomes. For example, when analyzing hospital-diagnosed conditions, the prevalence appears low from 2008 to 2018, with a sharp increase in incidence in 2018.
13	Main references	Trinh NT, Houghtaling J,Bernal FL,Hayati S,Maglanoc LA,Lupattelli A,Halvorsen L,Nordeng HM "Harmonizing Norwegian registries onto OMOP common data model: Mapping challenges and opportunities for pregnancy and COVID-19 research." International journal of medical informatics (2024): 39153282
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1000000409 Website: https://www.mn.uio.no/farmasi/english/research/groups/pharma-safe/

Spain: The Information System for Research in Primary Care (SIDIAP)

#	Section	Description
1	Data source identification and country	SIDIAP (The Information System for Research in Primary Care) Catalunya, Spain
2	Data partner information section	Institut d'Investigació en Atenció Primària de Salut Jordi Gol (IDIAPJGol)
3	Coverage and timespan	Data collection since: 2006 Extent: Regional. SIDIAP is a database of primary care electronic health records of the population of Catalonia, North-East Spain. It contains pseudo-anonymised records of more than 8 million people , of which 6.1 million are active as of 2022, representing around 76% of the Catalan population.
4	Healthcare setting / type of data	Primary care – General Practitioner, and hospital inpatient care. SIDIAP captured data includes routine visits, socio-demographics, diagnoses, laboratory tests, drugs (prescribed and dispensed), referrals, sick leaves, and lifestyle information.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records. Data is entered by primary care physicians upon healthcare contact, supplemented with hospital discharge records. The Institut Catala de la Salut (Catalan Health Institute) is the data controller.
6	General representativeness	It was previously shown that the captured SIDIAP population is highly representative of the entire Catalan region in terms of geographic, age, and sex distributions.
7	Data content / source coding	SIDIAP data covers all services that occur at the Primary Care Centres, as well as support services, such as sexual and reproductive health or home end-of-life care. Drugs are coded in ATC-WHO terminology in the source data. Health outcomes are captured in ICD-10CM codes. The SIDIAP contains all laboratory tests and results performed in primary health centres. Demographics, geographical, as well as socio-economic factors are recorded for each patient.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. A patient has a unique id, also upon changing practice or re-registration.
9	Quality control (data source specific)	Internal and external validation processes are carried out to determine the data quality of the SIDIAP information at each data update. These include stratifying the data by geographical regions and year in order to identify differences in data collection that need to be harmonized (e.g. recording of specific information under different codes).

#	Section	Description
		The measurement units of variables measuring one characteristic are also homogenized (e.g. transformation of the data from every laboratory that measures haemoglobin to grams per decilitre). Visual inspection of all data included in the database by week is also conducted, allowing one to see temporal patterns in the registry of a certain variable. With this information, the SIDIAP team can issue recommendations to researchers about the most common variable(s) where certain information is recorded (e.g., there are several variables with information concerning the women's menopausal status and with these visual inspection tools the SIDIAP team can inform the researchers about which related variables have the largest number of records and could be more helpful to capture menopause). Data availability (longitudinally and reliability), plausibility (range checks and unusual values), and consistency are inspected through visualisation tools. In addition, before accessing the data for a requested project, research teams have access to a quality-control report. This document contains counts, years, percentiles, maximums and minimums, incidences, and prevalence of the data requested for the project, allowing detection of inconsistencies in the data extraction prior to data delivery. External validation processes of the SIDIAP database mainly include assessing the data recorded in SIDIAP through linkage to external gold standard data sources, by analysing free text, or by sending questionnaires to health professionals.
10	Linkage	SIDIAP is linked to a hospital discharge database, pharmacy dispensation, and primary care laboratories. It can also be linked to other registries in Catalonia on a project by project basis.
11	Vital status	Mortality is fully captured in SIDIAP. The cause of death is not available, but can be linked to the Spanish death registry on a project by project basis.
12	Limitations	The SIDIAP data is not representative of individuals not using public primary care, and conditions that are usually followed by specialist care might not be properly captured. In addition, there is limited information on lifestyle variables. Patients are followed until Death or when transferring to another primary health care centre that does not contribute to SIDIAP.
13	Main references	Recalde M, Rodríguez C,Burn E,Far M,García D,Carrere-Molina J,Benítez M,Moleras A,Pistillo A,Bolíbar B,Aragón M,Duarte-Salles T "Data Resource Profile: The Information System for Research in Primary Care (SIDIAP)." International journal of epidemiology (2022): 35415748
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/50190 Website: https://www.sidiap.org/index.php/en

