



Study Protocol

P5-C1-013

DARWIN EU® - Incidence and characterisation of polyendocrine metabolic ovarian syndrome in Europe

23/09/2026

Version 3.0

Authors: Anna Saura-Lazaro, Amy Lam

Public

CONTENTS

LIST OF ABBREVIATIONS.....	5
1. TITLE	7
2. DESCRIPTION OF THE STUDY TEAM	7
3. ABSTRACT	8
4. AMENDMENTS AND UPDATES	11
5. MILESTONES.....	11
6. RATIONALE AND BACKGROUND	11
7. RESEARCH QUESTION AND OBJECTIVES.....	12
8. RESEARCH METHODS	12
8.1. Study design	12
Figure 1. Graphical depiction of the study design for Objective 1 (incidence of recorded PMOS diagnosis).....	13
Figure 2. Graphical depiction of the study design for Objective 2 (characterisation of individuals with PMOS on comorbidities).....	14
Figure 3. Graphical depiction of the study design for Objective 3 (characterisation of Individuals with PMOS on non-infertility drug treatments).	15
Figure 4. Graphical depiction of the study design for Objective 4 (incidence rates of health outcomes in individuals with PMOS) and Objective 5 (cumulative incidence of health outcomes in individuals with PMOS).....	16
8.2. Follow-up	17
Figure 5. Included observation time for the denominator population (Objective 1).	18
Figure 6. Included observation time for the denominator population with a washout period (Objectives 4 and 5).....	18
8.3. Study population with inclusion and exclusion criteria.....	19
8.4. Study setting and data sources	19
Table 1. Data sources.....	20
8.5. Study period	20
8.6. Variables	21
8.6.1. Outcome	21
8.6.2. Covariates	21
8.7. Study size	23
8.8. Analysis	23
8.8.1. Federated network analyses	23
8.8.2. Data privacy protection	23
8.8.3. Statistical model specification and assumptions of the analytical approach considered	24
Objective 1.....	24
Objectives 2 and 3	24
Objective 4.....	24
Objective 5.....	25
Sensitivity analysis	25
Table 2. Sensitivity analyses – rationale, strengths, and limitations.....	25
8.8.4. Output	28
8.9. Evidence synthesis.....	28
9. STRENGTHS AND LIMITATIONS.....	28

10. REFERENCES	30
11. ANNEXES	31
ANNEX I. Description of data sources.....	31
Danish Data Health Registries (DK-DHR)	31
Finnish Care Register for Health Care (FinOMOP-THL)	32
The Information System for Research in Primary Care (SIDIAP)	34
Integrated Primary Care Information (IPCI)	35
Clinical Practice Research Datalink GOLD (CPRD GOLD)	36
ANNEX II. Fitness for use assessment.....	39
Table S1. Fitness-for-use assessment of data sources.	41
ANNEX III. Operational and reporting considerations.....	44
ANNEX IV. Lists with concept definitions	46
Table S2. Preliminary list of definitions for polyendocrine metabolic ovarian syndrome.	46
Table S3. Preliminary list of definitions for comorbidities.	46
Table S4. Preliminary list of definitions for non-infertility drug treatments (ingredients).	46
Table S5. Preliminary list of definitions for cardiovascular and mental outcomes.	49
Table S6. Preliminary list of definitions for infertility.	49
ANNEX V. Mock tables and figures.....	51
Overall	51
MOCK Table 1. Number of individuals with polyendocrine metabolic ovarian syndrome (PMOS) during the study period, by study objective and data source.	51
Objective 1.....	52
MOCK Figure 1. Annual incidence of recorded polyendocrine metabolic ovarian syndrome (PMOS) diagnosis, overall and stratified by age group.....	52
Objective 2.....	53
MOCK Table 2. Characterisation of polyendocrine metabolic ovarian syndrome (PMOS) individuals on comorbidities.....	53
MOCK Table 3. Characterisation of polyendocrine metabolic ovarian syndrome (PMOS) individuals on comorbidities, stratified by age group.	55
Objective 3.....	61
MOCK Table 4. Characterisation of polyendocrine metabolic ovarian syndrome (PMOS) individuals on non-infertility drug treatments.	61
MOCK Table 5. Characterisation of polyendocrine metabolic ovarian syndrome (PMOS) individuals on non-infertility drug treatments, stratified by age group.	64
Objective 4.....	75
MOCK Figures 2–9. Annual incidence of health outcomes in polyendocrine metabolic ovarian syndrome (PMOS) individuals, overall and stratified by age group.	75
Objective 5.....	76
MOCK Figure 10–17. 5-year cumulative incidence of health outcomes in polyendocrine metabolic ovarian syndrome (PMOS) individuals, overall and stratified by age group.	76
MOCK Figure 18–25. 10-year cumulative incidence of health outcomes in polyendocrine metabolic ovarian syndrome (PMOS) individuals, overall stratified by age group.	77
Shinyapp	78
MOCK Figure S1. Median age at index date by calendar year.	78
MOCK Figure S2. Median time from the first recorded occurrence of each pre-defined comorbidity of interest to the first recorded PMOS diagnosis.	79
ANNEX VI. ENCePP checklist for study protocols	80
ANNEX VII. Glossary.....	86

Study title	DARWIN EU® - Incidence and characterisation of polyendocrine metabolic ovarian syndrome in Europe
Protocol version	V3.0
Date	23/09/2026
EUPAS number	EUPAS1000001090
Active Substance	Not applicable
Medicinal Product	Not applicable
Research question and objectives	The research questions of this study are: 1. How has the incidence of polyendocrine metabolic ovarian syndrome (PMOS) diagnosis changed over time? 2. What are the characteristics of individuals with PMOS at the time of diagnosis? 3. What non-infertility drug treatments are commonly used in individuals with PMOS following diagnosis? 4. What is the incidence of infertility, cardiovascular, and mental health outcomes in individuals with PMOS? This study aims to: 1. To estimate the incidence of recorded PMOS diagnoses over time. 2. To characterise individuals with PMOS at the time of recorded diagnosis in terms of age and pre-defined comorbidities. 3. To characterise individuals with PMOS in terms of pre-defined non-infertility drug treatments received following diagnosis. 4. To estimate the incidence rate of infertility, cardiovascular, and mental health outcomes in individuals with PMOS. 5. To estimate the 5-year and 10-year cumulative incidence of infertility, cardiovascular, and mental health outcomes in individuals with PMOS.
Countries of study	Denmark, Finland, Spain, the Netherlands, the United Kingdom
Authors	Anna Saura-Lazaro (a.sauralazaro@darwin-eu.org) Amy Lam (a.lam@darwin-eu.org)

LIST OF ABBREVIATIONS

Acronyms/terms	Description
ATC	Anatomical Therapeutic Chemical
CC	Coordination Centre
CDM	Common Data Model
CPRD GOLD	Clinical Practice Research Datalink GOLD
DARWIN EU®	Data Analysis and Real-World Interrogation Network
DK-DHR	Danish Data Health Registries
DKMA	Danish Medicines Agency
DPP-4	Dipeptidyl peptidase-4
DRE	Digital Research Environment
DTZ	Data Transfer Zone
EHR	Electronic Health Records
EMA	European Medicines Agency
EMC	Erasmus University Medical Center
ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
EU	European Union
EUPAS	EU Post-Authorisation Studies Register
FinOMOP-THL	Finnish Care Register for Health Care
GDPR	General Data Protection Regulation
GIP	Glucose-dependent insulintropic polypeptide
GLP-1	Glucagon-like peptide-1
GP	General Practitioner
ICD	International Classification of Diseases
IDIAPJGol	Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina
IPCI	Integrated Primary Care Information
IQR	Interquartile range
IRB	Institutional Review Board
M	Million
MDSLD	Metabolic dysfunction-associated steatotic liver disease
N/A	Not applicable
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
Oxford NDORMS	University of Oxford - Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences
PCOS	Polycystic ovary syndrome
PMOS	Polyendocrine metabolic ovarian syndrome
RxNorm	Medical prescription normalised

Acronyms/terms	Description
SGLT2	Sodium-glucose co-transporter 2
SIDIAP	The Information System for Research in Primary Care
SNOMED	Systemized Nomenclature of Medicine
T2D	Type 2 diabetes
VTE	Venous thromboembolism
WHO	World Health Organization

1. TITLE

DARWIN EU® - Incidence and characterisation of polyendocrine metabolic ovarian syndrome in Europe

2. DESCRIPTION OF THE STUDY TEAM

Study team role	Names	Organisation
Principal Investigator	Anna Saura-Lazaro Amy Lam	University of Oxford
Data Scientist	Marta Alcalde-Herraiz Edward Burn	University of Oxford
Clinical Domain Expert	Anna Saura-Lazaro Alejandro Ballve	University of Oxford
Study Manager	Natasha Yefimenko Nosova	Erasmus MC
Data source	Names	Data Partner Organisation*
Danish Data Health Registries (DK-DHR)	Elvira Bräuner Susanne Bruun	Danish Medicines Agency (DKMA)
Finnish Care Register for Health Care (FinOMOP-THL)	Toni Lehtonen Laura Salonen	FinOMOP
The Information System for Research in Primary Care (SIDIAP)	Elena Roel Herranz Laura Granés González Agustina Giuliodori Picco Talita Duarte Salles	Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol)
Integrated Primary Care Information (IPCI)	Katia Verhamme Mees Mosseveld Guido van Leeuwen	Erasmus University Medical Center (EMC)
Clinical Practice Research Datalink GOLD (CPRD GOLD)	Antonella Delmestri Mandickel Kamtengeni Marta Pineda Moncusí	University of Oxford - Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences (Oxford NDORMS)

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

3. ABSTRACT

Title

DARWIN EU® - Incidence and characterisation of polyendocrine metabolic ovarian syndrome in Europe

Rationale and background

Polyendocrine metabolic ovarian syndrome (PMOS), previously named polycystic ovary syndrome (PCOS), is a common, heterogeneous endocrine disorder affecting females. PMOS is characterised by substantial heterogeneity in its clinical presentation, diagnosis, and management across healthcare settings. A clear understanding of real-world pharmacologic management is needed to characterise how PMOS is treated in routine clinical practice. However, analyses describing the use of pharmacological therapies following PMOS diagnosis and Real-World Evidence on the incidence of various health outcomes among individuals with PMOS remain limited.

Research questions and objectives

Research questions

1. How has the incidence of PMOS diagnosis changed over time?
2. What are the characteristics of individuals with PMOS at the time of diagnosis?
3. What non-infertility drug treatments are commonly used in individuals with PMOS following diagnosis?
4. What is the incidence of infertility, cardiovascular, and mental health outcomes in individuals with PMOS?

Objectives

The objectives of this study are:

1. To estimate the incidence of recorded PMOS diagnoses over time.
2. To characterise individuals with PMOS at the time of recorded diagnosis in terms of age and pre-defined comorbidities.
3. To characterise individuals with PMOS in terms of pre-defined non-infertility drug treatments received following diagnosis.
4. To estimate the incidence rate of infertility, cardiovascular, and mental health outcomes in individuals with PMOS.
5. To estimate the 5-year and 10-year cumulative incidence of infertility, cardiovascular, and mental health outcomes in individuals with PMOS.

Methods

Study design

- Population-level descriptive disease epidemiology study (Objectives 1, 4, and 5)
- Patient-level characterisation (Objectives 2 and 3)

The index date will be the date of first PMOS diagnosis, and individuals will be followed up until 31/12/2024 or the end of the observation period, whichever occurs first.

Population

The study population will include all female individuals with the date of first PMOS diagnosis (index date) within the study period from 01/01/2010 to 31/12/2024. Other eligibility criteria are:

- Aged ≥ 10 years
- At least 365 days of data source history prior to the index date

Variables

Outcome:

- (Objective 1) PMOS
- (Objectives 4 and 5) Infertility, cardiovascular outcomes (hypertension, acute myocardial infarction, ischaemic stroke, stroke, venous thromboembolism), mental health outcomes (anxiety, depression)

Relevant covariates:

- Age at diagnosis
- Pre-defined comorbidities:
 - Cardiometabolic/vascular: obesity, dyslipidaemia, hypertension, type 2 diabetes, metabolic dysfunction-associated steatotic liver disease, obstructive sleep apnoea, myocardial infarction, stroke, venous thromboembolism
 - Hyperandrogenic/reproductive: acne, hirsutism, alopecia, oligomenorrhoea/amenorrhoea, infertility
 - Mental health: anxiety, depression, eating disorder
- Pre-defined non-infertility drug treatments:
 - Combined oral contraceptives, metformin, glucagon-like peptide-1 (GLP-1) receptor agonists, other antidiabetic medications (sulfonylureas, thiazolidinediones, dipeptidyl peptidase-4 inhibitors, dual glucose-dependent insulinotropic polypeptide/GLP-1 receptor agonist, sodium-glucose co-transporter 2 inhibitors), insulin, anti-androgen medications (spironolactone, flutamide, bicalutamide, finasteride, cyproterone), orlistat, inositol

Data sources

1. Denmark: Danish Data Health Registries (DK-DHR)
2. Finland: Finnish Care Register for Health Care (FinOMOP-THL)
3. Spain: The Information System for Research in Primary Care (SIDIAP)
4. The Netherlands: Integrated Primary Care Information (IPCI)
5. The United Kingdom: Clinical Practice Research Datalink GOLD (CPRD GOLD)

Study size

No sample size has been calculated, as this is an exploratory study which will not test a specific hypothesis. Based on a preliminary feasibility assessment, the expected number of person-counts for PMOS in the data sources included in this study range from 11,500 (IPCI) to 72,100 (CPRD GOLD).

Statistical analysis

Annual incidence rates will be described as the number of individuals with PMOS per 100,000 person-years among females for each calendar year. Incidence rates will be given together with 95% Poisson confidence intervals.

For patient characterisation, pre-defined comorbidities will be assessed at any time prior to the index date $[-\infty, 0]$, and pre-defined non-infertility drug treatments will be assessed as prescription or dispensation records within the time windows $[0, 30]$ days and $[31, 365]$ days relative to the index date.

The incidence of health outcomes will be provided as annual incidence rates and cumulative incidence. Annual incidence rates will be described as the number of individuals with a first occurrence of the outcome of interest after the first PMOS diagnosis per 100,000 person-years among PMOS females for each calendar year. Incidence rates will be provided together with the 95% Poisson confidence intervals. Cumulative incidence will be estimated separately for each outcome using the Kaplan-Meier estimator and will be presented at 5 and 10 years following the first recorded PMOS diagnosis, with corresponding 95% confidence intervals.

A minimum cell count of 5 will be used when reporting the results, with any smaller count reported as " <5 " and zero counts as "0".

4. AMENDMENTS AND UPDATES

None.

5. MILESTONES

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

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

6. RATIONALE AND BACKGROUND

Polyendocrine metabolic ovarian syndrome (PMOS), previously named polycystic ovary syndrome (PCOS), is a common, heterogeneous endocrine disorder affecting females.[1] In adults, PMOS is diagnosed, according to the Rotterdam criteria, when at least two of the following criteria are present: (1) oligo-anovulation, (2) clinical or biochemical hyperandrogenism, and (3) polycystic ovaries on ultrasound or elevated anti-Müllerian hormone.[2, 3] Based on these diagnostic criteria, a meta-analysis of 29 studies published between 1999 and 2024 estimated the pooled prevalence of PMOS at 12.1% globally.[4] In the adolescent population aged 10–19 years, diagnosis is based on the presence of the first two criteria.[2, 5]

PMOS is characterised by substantial heterogeneity in its clinical presentation, diagnosis, and management across healthcare settings. Management is individualised and primarily aims to alleviate symptoms while preventing long-term complications. Clinical manifestations span multiple domains, including metabolic (obesity, dysglycaemia, type 2 diabetes, hypertension, dyslipidaemia, metabolic dysfunction-associated steatotic liver disease, cardiovascular disease, and obstructive sleep apnoea), reproductive (ovulatory dysfunction, irregular menstrual cycles, infertility, pregnancy complications, and endometrial cancer), psychological (depression, anxiety, impaired quality of life, and eating disorders), and dermatological (acne, alopecia, and hirsutism).

Despite its high prevalence and broad clinical impact, pharmacoepidemiologic evidence gaps remain regarding real-world trajectories from symptom onset to diagnosis, treatment initiation, and subsequent outcomes. Diagnostic delays are frequently reported in PMOS, with prolonged time between initial symptoms and formal diagnosis potentially influencing treatment choices, timing of therapy initiation, and opportunities for symptom improvement.

A clear understanding of real-world pharmacological management is needed to characterise how PMOS is treated in routine clinical practice. However, evidence describing the use of pharmacological therapies following PMOS diagnosis remains limited. In addition, there is limited evidence on the incidence of reproductive, mental health, and cardiometabolic outcomes among individuals with PMOS. Describing medication use following diagnosis and the occurrence of these outcomes will provide important information for future research on the management and long-term burden of PMOS.

