Study Protocol

P5-C1-022

DARWIN EU® - Characterisation of pharmacogenetic testing data in routine healthcare

13/10/2026

Version 2.0

EUPAS1000001139

Authors: Cindy Porras, Nicholas Hunt, Natasha Yefimenko, Katia Verhamme, Ross Williams, Cesar Barboza

Public

This document represents the views of the DARWIN EU® Coordination Centre only and cannot be interpreted as reflecting those of the European Medicines Agency or the European Medicines Regulatory Network.

CONTENTS

LIST OF ABBREVIATIONS **4**

1. ABSTRACT **5**

2. MILESTONES **7**

3. RATIONALE AND BACKGROUND **7**

4. OBJECTIVES **7**

5. RESEARCH METHODS **8**

5.1. Study design 8

Figure 1. Graphical depiction of the study design. 8

5.2. Follow-up 8

5.3. Study population with inclusion and exclusion criteria 9

5.4. Study setting and data sources 9

Table 1. Data sources 10

5.5. Study period 10

5.6. Variables 11

5.6.1. Exposure 11

5.6.2. Outcome 11

5.6.3. Variables of interest 11

5.7. Study size 14

5.8. Analysis 14

5.8.1. Federated network analyses 14

5.8.2. Data privacy protection 15

5.8.3. Statistical model specification and assumptions of the analytical approach considered 15

5.8.4. Sensitivity analysis 16

5.8.5. Output 16

MOCK Table 1. Frequency of pharmacogenetic testing, overall, by pharmacogenetic test type, and data source. 17

MOCK Table 2a. Availability of pharmacogenetic test results, overall, by pharmacogenetic test, and by data source. 18

MOCK Table 2b. Summarised pharmacogenetic test results, by pharmacogenetic test and by data source. 20

MOCK Table 3. Demographic characteristics of individuals with a recorded pharmacogenetic test, by pharmacogenetic test and data source. 21

MOCK Table 4. Comorbidities of individuals with a recorded pharmacogenetic test, by pharmacogenetic test and data source. 21

MOCK Table 5. Test-related medication recorded before and after pharmacogenetic test (-60 to -1 days and 0 to 60 days), by pharmacogenetic test and data source. 22

MOCK Table 6. Sensitivity analysis: Test-related medication recorded before and after pharmacogenetic test (-120 to -1 days and 0 to +120 days), by pharmacogenetic test and data source. 23

MOCK Figure 1. Annual frequency of pharmacogenetic test by data source. 25

5.9. Evidence synthesis 25

6. STRENGTHS AND LIMITATIONS **25**

7. REFERENCES **27**

8. ANNEXES **28**

ANNEX I. Description of the Study Team 28
ANNEX II. Protocol Amendments 29
ANNEX III. Description of the data sources 30
Croatia: Croatian National Public Health Information System (NAJS) 30
Estonia: Estonian BioBank (EBB) 33
Spain: Institut Municipal Assistència Sanitària Information System (IMASIS) 36
The United Kingdom: Clinical Practice Research Datalink Aurum (CPRD Aurum) 38
The United Kingdom: UK BioBank (UKBB) 41
The United States: IQVIA US - Ambulatory EMR (IQVIA US - AmbEMR) 43
ANNEX IV. Fitness for use assessment 45
ANNEX V. Operational and reporting considerations 47
ANNEX VI. List of stand-alone documents 48
Table S2. Preliminary list of conditions definitions. 48
Table S3. Preliminary list of medicines definitions. 49
ANNEX VII. ENCePP checklist for study protocols 51

LIST OF ABBREVIATIONS

Acronyms/terms	Description
ATC	Anatomical Therapeutic Chemical
CC	Coordination Centre
CDM	Common Data Model
CI	Confidence Interval
CIPH	Croatian Institute of Public Health
CPRD Aurum	Clinical Practice Research Datalink
DARWIN EU®	Data Analysis and Real-World Interrogation Network
DRE	Digital Research Environment
DTZ	Data Transfer Zone
EBB	Estonian BioBank
EHR	Electronic Health Records
EMA	European Medicines Agency
ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
Erasmus MC	Erasmus University Medical Center
EU	European Union
EUPAS	EU Post-Authorisation Studies Register
GDPR	General Data Protection Regulation
GP	General Practitioner
ICD	International Classification of Diseases
IMASIS	Institut Municipal Assistència Sanitària Information System
IQR	Interquartile range
IRB	Institutional Review Board
IQVIA US - AmbEMR	IQVIA US - Ambulatory EMR
N	Total number of individuals included for the study in each data source
n	number of individuals with the characteristic of interest
NAJS	Croatian National Public Health Information System
MHRA	Medicines and Healthcare products Regulatory Agency.
OMOP	Observational Medical Outcomes Partnership
Oxford NDORMS	University of Oxford - Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences
PGx	Pharmacogenetic
RWD	Real-World Data
RxNorm	Medical prescription normalised
SD	Standard deviation
SmPC	Summary of Product Characteristics
SNOMED	Systematized Nomenclature of Medicine

1. ABSTRACT

Rationale

Pharmacogenetic testing is most commonly used when genetic variation may significantly influence the efficacy or safety of a drug, potentially impacting dosing. The increasing use of genetic testing in clinical practice has the potential to inform treatment decisions and support more personalised patient care. However, it remains unclear to what extent genetic test orders and test results are captured in Real-World data (RWD) sources that reflect routine clinical practice. Evaluating the availability of these data, along with information on medications prescribed around the time of testing, is an important step in assessing the potential of RWD to support research on the clinical use and impact of genetic testing.

Objectives

The aim of this study is to characterise pharmacogenetic testing data in routine healthcare.

The specific objectives are:

1. To describe the frequency of pharmacogenetic testing within the DARWIN EU® Data Partner Network, by data source and calendar year.
2. To describe the availability and summarised results of pharmacogenetic tests, by type of pharmacogenetic test and data source.
3. To characterise individuals with a record of a pharmacogenetic test in terms of demographic characteristics, comorbidities, and drugs prescribed before and after testing, overall and by pharmacogenetic test.

Methods

Study design

A retrospective cohort study will be conducted using data from the DARWIN EU[å] Network.

Population

For Objective 1, the study population will include all individuals contributing observable person-time in participating DARWIN EU® Data Sources during the study period. For Objective 2, all recorded pharmacogenetic tests of interest will be included. For Objective 3, individuals with at least one recorded pharmacogenetic test of interest will be characterised in terms of demographics, comorbidities, and drug prescribed before and after testing.

Variables

Pharmacogenetic tests of interest

Covariates for characterisation:

- Sex
- Age (categorised as follows: 0–£18 years, >18–£65 years, >65 years)
- Comorbidities and drugs at the time of pharmacogenetic testing

Statistical analyses

The number and type of pharmacogenetic tests and the number of individuals undergoing testing will be reported using counts and proportions, overall (i.e. for all the tests) and by pharmacogenetic test type. The frequency of pharmacogenetic testing will be expressed as the number of tests per 1,000 person-years, with 95% confidence intervals, overall and by test type (Objective 1).

The availability of test results and the distribution of recorded results will be summarised by test type and data source (Objective 2). Individuals undergoing testing will be characterised in terms of age, sex, and comorbidities, and medication will be reported at the index date, overall, by pharmacogenetic test type, and data source (Objective 3).

Data sources

Croatian National Public Health Information System (NAJS); Estonian BioBank (EBB); Institut Municipal Assistència Sanitària Information System (IMASIS); Clinical Practice Research Datalink Aurum (CPRD Aurum); UK BioBank (UKBB); IQVIA US - Ambulatory EMR (IQVIA US - AmbEMR)

Study size

No sample size will be calculated, as this is an exploratory study which will not test a specific hypothesis.

2. MILESTONES

Study milestones and deliverables	Planned dates*
Final Study Protocol	6 October 2026
Creation of Analytical code	October 2026
Execution of Analytical Code on the data	November 2026
Draft Study Report	15 January 2027
Final Study Report	January 2027

*Planned dates are dependent on obtaining approval from the internal review boards of the data sources, code execution by the data sources, and deliverable review cycles.

3. RATIONALE AND BACKGROUND

Drug response is influenced by multiple factors, including genetic variations affecting drug metabolism, transport, and immune-mediated reactions.[1] Pharmacogenetic testing may be particularly relevant when altered drug exposure or response could result in serious adverse reaction, therapeutic failure, or the need of dose adjustment.[2] Regulatory agencies and expert groups, including the Dutch Pharmacogenetics Working Group, have developed recommendations for a growing number of gene-drug associations. However, the strength of evidence and the recommended clinical action vary between gene-drug pairs, and the presence of an association does not necessarily imply that testing should be performed before prescribing.[3, 4]

As pharmacogenetic testing becomes increasingly available,[5, 6] Real-World Data (RWD) could support the evaluation of how testing is used in clinical practice, which patients undergo testing, and whether prescribing patterns change around the time of testing. However, the extent to which test orders and results are captured in RWD sources remains unclear. Assessing their availability and linking them to medications prescribed before and after testing may help differentiate between testing performed before treatment initiation and testing prompted by treatment failure, adverse effects, or other clinical concerns.[7-9] This study may help identify current data gaps and inform the methodological and infrastructural improvements needed to investigate the clinical use and impact of pharmacogenetic testing using RWD.

4. OBJECTIVES

The aim of this study is to characterise pharmacogenetic testing data in routine healthcare.

The specific objectives are:

1. To describe the frequency of pharmacogenetic testing within the DARWIN EU® Data Partner Network, by data source and calendar year.
2. To describe the availability and summarised results of pharmacogenetic tests, by type of pharmacogenetic test and data source.
3. To characterise individuals with a record of pharmacogenetic test in terms of demographic characteristics, comorbidities, and drugs prescribed before and after testing, overall and by pharmacogenetic test.

5. RESEARCH METHODS

5.1. Study design

A retrospective cohort study will be conducted using routinely collected health data from 6 data sources from 4 countries across Europe and the United States. The study will comprise:

- A description of the frequency of pharmacogenetic testing within the DARWIN EU® Data Network.
- A description of the availability and summarised results of pharmacogenetic tests.
- A characterisation of individuals undergoing pharmacogenetic testing.