United Kingdom: Clinical Practice Research Datalink GOLD (CPRD GOLD)

#	Section	Description
1	Data source identification and country	CPRD GOLD (Clinical Practice Research Datalink GOLD) the United Kingdom
2	Data partner information section	University of Oxford - Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences (Oxford NDORMS) NDORMS
3	Coverage and timespan	Data collection since: 1987 Extent: Nation-wide. CPRD GOLD consists of patients in contributing practices using Vision software. Historically this covered the whole of the UK, but the number of contributing practices in the England is dropping. In January 2025 only 3 practices from England were a part of CPRD GOLD, while historical patient data were from the whole of the UK, and will continue to be so. In the future, no practices from England will be present, only practices from Scotland, Wales, and Northern Ireland.
4	Healthcare setting / type of data	Primary care – General Practitioner, and primary care specialists (e.g. paediatricians), and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care.

#	Section	Description
		CPRD GOLD data include patient demographics, biological measurements, clinical symptoms and diagnoses, referrals to specialist/hospital and their outcome, laboratory tests/results, and prescribed medications.
5	Data collection process	Outpatient electronic health records. Data are entered by clinicians into the EHR. Data is processed by CPRD that provides data releases for research.
6	General representativeness	In the last 10 years, the CPRD GOLD regional distribution of currently contributing general practitioner (GP) practices has significantly shifted, resulting in many new practices joining from Scotland, Wales, and Northern Ireland, and fewer participating from England. These changes have affected the CPRD GOLD population size, regional coverage, and eligibility for data linkages. CPRD GOLD January 2024 contains >21.3 million historical and current patients (12.9 in England, 3.1 in Wales, 4.7 in Scotland, 0.7 in Northern Ireland). Of these, nearly 3 million are currently registered in a GP practice and represent ~4.3% of the estimated current UK population (0.1% in England, 32.3% in Wales, 28.6% in Scotland, 16.2% in Northern Ireland). Patients currently registered in CPRD GOLD January 2024 are broadly representative of the UK population with respect to age and sex. Reference: https://doi.org/10.1093/ije/dyaf077
7	Data content / source coding	Gemscript, Read, dm+d
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. In GOLD, a patient can be registered under different ID numbers upon changing practice or re-registration. Researchers are not able to identify these patients, as the data are anonymised. However, GOLD covers less than 5% of the current UK GP practices and it is unlikely that an individual who does change GP practice ends up in another GP practice which uses the Vision software and accepts the CPRD data collection agreement. The very small number of duplicated IDs will have different observation periods and should not have an impact on the data analyses.
9	Quality control (data source specific)	CPRD GOLD only includes practices whose data quality is assessed to be up-to-standard (UTS). Each practice is associated to an UTS date set when the data quality standards become satisfactory, and CPRD recommend using only longitudinal data starting from this UTS date. Every time CPRD collect the EHR from a practice, checks are run for the data quality standards, and if they are not adequate, the EHR is not accepted. When the data quality becomes acceptable again, CPRD updates the practice UTS date. CPRD also checks data quality standards at the patient level, and associates each patient with a flag, reporting if its data are acceptable for clinical research. Only patients with acceptable data quality are included in the population to be mapped to CDM.
10	Linkage	CPRD GOLD can be linked to several sources, however our Oxford OMOP CDM is only linked to the CPRD GOLD Ethnicity Record and to the CPRD Townsend Deprivation Index at the Practice Level.
11	Vital status	The date of death in CPRD GOLD has been validated against the Population registry (ONS) mortality data. Reference: https://doi.org/10.1002/pds.4747
12	Limitations	The main limitation is due to the fact that CPRD GOLD is limited to GP records, and although it contains information on referrals and discharge letters, it may not fully capture specific hospital information. Events from hospital and specialist care are not covered.
13	Main references	Sanchez-Santos MT, Axson EL,Dedman D,Delmestri A "Data Resource Profile Update: CPRD GOLD." International journal of epidemiology (2025): 40499193
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111113 Website: https://www.cprd.com/data/primary-care-data/cprd-gold

ANNEX II. Fitness for use assessment

Norway: Norwegian Linked Health Registry data (NLHR)

The inclusion of NLHR in this study was based on its relevance to the perinatal field. NLHR is a registry data source that has been recognised for its value in providing reliable and comprehensive data for pregnancy research. It includes detailed information related to all pregnancies in Norway, including approximately 60,000 deliveries per year from the Medical Birth Registry of Norway (MBRN).[1]