7. RESEARCH QUESTION AND OBJECTIVES

Research questions

1. How has the incidence of PMOS diagnosis changed over time?
2. What are the characteristics of individuals with PMOS at the time of diagnosis?
3. What non-infertility drug treatments are commonly used in individuals with PMOS following diagnosis?
4. What is the incidence of infertility, cardiovascular, and mental health outcomes in individuals with PMOS?

Research objectives

The objectives of this study are:

1. To estimate the incidence of recorded PMOS diagnosis over time.
2. To characterise individuals with PMOS at the time of recorded diagnosis in terms of age and pre-defined comorbidities.
3. To characterise individuals with PMOS in terms of pre-defined non-infertility drug treatments received following diagnosis.
4. To estimate the incidence rate of infertility, cardiovascular, and mental health outcomes in individuals with PMOS.
5. To estimate the 5-year and 10-year cumulative incidence of infertility, cardiovascular, and mental health outcomes in individuals with PMOS.

8. RESEARCH METHODS

8.1. Study design

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

- A population-level descriptive disease epidemiology study, which will be conducted to address Objective 1 (incidence of recorded PMOS diagnosis), Objective 4 (incidence rates of health outcomes in Individuals with PMOS), and Objective 5 (cumulative incidence of health outcomes in Individuals with PMOS)
- A patient-level characterisation study, which will be conducted to address Objective 2 (characterisation of Individuals with PMOS on comorbidities) and Objective 3 (characterisation of Individuals with PMOS on non-infertility drug treatments)

Detailed study designs for Objectives 1–5 are illustrated in [Figures 1–4](#).

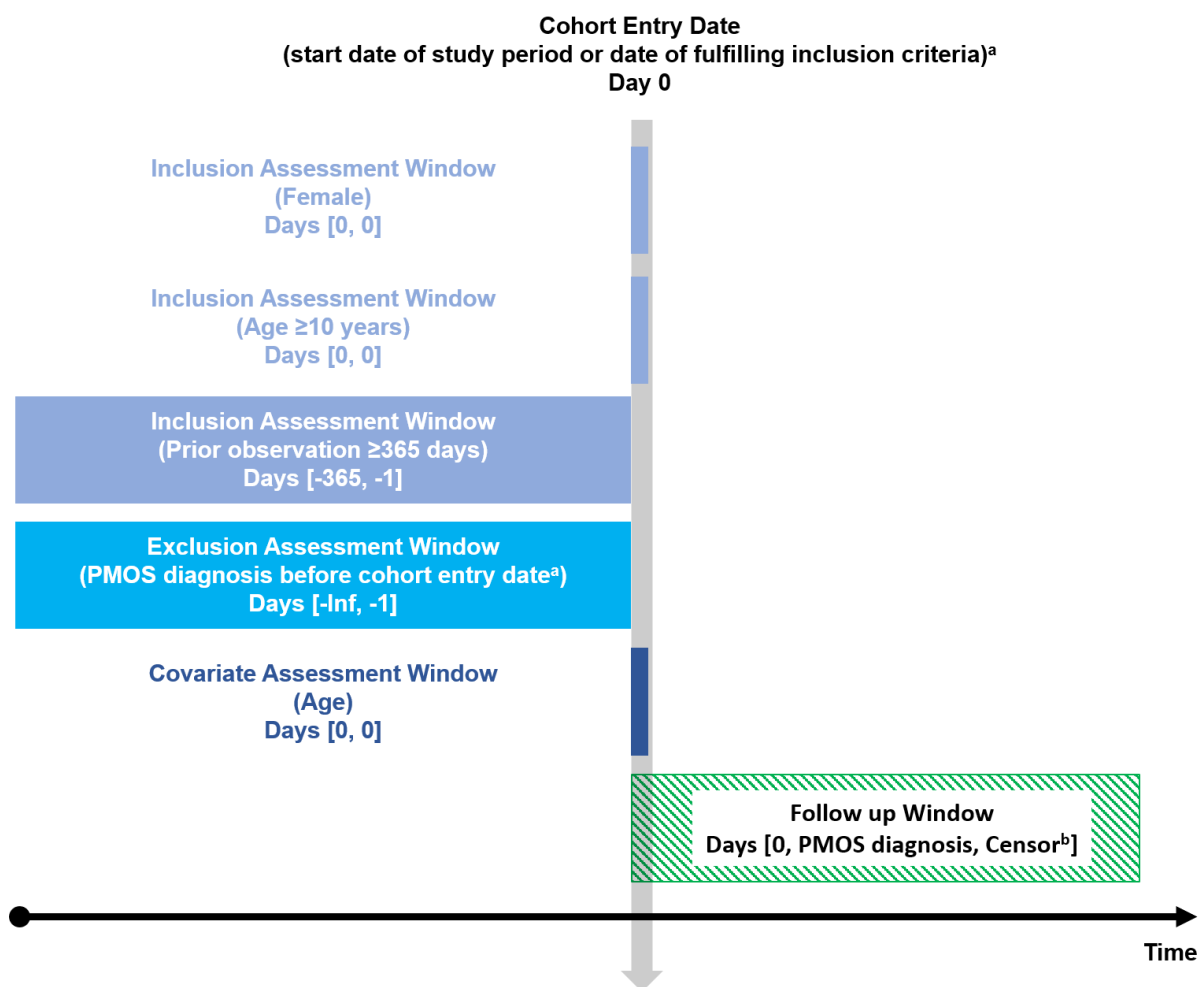


Figure 1. Graphical depiction of the study design for Objective 1 (incidence of recorded PMOS diagnosis).

- Cohort entry date will be the latest of: the start of the study period (01/01/2010), having at least 365 days of prior observation, or fulfilling the age criteria of at least 10 years old.
- Individuals will be censored at death, end of observation, or end of the study period (31/12/2024).

PMOS = polyendocrine metabolic ovarian syndrome.

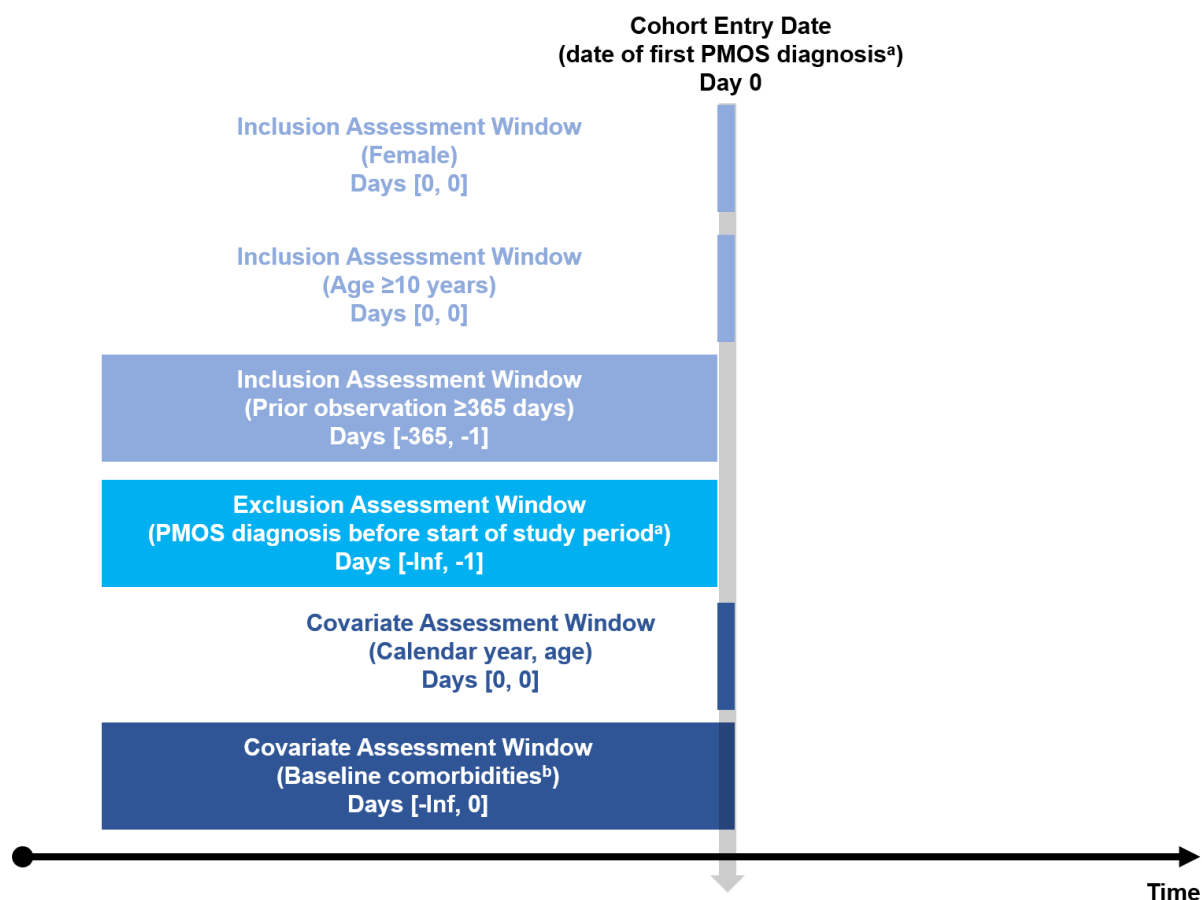


Figure 2. Graphical depiction of the study design for Objective 2 (characterisation of individuals with PMOS on comorbidities).

- a. Cohort entry date will be the date of first PMOS diagnosis during the study period (01/01/2010 to 31/12/2024).
- b. Baseline comorbidities include cardiometabolic/vascular conditions (obesity, dyslipidaemia, hypertension, type 2 diabetes, metabolic dysfunction-associated steatotic liver disease, obstructive sleep apnoea, myocardial infarction, stroke, and VTE), hyperandrogenic/reproductive conditions (acne, hirsutism, alopecia, oligomenorrhoea/amenorrhoea, and infertility), and mental health conditions (anxiety, depression, and eating disorder).

PMOS = polyendocrine metabolic ovarian syndrome, VTE = venous thromboembolism.

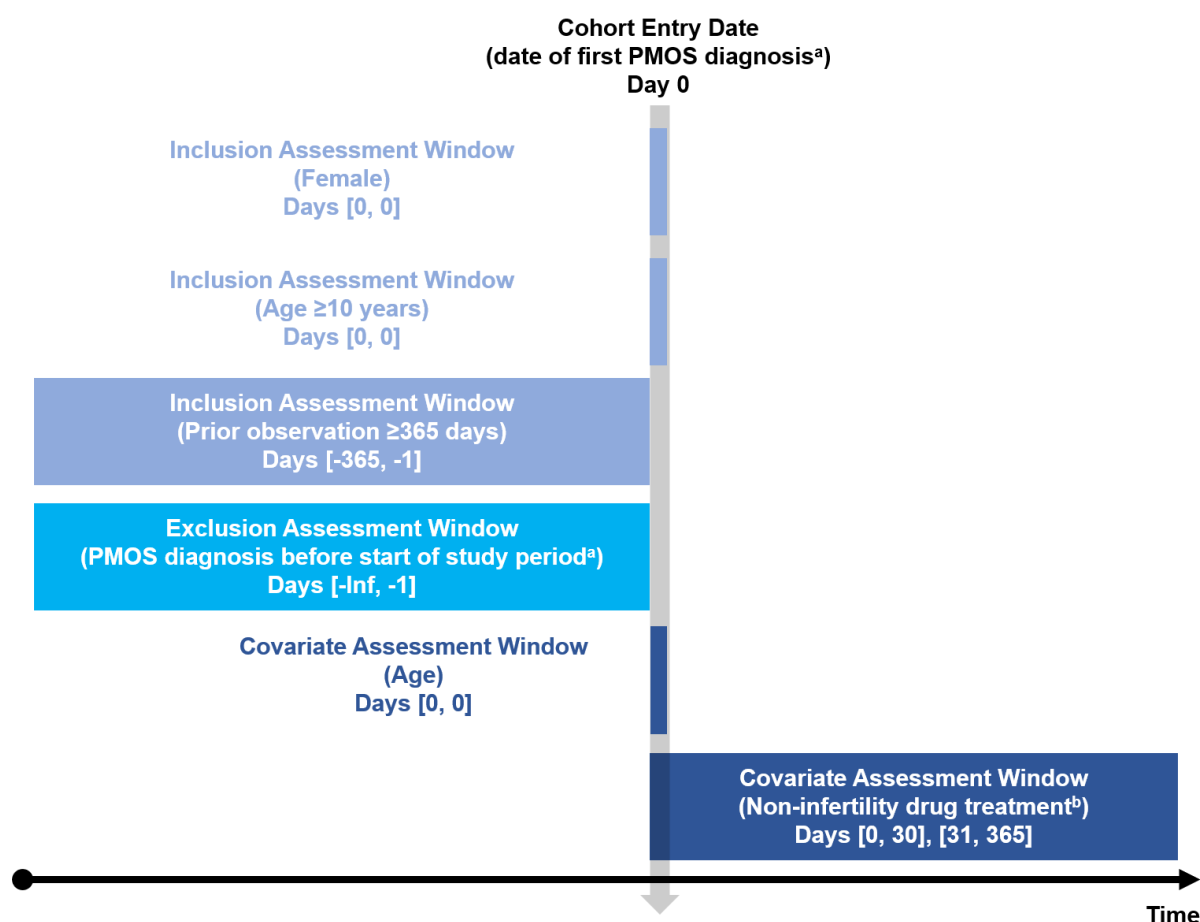


Figure 3. Graphical depiction of the study design for Objective 3 (characterisation of Individuals with PMOS on non-infertility drug treatments).

- a. Cohort entry date will be the date of first PMOS diagnosis during the period from the start of the study period (01/01/2010) to the earliest of the end of the study period (31/12/2024) or 365 days before the end of data source observation.
- b. Non-infertility drug treatments include combined oral contraceptives, metformin, GLP-1 receptor agonists (lixisenatide, albiglutide, liraglutide, semaglutide, dulaglutide, exenatide), other antidiabetic medication (sulfonylureas, thiazolidinediones, DPP-4 inhibitors, dual GIP/GLP-1 receptor agonist, SGLT2 inhibitors), insulin, anti-androgen medications (spironolactone, flutamide, bicalutamide, finasteride, cyproterone), orlistat, and inositol.

DPP-4 = dipeptidyl peptidase-4, GIP = glucose-dependent insulintropic polypeptide, GLP-1 = glucagon-like peptide-1, PMOS = polyendocrine metabolic ovarian syndrome, SGLT2 = sodium-glucose co-transporter 2.

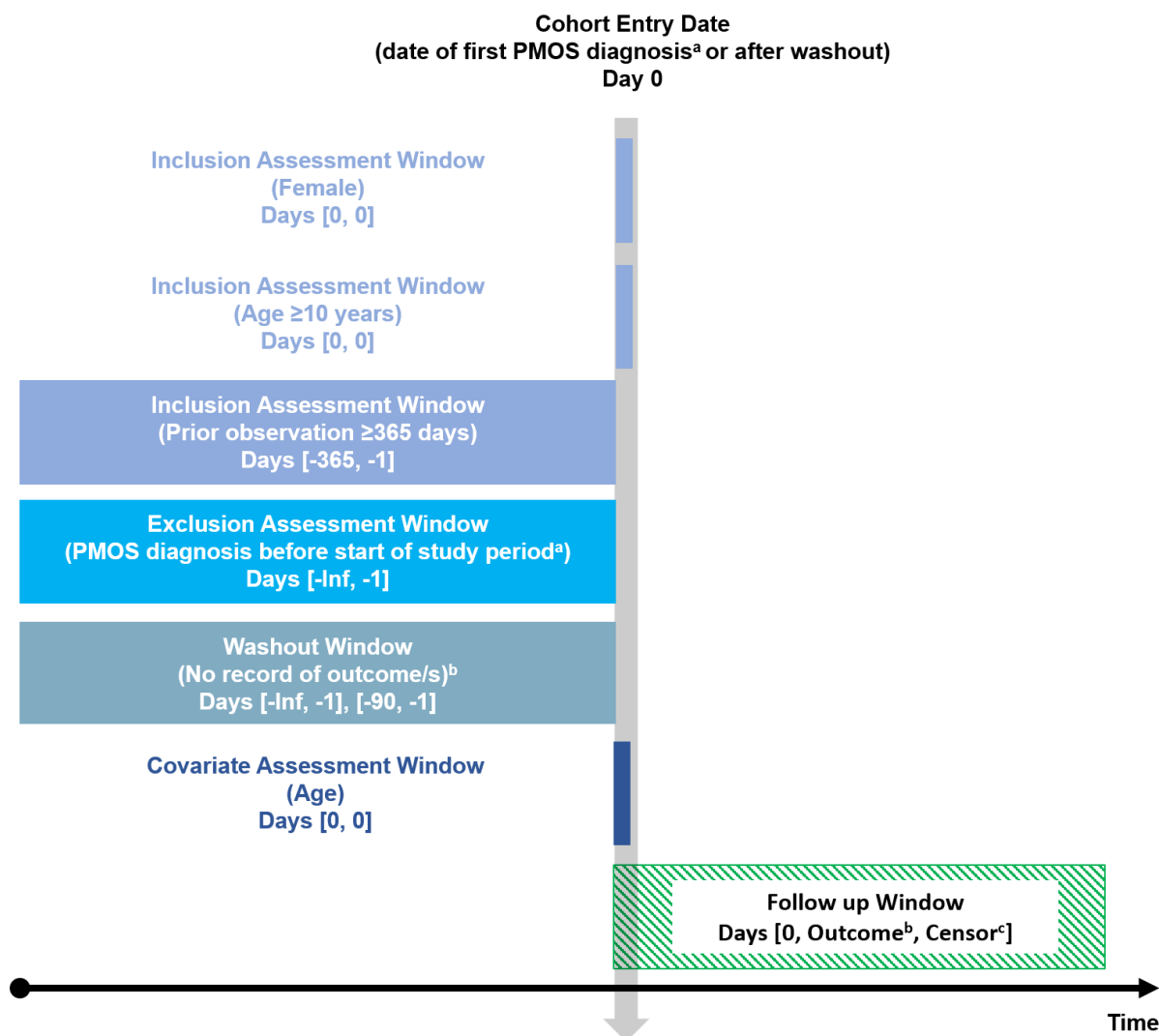


Figure 4. Graphical depiction of the study design for Objective 4 (incidence rates of health outcomes in individuals with PMOS) and Objective 5 (cumulative incidence of health outcomes in individuals with PMOS).

