The OMOPCAN Project: Protocol for assessing cancer epidemiology and patient characteristics in a multinational network of heterogeneous real-world data

Background:

Cancer registries remain the *gold standard* for cancer monitoring; however, they often lack comprehensive patient information beyond cancer-specific data. Complementing their strengths with real-world evidence from routine clinical care, which offers timely insights into cancer outcomes and additional medical history, would further improve our understanding of the cancer burden.

Aim:

Serve as a **proof-of-concept** using **real-world data mapped to the OMOP CDM** to support **monitoring of cancer epidemiology** and characterising patients' clinical histories and outcomes.

Methods:

- Multinational network study across primary care, secondary care, health insurance claims, and cancer registry data.
- From 1st January 2000 to 31st December 2024, or maximum available period per database
- We will study 36 Cancer types (ICD-10 C00-97) using SNOMED CT and ICDO-3 vocabularies, developed with CodelistGenerator and ATLAS.
- We will estimate incidence and survival, including their temporal trends; 5-year partial prevalence, and patient characteristics.
- In primary care and claims databases, individuals will be required to have 1 year of prior observation before cohort entry.

Database	Country	Type of data	Healthcare setting	Observation start in data source	Persons in data source
Belgian Cancer Registry - Use case breast and Lung cancer (BCR – BALC)	Belgium	Registry	Population-based Cancer Registry	2014	210.9k
Clinical Practice Research Datalink (CPRD GOLD)	UK	EHR	Primary Care	1993	17.52M
Croatian National public health information system (NAJS)	Croatia	Integrated EHR & Registry	Primary Care + Secondary care (Outpatient specialist care, and Inpatient Care) + Population-based Cancer Registry	2014	4.85M
EST-Health-30 (EH30)	Estonia	Integrated EHR & Registry	Population-based Cancer Registry + Secondary Care + Primary Care	2012	488k
Finnish Institute for Health and Welfare (THL)	Finland	Integrated EHR & Registry	Primary care + Secondary care	2011	6.61M
Geneva Cancer Registry (GCR)	Switzerland	Registry	Population-based Cancer Registry	1970	123k
HealthPartners	USA	Claims	Secondary Care + Clinics	1994	3.40M
Integrated Primary Care Information (IPCI)	The Netherlands	EHR	Primary Care	2006	8.55M
Netherlands Cancer Registry (NCR)	The Netherlands	Registry	Population-based Cancer Registry	1993	2.63M
OncEMR	USA	Claims	Secondary Care + Clinics	1939	1.88M
Semmelweis University Clinical Data (SUCD)	Hungary	EHR	Secondary Care	1998	2.49M
Taipei Medical University (TMUCRD)	Taiwan	EHR	Secondary Care	2001	3.39M
The Cancer Registry of Norway (CRN)	Norway	Registry	Population-based Cancer Registry	1953	1.22M
The Health Improvement Network® Belgium (THIN® BE)-	Belgium	EHR	Primary Care	2008	2.02M
THIN® Spain (THIN® ES)	Spain	EHR	Primary Care	2010	1.89M
THIN® Italy (THIN® - IT)	Italy	EHR	Primary Care	2008	1.22M
THIN® Romania (THIN®- RO)	Romania	EHR	Primary Care	2008	1.13M
The Hospital for Sick Children (SickKids)	Canada	EHR	Paediatric Tertiary Care (Inpatient and Specialist Outpatient)	2018	1.12M
The Information System for Research in Primary Care (SIDIAP)	Spain	EHR	Primary Care + Secondary Care Discharge Records	2006	5.95M
Yonsei University Health System (YUHS)	South Korea	EHR	Secondary Care	1993	6.66M

Table 1. Participant databases as of February 2026

Workflow:

✓ Phenotyping based on GLOBAL CANCER OBSERVATORY

1. Preliminary code list using CODELIST GENERATOR
2. Code list refinement by clinical experts
3. Cohort generation through **ATLAS**
4. Cohort evaluation through ***CohortDiagnostics****
5. Cohort refinement by clinical experts

Already runned by all datapartners involved as of February 2026

Statistical plan

1. **Initial Data Processing** (To be run by all datapartners)

   RUN ANALYSIS FOR X CANCER (reiterative process)
   - Step 1: Create Outcome Cohort
   - Step 2: Cohort Characterisation
   - Step 3: Crude Incidence
   - Step 4: 5-year Partial Prevalence
   - Step 5: Crude Survival

2. **Post Data Processing** (To be processed by each paper's lead researcher)
   - **Age standarisation** of rates
   - **Jointpoint modelling** of Incidence and Survival using Jointpont Trend Analysis Software and JPSurv Software from NCI

Irene López-Sánchez (Universitat Autònoma de Barcelona, Spain; IDIAP Jordi Gol, Spain), Anna Palomar-Cros (IDIAP Jordi Gol, Spain), Agustina Giuliodori (IDIAP Jordi Gol, Spain), Laura Granés (IDIAP Jordi Gol, Spain), Laura Pérez-Crespo (IDIAP Jordi Gol, Spain), Berta Raventós (Erasmus MC, Netherlands), Edward Burn (University of Oxford, UK), Anton Barchuk (Erasmus MC, Netherlands), Laia Peruchet-Noray (Imperial Kings College, UK), Raivo Kolde (University of Tartu, Estonia), Sulev Reisberg (University of Tartu, Estonia), Lillian Sung (Hospital of Sick Children, Canada), Elin J Rowlands (University of Oxford, UK), Danielle Newby (University of Oxford, UK), Annelies Verbiest (Antwerp University Hospital, Belgium), Seng Chan You (Yonsei University, Republic of Korea), Subin Kim (Yonsei University, Republic of Korea), Maaike van Swieten (NCR, Netherlands), Jelle Evers (NCR, Netherlands), Tor Åge Myklebust (CRN, Norway), Espen Enerly (CRN, Norway), Josip Vrancic (General Hospital Varazdin, Croatia), Mario Šekerija (Croatian Institute of Public Health, Croatia), Zsolt Bagyura (Semmelweis University, Hungary), Loretta Kiss (Semmelweis University, Hungary), Jason C. Hsu (Taipei Medical University, Taiwan), Thi Kim Hien Nguyen (Taipei Medical University, Taiwan), Patricia L. Mabry (HealthPartners Institute, USA) , Samuel Patnoe (HealthPartners Institute, USA), Asieh Golozar (Nemesis Health, USA) , Elena Roel (IDIAP Jordi Gol, Spain), Talita Duarte-Salles (IDIAP Jordi Gol, Spain; Erasmus MC, Netherlands)

IDIAP JORDI GOL