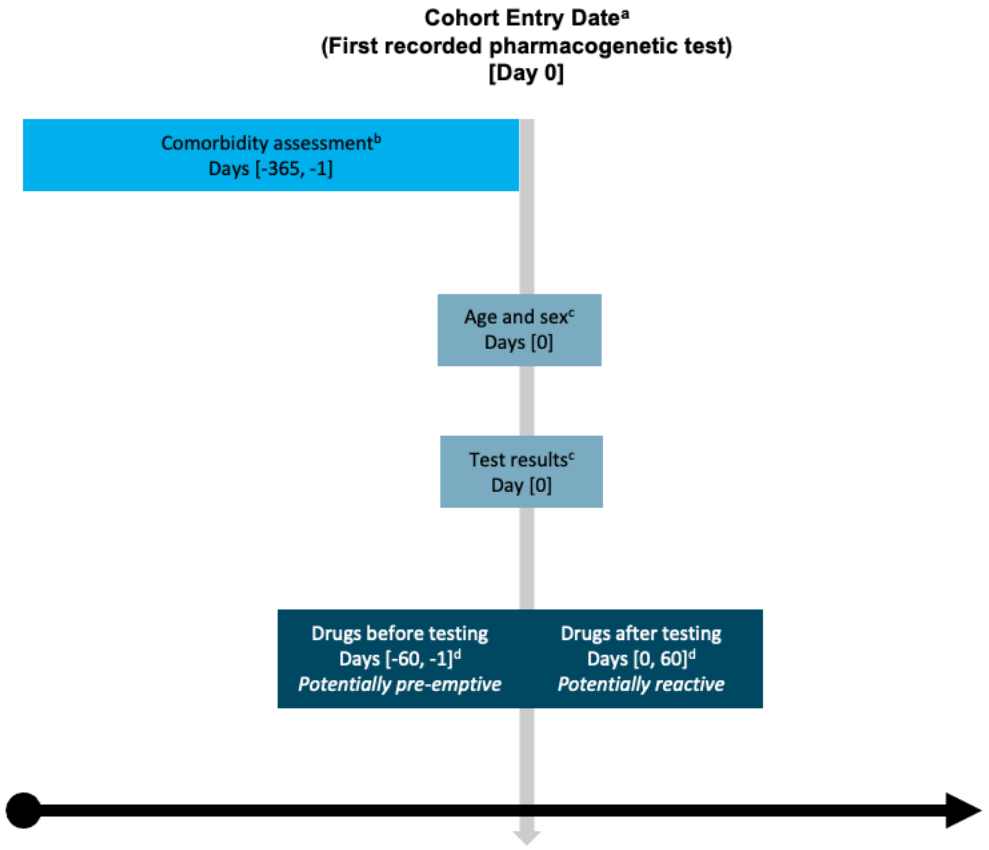

Figure 1. Graphical depiction of the study design.

a. Cohort entry is defined as the date of the first recorded pharmacogenetic test of interest during the study period (index date).
b. Comorbidities of interest will be assessed during the 365 days of continuous observation before cohort entry.
c. Age, sex, and test results will be recorded at cohort entry.
d. For the sensitivity analysis, the medication assessment window will be extended to days -120 to -1 (prescriptions before testing) and day 0 to 120 (prescriptions after testing).

5.2. Follow-up

Follow-up will vary according to the study objective.

For Objective 1, individuals in the source population will contribute person-time from the latest of the start of the study period or the start of their observation period until the earliest of death, the end of their

observation period, the end of data availability, or the end of the study period. Pharmacogenetic tests recorded during eligible person-time will contribute to the numerator used to estimate testing frequency per 1,000 person-years.

For Objective 2, cohort entry will be defined as the date of the first recorded pharmacogenetic test of interest (index date). The availability and results of pharmacogenetic tests will be assessed at the index date. No follow-up will be required for this objective.

For Objective 3, age and test results will be assessed at the index date. Comorbidities will be assessed during the 365-day period before the index date. No post-index follow-up is required for these characteristics. Drug prescribing will be assessed during the 60 days before and 60 days after the index date. For post-index drug prescription, individuals will be followed from the index date until the earliest of 60 days after the index date, death, the end of the individual's observation period, the end of data availability, or the end of the study period.

5.3. Study population with inclusion and exclusion criteria

Objective 1

The study population will include all individuals contributing eligible person-time during the study period and the data source specific period of reliable pharmacogenetic data capture.

Inclusion criteria:

- At least one day of observation overlapping with the study period and the data source-specific period of reliable pharmacogenetic data capture.

Objective 2

Separate cohorts will be defined for each pharmacogenetic test type.

Inclusion criteria:

- At least one pharmacogenetic test of interest recorded during the eligible study period.

Objective 3

The study population will include individuals with at least one pharmacogenetic test of interest recorded during the study period.

Inclusion criteria:

- At least 365 days of prior observation for the comorbidity assessment.
- At least 60 days of prior observation for pre-index drug prescription assessment.
- At least 60 days of post-index observation for the post-index drug prescription assessment.

5.4. Study setting and data sources

This study will be conducted using routinely collected data from 6 data sources in the DARWIN EU® Network of data partners from 4 European countries and the United States. 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	Healthcare setting	Type of data	Number of active individuals	Calendar period covered by each data source	Contributing to
Croatia	Croatian National Public Health Information System (NAJS)	Primary care, outpatient specialist care, and inpatient care	Registry	4,835,791	2014–2025	All objectives*
Estonia	Estonian BioBank (EBB)	Primary and secondary care	Registry	212,300	2020–2024	All objectives
Spain	Institut Municipal Assistència Sanitària Information System (IMASIS)	Secondary care	Inpatient (Hospital) Care	169,560	1990–2025	All objectives
United Kingdom	Clinical Practice Research Datalink (CPRD Aurum)	Primary care	Outpatient General Practitioner Care	17,438,460	1997–2025	All objectives
United Kingdom	UK BioBank (UKBB)	Primary and secondary care	Registry + GP + Hospital	499,900	1985–2017	All objectives
United States	IQVIA US - Ambulatory EMR (IQVIA US - AmbEMR)	Primary and secondary care	Outpatient General Practitioner Care	9,132,000	2006–2025	All objectives

*NAJS cannot provide summarised test results for Objective 1.
GP: general practitioner.

Data sources selection

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

5.5. Study period

The study period is from 01/01/2015 (or the earliest data available) to 31/12/2025 (or the most recent data available) for each contributing data source.

It should be noted that for UKBB, the study period will end in 2017, as the primary care data, including General Practitioner (GP) prescriptions, are only updated up to that year. For EBB, the study period will

begin in 2020 or 2021, depending on the confirmed date from which pharmacogenetic tests are reliably available.

5.6. Variables

5.6.1. Exposure

There are no exposures of interest for any of the objectives.

5.6.2. Outcome

For Objective 1, the outcomes of interest are the following pharmacogenetic tests:

- ABCG2/SLCO1B1
- CFTR
- CYP2B6
- CYP2C19
- CYP2C19/CYP2D6
- CYP2C19/CYP3A4
- CYP2C9/VKORC
- CYP2D6
- CYP3A5
- DPYD
- F5
- G6PD
- HLA-B*15:02
- HLA-B*15:02/HLA-A*31:01
- HLA-B*57:01
- HLA-B*58:01
- SLCO1B1
- TPMT/NUDT15
- UGT1A1

For Objectives 2 and 3, there are no outcomes of interest.

5.6.3. Variables of interest

Objective 1:

The following variables will be used to describe the frequency of pharmacogenetic testing:

- Data source
- Calendar year
- Type of pharmacogenetic test
- Number of pharmacogenetic tests recorded

- Number of individuals with a recorded pharmacogenetic test
- Eligible observable person-time

Objective 2:

The following variables will be used to describe the availability and results of pharmacogenetic tests:

- Type of pharmacogenetic test
- Availability of a recorded test result
- Result value or category, if available
- Data source

Objective 3:

The following individual characteristics will be described overall and by pharmacogenetic test type.

- Sex
    - Male/Female
- Age groups, defined as follows:
    - 0–≤18 years
    - >18–≤65 years
    - >65 years

The following covariates will be used to characterise individuals receiving pharmacogenetic tests in terms of comorbidities:

- Liver disease: chronic liver disease, chronic hepatitis, viral hepatitis, steatotic liver disease, hepatic fibrosis, cirrhosis, portal hypertension, hepatic failure, autoimmune liver disease.
- Kidney disease: acute kidney injury, chronic kidney disease, end-stage kidney disease, dialysis.
- Gastrointestinal disease: gastro-oesophageal reflux disease, gastritis, duodenitis, peptic ulcer disease, inflammatory bowel disease, coeliac disease, intestinal malabsorption, pancreatitis.
- Cardiovascular disease: hypertensive disorder, ischemic heart disease, heart failure, cardiac arrhythmias, cerebrovascular disease, peripheral arterial disease, heart valve disorder.
- Inflammatory disease: rheumatoid arthritis, psoriatic arthritis, vasculitis, psoriasis, ankylosing spondylitis, systemic lupus erythematosus.
- Psychiatric condition: depressive disorders, anxiety disorders, bipolar disorder, psychotic disorders, substance use disorders, personality disorders.
- Neurological condition: Epilepsy, Parkinson disease, dementia, multiple sclerosis, neuropathic pain, neuropathy.
- Any cancer: any malignant solid tumour, haematological malignancies, metastatic cancer, other malignant neoplasms.

The following drugs will be used to characterise individuals receiving pharmacogenetic tests in terms of drugs prescribed before and after testing:

- ABCG2/SLCO1B1:

  - Rosuvastatin

- CFTR:
  - Ivacaftor
- CYP2B6:
  - Efavirenz
- CYP2C19:
  - Clopidogrel
  - Omeprazole
  - Voriconazole
  - Citalopram
  - Escitalopram
  - Mavacamten
  - Sertraline
- CYP2C19/CYP2D6:
  - Amitriptyline
- CYP2C19/CYP3A4:
  - Lansoprazole
- CYP2C9/VKORC1:
  - Warfarin
- CYP2D6:
  - Codeine
  - Tramadol
  - Tamoxifen
  - Atomoxetine
  - Ondansetron
- CYP3A5:
  - Tacrolimus; genetic impact not described in the Summary of Product Characteristics (SmPC)[10]
- DPYD:
  - Fluorouracil
  - Capecitabine
  - Tegafur
- F5:
  - Levonorgestrel and ethinylestradiol
  - Desogestrel and ethinylestradiol

- G6PD:

  - Rasburicase

- HLA-B*15:02:

  - Oxcarbazepine

  - Phenytoin

- HLA-B*15:02/HLA-A*31:01:

  - Carbamazepine

- HLA-B*57:01:

  - Abacavir

- HLA-B*58:01:

  - Allopurinol

- SLCO1B1:

  - Simvastatin

  - Atorvastatin

- TPMT/NUDT15:

  - Azathioprine

  - Mercaptopurine

  - Thioguanine

- UGT1A1:

  - Irinotecan

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.

5.7. Study size

No sample size has been calculated, as this is a characterisation study which will not test a specific hypothesis.

5.8. Analysis

5.8.1. 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 V**), the final package will be released in a version-controlled study repository for execution against all the participating data sources.

5.8.2. 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 will be made available to the team, and the Study Dissemination Phase can start. All analyses will be conducted separately for each data source and will be carried out in a federated manner, allowing analyses to be run locally without sharing patient-level data. Cell counts <5 will be suppressed when reporting results to comply with the data source's privacy protection regulations.