- Cohort entry date will be the date of first PMOS diagnosis during the study period (01/01/2010 to 31/12/2024).
- Outcomes of interest include infertility, cardiovascular outcomes (hypertension, acute myocardial infarction, ischaemic stroke, stroke, VTE), and mental outcomes (anxiety, depression). For all outcomes, only the first occurrence during the follow-up period will be considered. For hypertension, infertility, and mental outcomes, given that they are chronic conditions in nature, a washout period of any time before the first PMOS diagnosis during the study period will be applied. For cardiovascular outcomes other than hypertension (acute myocardial infarction, ischaemic stroke, stroke, VTE), given they are recurrent condition in nature, a washout period of 90 days will be applied.
- Individuals will be censored at death, end of observation, or end of the study period (31/12/2024).

PMOS = polyendocrine metabolic ovarian syndrome, VTE = venous thromboembolism.

8.2. Follow-up

For Objective 1 (incidence of recorded PMOS diagnoses), index date will be the latest of: the study start date (01/01/2010), having at least 365 days of prior observation, or fulfilling the age criteria of at least 10 years old. Follow-up will end on the earliest of: death, end of the observation period (the latest available data), the study end date (31/12/2024), or first occurrence of outcome of interest (record of PMOS diagnosis).

For Objective 2 (characterisation of individuals with PMOS on comorbidities), index date will be the date of first PMOS diagnosis. Age and comorbidities will be assessed on or before index date, and therefore there will be no follow-up period.

For Objective 3 (characterisation of individuals with PMOS on non-infertility drug treatments), index date will be the date of first PMOS diagnosis. Follow-up will end on the earliest of: death, end of the observation period (the latest available data), the study end date (31/12/2024), or 365 days after PMOS diagnosis.

For Objective 4 (incidence rates of health outcomes in Individuals with PMOS) and Objective 5 (cumulative incidence of health outcomes in Individuals with PMOS), index date will be the date of first PMOS diagnosis. Follow-up will end on the earliest of: death, end of the observation period (the latest available data), the study end date (31/12/2024), or occurrence of the outcome of interest.

Incidence requires an appropriate denominator population and contributed observation time to first be identified. Study participants in the denominator population will begin contributing person-time-at-risk from the index date. Entry into and exit from the denominator population will depend on the outcome of interest and the washout period for each outcome, as defined.

For Objective 1 (incidence of recorded PMOS diagnosis), only the first PMOS diagnosis during the study period will be considered as an outcome, and only one outcome event (PMOS diagnosis) will be allowed.

An example of entry and exit into the denominator population is shown in [Figure 5](#). In this example, person ID 1 already has sufficient prior history before the study start date, and the observation period ends after the study end date, so this person will contribute during the complete study period. Person IDs 2 and 4 enter the study only when they have sufficient prior history. Person ID 3 leaves when exiting the data source (the end of the observation period). Lastly, person ID 5 has two observation periods in the data source. The first period contributes time from the study start until the end of the observation period. The second starts contributing time again once sufficient prior history is reached and exits at the study end date.

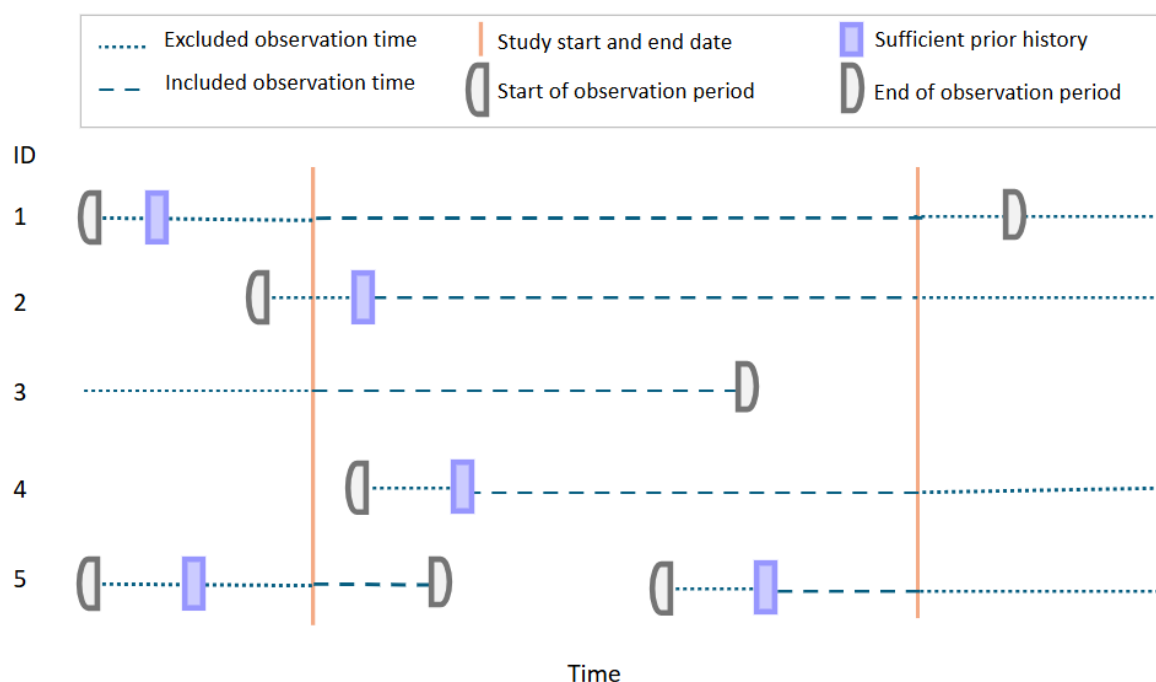


Figure 5. Included observation time for the denominator population (Objective 1).

For Objective 4 (incidence rates of health outcomes in Individuals with PMOS) and Objective 5 (cumulative incidence of health outcomes in Individuals with PMOS), a washout period of 90 days will be applied for cardiovascular outcomes other than hypertension (acute myocardial infarction, stroke, VTE). Similarly, a washout period of any time before will be applied for hypertension, mental outcomes (anxiety, depression), and infertility. Entry into and exit from the denominator population should take a washout period into account.

In addition to the entry and exit rule explained in Figure 5, a washout period after outcome of interest will be considered, as shown in Figure 6. Person ID 1 has an outcome of interest before the study period. Observation time will only be included for person ID 1 once the washout period has been completed.

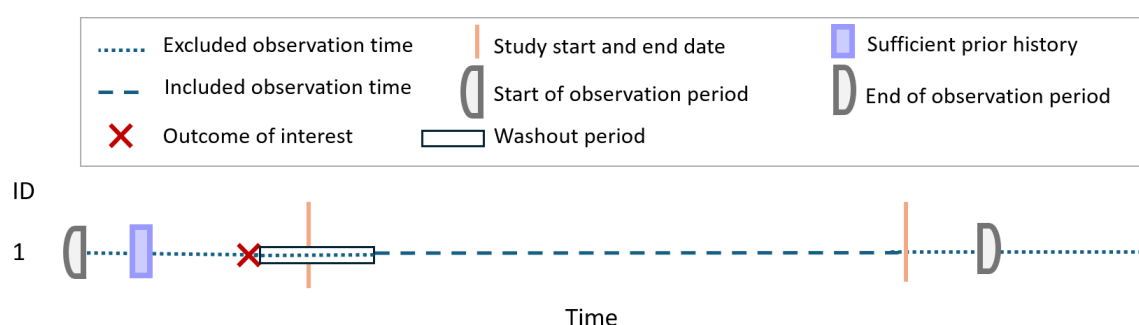


Figure 6. Included observation time for the denominator population with a washout period (Objectives 4 and 5).

8.3. Study population with inclusion and exclusion criteria

Objective 1 (incidence of recorded PMOS diagnosis)

Inclusion criteria

- Present in the period 01/01/2010 to 31/12/2024 (or the latest available date)
- Female individuals
- Aged ≥ 10 years
- Minimum 365 days of prior observation

Exclusion criteria

- Record of PMOS before the start of the study period (01/01/2010)

Objective 2 (characterisation of individuals with PMOS on comorbidities), 4 (incidence of health outcomes in individuals with PMOS), and 5 (cumulative incidence of health outcomes in individuals with PMOS)

Inclusion criteria

- Record of PMOS diagnosis in the period 01/01/2010 to 31/12/2024 (or the latest available date)
- Female individuals
- Aged ≥ 10 years
- Minimum 365 days of prior observation

Exclusion criteria

- Record of PMOS before the start of the study period (01/01/2010)

Objective 3 (characterisation of individuals with PMOS on non-infertility drug treatments)

Inclusion criteria

- Record of PMOS diagnosis in the period 01/01/2010 to 31/12/2024 (or 365 days before the end of data source observation)
- Female individuals
- Aged ≥ 10 years
- Minimum 365 days of prior observation

Exclusion criteria

- Record of PMOS before the start of the study period (01/01/2010)

8.4. Study setting and data sources

This study will be conducted using routinely collected data from five nationwide linked registries or primary care data sources in the DARWIN EU® Network of Data Partners from five European countries ([Table 1](#)). All data were *a priori* mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).

Table 1. Data sources.

Country	Name of Data source	Health Care setting	Type of Data	Number of active individuals	Calendar period covered by each data source	Contributing to
Denmark	Danish Data Health Registries (DK-DHR)	Secondary care specialists, hospital inpatient care, community pharmacy	EHR, registries	6.02M	01/1995–10/2025	All objectives
Finland	Finnish Care Register for Health Care (FinOMOP-THL)	Primary care GP and specialists, secondary care specialists, hospital inpatient care	EHR, registries	6.15M	01/2015–12/2025	All objectives
Spain	The Information System for Research in Primary Care (SIDIAP)	Primary care GP, hospital inpatient care	EHR	6.07M	01/2006–12/2024	All objectives
The Netherlands	Integrated Primary Care Information (IPCI)	Primary care GP	EHR	1.34M	01/2006–12/2025	All objectives
The United Kingdom	Clinical Practice Research Datalink GOLD (CPRD GOLD)	Primary care GP	EHR	2.37M	11/1987–10/2025	All objectives

EHR = electronic health records; GP = general practitioner; M = million.

Data sources selection

These data sources fulfil the criteria required in terms of data quality, completeness, timeliness, and representativeness for population-level descriptive epidemiology study while covering different regions of Europe ([Annex II](#)).

8.5. Study period

The study period will be from 01/01/2010 to 31/12/2024.

Due to data availability, the study period for FinOMOP-THL will start from 01/2016 to ensure one year of data availability in the data source.

8.6. Variables

8.6.1. Outcome

Objective 1 (incidence of recorded PMOS diagnosis)

The outcome for this objective will be PMOS, which will be defined using condition and/or observation codes.

Objective 4 (incidence rates of health outcomes in individuals with PMOS) and 5 (cumulative incidence of health outcomes in individuals with PMOS)

The outcomes for this objective will be as follows:

- Infertility
- Cardiovascular outcomes:
 - Hypertension, including hypertensive disorders of pregnancy, such as gestational hypertension, pre-eclampsia, and eclampsia
 - Acute myocardial infarction
 - Ischaemic stroke
 - Stroke (including ischaemic stroke, haemorrhagic stroke, unspecified stroke)
 - Venous thromboembolism (VTE)
- Mental health outcomes:
 - Anxiety
 - Depression

Hypertension, acute myocardial infarction, ischaemic stroke, stroke, VTE, anxiety, and depression will be defined using condition and/or observation codes. Infertility will be identified by the earliest occurrence of at least one of the following: (i) a diagnosis code for female infertility, including broad female infertility concepts and anovulatory infertility (the subtype most commonly associated with PMOS), excluding diagnosis codes related exclusively to non-anovulatory causes of infertility; or (ii) a prescription for pharmacological treatments commonly used for anovulatory infertility associated with PMOS, including letrozole, clomiphene, or gonadotropins; or (iii) a procedure code for assisted reproductive technologies or surgical interventions, including in vitro fertilisation or ovarian drilling, based on the recommendations of the 2023 International Evidence-based Guideline for the Assessment and Management of Polycystic Ovary Syndrome.[2]

All outcomes will be defined as incident outcomes.

The preliminary concept sets used for the identification of outcomes are described in [Annex IV](#). These codes will be refined during 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.

8.6.2. Covariates

All objectives

The following covariate, assessed on the index date, will be used for stratification:

- Age groups:
 - 10–17 years

- 18–29 years
- 30–39 years
- 40–49 years
- 50 years and above

Objective 2 (characterisation of individuals with PMOS on comorbidities)

- Age, assessed at index date
- Calendar year, assessed at index date
- Pre-defined comorbidities representing the most common clinical manifestations and disease burden associated with PMOS [2] assessed any time prior to index date:
 - Cardiometabolic/vascular:
 - Obesity
 - Dyslipidaemia
 - Hypertension
 - Type 2 diabetes
 - Metabolic dysfunction-associated steatotic liver disease
 - Obstructive sleep apnoea
 - Myocardial infarction
 - Stroke
 - VTE
 - Hyperandrogenic/reproductive:
 - Acne
 - Hirsutism
 - Alopecia
 - Oligomenorrhoea/amenorrhoea
 - Infertility
 - Mental health:
 - Anxiety
 - Depression
 - Eating disorder, including anorexia nervosa, bulimia nervosa, binge eating disorder, and avoidant restrictive food intake disorder.

Objective 3 (characterisation of PMOS on non-infertility drug treatments)

Pre-defined non-infertility drug treatments prescription or dispensation for the management of PMOS[2] assessed on the time windows [0,30] days and [31, 365] days:

- Combined oral contraceptives.
- Metformin.

- Glucagon-like peptide-1 (GLP-1) receptor agonists (lixisenatide, albiglutide, liraglutide, semaglutide, dulaglutide, exenatide). They will be described both as a drug class and as individual active ingredients.
- Dual glucose-dependent insulinotropic polypeptide (GIP)/GLP-1 receptor agonist.
- Other antidiabetic medications (sulfonylureas, thiazolidinediones, dipeptidyl peptidase-4 (DPP-4) inhibitors, sodium-glucose co-transporter 2 (SGLT2) inhibitors). They will be described separately for each class.
- Insulin.
- Anti-androgen medications (spironolactone, flutamide, bicalutamide, finasteride, cyproterone). They will be described both as a drug class and as individual active ingredients.
- Orlistat.
- Inositol.

In addition to the above-mentioned pre-defined non-infertility drug treatments for the management of PMOS, large-scale characterisation on treatment use in individuals with PMOS will be conducted as a standard practice during the phenotyping process. If commonly used treatments are identified that were not included in the pre-defined list, they may be considered for inclusion in the treatment characterisation analyses.

The preliminary concept sets used for the identification of covariates are described in [Annex IV](#). These codes will be refined during 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.

8.7. Study size

No sample size has been calculated, as this is a descriptive disease epidemiology study which will not test a specific hypothesis. In addition, already collected available data will be used to estimate incidence of PMOS diagnosis. Thus, the sample size is driven by the availability of data for individuals with a PMOS diagnosis. Based on a preliminary feasibility assessment, the expected number of person-counts for PMOS in the data sources included in this study range from 11,500 (IPCI) to 72,100 (CPRD GOLD).

8.8. Analysis

8.8.1. Federated network analyses

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

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

8.8.2. Data privacy protection

The data partners will locally execute the analytics against the OMOP CDM in R Studio and review and approve the default aggregated results. They will then be made available to the Principal Investigators and study team in a secure online repository of the Data Transfer Zone (DTZ). 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 (Objectives 1–5) will be conducted as overall and stratified by age group.

