Estimating prevalence and incidence of acute myocardial infarction in a set of heterogeneous sources of observational health data collaborating in the EMIF Platform

Protocol version: 1.2

Title	Estimating prevalence and incidence of acute myocardial infarction in a set heterogeneous sources of observational health data collaborating in the EMIF Platform
Medicinal product(s) / Device(s)	Not applicable
Event(s) of interest	Acute myocardial infarction
Research question and objectives	A novel standard data derivation procedure for the execution of multi-national, multi-data source studies was specifically developed within the EMIF project. In this proof-of-concept study, the standard procedure will be applied for the identification of patients with acute myocardial infarction (AMI) in a set of heterogeneous sources of observational health data. Prevalence and incidence of AMI will be estimated from the participating data sources.
Country(ies) of study	Denmark, Estonia, Italy, The Netherlands, Spain, UK
Protocol author(s)	Giuseppe Roberto (ARS), Rosa Gini (ARS)

TABLE OF CONTENTS

LIST OF ABREVIATIONS:

AMI	Acute myocardial infarction
AUH	Aarhus University Hospital
ARS	Agenzia Regionale di Sanità della Toscana
DB	Database
DC	Data custodian
EGCUT	Estonian Genome Center of University of Tartu
EHR	Electronic Healthcare Records
EMC	Erasmus University Medical Center
EMIF	European Medical Information Framework
GSK	GlaxoSmithKline
HSD	Health Search IMS HEALTH LPD Database
IPCI	Integrated Primary Care Database
ICD-9CM	International Classification of Diseases version 9, clinical modification
ICD10	International Classification of Diseases version 10
ICPC	International Classification of Primary Care
IMASIS	Information System of Parc de Salut Mar Barcelona
READ	READ clinical terminology system
PHARMO	PHARMO Institute for Drug Outcomes Research
PRRE	Private Remote Research Environment
SIDIAP	The Information System for the Development of Research in Primary Care
THIN	The Health Improvement Network Database
UNIMAN	University of Manchester
UPF	Universitat Pompeu Fabra

RESPONSIBLE PARTIES

Name	Institution	Activity
Rosa Gini	ARS	Data custodian / Researcher
Giuseppe Roberto	ARS	Principal Investigator / Researcher
Maria Garcia	SIDIAP	Clinical expert
Talita Duarte	SIDIAP	Data Custodian/Researcher
Paul Avillach	EMC	Researcher
Rients van Wijngaarden	PHARMO	Data custodian / Researcher
Sulev Reisberg	University of Tartu	Data custodian / Researcher
Alessandro Pasqua	GENOMEDICS – Health Search HSD	Data custodian / Researcher
Lars Pedersen	AUH	Data custodian / Researcher
Lara Tramontan	Pedianet	Data custodian / Researcher
Miguel Angel Mayer	IMASIS	Data custodian / Researcher
Ron Herings	PHARMO	Data custodian / Researcher
Miriam Sturkenboom	EMC	Data custodian / Researcher
Johan van der Lei	EMC	Platform leader
Irene Bezemer	PHARMO	Data custodian / Researcher
Peter Rijnbeek	EMC	Work package leader

DOCUMENT HISTORY

Name	Date	Version	Description
G. Roberto, R. Gini	May 12, 2016	1.0	First draft
Giuseppe Roberto, Rosa Gini, Maria Garcia, Talita Duarte, Paul Avillach, Rients van Wijngaarden, Sulev Reisberg, Alessandro Pasqua, Lars Pedersen, Lara Tramontan, Miguel Angel Mayer, Ron Herings, Miriam Sturkenboom, Johan van der Lei, Martijn Schuemie, Peter Rijnbeek	July 7, 2016	1.1	Protocol revision and approval from all responsible parties

AMENDMENTS AND UPDATES

Version	Description of changes	Study protocol section	Date of effectiveness
1.2	Martijn Schuemie was removed from and Irene Bezemer was added to the responsible partties list.	Responsible parties	27/09/2016