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

The following R packages will be used: *IncidencePrevalence* (developed by DARWIN EU®) for Objective 1, *MeasurementDiagnostics* for Objective 2, and *CohortCharacteristics* (developed by DARWIN EU®) for Objective 3.[11-13]

Objective 1

For each data source, the number of individuals contributing person-time, the total person-years of observation, the number of individuals with at least one pharmacogenetic test, and the number of pharmacogenetic tests will be reported.

Eligible person-time will be restricted to the data source-specific period during which pharmacogenetic testing is considered reliably captured. The frequency of pharmacogenetic testing will be calculated as the number of pharmacogenetic tests divided by the total eligible person-time and expressed as the number of tests per 1,000 person-years, with the corresponding 95% confidence interval (CI). Estimates will be reported overall and by pharmacogenetic test type, data source, and calendar year.

An individual may contribute multiple test events during the study period. Different pharmacogenetic test types recorded on the same date will be retained as separate test events. Multiple records of the same test type for the same individual on the same date will be considered as a single test event.

$$\text{Testing frequency} = \frac{\text{Number of pharmacogenetic tests recorded during eligible person time}}{\text{Total eligible person time in years}} \text{ X } 1000$$

Objective 2

For each pharmacogenetic test type and data source, the total number of recorded test events, number and proportion of test events with available results, and result recorded, including numeric values and categorical variables, will be reported.

Numeric results will be summarised using median and interquartile range (IQR). Categorical results will be summarised using counts and percentages.

Objective 3

Descriptive characterisation will be performed at patient level, overall and by pharmacogenetic test type. The index date will be defined as the date of the first recorded pharmacogenetic test of interest. Age and sex will be assessed at the index date. Age will be summarised using mean and standard deviation (SD), and using counts and percentages across 10-year age groups.

Comorbidities will be assessed within the 365 days prior to the index date. The comorbidities of interest are described in **Section 5.6.3**.

The number and percentage of individuals with a prescription of any of the drugs of interest will be assessed during day -60 to 1, and days 0 to 60 relative to the index date. Prescriptions reported before

testing will be classified as potentially pre-emptive, while prescriptions recorded after testing will be classified as potentially reactive.

5.8.4. Sensitivity analysis

A sensitivity analysis will assess associated medication use over a wider window around the first recorded pharmacogenetic test 120 days before testing (−120 to −1 days) and 120 days after testing (0 to 120 days). Results will be compared with those from the main analysis using the 60-day windows to assess whether the choice of window affects the observed medication-use patterns.

5.8.5. Output

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

- Table 1. Frequency of pharmacogenetic testing, overall, by pharmacogenetic test type, and data source.
- Table 2a. Availability of pharmacogenetic test results, overall, by pharmacogenetic test, and by data source.
- Table 2b. Summarised pharmacogenetic test results, by pharmacogenetic test and by data source.
- Table 3. Demographic characteristics of individuals with a recorded pharmacogenetic test, by pharmacogenetic test and data source.
- Table 4. Comorbidities among individuals with a recorded pharmacogenetic test, by pharmacogenetic test and data source.
- Table 5. Test-related medication recorded before and after pharmacogenetic test (-60 to -1 days and 0 to 60 days), by pharmacogenetic test and data source.
- Table 6. Sensitivity analysis: Test-related medication recorded before and after pharmacogenetic test (-120 to -1 days and 0 to 120 days, by pharmacogenetic) test and data source.
- Figure 1. Annual frequency of pharmacogenetic testing, by data source.

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

MOCK Table 1. Frequency of pharmacogenetic testing, overall, by pharmacogenetic test type, and data source.

Data source	PGx test	Individuals contributing eligible person-time, N	Eligible person-years	Individuals tested, N	Test events, N	Tests per 1,000 person-years (95%CI)
NAJS						
	Any PGx test					
	ABCG2 / SLCO1B1					
	CFTR					
	CYP2B6					
	CYP2C19					
	CYP2C19 / CYP2D6					
	CYP2C19 / CYP3A4					
	CYP2C9 / VKORC1					
	CYP2D6					
	CYP3A5					
	DPYD					
	F5					
	G6PD					
	HLA-B*15:02					
	HLA-B*15:02 / HLA-A*31:01					
	HLA-B*57:01					
	HLA-B*58:01					
	SLCO1B1					
	TPMT / NUDT15					
	UGT1A1					
EBB						
	Any PGx test					
	ABCG2 / SLCO1B1					
	CFTR					
	CYP2B6					
	CYP2C19					
	CYP2C19 / CYP2D6					
	CYP2C19 / CYP3A4					
	CYP2C9 / VKORC1					
	CYP2D6					
	CYP3A5					
	DPYD					

Data source	PGx test	Individuals contributing eligible person-time, N	Eligible person-years	Individuals tested, N	Test events, N	Tests per 1,000 person-years (95%CI)
	F5					
	G6PD					
	HLA-B*15:02					
	HLA-B*15:02 / HLA-A*31:01					
	HLA-B*57:01					
	HLA-B*58:01					
	SLCO1B1					
	TPMT / NUDT15					
	UGT1A1					
…						

N: total number of individuals included for the study in each data source; CI: confidence interval
PGx: pharmacogenetic; NAJS: Croatian National Public Health Information System; EBB: Estonian BioBank; IMASIS: Institut Municipal Assistència Sanitària Information System; CPRD Aurum: Clinical Practice Research Datalink Aurum; UKBB: UK BioBank; IQVIA US - AmbEMR: IQVIA US - Ambulatory EMR.

MOCK Table 2a. Availability of pharmacogenetic test results, overall, by pharmacogenetic test, and by data source.

Data Source	PGx test	Eligible test records, N	Results available, n(%)	Numeric result, (%)	Categorical result, (%)
NAJS					
	ABCG2 / SLCO1B1				
	CFTR				
	CYP2B6				
	CYP2C19				
	CYP2C19 / CYP2D6				
	CYP2C19 / CYP3A4				
	CYP2C9 / VKORC1				
	CYP2D6				
	CYP3A5				
	DPYD				
	F5				
	G6PD				
	HLA-B*15:02				
	HLA-B*15:02 / HLA-A*31:01				
	HLA-B*57:01				
	HLA-B*58:01				
	SLCO1B1				

Data Source	PGx test	Eligible test records, N	Results available, n(%)	Numeric result, (%)	Categorical result, (%)
	TPMT / NUDT15				
	UGT1A1				
EBB					
	ABCG2 / SLCO1B1				
	CFTR				
	CYP2B6				
	CYP2C19				
	CYP2C19 / CYP2D6				
	CYP2C19 / CYP3A4				
	CYP2C9 / VKORC1				
	CYP2D6				
	CYP3A5				
	DPYD				
	F5				
	G6PD				
	HLA-B*15:02				
	HLA-B*15:02 / HLA-A*31:01				
	HLA-B*57:01				
	HLA-B*58:01				
	SLCO1B1				
	TPMT / NUDT15				
	UGT1A1				
…					

*NAJS cannot contribute to the analysis of PGx test results because genetic measurement results are not mapped to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM).
N: total number of individuals included for the study in each data source; n: number of individuals with the comorbidity of interest
PGx: pharmacogenetic; NAJS: Croatian National Public Health Information System; EBB: Estonian BioBank; IMASIS: Institut Municipal Assistència Sanitària Information System; CPRD Aurum: Clinical Practice Research Datalink Aurum; UKBB: UK BioBank; IQVIA US - AmbEMR: IQVIA US - Ambulatory EMR.

MOCK Table 2b. Summarised pharmacogenetic test results, by pharmacogenetic test and by data source.

Data source	Pharmacogenetic test	Result variable/category	Unit	Result, N (%)	Median (IQR)
NAJS					
	ABCG2 / SLCO1B1				
	CFTR				
	CYP2B6				
	CYP2C19				
	CYP2C19 / CYP2D6				
	CYP2C19 / CYP3A4				
	CYP2C9 / VKORC1				
	CYP2D6				
	CYP3A5				
	DPYD				
	F5				
	G6PD				
	HLA-B*15:02				
	HLA-B*15:02 / HLA-A*31:01				
	HLA-B*57:01				
	HLA-B*58:01				
	SLCO1B1				
	TPMT / NUDT15				
	UGT1A1				
EBB					
	ABCG2 / SLCO1B1				
	CFTR				
	CYP2B6				
	CYP2C19				
	CYP2C19 / CYP2D6				
	CYP2C19 / CYP3A4				
	CYP2C9 / VKORC1				
	CYP2D6				
	CYP3A5				
	DPYD				
	F5				
	G6PD				
	HLA-B*15:02				
	HLA-B*15:02 / HLA-A*31:01				

Data source	Pharmacogenetic test	Result variable/category	Unit	Result, N (%)	Median (IQR)
	HLA-B*57:01				
	HLA-B*58:01				
	SLCO1B1				
	TPMT / NUDT15				
	UGT1A1				
…					

*NAJS cannot contribute to the analysis of PGx test results because genetic measurement results are not mapped to the OMOP CDM.
N: total number of individuals included for the study in each data source; IQR: interquartile range.
NAJS: Croatian National Public Health Information System; EBB: Estonian BioBank; IMASIS: Institut Municipal Assistència Sanitària Information System; CPRD Aurum: Clinical Practice Research Datalink Aurum; UKBB: UK BioBank; IQVIA US - AmbEMR: IQVIA US - Ambulatory EMR.

MOCK Table 3. Demographic characteristics of individuals with a recorded pharmacogenetic test, by pharmacogenetic test and data source.

		NAJS	EBB	IMASIS	CPRD Aurum	UKBB	IQVIA US - AmbEMR
PGx test	Individual, N						
	Age, mean (SD)						
	0–≤18 years, n (%)						
	>18–≤65 years, n (%)						
	>65 years, n (%)						

PGx: pharmacogenetics
SD: standard deviation; N: total number of individuals included for the study in each data source; n: number of individuals with the characteristic of interest.
NAJS: Croatian National Public Health Information System; EBB: Estonian BioBank; IMASIS: Institut Municipal Assistència Sanitària Information System; CPRD Aurum: Clinical Practice Research Datalink Aurum; UKBB: UK BioBank; IQVIA US - AmbEMR: IQVIA US - Ambulatory EMR

MOCK Table 4. Comorbidities of individuals with a recorded pharmacogenetic test, by pharmacogenetic test and data source.