Objective 1

The incidence of PMOS will be calculated based on OMOP CDM mapped data using the *IncidencePrevalence* R package, developed by DARWIN EU®.[6]

Annual incidence rates will be described as the number of individuals with the outcome of interest per 100,000 person-years among females for each calendar year. As PMOS is considered a chronic condition, only the first occurrence of the outcome will be counted, and individuals will not re-enter the at-risk population following diagnosis. Participants without outcome exposure will contribute time-at-risk continuously until censoring. Time-at-risk of individuals who die will be censored at the time of death. Similarly, time-at-risk of individuals who are lost to follow-up will be censored at the time of loss to follow-up (i.e. last contact). Incidence rates will be given together with 95% Poisson confidence intervals.

Estimates will only be calculated for study periods with complete calendar-year data available in each data source.

Objectives 2 and 3

Patient characteristics of individuals with a diagnosis of PMOS will be summarised based on OMOP CDM mapped data using the *CohortCharacteristics* R package, developed by DARWIN EU®.[7]

Patient characterisation will be provided by pre-defined comorbidities and pre-defined non-infertility drug treatments. Pre-defined comorbidities will be assessed at any time prior to the index date [-Inf, 0] days. Pre-defined non-infertility drug treatments will be assessed as prescription or dispensation records within the time windows [0, 30] days and [31, 365] days relative to the index date.

A list of pre-defined covariates of interest is given in [Section 8.6.2](#).

Age at the index date will also be described for each calendar year using the median and interquartile range (IQR).

As an exploratory analysis, the median time from the first recorded occurrence of each pre-defined comorbidity of interest to the first recorded PMOS diagnosis will be estimated, to characterise the interval between the recorded onset of each comorbidity and the initial recorded diagnosis of PMOS.

Objective 4

The incidence of infertility, hypertension, acute myocardial infarction, ischaemic stroke, stroke, VTE, anxiety, and depression will be calculated based on OMOP CDM mapped data using the *IncidencePrevalence* R package, developed by DARWIN EU®.[6]

Annual incidence rates will be described as the number of individuals with a first occurrence of the outcome of interest after the first diagnosis of PMOS per 100,000 person-years among PMOS females for each calendar year. Incident outcomes will be identified using outcome-specific washout periods: a 90-day washout period before the index date will be applied for cardiovascular outcomes other than hypertension (acute myocardial infarction, ischaemic stroke, stroke, and VTE), whereas all available prior observation time before the index date will be used as the washout period for hypertension, infertility, anxiety, and depression. Study participants who enter the denominator population will contribute person-time from the start of follow-up (PMOS diagnosis) until the earliest of the following: occurrence of the outcome, death, end of observation, or end of the study period. As only the first occurrence of each outcome after the

PMOS diagnosis will be considered, individuals will not re-enter the at-risk population following occurrence of the outcome. Participants without outcome exposure will contribute time-at-risk continuously until censoring. Time-at-risk of individuals who die will be censored at the time of death. Similarly, time-at-risk of individuals who are lost to follow-up will be censored at the time of loss to follow-up (i.e. last contact). Incidence rates will be given together with 95% Poisson confidence intervals.

Estimates will only be calculated for study periods with complete calendar-year data available in each data source.

Objective 5

The cumulative incidence of infertility, hypertension, acute myocardial infarction, ischaemic stroke, stroke, VTE, anxiety, and depression will be calculated based on OMOP CDM mapped data using the *CohortSurvival* R package, developed by DARWIN EU®.

The 5-year and 10-year cumulative incidence of each outcome will be estimated among females with a diagnosis of PMOS. Participants will enter the risk set on the date of their first recorded PMOS diagnosis. Incident outcomes will be defined using the same outcome-specific washout periods as described above: a 90-day washout period before the index date for cardiovascular outcomes other than hypertension and all available prior observation time before the index date for hypertension, infertility, anxiety, and depression. Only the first occurrence of each outcome after PMOS diagnosis will be considered. Follow-up will continue until the earliest of the following: occurrence of the outcome of interest, death, end of observation, or end of the study period. Individuals who do not experience the outcome will be censored at the earliest of death, end of observation, or end of the study period.

Cumulative incidence will be estimated separately for each outcome using the Kaplan-Meier estimator and will be presented at 5 and 10 years following the first recorded PMOS diagnosis with corresponding 95% confidence intervals.

Sensitivity analysis

Descriptions of sensitivity analyses are presented by means of **Table 2**.

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

	What is being varied? How?	Why? (What do you expect to learn?)	Strengths of the sensitivity analysis compared to the primary	Limitations of the sensitivity analysis compared to the primary
Objective 4 (incidence rates of health outcomes in Individuals with PMOS) and Objective 5 (cumulative incidence of health outcomes in Individuals with PMOS)	<u>Main analysis:</u> Include all females with PMOS regardless of prior history of the outcome before the index date (first recorded PMOS diagnosis), with a washout period of 90 days applied for cardiovascular outcome other than hypertension, and a washout period of any time before for hypertension, mental outcome, and infertility.	To assess whether the estimated incidence is influenced by including individuals with a prior history of the outcome before the index date.	Restricting the analysis to individuals without a prior history of the outcome provides estimates that are more readily interpreted as the risk of a first-ever occurrence of the outcome following PMOS diagnosis.	Reduces the study population and the number of events. May also reduce generalisability.

	What is being varied? How?	Why? (What do you expect to learn?)	Strengths of the sensitivity analysis compared to the primary	Limitations of the sensitivity analysis compared to the primary
	<u>Sensitivity analysis:</u> Exclude females with a history of the outcome of interest before the index date (first recorded PMOS diagnosis)			
Objective 1 (incidence of recorded PMOS diagnosis)	<u>Main analysis:</u> Age groups used: • 10–17 years • 18–29 years • 30–39 years • 40–49 years • 50 years and above <u>Sensitivity analysis:</u> 10-year band age groups: • 10–19 years • 20–29 years • 30–39 years • 40–49 years • 50 years and above	To facilitate comparison with data sources using standard 10-year age bands.	Facilitates age standardisation of estimates.	Adolescents and adults are not clearly distinguished in the age stratification.
Objective 4 (incidence rates of health outcomes in Individuals with PMOS) and Objective 5 (cumulative incidence of health outcomes in Individuals with PMOS)	<u>Main analysis:</u> Age groups used: • 10–17 years • 18–29 years • 30–39 years • 40–49 years • 50 years and above <u>Sensitivity analysis:</u> 10-year band age groups: • 10–19 years • 20–29 years • 30–39 years • 40–49 years	To facilitate comparison with data sources using standard 10-year age bands.	Facilitates age standardisation of estimates.	Adolescents and adults are not clearly distinguished in the age stratification.

	What is being varied? How?	Why? (What do you expect to learn?)	Strengths of the sensitivity analysis compared to the primary	Limitations of the sensitivity analysis compared to the primary
	50 years and above			
Objective 4 (incidence rates of health outcomes in individuals with PMOS)	<u>Main analysis:</u> Incidence rates are estimated using all available follow-up time after the index date (first recorded PMOS diagnosis). <u>Sensitivity analysis:</u> Restrict follow-up to the first five years after the index date when calculating person-time for the incidence rate denominator.	To assess whether observed time trends are influenced by time since diagnosis across calendar years, given that individuals diagnosed earlier contribute more long-term follow-up than those diagnosed later.	Makes the population more comparable across calendar years and reduces distortion from differing time since diagnosis. Helps determine whether trends are driven by follow-up structure rather than true changes in risk.	Reduces person-time and number of events, so estimates may be less precise. It may under-represent outcomes that occur later after diagnosis and may not reflect the full burden of disease over time.

PMOS = Polyendocrine metabolic ovarian syndrome.

8.8.4. Output

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

- Table 1. Number of individuals with polyendocrine metabolic ovarian syndrome (PMOS) during the study period, by study objective and data source.
- Table 2. Characterisation of polyendocrine metabolic ovarian syndrome (PMOS) individuals on comorbidities.
- Table 3. Characterisation of polyendocrine metabolic ovarian syndrome (PMOS) individuals on comorbidities, stratified by age group.
- Table 4. Characterisation of polyendocrine metabolic ovarian syndrome (PMOS) individuals on non-infertility drug treatments.
- Table 5. Characterisation of polyendocrine metabolic ovarian syndrome (PMOS) individuals on non-infertility drug treatments, stratified by age group.
- Figure 1. Annual incidence of recorded polyendocrine metabolic ovarian syndrome (PMOS) diagnosis, overall and stratified by age group.
- Figure 2–9. Annual incidence of health outcomes in polyendocrine metabolic ovarian syndrome (PMOS) individuals, overall and stratified by age group.
- Figure 10–17. 5-year cumulative incidence of health outcomes in polyendocrine metabolic ovarian syndrome (PMOS) individuals, overall and stratified by age.
- Figure 18–25. 10-year cumulative incidence of health outcomes in polyendocrine metabolic ovarian syndrome (PMOS) individuals, overall stratified by age group.

Mock tables and figures are available in [Annex V](#).

An interactive dashboard will be generated by incorporating all the results (tables and figures) included in the PDF report mentioned above.

8.9. Evidence synthesis

Results from analyses described in [Section 8.8.3](#) will be presented separately for each data source. No meta-analysis of results will be conducted.

9. STRENGTHS AND LIMITATIONS

The study will be informed by routinely collected health care data, and so, data quality issues must be considered.

Delayed diagnosis of PMOS is common in routine clinical practice and may affect the accurate classification of cardiometabolic disorders and cardiovascular outcomes.[8] Cardiovascular and metabolic conditions may be diagnosed before the formal diagnosis of PMOS and will therefore be classified as baseline comorbidities in Objective 2 rather than incident health outcomes in Objective 4. Consequently, the incidence of these outcomes following PMOS diagnosis may be underestimated.

In Objective 4, depression and anxiety will be evaluated as chronic conditions, with only the first recorded event after PMOS diagnosis being considered. Recurrent episodes will not be assessed. Although depression and anxiety are characterised by periods of remission and recurrence, reliably distinguishing resolved episodes from recurrent events using routinely collected healthcare data is challenging. Therefore, only the first occurrence of each mental health outcome following PMOS diagnosis will be included in the analyses.

Infertility will also be evaluated as a health outcome in Objective 4. However, there is uncertainty regarding the completeness of infertility ascertainment in routinely collected healthcare data. In this study, infertility will be identified using infertility-related diagnoses, medications, and procedures as proxy indicators. Feasibility assessments suggest that sufficient data are available to support an exploratory descriptive analysis using these proxy indicators. Infertility-related diagnoses, medications, and procedures will be included to provide a comprehensive representation of infertility-related healthcare encounters within the available data.

The documentation of comorbidities required for population characterisation may vary across data sources. To ensure consistent definitions of comorbidities, a careful phenotyping process will be conducted for the selected comorbidities using the DARWIN EU® - Population demographics and disease frequency across the DARWIN EU® Network study ([EUPAS1000000962](#)).

Additionally, the results estimated from this study will only reflect the populations from the included data sources. Electronic health records have certain inherent limitations because they were collected for clinical purposes rather than primarily for research use. Consequently, using five primary care or nationwide registries data sources from Denmark, Finland, Spain, the Netherlands, and the United Kingdom limits generalisability to those countries.

10. REFERENCES

1. Teede HJ, Khomami MB, Morman R, Laven JSE, Joham AE, Costello MF, et al. Polyendocrine metabolic ovarian syndrome, the new name for polycystic ovary syndrome: a multistep global consensus process. *Lancet*. 2026;407(10545):2329-39.
2. Teede HJ, Tay CT, Laven JJE, Dokras A, Moran LJ, Piltonen TT, et al. Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome. *Eur J Endocrinol*. 2023;189(2):G43-g64.
3. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertil Steril*. 2004;81(1):19-25.
4. Neven ACH, Forslund M, Ranasinha S, Sethi P, Dhungana RR, Mousa A, et al. Prevalence of polycystic ovary syndrome: a global and regional systematic review and meta-analysis. *Hum Reprod Update*. 2026;32(3):277-312.
5. Peña AS, Witchel SF, Boivin J, Burgert TS, Ee C, Hoeger KM, et al. International evidence-based recommendations for polycystic ovary syndrome in adolescents. *BMC Med*. 2025;23(1):151.
6. Raventós B, Català M, Du M, Guo Y, Black A, Inberg G, et al. IncidencePrevalence: An R package to calculate population-level incidence rates and prevalence using the OMOP common data model. *Pharmacoepidemiol Drug Saf*. 2024;33(1):e5717.
7. Du M, Prats-Urbe 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. *Eur J Epidemiol*. 2026;41(4):507-17.
8. Gibson-Helm M, Teede H, Dunaif A, Dokras A. Delayed Diagnosis and a Lack of Information Associated With Dissatisfaction in Women With Polycystic Ovary Syndrome. *J Clin Endocrinol Metab*. 2017;102(2):604-12.

11. ANNEXES

ANNEX I. Description of data sources

Danish Data Health Registries (DK-DHR)

#	Section	Description
1	Data source identification and country	DK-DHR (Danish Data Health Registries) Denmark
2	Data partner information section	Danish Medicines Agency (DKMA) Data & Data Analysis (DAD)
3	Coverage and timespan	Data collection since: 1995 Extent: Nation-wide. The data is representative of the entire Danish population.
4	Healthcare setting / type of data	Community pharmacists, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care. The following data elements are collected: diagnosis (including rare diseases and pregnancy data), hospital admissions, discharge and ICU data, Cause of death, Drug prescriptions, dispensing, vaccination and contraception, Procedures (surgical and non-surgical hospital), and Sociodemographic information (sex and age only).
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Registries, and Other. All causes of deaths, all retrieved drug prescriptions, all records of vaccinations, all hospital inpatient and outpatients contacts including disease diagnoses and hospital surgical and non-surgical procedures, histologically confirmed incident cancers, laboratory test results for the entire Danish population from 1/1/1995 onwards.
6	General representativeness	The data is representative of the entire Danish population. Healthcare is free in Denmark, so we do not expect any bias in data collection based on socio-economic status.
7	Data content / source coding	Diagnoses and causes of death are collected using the ICD-10 vocabulary. ATC and RxNorm are used for Drugs. SNOMED codes are used for Procedures.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Patients have unique identifiers used to link datasets.
9	Quality control (data source specific)	The data we have received relating to nationwide Danish Health Data registries offer an opportunity for large-scale, population-based studies with several advantages 1) Their large size improves the precision of estimates and enables the study of rare exposures and outcomes with long-term latency, 2) Inclusion of nearly all individuals in the target population ensures that the data reflect routine clinical care and all clinical segments of the source population, 3) Data are collected independently of each research study, thus minimising certain types of bias, e.g., non-response, and the influence from attention to the research question on the diagnostic process. Before the source data is sent to us, the Danish Health Data Authority runs and does comprehensive checks of the registry table data validity of the variables, breaks in data, changes in variable coding, missingness, etc. We perform checks of missingness/completeness in relation to requested variables. In essence, we are receiving a dump of a mirror of the data that is controlled by the SDS. The documentation performed by SDS is available online, in Danish primarily https://www.esundhed.dk/Dokumentation (all variables), but also in English https://sundhedsdatastyrelsen.dk/da/english/health_data_and_registers/national_health_registers
10	Linkage	There is no linkage in this data source.
11	Vital status	The Cause of Death registry (DAR) is used, the cause of death is collected using ICD-10 codes.

#	Section	Description
12	Limitations	- Lifestyle factors (such as smoking status and weight) are only complete for pregnant women. - Stillborn children will not have any records in our CDM. This means that e.g. birth weight and birth length of stillborns is not recorded. - There is no information on key socio-economic status (SES) factors, such as occupation, education, and income. - Patient contacts in primary care (visits to the GP) are not included. Consequently, the incidence of chronic diseases like Type 2 Diabetes (T2D) and asthma must be determined using drug prescriptions as a proxy.
13	Main references	Schmidt M, Schmidt SAJ, Adelborg K, Sundbøll J, Laugesen K, Ehrenstein V, Sørensen HT "The Danish health care system and epidemiological research: from health care contacts to database records." Clinical epidemiology (2019): 31372058
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111217 Website: https://sundhedsdatastyrelsen.dk/da/english/health_data_and_registers/healthdatadenmark

Finnish Care Register for Health Care (FinOMOP-THL)

#	Section	Description
1	Data source identification and country	FinOMOP-THL (Finnish Care Register for Health Care) Finland
2	Data partner information section	FinOMOP (FinOMOP) The Department of Data and Analytics
3	Coverage and timespan	Data collection since: 1998 Extent: Nation-wide. The current CDM population comprises all persons having been alive and residing in Finland since the beginning of 2011.
4	Healthcare setting / type of data	Primary care – General Practitioner, and primary care specialists (e.g. paediatricians), and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and other (specify). THL maintains health registers that cover both public and private, primary, and specialised inpatient, urgent and outpatient health care encounters in Finland, starting from 2011. The entire public sector and private inpatient encounters have been included since 2011, while private outpatient encounters, including occupational care, are included since 2020. Since 1998, the Care Register for Health Care (TerveysHilmo) has covered both public outpatient and inpatient specialized care and private inpatient care. Since 2009, the Finnish National Vaccination Register has covered all vaccinations from the public sector and from a large part of private vaccination providers, with the data coverage from both sections being very good to complete from 2020 onwards. Since 2011, the Register of Primary Health Care Visits (AvoHilmo) has covered public primary care. Since 2020, the register has also covered private outpatient care and occupational care. In addition, the CDM also contains positive COVID-19 test results from the Finnish National Infectious Diseases Register. The register itself covers all laboratory confirmations for around 70 specific microbes from 1995 onwards, but only COVID-19 has currently been mapped to the CDM. The CDM also includes prescription records from multiple different sources. Both Care Register for Health Care and Register of Primary Health Care Visits contain very basic prescription data recorded during health care encounters that include just the ATC-code and trade name of medication. More comprehensive prescription data from Kanta Prescription Centre, maintained by Social Insurance Institution of Finland (Kela), has been integrated into the CDM since its 2024 release. The Kanta Prescription records are based electronic prescriptions, which were adopted by most public health care providers in 2010 and by most private providers by 2017.
5	Data collection process	Outpatient electronic health records, and Inpatient hospital electronic health records, and Registries. Data is entered by clinicians upon healthcare contact and processed by THL (Kela in the case of Kanta Prescription Centre).