There were study-specific limitations present in NLHR. Notifications of pregnancies and pregnancy outcomes are mandatory in Norway for all pregnancies past gestational week 12, and therefore, those ending before that threshold may be missed. The UiO Pregnancy algorithm improves the detection and dating of early non-live pregnancy outcomes that would have gone unnoticed if relying solely on the medical birth registry information by linking the MBRN records with hospital and primary care records.[2] The algorithm allow inferring gestational age for 66.3% of pregnancies ending in miscarriage and 47.2% ending in elective termination. For the remaining pregnancies, gestational age is assigned using pre-specified durations of 10 and 8 weeks, respectively; therefore, the estimated pregnancy start date may be less precise for these cases.[2] Lastly, it should be noted that secondary care data is available from 2018 onwards, which may influence diagnostic capture over the study period (e.g. diagnoses recorded after 2018 may include greater granularity and detail than those captured with primary care data only).

This data source has 2 available data sources onboarded into DARWIN EU®, which are: the MBRN (NLHR@University of Oslo (UiO):PERINATAL CDM) and NLHR (NLHR@UiO CDM dataset). Both data sources use the UiO pregnancy algorithm and cover the entire Norwegian population. However, they have different source populations, based on the data source that provides the linkage identification numbers (MBRN vs. national registry) and the data period covered. Specifically:

- **NLHR@UiO:PERINATAL CDM**, IDs come from the medical birth registry, covering all individuals with a pregnancy event >12 gestational weeks, and covers the period 2008–2019.
- **NLHR@UiO CDM**, IDs come from the National Registry covering the entire Norwegian population and covers the period 2008–2023.

Overall, both can identify similar pregnancy events. However, the NLHR@UiO data source has a greater ability to detect early pregnancies (<12 gestational weeks), as its source population includes the entire Norwegian population. In contrast, the source population in the NLHR@UiO:PERINATAL is based on the MBRN, which only includes individuals with a pregnancy episode lasting beyond gestational week 12. Although the UiO pregnancy algorithm improves the identification of pregnancy events <12 gestational weeks, in this data source it can only improve detection for individuals present in the MBRN. Consequently, differences between the two data sources are likely driven by the NLHR@UiO:PERINATAL CDM's lower ability to detect elective terminations, particularly among younger individuals, who are less likely to have had a prior pregnancy extending beyond 12 gestational weeks, and therefore less likely to appear in the MBRN.

Important: Of the two available data sources, we selected the NLHR@UiO data source for this study because it has a greater ability to detect early pregnancies and its more granular information on drug exposures, making it more suitable for use in future studies. For simplicity, NLHR@UiO is referred as NLHR in this report. In this data source, data expands up to December 2023, and the median length of observation period is 5,820 days (4,200–5,820).

Spain: The Information System for Research in Primary Care (SIDIAP)

SIDIAP was included in this study because it contains detailed pregnancy information from reproductive health clinics in primary care, including information related to the perinatal period for approximately 40,000 deliveries per year.[3] In addition, SIDIAP can be linked to hospital discharge data, which was requested as part of this study.[4] These data will be requested based on findings from a previous validation study of cancer diagnoses conducted within this data source, which showed that incorporating these data increased sensitivity for all cancers.[5]

Data availability in SIDIAP was considered sufficient for this study, as data collection began in 2006, and the most recent data extraction date was December 2024. The median length of the observation period was 5,760 days (IQR: 2,170–6,940).

There are study-specific limitations present in SIDIAP. First, SIDIAP does not include data from private healthcare providers.[4] Therefore, underreporting of health events may occur among individuals who choose to receive antenatal care through private healthcare settings. While reproductive health clinics in primary care are covered, individuals considered to have high-risk pregnancies are referred to outpatient hospital care. Recording of diagnoses and prescribed medications in outpatient hospital care may be underreported, as hospital data is only available for inpatient diagnoses. Lastly, voluntary pregnancy losses may be underreported.

The United Kingdom: Clinical Practice Research Datalink GOLD (CPRD GOLD)

CPRD GOLD was included in this study because it contains all clinical and referral events in primary care. Pregnancy episodes in CPRD GOLD are inferred using a previously validated algorithm,[6] which was validated for CPRD GOLD data mapped to the OMOP CDM.