ABSTRACT

The European Medical Information Framework (EMIF) project has the main objective of building an infrastructure for the efficient re-use of existing health care data for epidemiological research. Within the project, the EMIF-Platform represents a federation of heterogeneous sources of health data (e.g. administrative, hospital or primary care databases, disease registries, biobanks). One of the major challenges for the EMIF project is to deal with the different characteristics of the participating data sources in order to facilitate the execution of large multi-national, multi-data source observational studies and generate high quality evidence. For this purpose, a template data derivation procedure was specifically developed. In this proof-of-concept study, the standard procedure will be applied for the identification of patients with acute myocardial infarction (AMI) in a set of heterogeneous sources of observational health data. Validity indices (sensitivity, PPV) of the data source-tailored case-finding algorithms will be estimated from available evidence, and adjusted prevalence and incidence of AMI will be estimated from the participating data sources

BACKGROUND AND STUDY OBJECTIVE

The European Medical Information Framework (EMIF) project was launched at the end of 2012 with the objective of developing an infrastructure for the efficient re-use and exploitation of different types of existing observational health datan source (http://www.emif.eu/). Within the project, The EMIF-Platform represents a federation of observational healthcare data sources with heterogeneous characteristics (e.g. administrative, hospital or primary care databases, disease registries, biobanks) that currently collects real world data on over 50 million European citizens.

The data sources collaborating in the Platform may differ in terms of database structure, contents, reasons for recording, language, coding terminologies and healthcare system organization, so that each of them may have different strengths and limitations with respect to the identification of a specific event, population or study variable of interest[1-3]. Although heterogeneity of data sources represents a huge resource for the Platform, when multi-national, multi-data source studies are performed, the correct interpretation of results and their benchmarking across data sources might result problematic. In this contest, where data source-tailored case-finding algorithms are needed [4;5], the full and transparent documentation of the case identification process become also an issue[6;7].

In order to deal with the heterogeneity of the data sources collaborating in the Platform and facilitate the execution of high quality, multi-national, multi-data source studies, a novel standard data derivation procedure was specifically developed in EMIF. The advantages from the application of this procedure, which allows building data source-tailored case-finding algorithms in a standardized fashion, were demonstrated by the results of the first proof-of-concept study[8]: prevalent cases of a chronic clinical condition, type 2 diabetes, were identified in eight different data sources from six European countries, providing insight into both the strengths and limitations of each participating data source as well as the population of cases respectively identified.

As the second proof-of-concept study for the standard data derivation procedure developed in EMIF, we will identify patients with acute myocardial infarction (AMI) in a set of heterogeneous sources of observational health data in order to estimate the prevalence and incidence of AMI occurrence in the participating data sources.

MATERIALS AND METHODS

Data sources

Nine data sources from six different European countries (Italy, Denmark, Estonia, The Netherlands and Spain) will participate to the study. A brief description of the participating data sources provided by the relevant local data source expert is reported below.

- Agenzia regionale di sanità della Toscana (ARS)

Database description: The Italian National Healthcare System is organized at regional level: each region is responsible for providing to all their inhabitants a prespecified level of assistance through a national tax-based funding. The ARS data source comprises all the tables that are collected by the Tuscany Region to account for the healthcare services delivered to all the persons that are officially resident in the region. Moreover, ARS collects tables from regional initiatives. A unique anonymized person identifier code allows the linkage of patient-level information from different data tables. ARS data have been extensively used and validated for epidemiologic research purposes[1;3]. The collection of data into the ARS database started in 1996. Currently the database contains information from over 5 millions subjects with an average follow-up time of 9 years: they are all the subjects who have lived in Tuscany for at

least some time from 2003 on, except those who have never requested to be listed in the National Healthcare Service (a negligible part).

Database updates and data time lag: The database is updated every 15 days and consolidated every year.

Data subsets and variables: The database collects demographic information (birthdate, sex, citizenship, residence, data), diagnoses from hospital discharge records, death registry and registry of exemption from copayment (ICD9-CM), drug prescriptions dispensed for outpatient use (ATC), healthcare procedures from inpatient (ICD9-CM) and outpatient setting (local coding system)

Limitations of the database: Diagnoses are only recorded in inpatient setting. Primary care diagnoses are not available. Diagnoses of certain chronic and/or invalidating conditions may be recorded in outpatient setting as the reason for the exemption for copayment, although with low sensitivity and granularity. The prescription database captures medications dispensed for outpatient use only, which includes those prescribed by GP, during ambulatory care and upon hospital discharge. Medications used in inpatient setting are not available. The indications of use are not available as well.