#	Section	Description
6	General representativeness	The THL data has national coverage and is therefore well representative of the Finnish population. Using the complete population as a basis for the person table also serves to facilitate calculations on a population level, e.g. incidence rates.
7	Data content / source coding	The following coding systems have been OMOP-mapped, typically to a good level of completeness: ICD10fi Finnish Extension, ICPC-2, ATC, Toimenpideluokitus (procedure classification adapted from the Nordic Classification of Surgical Procedures (NCSP)), Terveystieteiden tutkimuskeskus (Hilmo specific provider speciality), Rokatustapa (AR/YDIN National classification for vaccine administration), Tupakointistatus (AR/YDIN National classification for smoking status). Vaccinations are identified on product level based on batch number, trade name, vaccine title, and ATC-code. This is mapped on brand and type in the OMOP CDM.
8	Data Harmonisation	The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network. Each patient in THL has a unique identifier.
9	Quality control (data source specific)	The source data collection undergoes a structural and semantic validation before entry into the source database. Additionally, some coded variables undergo quality assessment against the respective code systems post entry into the database. The source registers are also assessed for completeness and coverage, with the aim of improving future collection in the areas where data is lacking.
10	Linkage	THL is already a linkage of multiple Finnish registries (see above).
11	Vital status	The National Population registry data forms the basis for forming the patient population. This ensures an up-to-date location (municipality of residence) of patients, as well as complete death occurrences (although not the cause of death).
12	Limitations	All drug records in the CDM are currently based on prescriptions. Kanta Prescription Centre also includes information on drug dispensings, but these have not currently been converted into OMOP CDM. Depending on the type of the medication, a single prescription can be for up to two years dosage of the drug. This means a patient can have up to two-year breaks between observations while actively using the drug. Observation of a prescription also does not mean that the patient necessarily bought or used the medication. The CDM does not currently have meaningful end dates or days of supply for drug exposure. This information is not available for Care Register for Health Care or Register of Primary Health Care Visits, and for Kanta Prescription Centre it is only available in unstructured, free-form format that has not been converted into OMOP CDM. The source system for prescription drug data is only available to the DP until Dec-2023, therefore drug_exposure records from this source are only available in that time range. Some vital signs, like height and weight measurements, are recorded even if the person has not been actually measured, just that the measurement is deemed relevant to the condition or it's treatment. This may lead to duplicate measurements as the value from previous visit might be taken. Not all private health care records are covered for the CDM's entire follow-up time from 2011 onwards. For Register of Primary Health Care Visits and Finnish National Vaccination Register, records from private health care have been available from 2020 onwards. For Kanta Prescription Centre, the coverage of private health care records has been good from 2017 onwards. The inclusion of private health care mainly presents itself as an increase in the number of observations, meaning that it has to be accounted for when interpreting any time series data from the CDM.
13	Main references	Häkkinen, Pirjo; Mölläri, Kaisa; Saukkonen, Sanna-Mari; Väyrynen, Riikka; Mielikäinen, Lasse; Järvelin, Jutta "Hilmo - Sosiaali- ja terveydenhuollon hoitoilmoitus 2020 : Määrittelyt ja ohjeistus : Voimassa 1.1.2020 alkaen" Terveystieteiden tutkimuskeskus ja hyvinvoinnin laitos (2019):
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111187 ; https://catalogues.ema.europa.eu/data-source/1111191 Website: https://thl.fi/en/statistics-and-data/data-and-services/register-descriptions ; https://www.kanta.fi/en/research-and-knowledge-management

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 Català 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). 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.

#	Section	Description
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, Bolívar 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

Integrated Primary Care Information (IPCI)

#	Section	Description
1	Data source identification and country	IPCI (Integrated Primary Care Information) the Netherlands
2	Data partner information section	Erasmus University Medical Center (EMC) Department of Medical Informatics
3	Coverage and timespan	Data collection since: 2006 Extent: Nation-wide. IPCI is a Dutch database that contains patient records from 2006 onwards. However, it mainly covers the central part of the country, including the most densely populated area (the 'Randstad') and non-urban areas. IPCI contains information on all patients registered with GPs responsible for non-emergency care and referrals. A patient is registered at birth or at first encounter with the GP.
4	Healthcare setting / type of data	Primary care – General Practitioner. Data is collected from primary care EHR. This includes demographic information, complaints and symptoms, diagnoses, laboratory test results, lifestyle factors (in limited amount), and correspondence with secondary care, such as referral and discharge letters.
5	Data collection process	Outpatient electronic health records. Data is entered into the EHR system by the GPs, during or after the visit. The patient dossiers are collected by Erasmus MC data managers and combined in one harmonized database. Several checks are done on this database to ensure correct data processing. Persons can have dossiers at multiple GPs.
6	General representativeness	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.
7	Data content / source coding	Dutch GPs use mainly Dutch standard codes, like ICPC-1 and Diagnostische Bepalingen maintained by NHG. And for therapy the G-Standard is used, maintained by ZIndex.
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. Patients can be registered under different IDs, but since a patient can only be registered at one GP at a time, the observations periods will not overlap.

#	Section	Description
9	Quality control (data source specific)	Prior to each data release, extensive quality control steps are performed, e.g. comparison of patient characteristics between practices, and checks to identify abnormal temporal data patterns in practices. For each practice, around 200 quality indicators are obtained. Of these indicators, a quarter refer to population characteristics, e.g. number of birth and mortalities relative to practice size, temporal consistency. The other indicators are based on medical data, e.g. distribution of measurement values, frequencies of diagnoses and procedures relative to age, completeness of data. The indicators are combined in a couple of quality scores for each practice. For these scores, cut-off values for acceptable quality have been defined. Practices with a score below a cut-off are excluded for research. This approach has shown to be very important, for example to check if data from practices that just joined the database are at an acceptable level of quality. The details of the approach, like the cut-off values for acceptance, are based on years of experience. In addition, trends are compared with the previous database release. 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 and completeness of laboratory results). Records of low quality are excluded from the database.
10	Linkage	Linkage requires additional approval steps and needs to be assessed on a case-by-case basis. IPCI is not routinely linked with other databases.
11	Vital status	Vital status (death date and cause) is collected based on GP records.
12	Limitations	The main limitation comes with the fact that IPCI is limited to GP records, and although it contains information on referrals and discharge letters, it may not fully capture specific hospital information. Capture of oral contraceptives in IPCI is incomplete. The first prescription is recorded within the patient's medical file but not the subsequent prescriptions. If you want to use IPCI data for research questions on oral contraceptives, you can do that but then you need to select those GPs that use MEDICOM as software system (where you have prescribing and dispensing). However, of hormone substitution therapy, the information on prescribing is COMPLETE. IPCI does not include coded/detailed data about medications/procedures/test results from the hospital or other care-providers.
13	Main references	de Ridder MAJ, de Wilde M, de Ben C, Leyba AR, Mosseveld BMT, Verhamme KMC, van der Lei J, Rijnbeek PR "Data Resource Profile: The Integrated Primary Care Information (IPCI) database, The Netherlands." International journal of epidemiology (2022): 35182144
14	Link to HMA-EMA catalogue and data source webpage	HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/42618 Website: http://www.ipci.nl

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,

#	Section	Description
		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. 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

#	Section	Description
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

Danish Data Health Registries (DK-DHR)

DK-DHR will be included in this study because it is a nationwide linked registries data source that provides relevant information on polyendocrine metabolic ovarian syndrome (PMOS) in the general female population.

Based on a preliminary feasibility assessment, the expected number of person-counts for PMOS in DK-DHR will be 27,000.

Moreover, data availability and follow-up in DK-DHR is sufficient, as data availability starts in 1995, and the date of the most recent data extraction (as of July 2026) is 10/2025, which aligns with the study period. The median follow-up of the first observation period in DK-DHR is 7,930 days (interquartile range IQR 2,570–11,300).

There are some study specific limitations present in DK-DHR, namely: absence of data from primary care settings.

Lastly, DK-DHR has blanket Institutional Review Board (IRB) approval, which makes the execution of this study feasible within the current study timelines.

Finnish Care Register for Health Care (FinOMOP-THL)

FinOMOP-THL will be included in this study because it is a nationwide linked registries data source that provides relevant information on PMOS in the general female population.

Based on a preliminary feasibility assessment, the expected number of person-counts for PMOS in FinOMOP-THL will be 18,700.

Moreover, data availability and follow-up in FinOMOP-THL is sufficient, as data availability starts in 2015, and the date of the most recent data extraction (as of July 2026) is 12/2025, which aligns with most of the study period. The median follow-up of the first observation period in FinOMOP-THL is 4,020 days (IQR 4,020–4,020).

There are some study specific limitations present in FinOMOP-THL, namely: due to data availability, the study period for FinOMOP-THL will start from 01/2016 to ensure one year of data availability in the data source.

Lastly, FinOMOP-THL has blanket Institutional Review Board (IRB) approval, which makes the execution of this study feasible within the current study timelines.

The Information System for Research in Primary Care (SIDIAP)

SIDIAP will be included in this study because it is a primary care data source with linked hospital data that provides relevant information on PMOS in the general female population.

Based on a preliminary feasibility assessment, the expected number of person-counts for PMOS in SIDIAP will be 45,500.

Moreover, data availability and follow-up in SIDIAP is sufficient, as data availability starts in 2006, and the date of the most recent data extraction (as of July 2026) is 12/2024, which aligns with the study period. The median follow-up of the first observation period in SIDIAP is 5,760 days (IQR 2,170–6,940).

There are no study specific limitations present in SIDIAP.

Lastly, IRB approval for SIDIAP is estimated to take 2–3 months, which makes the execution of this study feasible within the current study timelines.

Integrated Primary Care Information (IPCI)

IPCI will be included in this study because it is a primary care data source that provides relevant information on PMOS in the general female population.

Based on a preliminary feasibility assessment, the expected number of person-counts for PMOS in IPCI will be 11,500.

Moreover, data availability and follow-up in IPCI is sufficient, as data availability starts in 2006, and the date of the most recent data extraction (as of July 2026) is 12/2025, which aligns with the study period. The median follow-up of the first observation period in IPCI is 1,830 days (IQR 822–3,200).

There are no study specific limitations present in IPCI.

Lastly, IPCI has blanket IRB approval, which makes the execution of this study feasible within the current study timelines.

Clinical Practice Research Datalink GOLD (CPRD GOLD)

CPRD GOLD will be included in this study because it is a primary care data source that provides relevant information on PMOS in the general female population.

Based on a preliminary feasibility assessment, the expected number of person-counts for PMOS in CPRD GOLD will be 72,100.

Moreover, data availability and follow-up in CPRD GOLD is sufficient, as data availability starts in 1987, and the date of the most recent data extraction (as of July 2026) is 10/2025, which aligns with the study period. The median follow-up of the first observation period in CPRD GOLD is 2,160 days (IQR 729–4,940).

There are no study specific limitations present in CPRD GOLD.

Lastly, CPRD GOLD has blanket IRB approval, which makes the execution of this study feasible within the current study timelines.

Table S1. Fitness-for-use assessment of data sources.

Design elements	Operational definition	Data elements for valid capture	Criticality of the quality of the element, including justification where relevant	Suggested extensiveness assessment	Suggested assessment of other quality dimensions	Suggested substantiation by documentation
Study population	PMOS assessment will be done in the general female population aged 10 or above.	Data capture on the general female population depends on the valid records on age and sex.	High	N/A	N/A	N/A
Treatment/exposure	N/A	Data capture on PMOS depends on the availability of condition records.	High	We will undergo the phenotyping review process for PMOS based on conditions.	N/A	N/A
Outcomes (if relevant)	PMOS will be defined using condition and/or observation codes. Hypertension, myocardial infarction, stroke, ischemic stroke, VTE, anxiety, and depression will be defined using condition and/or observation codes. Infertility will be identified by the presence of at least one of the following: (i) a diagnosis code for female infertility, including broad female infertility concepts and anovulatory infertility (the subtype most commonly associated with PMOS), excluding diagnosis codes related exclusively to non-anovulatory causes of infertility; or (ii) a prescription for pharmacological treatments commonly used for anovulatory infertility associated with PMOS, including letrozole, clomiphene, or gonadotropins; or (iii)	Cardiovascular and mental outcomes are commonly captured in other DARWIN EU® studies. However, infertility is less commonly characterised. Infertility will be defined by related conditions/drugs/procedures as proxy.	Medium	Based on the feasibility assessment on infertility-related conditions/drugs/procedures, it is possible to conduct an exploratory descriptive analysis to use related available data as a proxy to define infertility in PMOS population.	N/A	N/A

Design elements	Operational definition	Data elements for valid capture	Criticality of the quality of the element, including justification where relevant	Suggested extensiveness assessment	Suggested assessment of other quality dimensions	Suggested substantiation by documentation
	a procedure code for assisted reproductive technologies or surgical interventions, including in vitro fertilisation or ovarian drilling.					
Covariates (including confounders if relevant)	Pre-defined comorbidities - Cardiometabolic/vascular: obesity, dyslipidaemia, hypertension, type 2 diabetes, metabolic dysfunction-associated steatotic liver disease, obstructive sleep apnoea, myocardial infarction, stroke, VTE - Hyperandrogenic/reproductive: acne, hirsutism, alopecia, oligomenorrhoea/amenorrhoea, infertility - Mental health: anxiety, depression, eating disorder Pre-defined treatments - Combined oral contraceptives - Metformin - GLP-1 receptor agonists (lixisenatide, albiglutide,	Data capture on comorbidities and drug treatments depends on the availability of condition records and drug exposure records.	Medium	Complete phenotyping process has been conducted in another DARWIN EU® study (Population demographics and disease frequency across the DARWIN EU® Network, EUPAS1000000962).	N/A	N/A

Design elements	Operational definition	Data elements for valid capture	Criticality of the quality of the element, including justification where relevant	Suggested extensiveness assessment	Suggested assessment of other quality dimensions	Suggested substantiation by documentation
	liraglutide, semaglutide, dulaglutide, exenatide) - Other antidiabetic medications (sulfonylureas, thiazolidinediones, DPP-4 inhibitors, dual GIP/GLP-1 receptor agonist, SGLT2 inhibitors) - Insulin - Anti-androgen medications (spironolactone, flutamide, bicalutamide, finasteride, cyproterone) - Orlistat - Inositol					

DPP-4 = dipeptidyl peptidase-4; GIP = glucose-dependent insulintropic polypeptide; GLP-1 = glucagon-like peptide-1; N/A = not applicable; PMOS = polyendocrine metabolic ovarian syndrome; SGLT2 = sodium-glucose co-transporter 2; VTE = venous thromboembolism. The framework for fitness-for-use assessment is available at EMA Data Quality Framework for EU medicines regulation: application to Real-World Data for more information (https://www.ema.europa.eu/system/files/documents/other/data-quality-framework-eu-medicines-regulation-application-real-world-data_en.pdf).

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 standardized 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) 679/2016 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 *PhenotypeR* (<https://github.com/OHDSI/PhenotypeR>) R packages will be run, if needed, to assess the use of different codes across the data sources contributing to the study and identify any codes potentially omitted in error.

The study code will be based on DARWIN EU® R packages: *IncidencePrevalence* to estimate incidence of PMOS in the general female population and health outcomes in the PMOS population, *CohortCharacteristics* to characterise the PMOS population, and *CohortSurvival* to estimate cumulative incidence of health outcomes in the PMOS population. These packages will include numerous automated unit tests to ensure the validity of the codes, alongside software peer review and user testing. The R package will be made publicly available via GitHub.

PLANS FOR DISSEMINATING AND COMMUNICATING STUDY RESULTS

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

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

ANNEX IV. Lists with concept definitions

Table S2. Preliminary list of definitions for polyendocrine metabolic ovarian syndrome.