		NAJS	EBB	IMASIS	CPRD Aurum	UKBB	IQVIA US - AmbEMR
PGx test	Individual, N						
	Liver disease, n (%)						
	Kidney disease, n (%)						
	Gastrointestinal disease, n (%)						
	Cardiovascular disease, n (%)						
	Inflammatory disease, n (%)						

		NAJS	EBB	IMASIS	CPRD Aurum	UKBB	IQVIA US - AmbEMR
	Psychiatric condition, n (%)						
	Neurological disease, n (%)						
	Cancer, n (%)						

PGx: pharmacogenetics
N: total number of individuals included for the study in each data source; n: number of individuals with the comorbidity of interest; NAJS: Croatian National Public Health Information System; EBB: Estonian BioBank; IMASIS: Institut Municipal Assistència Sanitària Information System; CPRD Aurum: Clinical Practice Research Datalink Aurum; UKBB: UK BioBank; IQVIA US - AmbEMR: IQVIA US - Ambulatory EMR

MOCK Table 5. Test-related medication recorded before and after pharmacogenetic test (-60 to -1 days and 0 to 60 days), by pharmacogenetic test and data source.

Data source	Pharmacogenetic test	Drug	Before testing, days −60 to −1, n (%)	After testing, days 0 to 60, n (%)
NAJS	ABCG2 / SLCO1B1	Rosuvastatin		
	CFTR	Ivacaftor		
	CYP2B6	Efavirenz		
	CYP2C19	Clopidogrel		
		Omeprazole		
		Voriconazole		
		Citalopram		
		Escitalopram		
		Mavacamten		
		Sertraline		
	CYP2C19 / CYP2D6	Amitriptyline		
	CYP2C19 / CYP3A4	Lansoprazole		
	CYP2C9 / VKORC1	Warfarin		
	CYP2D6	Codeine		
		Tramadol		
		Tamoxifen		
		Atomoxetine		
		Ondansetron		
	CYP3A5	Tacrolimus[1]		
	DPYD	Fluorouracil		
		Capecitabine		
		Tegafur		
	F5	Levonorgestrel and ethinylestradiol		
		Desogestrel and ethinylestradiol		

Data source	Pharmacogenetic test	Drug	Before testing, days −60 to −1, n (%)	After testing, days 0 to 60, n (%)
	G6PD	Rasburicase		
	HLA-B*15:02	Oxcarbazepine		
		Phenytoin		
	HLA-B15:02 / HLA-A31:01	Carbamazepine		
	HLA-B*57:01	Abacavir		
	HLA-B*58:01	Allopurinol		
	SLCO1B1	Simvastatin		
		Atorvastatin		
	TPMT / NUDT15	Azathioprine		
		Mercaptopurine		
		Thioguanine		
	UGT1A1	Irinotecan		

N: total number of individuals in the corresponding test-specific cohort; n: number of individuals with at least one record of the specified medication during the assessment window.
[1]Genetic impact not described in the Summary of Product Characteristics (SmPC)
NAJS: Croatian National Public Health Information System

MOCK Table 6. Sensitivity analysis: Test-related medication recorded before and after pharmacogenetic test (-120 to -1 days and 0 to +120 days), by pharmacogenetic test and data source.

Data source	Pharmacogenetic test	Drug	Before testing, days −120 to −1, n (%)	After testing, days 0 to 120, n (%)
NAJS	ABCG2 / SLCO1B1	Rosuvastatin		
	CFTR	Ivacaftor		
	CYP2B6	Efavirenz		
	CYP2C19	Clopidogrel		
		Omeprazole		
		Voriconazole		
		Citalopram		
		Escitalopram		
		Mavacamten		
		Sertraline		
	CYP2C19 / CYP2D6	Amitriptyline		
	CYP2C19 / CYP3A4	Lansoprazole		
	CYP2C9 / VKORC1	Warfarin		
	CYP2D6	Codeine		
		Tramadol		
		Tamoxifen		
		Atomoxetine		

Data source	Pharmacogenetic test	Drug	Before testing, days –120 to –1, n (%)	After testing, days 0 to 120, n (%)
		Ondansetron		
	CYP3A5	Tacrolimus[1]		
	DPYD	Fluorouracil		
		Capecitabine		
		Tegafur		
	F5	Levonorgestrel and ethinylestradiol		
		Desogestrel and ethinylestradiol		
	G6PD	Rasburicase		
	HLA-B*15:02	Oxcarbazepine		
		Phenytoin		
	HLA-B15:02 / HLA-A31:01	Carbamazepine		
	HLA-B*57:01	Abacavir		
	HLA-B*58:01	Allopurinol		
	SLCO1B1	Simvastatin		
		Atorvastatin		
	TPMT / NUDT15	Azathioprine		
		Mercaptopurine		
		Thioguanine		
	UGT1A1	Irinotecan		

N: total number of individuals in the corresponding test-specific cohort; n: number of individuals with at least one record of the specified medication during the assessment window.
[1]Genetic impact not described in Summary of Product Characteristics (SmPC)
NAJS: Croatian National Public Health Information System



MOCK Figure 1. Annual frequency of pharmacogenetic test by data source.

PGx: pharmacogenetic
Frequencies are the number of recorded pharmacogenetic tests divided by the total person-years of observation in the corresponding calendar year and data source, multiplied by 1,000. All prespecified test types are combined. An individual may contribute more than one test. Results are presented separately by data source.

5.9. Evidence synthesis

Results from analyses described in **Section 5.8.4** will be presented separately for each data source. No meta-analysis of results will be conducted.

6. STRENGTHS AND LIMITATIONS

This study will provide a large-scale, multi-country characterisation of pharmacogenetic testing in routine healthcare across the DARWIN EU® Network of Data Sources. The study will assess the availability of data on several clinically relevant pharmacogenetic tests and their results, where recorded, as well as the use of a prespecified selection of associated drugs. This will enable the exploration of differences in testing patterns, tested populations, and medication use across healthcare settings and a range of gene-drug pairs. The use of a common protocol, data mapped to the OMOP CDM, federated study execution, and standardised analytics will support the consistent implementation of the study definitions and analyses across participating data sources.

The study will be informed by routinely collected health care data, so data quality issues must be considered. Pharmacogenetic tests, medication use, and comorbidities recorded outside the healthcare setting covered by a data source may not be completely captured. Mapping source data to the OMOP CDM may result in loss of detail or incomplete identification of relevant records. This is particularly important for pharmacogenetic results, which may be recorded as variants, genotypes, diplotypes, activity scores, phenotypes, or free text and may not be directly comparable.[14] In NAJS, genetic measurement results are

not mapped to the OMOP CDM and, therefore, this data source will not contribute to analyses of test-result availability or distribution. In UKBB, general practice prescription data are only available until 2017, limiting the period which will be able to contribute to the medication analyses.

In addition, 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. Using electronic healthcare data from six countries, mostly European and one from the United States, will limit the generalisability.

7. REFERENCES

1.Roden DM, Wilke RA, Kroemer HK, Stein CM. Pharmacogenomics: the genetics of variable drug responses. Circulation. 2011;123(15):1661–70.

2.Goldspiel BR, Flegel WA, DiPatrizio G, Sissung T, Adams SD, Penzak SR, et al. Integrating pharmacogenetic information and clinical decision support into the electronic health record. Journal of the American Medical Informatics Association. 2014;21(3):522–8.

3.Hertz DL, Bousman CA, McLeod HL, Monte AA, Voora D, Orlando LA, et al. Recommendations for pharmacogenetic testing in clinical practice guidelines in the US. Am J Health Syst Pharm. 2024;81(16):672–83.

4.Swen JJ, Nijenhuis M, de Boer A, Grandia L, Maitland-van der Zee AH, Mulder H, et al. Pharmacogenetics: From Bench to Byte— An Update of Guidelines. Clinical Pharmacology & Therapeutics. 2011;89(5):662–73.

5.Hippman C, Nislow C. Pharmacogenomic Testing: Clinical Evidence and Implementation Challenges. J Pers Med. 2019;9(3).

6.Mosch R, van der Lee M, Guchelaar HJ, Swen JJ. Pharmacogenetic Panel Testing: A Review of Current Practice and Potential for Clinical Implementation. Annu Rev Pharmacol Toxicol. 2025;65(1):91–109.

7.Hu R, Weng K, Guo B, Zhuang S, Zhou J, Hong Z, et al. Implementation of preemptive pharmacogenetic testing: progress, puzzles and priorities from an implementation science perspective. Frontiers in Pharmacology. 2026;Volume 17 - 2026.

8.Stevenson JM, Smith DM, Tuteja S, Patel JN, the Pharmacogenomics Global Research Network Publications C. Navigating Pharmacogenomic Testing in Practice: Who to Test and When to Test. Clinical Pharmacology & Therapeutics. 2025;118(3):561–6.

9.Weitzel KW, Cavallari LH, Lesko LJ. Preemptive Panel-Based Pharmacogenetic Testing: The Time is Now. Pharm Res. 2017;34(8):1551–5.

10. Agency EM. Summary of Product Characteristics 2026 [Available from: http://ema.europa.eu/en/glossary-terms/summary-product-characteristics.

11. EU D. IncidencePrevalence 2025 [Available from: https://github.com/darwin-eu/IncidencePrevalence.

12. EU D. CohortCharacteristics 2026 [Available from: https://darwin-eu.github.io/CohortCharacteristics/.

13. OHDSI. MeasurementDiagnostics 2026 [Available from: https://github.com/OHDSI/MeasurementDiagnostics.

14. Caudle KE, Dunnenberger HM, Freimuth RR, Peterson JF, Burlison JD, Whirl-Carrillo M, et al. Standardizing terms for clinical pharmacogenetic test results: consensus terms from the Clinical Pharmacogenetics Implementation Consortium (CPIC). Genet Med. 2017;19(2):215–23.

8. ANNEXES

ANNEX I. Description of the Study Team

Study team role	Names	Organisation
Principal Investigator	Cindy Porras	Erasmus MC
Co-Principal Investigator	Nicholas Hunt	Erasmus MC
Epidemiologist	Katia Verhamme	Erasmus MC
Data Scientist	Ross Williams Cesar Barboza	Erasmus MC Erasmus MC
Study Manager	Natasha Yefimenko	Erasmus MC
Data source	**Names**	**Data Partner Organisation***
Croatian National Public Health Information System (NAJS)	Anamaria Jurčević Antea Jezidžić Ivan Pristaš Jakov Vuković	Croatian Institute of Public Health (CIPH)
Estonian BioBank (EBB)	Ami Sild Marek Oja	University of Tartu (TARTU)
Institut Municipal Assistència Sanitària Information System (IMASIS)	Juan Manuel Ramírez-Anguita Carla Arxé Angela Leis Miguel-Angel Mayer	Consorci Mar Parc de Salut de Barcelona (PSMAR) together with Fundació Institut Hospital del Mar d'investigacions Mèdiques
Clinical Practice Research Datalink Aurum (CPRD Aurum)	Zara Cuccu Adam Flynn Susan Hodgson	Clinical Practice Research Datalink (CPRD), Medicines and Healthcare products Regulatory Agency (MHRA), United Kingdom
UK BioBank (UKBB)	Antonella Delmestri Hezekiah Omulo Mandickel Kamtengeni	University of Oxford - Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences (Oxford NDORMS)
IQVIA US - Ambulatory EMR (IQVIA US - AmbEMR)	Hugo Vernooij	IQVIA

*Data Partners do not have an investigator role. The role of the Data Partners is to execute analytical code and review and approve outputs using their expert knowledge about their respective data sources.