Data availability and follow-up in CPRD GOLD were considered sufficient for this study. Data availability began in 1987, and the date of the most recent data extraction was December 2025. The median length of the observation period was 2,160 days (IQR: 729–4,940).

There are study-specific limitations present in CPRD GOLD. First, events occurring in hospital settings are likely underreported, as it covers only primary care data. However, given the outcomes assessed in this study, we expect most cancer cases to also be captured in primary care data. Second, the algorithm used to construct pregnancy episodes in CPRD GOLD does not adequately distinguish between induced and spontaneous abortions, and these will therefore likely be affected by misclassification.[17] Lastly, the coverage and population size of CPRD GOLD has substantially changed over the last decade, which should be considered when interpreting results stratified by study period.

Fitness for use assessment references:

1. Medical Birth Registry of Norway - NIPH. https://www.fhi.no/en/ch/medical-birth-registry-of-norway/ (accessed 16 Oct 2025).
2. Nordeng H, Lupattelli A, Hilde, Marleen. Detecting and dating early non-live pregnancy outcomes: generation of a novel pregnancy algorithm from Norwegian Linked Health Registries. *Pharmacoepidemiol Drug Saf* 2024; 33(9).
3. Abellan A, Burn E, Trinh NTH, et al. Expanding the OMOP Common Data Model to Support Perinatal Research in Network Studies. Pharmacoepidemiol Drug Saf. 2025;34(2):e70106. doi:10.1002/pds.70106.
4. Recalde M, Rodríguez C, Burn E, et al. Data resource profile: The Information System for Research in Primary Care (SIDIAP). *Int J Epidemiol* 2022; 51(6): e324–36.
5. Recalde M, Manzano-Salgado CB, Díaz Y, et al. Validation Of Cancer Diagnoses In Electronic Health Records: Results From The Information System For Research In Primary Care (SIDIAP) In Northeast Spain. *Clinical epidemiology*, 11, 1015–1024.
6. Matcho A, Ryan P, Fife D, Gifkins D, Knoll C, Friedman A. Inferring pregnancy episodes and outcomes within a network of observational databases. *PLOS ONE* 2018; 13(2): e0192033.

ANNEX III. Operational and reporting considerations

DATA MANAGEMENT

Data management

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

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

Data storage and protection

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

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

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

QUALITY CONTROL

Data source quality control

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

When defining cohorts for indications, a systematic search of possible codes for inclusion will be identified using the *CodelistGenerator* R package (https://github.com/darwin-eu/CodelistGenerator). This package allows the user to define a search strategy and will use this to query the vocabulary tables of the OMOP common data model so as to find potentially relevant codes. In addition, the *CohortDiagnostics* (https://github.com/OHDSI/CohortDiagnostics) R package or, alternatively, the PhenotypeR (https://github.com/OHDSI/PhenotypeR) R package, will be run to assess the use of different codes across the data sources contributing to the study and identify any codes potentially omitted in error.

The study code will be based on the DARWIN EU® R package *CohortCharacteristics*. This package will include numerous automated unit tests to ensure the validity of the codes, alongside software peer review and user testing. The R package will be made publicly available via GitHub.

PLANS FOR DISSEMINATING AND COMMUNICATING STUDY RESULT

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

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

ANNEX IV. List of stand-alone documents

Table S1. Pregnancy outcomes captured in DARWIN EU® PeriNet (Objective 4) in data sources participating in this study.

	NLHR	SIDIAP	CPRD GOLD
Number of pregnancies, N[1]	367,803	409,207	466,615
Number of pregnant individuals, N	257,402	294,381	352,508
Pregnancy outcome, N (%)[2]			
- Live birth	256,773 (69.8%)	369,803 (90.4%)	332,606 (71.3%)
- Miscarriage	53,141 (14.4%)	34,147 (8.3%)	129,464 (27.8%)
- Still birth	819 (0.2%)	2,242 (0.6%)	1,200 (0.3%)
- Elective termination	52,782 (14.3%)	2,125 (0.5%)	3,345 (0.7%)
- Ectopic	3,910 (1.1%)	–	–
- Molar	378 (0.1%)	–	–
- Discordant	–	–	–
- Unknown	–	890 (0.2%)	–

CPRD GOLD=Clinical Practice Research Datalink GOLD; NLHR=Norwegian Linked Health Registry data; PET=Perinatal Expansion Table; SIDIAP=The Information System for Research in Primary Care.
"–" Indicates that values for this concept or variable were not found in the PET.
[1]The enrolment period for pregnancies spanned from 01/01/2010 (01/01/2018 for NLHR) to one year prior to end of data availability. End of data availability differed across data sources: NLHR: 31/12/2023; SIDIAP: 31/12/2024; CPRD GOLD: 20/05/2025.
[2]Outcomes of the pregnancy based on the PET implementation included: live birth, miscarriage, still birth (≥20 weeks), elective termination of pregnancy, discordant (different outcomes in multiple pregnancies), unknown. Ectopic and molar pregnancies were not part of the original design of the PET but were recorded as outcomes in NLHR.