Concept ID (including descendants)	Concept name
36683296	Polycystic ovary
3654291	Polycystic left ovary
3654292	Polycystic right ovary
3654293	Polycystic bilateral ovaries
3654294	Polycystic ovary syndrome of right ovary
3654295	Polycystic ovary syndrome of left ovary
3654296	Polycystic ovary syndrome of bilateral ovaries
4300909	Androgenetic alopecia in association with polycystic ovaries
40443308	Polycystic ovary syndrome

Table S3. Preliminary list of definitions for comorbidities.

Phenotype	Concept ID (including descendants)	Concept name
Obesity	433736	Obesity
Dyslipidaemia	4159131	Dyslipidemia
	432867	Hyperlipidemia
Hypertension	316866	Hypertensive disorder
Type 2 diabetes	201826	Type 2 diabetes mellitus
Metabolic dysfunction-associated steatotic liver disease	37164766	Metabolic dysfunction-associated steatotic liver disease
Obstructive sleep apnoea	313459	Sleep apnea
Acne	141095	Acne
Hirsutism	134718	Hirsutism
Alopecia	133280	Alopecia
Oligomenorrhoea/amenorrhoea	442274	Oligomenorrhoea
	443800	Amenorrhoea
Eating disorder	439002	Eating disorder

Table S4. Preliminary list of definitions for non-infertility drug treatments (ingredients).

Class/Substance Name	ATC code	Ingredient Concept ID (including descendants)	Concept name
Metformin	A10BA02 metformin	1503297	metformin
Glucagon-like peptide-1 (GLP-1) analogues	A10BJ Glucagon-like peptide-1 (GLP-1) analogues	44506754	lixisenatide
		44816332	albiglutide
		40170911	liraglutide
		793143	semaglutide

Class/Substance Name	ATC code	Ingredient Concept ID (including descendants)	Concept name
Insulin	A10A INSULINS AND ANALOGUES	45774435	dulaglutide
		1583722	exenatide
		19013926	insulin, protamine zinc, beef-pork
		19013951	insulin, regular, beef
		19091621	insulin, protamine zinc, human
		19090180	insulin, protamine zinc, beef
		19090187	insulin, protamine zinc, pork
		19090204	lente insulin, beef
		19090221	ultralente insulin, beef-pork
		19090226	ultralente insulin, beef
		19090229	insulin, prompt zinc, beef-pork
		19090244	insulin, prompt zinc, beef
		19090247	insulin, prompt zinc, human
		19090249	insulin, prompt zinc, pork
		1502905	insulin glargine
		1513843	lente insulin, beef-pork
		1513876	insulin lispro protamine, human
		1516976	insulin detemir
		1531601	insulin aspart protamine, human
		1544838	insulin glulisine, human
		1550023	insulin lispro
		1562586	lente insulin, pork
		1567198	insulin aspart, human
		1586346	insulin, regular, pork
		1586369	ultralente insulin, human
		1588986	insulin human, rDNA origin
		1590165	insulin, regular, beef-pork
		1596977	insulin, regular, human
		35198096	insulin aspart
		35602717	insulin degludec
		42899447	insulin, pork
		46221581	insulin isophane
Other antidiabetic medications: sulfonylureas	A10BB Sulfonylureas	40798860	Glisoxepide
		1502809	tolazamide
		1502855	tolbutamide
		1530014	acetohexamide

Class/Substance Name	ATC code	Ingredient Concept ID (including descendants)	Concept name
		1559684	glyburide
		1560171	glipizide
		1594973	chlorpropamide
		1597756	glimepiride
		19001409	glibornuride
		19059796	gliclazide
		19097821	gliquidone
		19033498	carbutamide
		36861609	METAHEXAMIDE
Other antidiabetic medications: thiazolidinediones	A10BG Thiazolidinediones	1515249	troglitazone
		1525215	pioglitazone
		1547504	rosiglitazone
		43009055	lobeglitazone
Other antidiabetic medications: dipeptidyl peptidase 4 (DPP-4) inhibitors	A10BH Dipeptidyl peptidase 4 (DPP-4) inhibitors	40239216	linagliptin
		1580747	sitagliptin
		40166035	saxagliptin
		43009051	evogliptin
		43009070	teneligliptin
		43009089	gemigliptin
		43013884	alogliptin
		19122137	vildagliptin
Other antidiabetic medications: dual glucose-dependent insulinotropic polypeptide (GIP)/GLP-1 receptor agonist	N/A	779705	tirzepatide
Other antidiabetic medications: sodium-glucose co-transporter 2 (SGLT2) inhibitors	A10BK Sodium-glucose co-transporter 2 (SGLT2) inhibitors	43526465	canagliflozin
		1301517	bexagliflozin
		44785829	dapagliflozin
		45774751	empagliflozin
		35198021	luseogliflozin hydrate
		43009020	ipragliflozin
		793293	ertugliflozin
Combined oral contraceptives	G03A HORMONAL CONTRACEPTIVES FOR SYSTEMIC USE	19015533	nomegestrol
		19026733	chlormadinone
		19054876	dienogest
		19092358	lynestrenol
		19097684	gestodene

Class/Substance Name	ATC code	Ingredient Concept ID (including descendants)	Concept name
		1500211	medroxyprogesterone
		1512674	drospirenone
		1515774	norgestimate
		1518198	norelgestromin
		1519936	etonogestrel
		1521369	norethindrone
		1521592	norgestrel
		1588000	desogestrel
		1589505	levonorgestrel
		1595461	ethynodiol
		19128258	norgestrienone
		40225722	ulipristal
Anti-androgenic medications	N/A	970250	spironolactone
		1356461	flutamide
		1344381	bicalutamide
		996416	finasteride
		19010792	cypoterone
Orlistat	N/A	741530	orlistat
Inositol	N/A	1379525	inositol

ATC = anatomical therapeutic chemical; N/A = not applicable.

Table S5. Preliminary list of definitions for cardiovascular and mental outcomes.

Phenotype	Concept ID (including descendants)	Concept name
Hypertension	316866	Hypertensive disorder
Myocardial infarction	4329847	Myocardial infarction
Stroke	4310996	Ischemic stroke
	35609033	Haemorrhagic stroke
	373503	Transient cerebral ischemia
Venous thromboembolism	4327889	Thromboembolism of vein
Depression	440383	Depressive disorder
Anxiety	442077	Anxiety disorder

Table S6. Preliminary list of definitions for infertility.

Concept ID (including descendants)	Concept name
Conditions	
192370	Secondary female infertility
201909	Female infertility

Concept ID (including descendants)	Concept name
4113207	Primary anovulatory infertility
4113842	Secondary anovulatory infertility
4129548	Female infertility due to ovulatory disorder
4141641	Female infertility of anovulatory origin
4174007	Primary infertility
197044	Female infertility associated with anovulation
36713460	Supervision of pregnancy with history of infertility
4129550	Unexplained infertility
4309021	Complications of attempted introduction of fertilized ovum following in vitro fertilization
Drugs	
1315946	letrozole
1598819	clomiphene
1542948	follitropin alfa
36878940	corifollitropin alfa
36852653	menotropins
1563600	chorionic gonadotropin
1597235	follitropin beta
1543112	lutropin alfa
1515417	urofollitropin
43049220	follitropin delta
Procedures	
4199599	In vitro fertilization
4138633	In vitro fertilization with intracytoplasmic spermatozoa injection
4139355	Endoscopic drilling of ovary

ANNEX V. Mock tables and figures

Mock tables and figures are provided below for illustrative purposes only and do not contain any results. They are presented as table and figure shells.

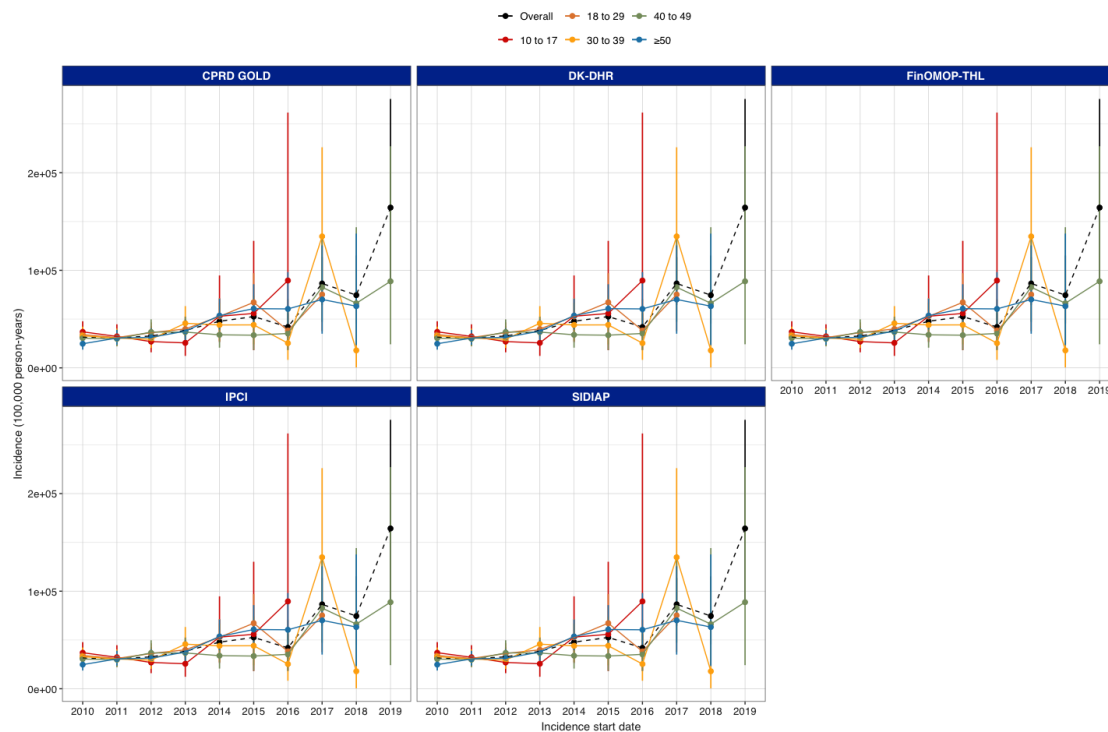
Overall

MOCK Table 1. Number of individuals with polyendocrine metabolic ovarian syndrome (PMOS) during the study period, by study objective and data source.

Reason	Study objective	Additional requirement	Data source				
			DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Total number of eligible population	Overall	N/A					
Total number of individuals with PMOS	Incidence, characterisation (comorbidities)	Restrict to first PMOS diagnosis					
	Characterisation (non-infertility drug treatments)	Full year contribution of data source					

CPRD GOLD = Clinical Practice Research Datalink GOLD; DK-DHR = Danish Data Health Registries; FinOMOP-THL = Finnish Care Register for Health Care; IPCI = Integrated Primary Care Information; N/A, not applicable; PMOS = Polyendocrine metabolic ovarian syndrome; SIDIAP = The Information System for Research in Primary Care.

Objective 1



MOCK Figure 1. Annual incidence of recorded polyendocrine metabolic ovarian syndrome (PMOS) diagnosis, overall and stratified by age group.

CPRD GOLD = Clinical Practice Research Datalink GOLD; DK-DHR = Danish Data Health Registries; FinOMOP-THL = Finnish Care Register for Health Care; IPCI = Integrated Primary Care Information; SIDIAP = The Information System for Research in Primary Care.

Objective 2

MOCK Table 2. Characterisation of polyendocrine metabolic ovarian syndrome (PMOS) individuals on comorbidities.

Variable name	Variable level	Estimate name	Data source				
			DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Demographic characteristics							
Number subjects	–	N					
Age	–	Median [Q25 – Q75]					
		Range					
Age group	10–17	N (%)					
	18–29	N (%)					
	30–39	N (%)					
	40–49	N (%)					
	≥50	N (%)					
Cardiometabolic/vascular comorbidities							
Hypertension	–	N (%)					
Type 2 diabetes	–	N (%)					
Dyslipidaemia	–	N (%)					
Metabolic dysfunction-associated steatotic liver disease	–	N (%)					
Obesity	–	N (%)					
Obstructive sleep apnoea	–	N (%)					
Myocardial infarction	–	N (%)					
Stroke	–	N (%)					
VTE	–	N (%)					
Hyperandrogenic/reproductive comorbidities							
Acne	–	N (%)					
Alopecia	–	N (%)					

Variable name	Variable level	Estimate name	Data source				
			DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Hirsutism	–	N (%)					
Oligomenorrhoea/amenorrhoea	–	N (%)					
Infertility	–	N (%)					
Mental health							
Anxiety	–	N (%)					
Depression	–	N (%)					
Eating disorder	–	N (%)					

CPRD GOLD = Clinical Practice Research Datalink GOLD; DK-DHR = Danish Data Health Registries; FinOMOP-THL = Finnish Care Register for Health Care; IPCI = Integrated Primary Care Information; SIDIAP = The Information System for Research in Primary Care; VTE = venous thromboembolism.

MOCK Table 3. Characterisation of polyendocrine metabolic ovarian syndrome (PMOS) individuals on comorbidities, stratified by age group.

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
10–17; Demographic characteristics						
Number subjects	N					
Age	Median [Q25 – Q75]					
	Range					
10–17; Cardiometabolic/vascular comorbidities						
Hypertension	N (%)					
Type 2 diabetes	N (%)					
Dyslipidaemia	N (%)					
Metabolic dysfunction-associated steatotic liver disease	N (%)					
Obesity	N (%)					
Obstructive sleep apnoea	N (%)					
Myocardial infarction	N (%)					
Stroke	N (%)					
VTE	N (%)					
10–17; Hyperandrogenic/reproductive comorbidities						
Acne	N (%)					
Alopecia	N (%)					
Hirsutism	N (%)					
Oligomenorrhoea/amenorrhoea	N (%)					
Infertility	N (%)					
10–17; Mental health						
Anxiety	N (%)					
Depression	N (%)					

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Eating disorder	N (%)					
18–29; Demographic characteristics						
Number subjects	N					
Age	Median [Q25 – Q75]					
	Range					
18–29; Cardiometabolic/vascular comorbidities						
Hypertension	N (%)					
Type 2 diabetes	N (%)					
Dyslipidaemia	N (%)					
Metabolic dysfunction-associated steatotic liver disease	N (%)					
Obesity	N (%)					
Obstructive sleep apnoea	N (%)					
Myocardial infarction	N (%)					
Stroke	N (%)					
VTE	N (%)					
18–29; Hyperandrogenic/reproductive comorbidities						
Acne	N (%)					
Alopecia	N (%)					
Hirsutism	N (%)					
Oligomenorrhoea/amenorrhoea	N (%)					
Infertility	N (%)					
18–29; Mental health						
Anxiety	N (%)					
Depression	N (%)					

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Eating disorder	N (%)					
30–39; Demographic characteristics						
Number subjects	N					
Age	Median [Q25 – Q75]					
	Range					
30–39; Cardiometabolic/vascular comorbidities						
Hypertension	N (%)					
Type 2 diabetes	N (%)					
Dyslipidaemia	N (%)					
Metabolic dysfunction-associated steatotic liver disease	N (%)					
Obesity	N (%)					
Obstructive sleep apnoea	N (%)					
Myocardial infarction	N (%)					
Stroke	N (%)					
VTE	N (%)					
30–39; Hyperandrogenic/reproductive comorbidities						
Acne	N (%)					
Alopecia	N (%)					
Hirsutism	N (%)					
Oligomenorrhoea/amenorrhoea	N (%)					
Infertility	N (%)					
30–39; Mental health						
Anxiety	N (%)					
Depression	N (%)					

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Eating disorder	N (%)					
40–49; Demographic characteristics						
Number subjects	N					
Age	Median [Q25 – Q75]					
	Range					
40–49; Cardiometabolic/vascular comorbidities						
Hypertension	N (%)					
Type 2 diabetes	N (%)					
Dyslipidaemia	N (%)					
Metabolic dysfunction-associated steatotic liver disease	N (%)					
Obesity	N (%)					
Obstructive sleep apnoea	N (%)					
Myocardial infarction	N (%)					
Stroke	N (%)					
VTE	N (%)					
40–49; Hyperandrogenic/reproductive comorbidities						
Acne	N (%)					
Alopecia	N (%)					
Hirsutism	N (%)					
Oligomenorrhoea/amenorrhoea	N (%)					
Infertility	N (%)					
40–49; Mental health						
Anxiety	N (%)					
Depression	N (%)					

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Eating disorder	N (%)					
≥50; Demographic characteristics						
Number subjects	N					
Age	Median [Q25 – Q75]					
	Range					
≥50; Cardiometabolic/vascular comorbidities						
Hypertension	N (%)					
Type 2 diabetes	N (%)					
Dyslipidaemia	N (%)					
Metabolic dysfunction-associated steatotic liver disease	N (%)					
Obesity	N (%)					
Obstructive sleep apnoea	N (%)					
Myocardial infarction	N (%)					
Stroke	N (%)					
VTE	N (%)					
≥50; Hyperandrogenic/reproductive comorbidities						
Acne	N (%)					
Alopecia	N (%)					
Hirsutism	N (%)					
Oligomenorrhoea/amenorrhoea	N (%)					
Infertility	N (%)					
≥50; Mental health						
Anxiety	N (%)					
Depression	N (%)					

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Eating disorder	N (%)					

CPRD GOLD = Clinical Practice Research Datalink GOLD; DK-DHR = Danish Data Health Registries; FinOMOP-THL = Finnish Care Register for Health Care; IPCI = Integrated Primary Care Information; SIDIAP = The Information System for Research in Primary Care; VTE = venous thromboembolism.