ANNEX II. Protocol Amendments

None.

ANNEX III. Description of the data sources

Croatia: Croatian National Public Health Information System (NAJS)

1. Data source identification

Croatian National Public Health Information System (NAJS)
Croatia

2. Data partner

Croatian Institute of Public Health (CIPH)
Department of Data Science and Analytics

3. Coverage and timespan

Data collection since: 1998
Extent: Nation-wide.
Geographic coverage covers whole Croatia, with various levels of resolution for different registries. Current estimates for the population in Croatia will be available at: https://podaci.dzs.hr/hr/podaci/stanovnistvo/procjena-stanovnistva/ for each year.
The total and active person-count in the NAJS data is larger than the current population of Croatia. This is explained by: a) the person table includes deceased and all previously insured people, b) there is no information about insurance ending, c) healthcare is also used by individuals with dual citizenship from neighbouring countries It is known that a lot of individuals emigrated (300k–400k) and were not included in the last population census but still are in the NAJS database. There is also an influx of immigrant workers who are insured and registered but were not included in the census.

4. Healthcare setting and type of data

Primary care – General Practitioner, and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care.
For both inpatient and outpatient setting diagnoses, medication, procedures, and measurements are captured. The year of availability of information depends on the setting:
• 2014–2025 for biochemical lab tests in primary care from EHR patients records (measurements with results)
• 2015–2025 for primary care data from EHR patient records (conditions, procedures, and drug prescriptions)
• 2015–2024 for inpatient hospital data from EHR administrative records (conditions, procedures, measurements without results and drug administrations)
• 2016–2025 for health risk assessment data entered by GPs (measurements with results: height, weight…)
• 2016–2022 for secondary conciliatory care data from EHR administrative records (conditions, procedures, measurements without results, and drug administrations)
• 2016–2022 for emergency care data from EHR patient records (conditions)
• 2017–2025 for hospital records from registry data (conditions and procedures)
• 2020–2025 for vaccination data from EHR patient records

5. Data collection process

Inpatient hospital electronic health records, and Inpatient hospital billing systems, and Registries, and Other.
Data is entered by clinicians at healthcare contact, then combined by CIPH into the NAJS database and integrated with registries for public health purposes.

6. General representativeness

The data is collected from the evidence of public health records collected for public health purposes, as the majority of health care in Croatia is public and under single health insurance provider. Personal details are collected to a better extent for insured individuals compared to uninsured patients, who are excluded in the ETL process.

7. Data content and source coding

Medication prescriptions are recorded with ATC codes with an additional 3-digit code denoting the package.
• Diagnoses with ICD10 codes (Australian modification).
• Procedures with local source codes.
• Lab results with local source codes.

8. Data Harmonisation

The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network.
Records from 2017 include insured patients with reliable IDs. Uninsured patients do not have reliable IDs. For example, if a patient changed their status from insured to uninsured, or vice versa, they could be counted several times, as could tracking records from before 2017 and after. By using the unique personal identifier for Croatian citizens, it can be checked and verified.

9. Quality control (data source specific)

There is a network of registry personnel (leaders, administrators, coders, sources) working on data coverage and other quality dimensions. An analytical team routinely checks for erroneous entries in hospital records, removing double entries, false dates, and overlapping stays. Entries without enough data or with obviously erroneous dates from primary care analysis are excluded.

10. Linkage

The national death registry is updated yearly, with a one-year lag, but individual death information (just the date) is updated daily, without the cause of death or any other additional details. Primary care is updated weekly and hospital level care monthly. Specific registries are included in NAJS (e.g. diabetes registry), where inclusion criteria vary across these registries.

11. Vital status

NAJS is linked to the national death registry.

12. Limitations

Hospital data is available from 2017 onwards. This is often used as start of data collection, while laboratory and GP data has been captured since before then (since 2014 and 2015 respectively).
Drug duration is often not available, and set to 1 day for administration and 30 days for prescription.
Some vital signs, such as blood pressure measurements, are coming from a preventive in GP that is usually done yearly. This means that these measurements do not represent the patient's longitudinal journey.
Hospital discharge summaries are currently not captured in NAJS. Hospital drug administration data is less reliable than prescription data from primary care, with some drugs (monoclonal antibodies/precision medicine drugs) which require additional approval not being recorded at all.

13. Main references

No main reference provided.

14. Link to HMA-EMA catalogue and data source webpage

HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111155
Website: https://www.hzjz.hr/nacionalni-javnozdravstveni-informacijski-sustav-najs/

Estonia: Estonian BioBank (EBB)

1. Data source identification

Estonian BioBank (EBB)
Estonia

2. Data partner

University of Tartu (TARTU)
Institute of Computer Science

3. Coverage and timespan

Data collection since: 2004
Extent: Nation-wide.
EBB is a nation-wide database containing records from 2004 onwards. Estonian population-based cohort size of 211,800 participants (01/01/2024) aged 18 years and older recruited at GP offices, private practices, and hospitals or in the recruitment offices of the Estonian Genome Center.

4. Healthcare setting and type of data

Primary care – General Practitioner, and community pharmacists, and primary care specialists (e.g. paediatricians), and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care.
Healthcare providers must report the data to national registries. The data includes claims, prescriptions, EHR, and two registries - the cause of death and cancer registry. The claims data covers diagnosis, procedures, and services performed. Claims data contains information about services paid by the Estonian Health Insurance Fund and not the services paid by the individual themself. The prescription data contains information on prescribed and dispensed prescription drugs. It does not contain information about over-the-counter drugs. The EHR data includes diagnoses, results of the procedures and lab measurements, and case summaries. The cause of death and cancer registry data is checked by the registry before it is entered to the registry. Furthermore, Estonian BioBank has collected genetic and questionnaire data when the individual gives consent. Questionnaire data includes personal data (e.g. nationality), genealogical data (family history of medical conditions), educational history, and lifestyle data (e.g. smoking, alcohol consumption etc).

5. Data collection process

Insurance/administrative claims, and Outpatient electronic health records, and Inpatient hospital electronic health records, and Registries, and BioBank, and Other.
Data is retrieved by Estonian BioBank once a year from national registries. The insurance claims are requested from Estonian Health Insurance Fund. The inpatient and outpatient electronic health records are requested from National Health Information System. The cancer registry and cause of death registry information is requested from The National Institute for Health Development. The data is sent to the national registry by the healthcare providers.

6. General representativeness

The age, sex, and geographical distributions closely reflect those of the Estonian adult population and encompass approximately 20% of adult population. Female participants are over-represented in EBB. Older individuals tend to participate less frequently; however, all age groups are well represented.

7. Data content and source coding

All participants have undergone a standardised health assessment, including provision of blood samples for purification of DNA, white blood cells, and plasma, and completed a questionnaire covering various health-

related topics, such as lifestyle, diet, and clinical diagnoses. Diseases and health problems are recorded as ICD-10 codes and prescribed medicine according to the ATC classification and local package codes. Procedures and services are coded with NOMESCO classifiers and local service codes. Lab measurements are coded with local codes and in later years with LOINC codes. In cancer registry also ICD-O-3 codes are used.
The data is available with different granularity. For claims data, we only know whether a procedure or service is done. The result of the procedure or measurement is recorded in EHR. Usually, the result is in unstructured text (except lab measurements) and needs specific pipelines for extraction.

8. Data Harmonisation

The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network.
There is one national identifier which allows linking all encounters across databases.

9. Quality control (data source specific)

The quality control procedures in the Estonian BioBank aim to remove the most obvious mistakes in the data, misspellings, impossible dates, duplicates.
Before performing the ETL, several problems are fixed on the source data. Since the ETL procedures are used for a number of different datasets (from the same national sources), we have a growing number of pre-processing steps that correspond to the issues we have previously discovered in the data, such as checking for the presence of critical values, harmonising date and unit of measurement formats, checking the validity of certain entries against classifiers, etc.

10. Linkage

Follow-up data are available via linkage with national health-related registries and via re-examination of participants. Furthermore, electronic health records are updated for phenotypic outcome information every year. The EBB database is regularly linked with national registries, hospital databases, and the databases of the Estonian Health Insurance Fund (EHIF) and the National Health Information System (NHIS).

11. Vital status

Vital status (death date and causes of death) is obtained from the Causes of Death Registry.

12. Limitations

Participation in the EBB cohort is voluntary; therefore, the biobank does not represent a random sample and could be subject to recruitment bias. Although recruitment was open to everyone, there is a disproportion between ethnic Estonians and ethnic Russians in the biobank, with Estonians being overrepresented.

13. Main references

Milani L, Alver M,Laur S,Reisberg S,Haller T,Aasmets O,Abner E,Alavere H,Allik A,Annilo T,Fischer K,Hofmeister R,Hudjashov G,Jõeloo M,Kals M,Karo-Astover L,Kasela S,Kolde A,Krebs K,Krigul KL,Kronberg J,Kruusmaa K,Kukuškina V,Kõiv K,Lehto K,Leitsalu L,Lind S,Luitva LB,Läll K,Lüll K,Metsalu K,Metspalu M,Mõttus R,Nelis M,Nikopensius T,Nurm M,Nõukas M,Oja M,Org E,Palover M,Palta P,Pankratov V,Pantiukh K,Pervjakova N,Pujol-Gualdo N,Reigo A,Reimann E,Smit S,Rogozina D,Särg D,Taba N,Talvik HA,Teder-Laving M,Tõnisson N,Vaht M,Vainik U,Võsa U,Yelmen B,Esko T,Kolde R,Mägi R,Vilo J,Laisk T,Metspalu A "The Estonian BioBank's journey from biobanking to personalized medicine." Nature communications (2025): 40188112

14. Link to HMA-EMA catalogue and data source webpage

HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111114
Website: https://genomics.ut.ee/en/content/estonian-biobank

Spain: Institut Municipal Assistència Sanitària Information System (IMASIS)

1. Data source identification

Institut Municipal Assistència Sanitària Information System (IMASIS)
Catalunya, Spain

2. Data partner

Parc de Salut Mar Barcelona (PSMAR)
Direcció de Control de Gestió (Management and Control Department)

3. Coverage and timespan

Data collection since: 1990
Extent: Regional.
The primary catchment area covers the Barcelona city area, in Catalonia (Spain). However, all recorded visits are included, which may also include patients residing outside this area.