ANNEX V. ENCePP checklist for study protocols

ENCePP Checklist for Study Protocols (Revision 4)

Study title: DARWIN EU® - Cancer occurrence during pregnancy

EU PAS Register® number: EUPAS1000001059 **Study reference number (if applicable):** P5-C1-007

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

Comments:

Section 2: Research question	**Yes**	**No**	**N/A**	**Section Number**
2.1 Does the formulation of the research question and objectives clearly explain:	☒	☐	☐	6,7
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?	☐	☐	☒	

Comments:

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

Section 3: Study design		**Yes**	**No**	**N/A**	**Section Number**
3.1	Is the study design described? (e.g. cohort, case-control, cross-sectional, other design)	☒	☐	☐	8.1
3.2	Does the protocol specify whether the study is based on primary, secondary or combined data collection?	☒	☐	☐	8.1
3.3	Does the protocol specify measures of occurrence? (e.g., rate, risk, prevalence)	☒	☐	☐	8.1
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))	☐	☐	☒	N/A
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)	☐	☐	☒	N/A

Comments:

Section 4: Source and study populations		**Yes**	**No**	**N/A**	**Section Number**
4.1	Is the source population described?	☒	☐	☐	8.3
4.2	Is the planned study population defined in terms of:				8.2, 8.3, 8.4
	4.2.1 Study time period	☒	☐	☐	
	4.2.2 Age and sex	☒	☐	☐	
	4.2.3 Country of origin	☒	☐	☐	
	4.2.4 Disease/indication	☒	☐	☐	
	4.2.5 Duration of 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)	☒	☐	☐	8.3

Comments:

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

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

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?	☒	☐	☐	8.6.2
6.2 Does the protocol describe how the outcomes are defined and measured?	☒	☐	☐	8.6.2
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)	☐	☒	☐	N/A
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)	☐	☐	☒	N/A

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

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)	☐	☐	☒	N/A

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)	☐	☐	☒	N/A
	9.1.2 Outcomes? (e.g. clinical records, laboratory markers or values, claims data, self-report, patient interview including scales and questionnaires, vital statistics)	☒	☐	☐	8.4, Annex I and II
	9.1.3 Covariates and other characteristics?	☒	☐	☐	8.4, Annex I and II
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)	☐	☐	☒	N/A
	9.2.2 Outcomes? (e.g. date of occurrence, multiple event, severity measures related to event)	☒	☐	☐	8.4, Annex I and II
	9.2.3 Covariates and other characteristics? (e.g. age, sex, clinical and drug use history, co-morbidity, co-medications, lifestyle)	☒	☐	☐	8.4, Annex I and II
9.3	Is a coding system described for:				
	9.3.1 Exposure? (e.g. WHO Drug Dictionary, Anatomical Therapeutic Chemical (ATC) Classification System)	☐	☐	☒	N/A
	9.3.2 Outcomes? (e.g. International Classification of Diseases (ICD), Medical Dictionary for Regulatory Activities (MedDRA))	☒	☐	☐	8.6.2
	9.3.3 Covariates and other characteristics?	☒	☐	☐	8.6.3
9.4	Is a linkage method between data sources described? (e.g. based on a unique identifier or other)	☒	☐	☐	Annex I and II

Comments:

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

Comments:

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

Comments:

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

Comments:

Section 13: Ethical/data protection issues	**Yes**	**No**	**N/A**	**Section Number**
13.1 Have requirements of Ethics Committee/ Institutional Review Board been described?	☒	☐	☐	
13.2 Has any outcome of an ethical review procedure been addressed?	☒	☐	☐	
13.3 Have data protection requirements been described?	☒	☐	☐	8.2, Annex III

Comments:

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

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)?	☒	☐	☐	8.8.4
15.2 Are plans described for disseminating study results externally, including publication?	☒	☐	☐	8.8.4

Comments:

Name of the main author of the protocol: Berta Raventós

Date: 24/June/2026

Signature: Berta Raventós