Objective 3

MOCK Table 4. Characterisation of polyendocrine metabolic ovarian syndrome (PMOS) individuals on non-infertility drug treatments.

Variable name	Variable level	Estimate name	Data source				
			DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Demographic characteristics							
Number subjects	–	N					
Age	–	Median [Q25 – Q75]					
		Range					
Age group	10–17	N (%)					
	18–29	N (%)					
	30–39	N (%)					
	40–49	N (%)					
	≥50	N (%)					
Non-infertility treatments (30 days after index date)							
Combined oral contraceptives	–	N (%)					
Metformin	–	N (%)					
GLP-1 receptor agonists	–	N (%)					
Albiglutide	–	N (%)					
Dulaglutide	–	N (%)					
Exenatide	–	N (%)					
Liraglutide	–	N (%)					
Lixisenatide	–	N (%)					
Semaglutide	–	N (%)					
Sulfonylureas	–	N (%)					
Thiazolidinediones	–	N (%)					
DPP-4 inhibitors	–	N (%)					

Variable name	Variable level	Estimate name	Data source				
			DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Dual GIP/GLP-1 receptor agonist	–	N (%)					
SGLT2 inhibitors	–	N (%)					
Insulin	–	N (%)					
Anti-androgen medications	–	N (%)					
Spironolactone	–	N (%)					
Flutamide	–	N (%)					
Finasteride	–	N (%)					
Bicalutamide	–	N (%)					
Cyproterone	–	N (%)					
Orlistat	–	N (%)					
Inositol	–	N (%)					
Non-infertility treatments (31 to 365 days after index date)							
Combined oral contraceptives	–	N (%)					
Metformin	–	N (%)					
GLP-1 receptor agonists	–	N (%)					
Albiglutide	–	N (%)					
Dulaglutide	–	N (%)					
Exenatide	–	N (%)					
Liraglutide	–	N (%)					
Lixisenatide	–	N (%)					
Semaglutide	–	N (%)					
Sulfonylureas	–	N (%)					
Thiazolidinediones	–	N (%)					
DPP-4 inhibitors	–	N (%)					
Dual GIP/GLP-1 receptor agonist	–	N (%)					

Variable name	Variable level	Estimate name	Data source				
			DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
SGLT2 inhibitors	–	N (%)					
Insulin	–	N (%)					
Anti-androgen medications	–	N (%)					
Spironolactone	–	N (%)					
Flutamide	–	N (%)					
Finasteride	–	N (%)					
Bicalutamide	–	N (%)					
Cyproterone	–	N (%)					
Orlistat	–	N (%)					
Inositol	–	N (%)					

CPRD GOLD = Clinical Practice Research Datalink GOLD; DK-DHR = Danish Data Health Registries; DPP-4 = dipeptidyl peptidase-4; FinOMOP-THL = Finnish Care Register for Health Care; GIP = glucose-dependent insulinotropic polypeptide; GLP-1 = glucagon-like peptide-1; IPCI = Integrated Primary Care Information; SGLT2 = sodium-glucose co-transporter 2; SIDIAP = The Information System for Research in Primary Care; VTE = venous thromboembolism.

MOCK Table 5. Characterisation of polyendocrine metabolic ovarian syndrome (PMOS) individuals on non-infertility drug treatments, stratified by age group.

Variable name	Estimate name	Data source				
		DK–DHR	FinOMOP–THL	SIDIAP	IPCI	CPRD GOLD
10–17; Demographic characteristics						
Number subjects	N					
Age	Median [Q25 – Q75]					
	Range					
10–17; Non-infertility treatments (30 days after index date)						
Combined oral contraceptives	N (%)					
Metformin	N (%)					
GLP-1 receptor agonists	N (%)					
Albiglutide	N (%)					
Dulaglutide	N (%)					
Exenatide	N (%)					
Liraglutide	N (%)					
Lixisenatide	N (%)					
Semaglutide	N (%)					
Sulfonylureas	N (%)					
Thiazolidinediones	N (%)					
DPP-4 inhibitors	N (%)					
Dual GIP/GLP-1 receptor agonist	N (%)					
SGLT2 inhibitors	N (%)					
Insulin	N (%)					
Anti-androgen medications	N (%)					
Spironolactone	N (%)					
Flutamide	N (%)					
Finasteride	N (%)					

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Bicalutamide	N (%)					
Cyproterone	N (%)					
Orlistat	N (%)					
Inositol	N (%)					
10–17; Non-infertility treatments (31 to 365 days after index date)						
Combined oral contraceptives	N (%)					
Metformin	N (%)					
GLP-1 receptor agonists	N (%)					
Albiglutide	N (%)					
Dulaglutide	N (%)					
Exenatide	N (%)					
Liraglutide	N (%)					
Lixisenatide	N (%)					
Semaglutide	N (%)					
Sulfonylureas	N (%)					
Thiazolidinediones	N (%)					
DPP-4 inhibitors	N (%)					
Dual GIP/GLP-1 receptor agonist	N (%)					
SGLT2 inhibitors	N (%)					
Insulin	N (%)					
Anti-androgen medications	N (%)					
Spironolactone	N (%)					
Flutamide	N (%)					
Finasteride	N (%)					
Bicalutamide	N (%)					
Cyproterone	N (%)					

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Orlistat	N (%)					
Inositol	N (%)					
18-29; Demographic characteristics						
Number subjects	N					
Age	Median [Q25 – Q75]					
	Range					
18-29; Non-infertility treatments (30 days after index date)						
Combined oral contraceptives	N (%)					
Metformin	N (%)					
GLP-1 receptor agonists	N (%)					
Albiglutide	N (%)					
Dulaglutide	N (%)					
Exenatide	N (%)					
Liraglutide	N (%)					
Lixisenatide	N (%)					
Semaglutide	N (%)					
Sulfonylureas	N (%)					
Thiazolidinediones	N (%)					
DPP-4 inhibitors	N (%)					
Dual GIP/GLP-1 receptor agonist	N (%)					
SGLT2 inhibitors	N (%)					
Insulin	N (%)					
Anti-androgen medications	N (%)					
Spironolactone	N (%)					
Flutamide	N (%)					
Finasteride	N (%)					

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Bicalutamide	N (%)					
Cyproterone	N (%)					
Orlistat	N (%)					
Inositol	N (%)					
18–29; Non-infertility treatments (31 to 365 days after index date)						
Combined oral contraceptives	N (%)					
Metformin	N (%)					
GLP-1 receptor agonists	N (%)					
Albiglutide	N (%)					
Dulaglutide	N (%)					
Exenatide	N (%)					
Liraglutide	N (%)					
Lixisenatide	N (%)					
Semaglutide	N (%)					
Sulfonylureas	N (%)					
Thiazolidinediones	N (%)					
DPP-4 inhibitors	N (%)					
Dual GIP/GLP-1 receptor agonist	N (%)					
SGLT2 inhibitors	N (%)					
Insulin	N (%)					
Anti-androgen medications	N (%)					
Spironolactone	N (%)					
Flutamide	N (%)					
Finasteride	N (%)					
Bicalutamide	N (%)					
Cyproterone	N (%)					

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Orlistat	N (%)					
Inositol	N (%)					
30–39; Demographic characteristics						
Number subjects	N					
Age	Median [Q25 – Q75]					
	Range					
30–39; Non-infertility treatments (30 days after index date)						
Combined oral contraceptives	N (%)					
Metformin	N (%)					
GLP-1 receptor agonists	N (%)					
Albiglutide	N (%)					
Dulaglutide	N (%)					
Exenatide	N (%)					
Liraglutide	N (%)					
Lixisenatide	N (%)					
Semaglutide	N (%)					
Sulfonylureas	N (%)					
Thiazolidinediones	N (%)					
DPP-4 inhibitors	N (%)					
Dual GIP/GLP-1 receptor agonist	N (%)					
SGLT2 inhibitors	N (%)					
Insulin	N (%)					
Anti-androgen medications	N (%)					
Spironolactone	N (%)					
Flutamide	N (%)					
Finasteride	N (%)					

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Bicalutamide	N (%)					
Cyproterone	N (%)					
Orlistat	N (%)					
Inositol	N (%)					
30–39; Non-infertility treatments (31 to 365 days after index date)						
Combined oral contraceptives	N (%)					
Metformin	N (%)					
GLP-1 receptor agonists	N (%)					
Albiglutide	N (%)					
Dulaglutide	N (%)					
Exenatide	N (%)					
Liraglutide	N (%)					
Lixisenatide	N (%)					
Semaglutide	N (%)					
Sulfonylureas	N (%)					
Thiazolidinediones	N (%)					
DPP-4 inhibitors	N (%)					
Dual GIP/GLP-1 receptor agonist	N (%)					
SGLT2 inhibitors	N (%)					
Insulin	N (%)					
Anti-androgen medications	N (%)					
Spironolactone	N (%)					
Flutamide	N (%)					
Finasteride	N (%)					
Bicalutamide	N (%)					
Cyproterone	N (%)					

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Orlistat	N (%)					
Inositol	N (%)					
40–49; Demographic characteristics						
Number subjects	N					
Age	Median [Q25 – Q75]					
	Range					
40–49; Non-infertility treatments (30 days after index date)						
Combined oral contraceptives	N (%)					
Metformin	N (%)					
GLP-1 receptor agonists	N (%)					
Albiglutide	N (%)					
Dulaglutide	N (%)					
Exenatide	N (%)					
Liraglutide	N (%)					
Lixisenatide	N (%)					
Semaglutide	N (%)					
Sulfonylureas	N (%)					
Thiazolidinediones	N (%)					
DPP-4 inhibitors	N (%)					
Dual GIP/GLP-1 receptor agonist	N (%)					
SGLT2 inhibitors	N (%)					
Insulin	N (%)					
Anti-androgen medications	N (%)					
Spironolactone	N (%)					
Flutamide	N (%)					
Finasteride	N (%)					

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Bicalutamide	N (%)					
Cyproterone	N (%)					
Orlistat	N (%)					
Inositol	N (%)					
40–49; Non-infertility treatments (31 to 365 days after index date)						
Combined oral contraceptives	N (%)					
Metformin	N (%)					
GLP-1 receptor agonists	N (%)					
Albiglutide	N (%)					
Dulaglutide	N (%)					
Exenatide	N (%)					
Liraglutide	N (%)					
Lixisenatide	N (%)					
Semaglutide	N (%)					
Sulfonylureas	N (%)					
Thiazolidinediones	N (%)					
DPP-4 inhibitors	N (%)					
Dual GIP/GLP-1 receptor agonist	N (%)					
SGLT2 inhibitors	N (%)					
Insulin	N (%)					
Anti-androgen medications	N (%)					
Spironolactone	N (%)					
Flutamide	N (%)					
Finasteride	N (%)					
Bicalutamide	N (%)					
Cyproterone	N (%)					

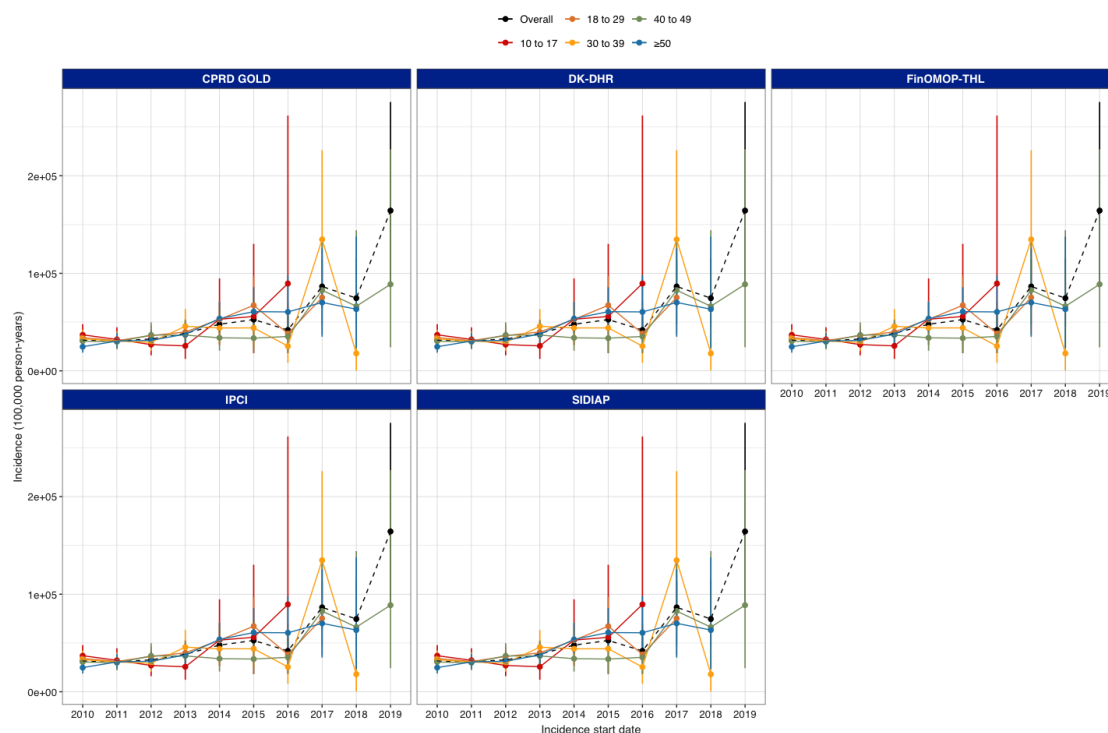
Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Orlistat	N (%)					
Inositol	N (%)					
≥50; Demographic characteristics						
Number subjects	N					
Age	Median [Q25 – Q75]					
	Range					
≥50; Non-infertility treatments (30 days after index date)						
Combined oral contraceptives	N (%)					
Metformin	N (%)					
GLP-1 receptor agonists	N (%)					
Albiglutide	N (%)					
Dulaglutide	N (%)					
Exenatide	N (%)					
Liraglutide	N (%)					
Lixisenatide	N (%)					
Semaglutide	N (%)					
Sulfonylureas	N (%)					
Thiazolidinediones	N (%)					
DPP-4 inhibitors	N (%)					
Dual GIP/GLP-1 receptor agonist	N (%)					
SGLT2 inhibitors	N (%)					
Insulin	N (%)					
Anti-androgen medications	N (%)					
Spironolactone	N (%)					
Flutamide	N (%)					
Finasteride	N (%)					

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Bicalutamide	N (%)					
Cyproterone	N (%)					
Orlistat	N (%)					
Inositol	N (%)					
≥50; Non-infertility treatments (31 to 365 days after index date)						
Combined oral contraceptives	N (%)					
Metformin	N (%)					
GLP-1 receptor agonists	N (%)					
Albiglutide	N (%)					
Dulaglutide	N (%)					
Exenatide	N (%)					
Liraglutide	N (%)					
Lixisenatide	N (%)					
Semaglutide	N (%)					
Sulfonylureas	N (%)					
Thiazolidinediones	N (%)					
DPP-4 inhibitors	N (%)					
Dual GIP/GLP-1 receptor agonist	N (%)					
SGLT2 inhibitors	N (%)					
Insulin	N (%)					
Anti-androgen medications	N (%)					
Spironolactone	N (%)					
Flutamide	N (%)					
Finasteride	N (%)					
Bicalutamide	N (%)					
Cyproterone	N (%)					

Variable name	Estimate name	Data source				
		DK-DHR	FinOMOP-THL	SIDIAP	IPCI	CPRD GOLD
Orlistat	N (%)					
Inositol	N (%)					

CPRD GOLD = Clinical Practice Research Datalink GOLD; DK-DHR = Danish Data Health Registries; DPP-4 = dipeptidyl peptidase-4; FinOMOP-THL = Finnish Care Register for Health Care; GIP = glucose-dependent insulinotropic polypeptide; GLP-1 = glucagon-like peptide-1; IPCI = Integrated Primary Care Information; SGLT2 = sodium-glucose co-transporter 2; SIDIAP = The Information System for Research in Primary Care; VTE = venous thromboembolism.