4. Healthcare setting and type of data

Secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care.
The available information includes diagnoses, procedures, drug administrations, dispensations and prescriptions, laboratory tests, visit types, cancer data, and vaccinations. These data are linked to primary care centers, which currently contribute information on diagnoses, and treatments, as well as to mortality data.

5. Data collection process

Outpatient electronic health records, and Inpatient hospital electronic health records, and Other.
The data collection process begins at the point of care, when a patient is attended in any facility within our hospital network or associated primary care centers. Clinical information is recorded directly into the electronic health records (EHRs) and subsequently complemented with data from external sources, including the regional vaccine registry, the cancer registry, the national mortality registry, and outpatient treatment records. All information is consolidated into a relational database using standardized formats and subjected to multiple quality control procedures to ensure completeness and consistency. The resulting datasets are then extracted, transformed, and loaded (ETL) into an OMOP Common Data Model (CDM) database on a quarterly to semi-annual basis.

6. General representativeness

The data source includes patients across all socioeconomic levels, ages, and genders, suggesting that the dataset is representative of the population of Barcelona requiring secondary (hospital-based) healthcare at any given time.

7. Data content and source coding

Health outcomes, diagnoses, and procedures are recorded using ICD-9-CM and ICD-10-CM classification. Laboratory tests are recorded in the source data using LOINC terminology and drugs using the national codes classification. Cancer registry is based on ICD-O-3 classification. Additionally, a set of demographics data are available such as age and sex.

8. Data Harmonisation

The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network.

A patient cannot be registered under different ID numbers. They are registered under the same ID, even when changing practice of re-registration or admission.

9. Quality control (data source specific)

Several quality control procedures have been implemented during the creation and update of IMASIS. The process includes a series of filters that exclude any data whose format or content do not match the specifications of the target variables.
Some examples of these actions include the following.

• Records containing codes that do not belong to the standardized vocabularies used for encoding are reviewed within their respective domains and, if necessary, discarded.

• Records with implausible values, such as dates preceding a patient's birthdate, are identified and corrected or removed.

• Before each IMASIS update, the number of records in the source files is verified. Record counts from the new IMASIS instance are compared with both the source files and previous IMASIS versions to detect potential inconsistencies or anomalies in the source data.

• In addition, if the Data Quality Dashboard (DQD) package identifies quality issues in the OMOP instance, for example diagnoses recorded for patients with an implausible gender, these discrepancies are reviewed and corrected in the source data.

10. Linkage

Information of diagnosis and treatments received in outpatient settings and in primary care as well as national mortality registry, cancer registry and regional vaccination registry.

11. Vital status

The hospital's death registry records the age at death for each patient when the event occurs within the hospital. In addition, deaths occurring outside the hospital are captured through linkage with the national mortality registry; however, in these cases, the specific cause of death is not currently available.

12. Limitations

No database-specific limitations are present but information coming from primary care are not complete. General limitations related to Real World Data can be considered.

13. Main references

Mayer MA, Furlong LI,Torre P,Planas I,Cots F,Izquierdo E,Portabella J,Rovira J,Gutierrez-Sacristan A,Sanz F "Reuse of EHRs to Support Clinical Research in a Hospital of Reference." Studies in health technology and informatics (2015): 25991136

14. Link to HMA-EMA catalogue and data source webpage

HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/104316
Website: https://www.parcdesalutmar.cat/en/

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

1. Data source identification

Clinical Practice Research Datalink Aurum (CPRD Aurum)
England, the United Kingdom

2. Data partner

Clinical Practice Research Datalink (CPRD)
Medicines and Healthcare products Regulatory Agency (MHRA) - Safety and Surveillance

3. Coverage and timespan

Data collection since: 1995
Extent: Nation-wide.
The CPRD Aurum database includes just over 17M currently registered patients, or around 25.5% of the UK population. All currently contributing practices (n= 1645) are in England, but coverage varies between English regions. The Aurum OMOP CDM database covers the period from 01/01/1995 to the most recent data collection (June-July 2025). For the Aurum + linked HES APC database, the coverage period is truncated to ensure that all patients are under observation in both the primary care and linked HES data, therefore the maximum allowable coverage period is currently 01/04/1997 to 31/03/2023 NB this will be updated in the next release, expected Jan 2026.
https://www.cprd.com/doi/cprd-aurum-september-2025

4. Healthcare setting and type of data

Primary care – General Practitioner, and secondary care – specialists (ambulatory or hospital outpatient care).
CPRD Aurum includes data captured in UK primary care. In the UK, almost all individuals are registered with a single general practitioner (GP), and around 90% of all healthcare contacts with the UK National Health Service (NHS) occur in primary care (also known as general practice). GPs have a central role as gatekeepers for access to secondary and tertiary services, and other community based services and intervention programmes.
Widespread adoption of dedicated computer software in UK primary care began in the late 1980s and was almost universal by 1995. GPs and general practice staff use these software to manage patients' electronic health records which include registration and demographic information, symptoms, diagnoses, clinical investigations and laboratory tests, prescriptions for medicines and appliances, a wide range of behavioural and social factors relevant to health and wellbeing, and administrative information. Some information is captured on contacts with other health services, including referral details and discharge summaries and other communications from those services. CPRD Aurum does not include any associated free text notes, scanned documents or images from the patient record.
Hospital Episode Statistics Admitted Patient Care (HES APC) data captures information on inpatient care delivered in NHS hospitals and independent treatment centres funded by the NHS. Key variables include age; sex; ethnicity, episode start/end date; a single primary diagnosis (ICD-10), with up to 19 other comorbidity diagnoses; up to 24 procedures and other interventions (with date), coded using the UK Office of Population Censuses and Surveys Classification of Interventions and Procedures classification (OPCS-4). Medical treatments (drugs and devices) and lab values are not recorded in HES APC.

5. Data collection process

Outpatient electronic health records, and Inpatient hospital electronic health records.
Patient EHR data are entered by practice staff using the EMIS Web GP practice management software. Data may also be captured automatically via electronic links from pathology and diagnostic laboratories. EMIS Web is a cloud based system, and practice data are uploaded to the vendor data centre. Pseudonymised

data is extracted from the vendor system and downloaded to CPRD via a dedicated cloud based interface. CPRD undertake additional checks to ensure data conform to specification, and comply with established data quality metrics, and apply additional pseudomymisation steps.
HES APC data are derived from hospital activity data which are abstracted from clinical records and discharge summaries by trained clinical coders according to centrally issued coding rules and standards. The data is collected centrally for monitoring, planning, and payment purposes. For secondary uses including research, additional cleaning and curation is undertaken to generate the HES APC data. This includes linking episodes for the same individual, and addition of a Master Person Service identifier (MPS ID) which allows linkage to primary care and other health datasets for research.

6. General representativeness

CPRD Aurum provides a purposive sample of English general practices using EMIS Web GP software. EMIS Web has an uneven regional coverage. Participation in the CPRD data collection is voluntary. While not a random sample, CPRD Aurum is broadly representative of the overall population in terms of age and gender (https://pubmed.ncbi.nlm.nih.gov/30859197/), with minor differences in terms of area based deprivation and urban-rural distribution. (https://pubmed.ncbi.nlm.nih.gov/35902219/)

7. Data content and source coding

Data is captured using READ, EMIS2, SNOMED, DM+D codes.

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 move and receive a different ID number upon changing practice, but a patient can only be registered at one practice at a time. Upon registering at a new practice, a patient's records will be transferred with them, and they will be deregistered at their previous practice. CPRD cannot track patients across practices as the data is de-identified and a patient key is generated for each patient at a practice level, not a country level. The patient key will remain the same for as long as the patient is registered at the practice.
The main impact of patients changing practice is to reduce the average follow up time available for each analytical unit (i.e. patient-within-practice). A secondary issue is that records for the same individual split across two or more practices are not strictly independent. The impact on inferential statistics is likely to be very limited, and analytical adjustment is typically not required (or applied).

9. Quality control (data source specific)

CPRD follows the same three-staged processing pipeline for all the data that it curates: curate individual data collections; integrate data into master database; and extract study specific datasets. For each part of the processing pipeline, CPRD applies checks to ensure that the quality of the data is maintained.
At the data curation stage, conducted at both a collection level and at a record level, the data is verified to ensure that the data conforms to an agreed specification. Collection level verification ensures that the collection contains only expected data files, that the collection files are from a known source, and that the collections' files are in the correct sequence. The record level verifications performed at Level 1 are structural in nature and are applied to every data element that CPRD receives without considering the context of the data item against the entirety of the master database. These checks include that the data item is of the correct type, that fields that should have values do indeed contain a value, and that, where specified, the format and length of the data is accurate. Any violations of these checks result in the collection being rejected, the data not being integrated into the master database, and further inspection or recollection of the data taking place.
At the data integration stage, the actual content of the data is validated, considering any existing content

already collected. These validations are contextual in nature and consider the values of multiple data items within the master database. As CPRD provide real world data no data is modified before being made available for research. The data is either rejected (as per the data curation checks) or integrated. Once integrated, data is not corrected, but patient, practice, and database level data quality metrics and markers are put in place, based on internal consistency of the data, to aid researchers in deciding if the data is suitable for research. There are some exceptions to this no modification principle. In the case that a patient is validated as being opted out, then their data is removed from future versions of the database to ensure their wishes are respected.
Further validation and cleaning of study specific datasets may be undertaken by researchers or as part of good data management, and is dependent upon the research question and the researchers performing the analysis.

10. Linkage

Linkage of CPRD Aurum to HES APC is undertaken by NHS England. Identifiers for patients in CPRD Aurum (NHS number, gender, full date of birth, residential postcode) flow directly from the vendor data centre to NHS England, and are linked to HES APC via the MPS ID. A bridging file of patient index numbers for each dataset is returned to CPRD to allow linkage of the clinical data.
Linkage is undertaken at approximately 6 month intervals, but this may be impacted by organisation factors external to CPRD.
Linkage to additional datasets is expected over the next 12-24 months, including linked death registrations (for accurate cause of death), hospital outpatient activity, and cancer registrations.

11. Vital status

The great majority of deaths are eventually captured in patients' primary care record. Published comparisons found that over 98% of national death registrations (the gold standard) were also captured in the CPRD GOLD primary care database https://pubmed.ncbi.nlm.nih.gov/30908785/, and results are similar for CPRD Aurum (unpublished). Estimated date of death in the primary care record shows good agreement with registered date of death (within 30 days in 98.8% of occurrences after 2013), but cause of death is captured inconsistently.

12. Limitations

CPRD Aurum uses data collected for the primary purpose of managing and delivering safe and effective patient care in general practice. There are requirements and incentives for practice staff to record high quality information, but there is heterogeneity between practices and over time. UK primary care databases have been extensively validated and found to be of sufficient quality to support observational research.
Information about care and treatments in settings other than primary care is likely to be captured only partially in CPRD Aurum. Practice staff may code key information about important diagnoses or treatments in other settings. However discharge summaries may simply be scanned or stored as electronic documents, or summarised as free text notes. None of these sources are available in CPRD Aurum.
CPRD Aurum captures information on prescriptions issued by practice staff, from which drug exposures can only be inferred. There is no information about drug exposures in secondary care via the HES APC.