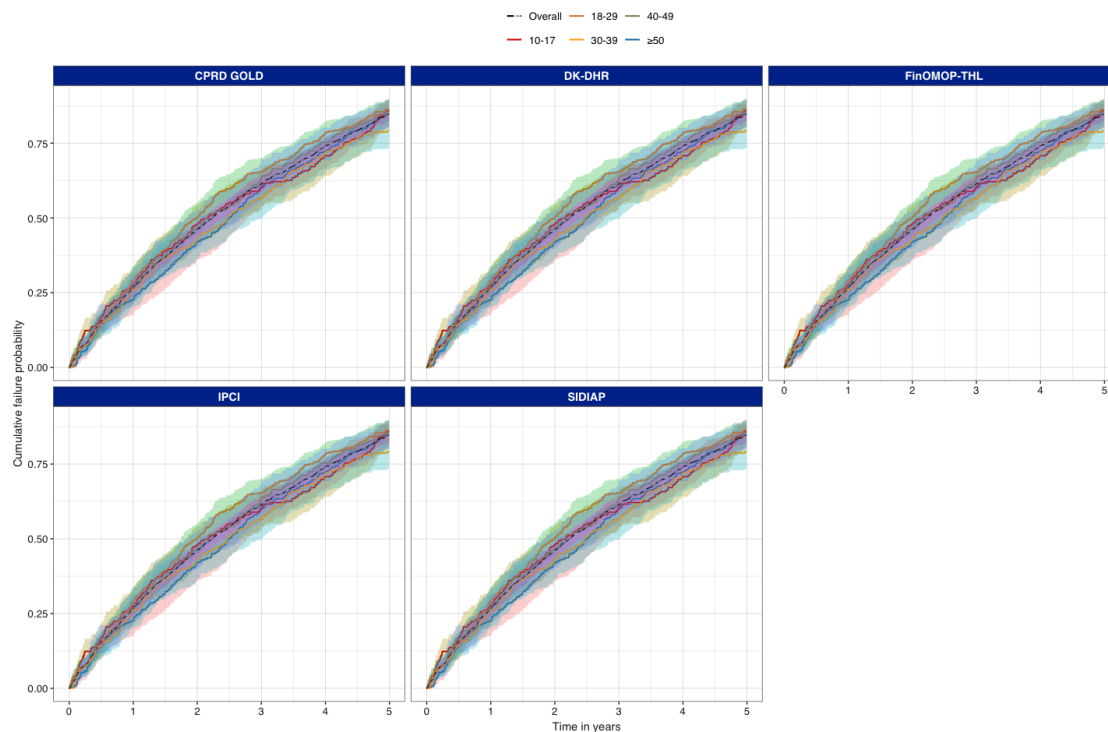
Objective 4



MOCK Figures 2–9. Annual incidence of health outcomes in polyendocrine metabolic ovarian syndrome (PMOS) individuals, overall and stratified by age group.

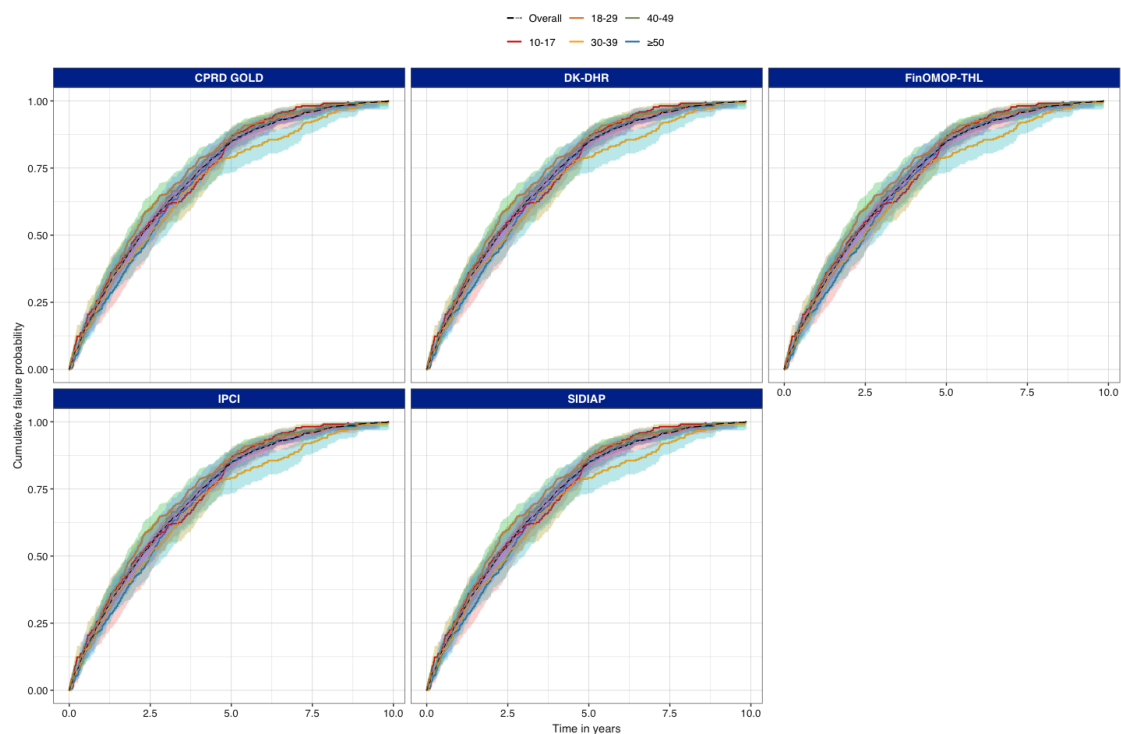
CPRD GOLD = Clinical Practice Research Datalink GOLD; DK-DHR = Danish Data Health Registries; FinOMOP-THL = Finnish Care Register for Health Care; IPCI = Integrated Primary Care Information; SIDIAP = The Information System for Research in Primary Care.

Objective 5



MOCK Figure 10–17. 5-year cumulative incidence of health outcomes in polyendocrine metabolic ovarian syndrome (PMOS) individuals, overall and stratified by age group.

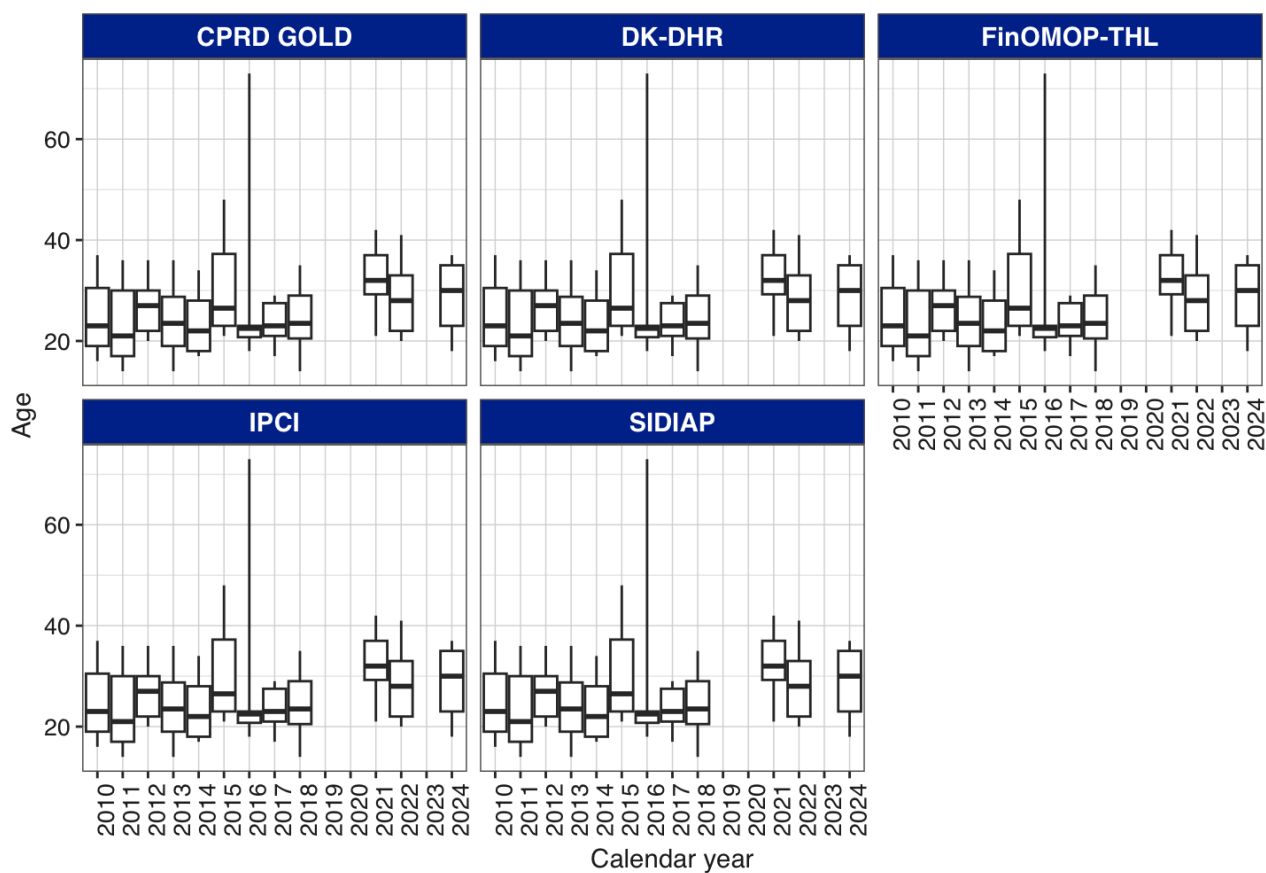
CPRD GOLD = Clinical Practice Research Datalink GOLD; DK-DHR = Danish Data Health Registries; FinOMOP-THL = Finnish Care Register for Health Care; IPCI = Integrated Primary Care Information; SIDIAP = The Information System for Research in Primary Care.



MOCK Figure 18–25. 10-year cumulative incidence of health outcomes in polyendocrine metabolic ovarian syndrome (PMOS) individuals, overall stratified by age group.

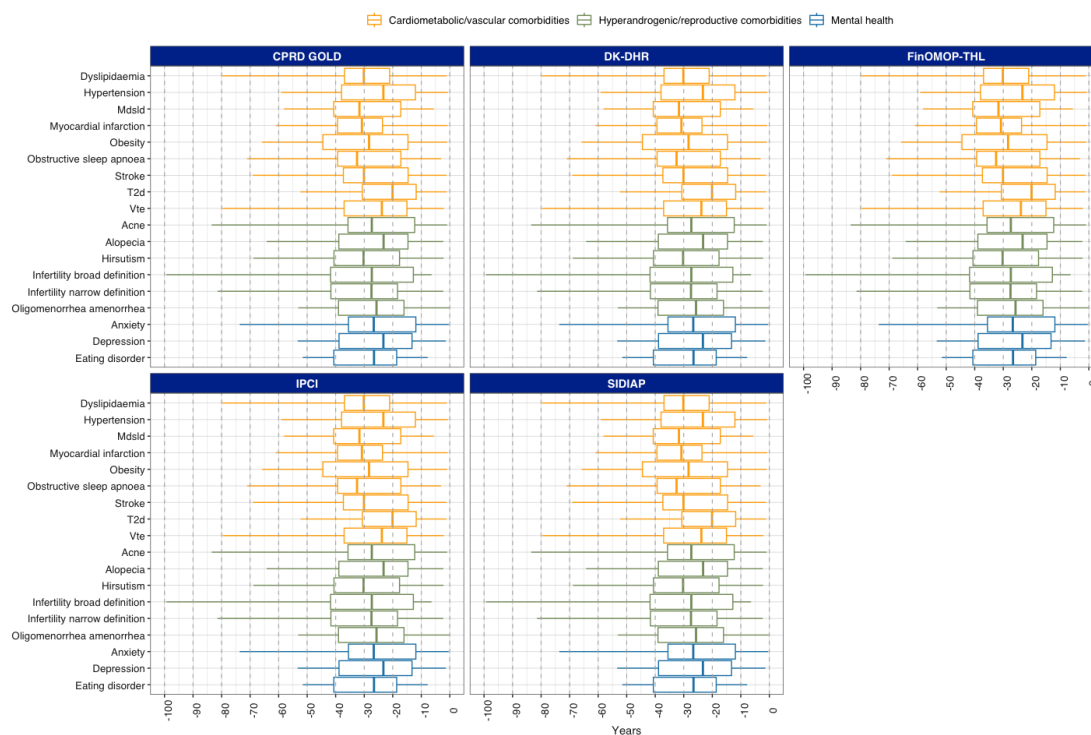
CPRD GOLD = Clinical Practice Research Datalink GOLD; DK-DHR = Danish Data Health Registries; FinOMOP-THL = Finnish Care Register for Health Care; IPCI = Integrated Primary Care Information; SIDIAP = The Information System for Research in Primary Care.

Shinyapp



MOCK Figure S1. Median age at index date by calendar year.

CPRD GOLD = Clinical Practice Research Datalink GOLD; DK-DHR = Danish Data Health Registries; FinOMOP-THL = Finnish Care Register for Health Care; IPCI = Integrated Primary Care Information; SIDIAP = The Information System for Research in Primary Care.



MOCK Figure S2. Median time from the first recorded occurrence of each pre-defined comorbidity of interest to the first recorded PMOS diagnosis.

CPRD GOLD = Clinical Practice Research Datalink GOLD; DK-DHR = Danish Data Health Registries; FinOMOP-THL = Finnish Care Register for Health Care; IPCI = Integrated Primary Care Information; MDSLD = metabolic dysfunction-associated steatotic liver disease; SIDIAP = The Information System for Research in Primary Care; T2d = type 2 diabetes; Vte= venous thromboembolism.

ANNEX VI. ENCePP checklist for study protocols

ENCePP Checklist for Study Protocols (Revision 4)

Study title: DARWIN EU® - Incidence and characterisation of polyendocrine metabolic ovarian syndrome in Europe

EU PAS Register® number: EUPAS1000001090

Study reference number (if applicable): P5-C1-013

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)	☐	☐	☒	
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 <i>a priori</i> 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.

<u>Section 3: Study design</u>		Yes	No	N/A	Section Number
3.1	Is the study design described? (e.g. cohort, case-control, cross-sectional, other design)	☒	☐	☐	8.1
3.2	Does the protocol specify whether the study is based on primary, secondary or combined data collection?	☒	☐	☐	8.4
3.3	Does the protocol specify measures of occurrence? (e.g., rate, risk, prevalence)	☒	☐	☐	8.6, 8.8
3.4	Does the protocol specify measure(s) of association? (e.g. risk, odds ratio, excess risk, rate ratio, hazard ratio, risk/rate difference, number needed to harm (NNH))	☐	☐	☒	
3.5	Does the protocol describe the approach for the collection and reporting of adverse events/adverse reactions? (e.g. adverse events that will not be collected in case of primary data collection)	☐	☐	☒	

Comments:

<u>Section 4: Source and study populations</u>		Yes	No	N/A	Section Number
4.1	Is the source population described?	☒	☐	☐	8.4
4.2	Is the planned study population defined in terms of:				
4.2.1	Study time period	☒	☐	☐	8.5
4.2.2	Age and sex	☒	☐	☐	8.3
4.2.3	Country of origin	☒	☐	☐	8.4
4.2.4	Disease/indication	☐	☐	☒	
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:

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

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

Comments:

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

Comments:

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

Comments:

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

Comments:

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

Comments:

Section 10: Analysis plan	Yes	No	N/A	Section Number
10.1 Are the statistical methods and the reason for their choice described?	☒	☐	☐	8.8
10.2 Is study size and/or statistical precision estimated?	☒	☐	☐	8.8
10.3 Are descriptive analyses included?	☒	☐	☐	8.8

<u>Section 10: Analysis plan</u>	Yes	No	N/A	Section Number
10.4 Are stratified analyses included?	☒	☐	☐	8.8
10.5 Does the plan describe methods for analytic control of confounding?	☐	☐	☒	
10.6 Does the plan describe methods for analytic control of outcome misclassification?	☐	☐	☒	
10.7 Does the plan describe methods for handling missing data?	☒	☐	☐	8.8
10.8 Are relevant sensitivity analyses described?	☐	☐	☒	

Comments:

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

Comments:

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

Comments:

<u>Section 13: Ethical/data protection issues</u>	Yes	No	N/A	Section Number
13.1 Have requirements of Ethics Committee/ Institutional Review Board been described?	☒	☐	☐	Annex II

<u>Section 13: Ethical/data protection issues</u>	Yes	No	N/A	Section Number
13.2 Has any outcome of an ethical review procedure been addressed?	☐	☐	☒	
13.3 Have data protection requirements been described?	☒	☐	☐	Annex III

Comments:

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

Comments:

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

Comments:

Name of the main author of the protocol:

Amy Lam, Anna Saura-Lazaro

Date: 24/07/2026

Signature:

A. Lam, Anna Saura-Lazaro

ANNEX VII. Glossary

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

Aggregated Data

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

Benefit-Risk Assessment

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

Common Data Model (CDM)

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

Complex Studies (C3)

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

Coordination Centre (CC)

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

Data Access

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

Data Quality Framework

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

Data Source

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

DARWIN EU®

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

EMA (European Medicines Agency)

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

Evidence Generation

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

Federated Network

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

GDPR (General Data Protection Regulation)

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

Health Technology Assessment (HTA)

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

Metadata

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

Off-the-Shelf Studies (OTS)

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

OHDSI (Observational Health Data Sciences and Informatics)

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

Patient-Level Data

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

OMOP (Observational Medical Outcomes Partnership)

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

Real-World Data (RWD)

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

Real-World Evidence (RWE)

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

Regulatory Decision-Making

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

Routine Repeated Studies (RR)

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

Study Protocol

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

Very Complex Studies (C4)

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