13. Main references

Wolf A, Dedman D,Campbell J,Booth H,Lunn D,Chapman J,Myles P "Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum." International journal of epidemiology (2019): 30859197

14. Link to HMA-EMA catalogue and data source webpage

HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/30008
Website: https://www.cprd.com/data/primary-care-data/cprd-aurum

The United Kingdom: UK BioBank (UKBB)

1. Data source identification

UK BioBank (UKBB)
the United Kingdom

2. Data partner

University of Oxford - Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences (Oxford NDORMS)
NDORMS

3. Coverage and timespan

Data collection since: 1996
Extent: Nation-wide.
People recruited from whole of the UK.

4. Healthcare setting and type of data

Primary care – General Practitioner, and primary care specialists (e.g. paediatricians), and secondary care – specialists (ambulatory or hospital outpatient care), and hospital inpatient care, and other (specify).
UK BioBank is made by a rich variety of data sources, which include genetic data, primary care data, hospital inpatient data, death data, and cancer registry.

5. Data collection process

Inpatient hospital electronic health records, and Registries, and BioBank.
The baseline assessment is consisting of both patient reported data (questionnaire) and physical measurements. GP and hospital data, as well as death and cancer registry records, are linked afterwards and are subject to data validation:
https://biobank.ctsu.ox.ac.uk/~bbdatan/Data_cleaning_overall_doc_showcase_v1.pdf

6. General representativeness

The database population consists of volunteers aged 40-69years. We can expect that volunteers might have been more willing to participate if living closer to one of the 22 recruitment centres or if more interested in health issues compared to the general population. These aspects might have introduced an unavoidable bias in the cohort.

7. Data content and source coding

READ2, READ3, DM+D, ICD9, ICD10, OPCS3, OPCS4, ICD-O-3 are used.

8. Data Harmonisation

The data has been mapped to the OMOP CDM v5 and the OMOP standard vocabularies. The format, structural and semantic conformance has been verified upon onboarding into the DARWIN EU® data network.
No.

9. Quality control (data source specific)

All UK BioBank data are provided already curated and each of the many datasets have specific curation algorithms and procedures. As always, primary care and hospital data, which come from real-world setting, need special attention regarding data quality. Please refer to the link below for specific details
https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/primary_care_data.pdf
https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/HospitalEpisodeStatistics.pdf

10. Linkage

The database contains liked data from death and cancer registry, GP and hospital for the participants.

11. Vital status

Linked to the national death registry.

12. Limitations

The UKBB source data will not be updated anymore and have no new records after December 2022. There is no day's supply information captured in the source. GP prescription data are available for 45% of the cohort. There is no information on dispensed medicines. GP laboratory tests and results are not available.

13. Main references

Hewitt J, Walters M,Padmanabhan S,Dawson J "Cohort profile of the UK Biobank: diagnosis and characteristics of cerebrovascular disease." BMJ open (2016): 27006341

14. Link to HMA-EMA catalogue and data source webpage

HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111233
Website: https://www.ukbiobank.ac.uk/

The United States: IQVIA US - Ambulatory EMR (IQVIA US - AmbEMR)

1. Data source identification

IQVIA US - Ambulatory EMR (IQVIA US - AmbEMR)
the United States

2. Data partner

IQVIA (IQVIA)

3. Coverage and timespan

Data collection since: 2006
Extent: Nation-wide.
Over 100 million patients, with data sourced from more than 800 ambulatory practices and 100,000+ physicians across the United States.

4. Healthcare setting and type of data

Primary care – General Practitioner, and primary care specialists (e.g. paediatricians).
The IQVIA Ambulatory EMR-US dataset primarily covers ambulatory care settings, which include: Outpatient physician practices, Primary care and specialty clinics, Independent and group practices, and non-hospital-based care environments.

5. Data collection process

Outpatient electronic health records.
Anonymized patient records collected from patient management software used by GPs and selected specialists to document patients' clinical encounters. These records are sent to a central location then processed and aggregated.

6. General representativeness

The data is representative of the US population, with a variety of socioeconomic status patients that is not limited to specific insurances.

7. Data content and source coding

The following source vocabularies are used: CPT, HCPCX, Dx, NDC.

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.
IQVIA's patient tokenization process generates a consistent, anonymous patient identifier by applying multi-layered, irreversible hashing to encrypted patient attributes, enabling privacy-protected linkage of de-identified data across care settings without exposing personally identifiable information. This enables longitudinal tracking of patients.

9. Quality control (data source specific)

Integrity constraints including
- 230+ IQVIA Internal QC checks
- DQD & Achilles review
- IQVIA internal business validation check

10. Linkage

Linkage to other IQVIA datasets is possible, e.g. PharmetricsPlus, but has to be done on a project basis.

11. Vital status

There is no linkage to a death registry.

12. Limitations

The main limitation is that no mortality information is available and there is no possibility to link a death registry.
Drug dosages cannot always be calculated as for half the drug exposures the quantity is missing.

13. Main references

No main reference provided.

14. Link to HMA-EMA catalogue and data source webpage

HMA-EMA Catalogue entry: https://catalogues.ema.europa.eu/data-source/1111118
Website: https://www.iqvia.com/locations/united-states/library/fact-sheets/iqvia-ambulatory-emr-us

ANNEX IV. Fitness for use assessment

Data source justification for inclusion and key characteristics

The selected data sources represent different healthcare settings, geographical regions, and populations and are expected to contain information relevant to the identification and characterisation of pharmacogenetic testing in routine clinical practice.

Croatian National Public Health Information System (NAJS)

NAJS will be included because it is a national registry covering primary care, outpatient specialist care, and inpatient care. It contains information relevant to the identification of pharmacogenetic tests and the characterisation of tested individuals, including demographic characteristics, comorbidities, and medication use. The data source includes approximately 4,835,791 active individuals and is expected to provide information on a broad healthcare population in Croatia.

Data are available from 2014 to 2025, covering almost the entire study period. Consequently, the data source can contribute to all study objectives, subject to confirmation of the period during which pharmacogenetic testing is reliably captured.

A data source specific limitation is that the results of genetic measurements are not mapped to the OMOP CDM. NAJS can therefore contribute to the estimation of testing frequency and the characterisation of tested individuals, comorbidities, and medication use, but currently cannot contribute to the analysis of pharmacogenetic test results.

NAJS has blanket Institutional Review Board (IRB) approval for studies conducted via the DARWIN EU® Data Network.

Estonian BioBank (EBB)

EBB will be included because it is a registry-based data source covering information from primary and secondary care. It contains information relevant to pharmacogenetic testing and the characterisation of tested individuals, including demographic characteristics, comorbidities, and medication use. EBB includes approximately 212,300 active individuals from the Estonian biobank population.

The reported availability of pharmacogenetic testing information starts from 2020–2021 onwards.

EBB can contribute to all objectives if pharmacogenetic test records, test results, and the required clinical characteristics are available and mapped. A data source limitation is that EBB is a biobank population, which may mean that its participants may not be fully representative of the general population of Estonia. IRB approval for EBB is estimated to take 2–4 weeks.

Institut Municipal Assistència Sanitària Information System (IMASIS)

IMASIS will be included because it is a secondary-care data source containing inpatient hospital information. It contains information relevant to pharmacogenetic testing and the characterisation of tested individuals, including demographic characteristics, comorbidities, and medications recorded within the hospital setting. The data source includes approximately 169,560 active individuals.

Data are available from 1990 to 2025, and therefore, cover the complete study period. IMASIS can contribute to all objectives, subject to confirmation of the period during which pharmacogenetic testing and test results are reliably recorded.

As a secondary-care data source, IMASIS may not completely capture pharmacogenetic tests performed outside the participating healthcare system or medications prescribed exclusively in primary care. Its results will therefore describe the population receiving care within the contributing hospital system rather than the Spanish general population.

IRB approval for IMASIS is estimated to take <3 months.

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

The Clinical Practice Research Datalink Aurum (CPRD Aurum) will be included because it is a primary-care electronic health-record data source containing outpatient general practitioner information. It contains information relevant to pharmacogenetic tests recorded in primary care and to the characterisation of tested individuals, including demographic characteristics, diagnoses, and primary-care prescriptions. CPRD Aurum includes approximately 17,438,460 active individuals.

Data are available from 1997 to 2025, and therefore, cover the complete study period. CPRD Aurum can contribute to all objectives.

A study-specific limitation is that pharmacogenetic tests performed in secondary care may only be identified when they are communicated to and recorded by the general practitioner. Medications prescribed or administered exclusively in secondary care may also be incompletely captured.

CPRD Aurum has blanket IRB approval for studies conducted via the DARWIN EU® Data Network.

United Kingdom: UK BioBank (UKBB)

UKBB will be included because it is a registry-based data source containing linked primary-care and hospital information. It contains information relevant to genetic testing and the characterisation of participants, including demographic characteristics, comorbidities, and medication use. UKBB includes approximately 499,900 active individuals.

The currently available data cover the period from 1985 to 2017. Therefore, medication analyses will be restricted accordingly.

UKBB has blanket IRB approval for studies conducted via the DARWIN EU® Data Network.

United States: IQVIA US - Ambulatory EMR (IQVIA US - AmbEMR)

IQVIA US - AmbEMR will be included because it is an outpatient electronic medical-record data source covering primary and specialist ambulatory care in the United States. It contains information relevant to pharmacogenetic testing recorded in ambulatory practice and to the characterisation of tested individuals, including demographic characteristics, comorbidities, and prescribed medications. The data source includes approximately 9,132,000 active individuals.

Data are available from 2006 to 2025, and therefore, cover the complete study period. The data source can contribute to all objectives, subject to confirmation of the period during which pharmacogenetic testing is reliably captured.

Because this is an ambulatory electronic medical-record data source, pharmacogenetic tests performed exclusively during inpatient care and medications administered in hospital may be incompletely captured. The population represents individuals receiving care from contributing ambulatory providers rather than a geographically defined general population.

IRB approval for IQVIA US - AmbEMR is estimated to take <1 week.

ANNEX V. Operational and reporting considerations

DATA MANAGEMENT

Data management

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

Data storage and protection

For this study, data sources will process personal data from individuals that is collected in national/regional electronic health record data sources.

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 regarding ethical use of the data and adequate privacy control are adhered to. In agreement with these regulations, rather than combining person level data and only performing 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 *CohortDiagnostics* (https://github.com/OHDSI/CohortDiagnostics) 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 and Prevalence, *MeasurementDiagnostics,* and *CohortCharacteristics,* to characterise the cohort by indication. 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 VI. List of stand-alone documents

Table S2. Preliminary list of conditions definitions.

Phenotype	Concept name	Concept ID (including descendants)	Exclude concept ID	Vocabulary
Liver disease	Viral hepatitis	4291005	193256	SNOMED
	Steatotic liver disease	4059290	37396531	
	Portal hypertension	192680		
	Hepatic fibrosis	4267417		
	Hepatic failure	4245975		
	Chronic liver disease	4212540		
	Autoimmune liver disease	4340391		
Kidney disease	End-stage renal disease	193782		SNOMED
	Dependence on renal dialysis	4019967		
	Chronic kidney disease	46271022		
	Acute kidney injury	197320		
Gastrointestinal disease	Peptic ulcer	4027663		SNOMED
	Pancreatitis	4192640		
	Intestinal malabsorption	194993		
	Inflammatory bowel disease	4074815		
	Gastroesophageal reflux disease	318800		
	Gastritis	201340		
	Duodenitis	433516		
	Celiac disease	194992		
Cardiovascular disease	Peripheral arterial disease	3654996	43021735	SNOMED
	Ischemic heart disease	4185932		
	Hypertensive disorder	316866		
	Heart valve disorder	4281749		
	Heart failure	316139		
	Cerebrovascular disease	381591		
	Cardiac arrhythmia	44784217		
Inflammatory disease	Vasculitis	4137275		SNOMED
	Spondyloarthritis	37205058		
	Rheumatoid arthritis	80809		
	Psoriasis	140168		
	Lupus erythematosus	255891		
	Ankylosing spondylitis	437082		
Psychiatric condition	Psychotic disorder	436073		SNOMED
	Psychoactive substance use disorder	4004672		
	Personality disorder	441838		

Phenotype	Concept name	Concept ID (including descendants)	Exclude concept ID	Vocabulary
	Depressive disorder	440383		
	Bipolar disorder	436665		
	Anxiety disorder	442077		
Neurological condition	Parkinson's disease	381270	4301699	SNOMED
	Neuropathic pain	4090564		
	Multiple sclerosis	374919		
	Epilepsy	380378		
	Dementia	4182210		
Any cancer	Malignant neoplastic disease	443392		SNOMED

Table S3. Preliminary list of medicines definitions.

Substance Name	Concept name	Class	Vocabulary ID	Ingredient Concept ID	Include descendants
Rosuvastatin	rosuvastatin	Ingredient	RxNorm	1510813	Yes
Ivacaftor	ivacaftor	Ingredient	RxNorm	42709323	Yes
Efavirenz	efavirenz	Ingredient	RxNorm	1738135	Yes
Clopidogrel	clopidogrel	Ingredient	RxNorm	1322184	Yes
Omeprazole	omeprazole	Ingredient	RxNorm	923645	Yes
Voriconazole	voriconazole	Ingredient	RxNorm	1714277	Yes
Citalopram	citalopram	Ingredient	RxNorm	797617	Yes
Escitalopram	escitalopram	Ingredient	RxNorm	715939	Yes
Mavacamten	mavacamten	Ingredient	RxNorm	779591	Yes
Sertraline	sertraline	Ingredient	RxNorm	739138	Yes
Amitriptyline	amitriptyline	Ingredient	RxNorm	710062	Yes
Lansoprazole	lansoprazole	Ingredient	RxNorm	929887	Yes
Warfarin	warfarin	Ingredient	RxNorm	1310149	Yes
Codeine	codeine	Ingredient	RxNorm	1201620	Yes
Tramadol	tramadol	Ingredient	RxNorm	1103314	Yes
Tamoxifen	tamoxifen	Ingredient	RxNorm	1436678	Yes
Atomoxetine	atomoxetine	Ingredient	RxNorm	742185	Yes
Ondansetron	ondansetron	Ingredient	RxNorm	1000560	Yes
Tacrolimus	tacrolimus1	Ingredient	RxNorm	950637	Yes
Fluorouracil	fluorouracil	Ingredient	RxNorm	955632	Yes
Capecitabine	capecitabine	Ingredient	RxNorm	1337620	Yes
Tegafur	tegafur	Ingredient	RxNorm	19056756	Yes

Substance Name	Concept name	Class	Vocabulary ID	Ingredient Concept ID	Include descendants
Levonorgestrel and ethinylestradiol	levonorgestrel and ethinylestradiol	ATC 5th	ATC	21602480	Yes
Desogestrel and ethinylestradiol	desogestrel and ethinylestradiol	ATC 5th	ATC	21602482	Yes
Rasburicase	rasburicase	Ingredient	RxNorm	1304565	Yes
Oxcarbazepine	oxcarbazepine	Ingredient	RxNorm	718122	Yes
Phenytoin	phenytoin	Ingredient	RxNorm	740910	Yes
Carbamazepine	carbamazepine	Ingredient	RxNorm	740275	Yes
Abacavir	abacavir	Ingredient	RxNorm	1736971	Yes
Allopurinol	allopurinol	Ingredient	RxNorm	1167322	Yes
Simvastatin	simvastatin	Ingredient	RxNorm	1539403	Yes
Atorvastatin	atorvastatin	Ingredient	RxNorm	1545958	Yes
Azathioprine	azathioprine	Ingredient	RxNorm	19014878	Yes
Mercaptopurine	mercaptopurine	Ingredient	RxNorm	1436650	Yes
Thioguanine	thioguanine	Ingredient	RxNorm	1437379	Yes
Irinotecan	irinotecan	Ingredient	RxNorm	1367268	Yes

ANNEX VII. ENCePP checklist for study protocols

ENCePP Checklist for Study Protocols (Revision 4)

Study title: DARWIN EU® - Characterisation of pharmacogenetic testing data in routine healthcare

EU PAS Register® number: EUPAS1000001139
Study reference number (if applicable): P5-C1-022

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

Comments:

Section 2: Research question	**Yes**	**No**	**N/A**	**Section Number**
2.1 Does the formulation of the research question and objectives clearly explain:	☒	☐	☐	3-4
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)	☒	☐	☐	3
2.1.2 The objective(s) of the study?	☒	☐	☐	4
2.1.3 The target population? (i.e. population or subgroup to whom the study results are intended to be generalised)	☒	☐	☐	5.3
2.1.4 Which hypothesis(-es) is (are) to be tested?	☐	☐	☒	
2.1.5 If applicable, that there is no *a priori* hypothesis?	☐	☐	☒	

Comments:

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

Section 3: Study design		**Yes**	**No**	**N/A**	**Section Number**
3.1	Is the study design described? (e.g. cohort, case-control, cross-sectional, other design)	☒	☐	☐	5.1
3.2	Does the protocol specify whether the study is based on primary, secondary or combined data collection?	☒	☐	☐	5.4
3.3	Does the protocol specify measures of occurrence? (e.g. rate, risk, prevalence)	☒	☐	☐	5.8.3
3.4	Does the protocol specify measure(s) of association? (e.g. risk, odds ratio, excess risk, rate ratio, hazard ratio, risk/rate difference, number needed to harm (NNH))	☐	☐	☒	
3.5	Does the protocol describe the approach for the collection and reporting of adverse events/adverse reactions? (e.g. adverse events that will not be collected in case of primary data collection)	☐	☐	☒	

Comments:

Section 4: Source and study populations		**Yes**	**No**	**N/A**	**Section Number**
4.1	Is the source population described?	☒	☐	☐	5.4
4.2	Is the planned study population defined in terms of:				
	4.2.1 Study time period	☒	☐	☐	5.5
	4.2.2 Age and sex	☒	☐	☐	5.6.3
	4.2.3 Country of origin	☒	☐	☐	5.4
	4.2.4 Disease/indication	☒	☐	☐	5.6.3
	4.2.5 Duration of follow-up	☒	☐	☐	5.2
4.3	Does the protocol define how the study population will be sampled from the source population? (e.g. event or inclusion/exclusion criteria)	☒	☐	☐	5.3

Comments:

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

Section 5: Exposure definition and measurement	**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:

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

Comments:

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

Comments:

This is a characterisation study

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)	☐	☐	☒	
9.1.2 Outcomes? (e.g. clinical records, laboratory markers or values, claims data, self-report, patient interview including scales and questionnaires, vital statistics)	☒	☐	☐	
9.1.3 Covariates and other characteristics?	☒	☐	☐	5.6.3
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)	☐	☐	☒	
9.2.2 Outcomes? (e.g. date of occurrence, multiple event, severity measures related to event)	☒	☐	☐	5.8.3
9.2.3 Covariates and other characteristics? (e.g. age, sex, clinical and drug use history, co-morbidity, co-medications, lifestyle)	☐	☐	☐	
9.3 Is a coding system described for:				
9.3.1 Exposure? (e.g. WHO Drug Dictionary, Anatomical Therapeutic Chemical (ATC) Classification System)	☐	☐	☒	
9.3.2 Outcomes? (e.g. International Classification of Diseases (ICD), Medical Dictionary for Regulatory Activities (MedDRA))	☒	☐	☐	Annex VI
9.3.3 Covariates and other characteristics?	☒	☐	☐	5.6.3
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?	☒	☐	☐	5.8
10.2 Is study size and/or statistical precision estimated?	☐	☐	☒	
10.3 Are descriptive analyses included?	☒	☐	☐	5.8.3

Section 10: Analysis plan	**Yes**	**No**	**N/A**	**Section Number**
10.4 Are stratified analyses included?	☐	☒	☐	
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?	☐	☐	☒	
10.8 Are relevant sensitivity analyses described?	☒	☐	☐	5.8.4

Comments:

Section 11: Data management and quality control	**Yes**	**No**	**N/A**	**Section Number**
11.1 Does the protocol provide information on data storage? (e.g. software and IT environment, database maintenance and anti-fraud protection, archiving)	☒	☐	☐	Annex V
11.2 Are methods of quality assurance described?	☒	☐	☐	Annex V
11.3 Is there a system in place for independent review of study results?	☒	☐	☐	5.6.3

Comments:

Section 12: Limitations	**Yes**	**No**	**N/A**	**Section Number**
12.1 Does the protocol discuss the impact on the study results of:				
12.1.1 Selection bias?	☒	☐	☐	6
12.1.2 Information bias?	☐	☐	☒	
12.1.3 Residual/unmeasured confounding? (e.g. anticipated direction and magnitude of such biases, validation sub-study, use of validation and external data, analytical methods).	☐	☐	☒	
12.2 Does the protocol discuss study feasibility? (e.g. study size, anticipated exposure uptake, duration of follow-up in a cohort study, patient recruitment, precision of the estimates)	☒	☐	☐	

Comments:

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

Section 13: Ethical/data protection issues	**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 V

Comments:

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

Comments:

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

Comments:

Name of the main author of the protocol: Cindy Porras

Date: 13/10/2026

Signature: 