- The Health Search IMS HEALTHCSD LPD (HSD)

Database description: The Health Search – IMS HEALTH/Longitudinal Patients Database (HSD) is a longitudinal observational database that is representative of the general Italian population. It was established in 1998 by the Italian College of General Practitioners. The HSD contains data from computer-based patient records from a select group of GPs (covering a total of 1.5 million patients) located throughout Italy who voluntarily agreed to collect data for the database and attend specified training courses. Turnover occurs as patients move and transfer to new practices. The records of 'transferred out' patients remain in the database and are available for retrospective studies with the appropriate time periods. The HSD complies with European Union, guidelines on the use of medical data for research. The HSD has been the data source for a number of peer-reviewed publications on the prevalence of disease conditions, drug safety and prescription patterns in Italian primary care. Approval for use of data is obtained from the Italian College of Primary Care Physicians. Data are in house, no ethical approval needed.

Data subset and variables: The database includes information on the age, gender, and identification of the patient, and GP registration information, which is linked to prescription information, clinical events and diagnoses, free text patients diary, hospital admission, and death. All diagnoses are coded according to the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM). Drug names are coded according to the ATC classification system. To be included in the study, GPs must have provided data for at least 1 year and meet standard quality criteria pertaining to: levels of coding, prevalence of well-known diseases, and mortality rates. At the time in which this study will initiate, 800 GPs homogenously distributed across all Italian areas, covering a patient population of around million patients, reached the standard quality criteria.

*Database updates and data time lag:*The database is updated continuously, every 6 months a data draw down is made for research purposes.

*Limitations of the database:*The main limitation is the difficulty to provide additional information from GPs since in such a case an ethical approval from all the local health authorities of the respective GP practice is needed. Medication

not reimbursed from the NHS are incomplete, as well as those prescribed by the specialists. Symptoms and diagnostic instrumental results are in free text form and are not necessarily complete.

Pubblications: http://www.healthsearch.it/pubblicazioni/articoli-pubblicati-su-riviste-indicizzate-su-pubmed/

- Integrated Primary Care Database (IPCI)

Database description: In 1992 the Integrated Primary Care Information Project (IPCI)[9] was started by the Department of Medical Informatics of the Erasmus University Medical School. IPCI is a longitudinal observational database that contains data from computer-based patient records of a selected group of general practitioners (GPs) throughout the Netherlands, who voluntarily chose to supply data to the database. GPs receive a minimal reimbursement for their data and completely control usage of their data, through the Steering Committee and are permitted to withdraw data for specific studies. Collaborating practices are located throughout the Netherlands and the collaborating GPs are comparable to other GPs in the country according to age and gender.

The database contains information on about 1.4 million patients. This is the cumulative amount of patients who have ever been part of the dynamic cohort of patients who have been registered. Turnover occurs as patients move and transfer to new practices. The records of 'transferred out' patients remain in the database and are available for retrospective studies with the appropriate time periods.

The system complies with European Union guidelines on the use of medical data for medical research and has been validated for pharmaco-epidemiological research. Approval for this study will be obtained from the 'Raad van Toezicht' an IPCI specific ethical review board.

Database updates and data time lag: The database is updated continuously, every 3 months a data draw down is made for research purposes.

Data subsets and variables: The database contains identification information (age, sex, patient identification, GP registration information), notes, prescriptions, physician-linked indications for therapy, physical findings, and laboratory values (e.g. potassium, creatinine). The International Classification of Primary Care (ICPC) is the coding system for patient complaints and diagnoses, but diagnoses and complaints can also be entered as free text. Prescription data such as product name, quantity dispensed, dosage regimens, strength and indication are entered into the computer. The National Database of Drugs, maintained by the Royal Dutch Association for the Advancement of Pharmacy, enables the coding of prescriptions, according to the Anatomical Therapeutic Chemical (ATC) classification scheme recommended by the WHO.

Limitations of the database: Limitations of the databases are that a lot of information is available in narratives, especially information from specialists and symptoms. Also specialist medications are not complete if the GP does not enter them. It is known, however, that this proportion is minor.

- The Health Improvement Network (THIN)

Database description: Pseudo-anonymised patient data are collected by THIN in a non-interventional way from the daily record keeping of general practices which use the Vision practice management software and have agreed to contribute to the scheme. As of May 2015, the THIN database contains primary care medical records from over 12 million patients, of which over 3.5 million are actively registered. IMS Health has a licence to facilitate access to THIN

Data for the purposes of medical research. IMS Health and researchers do not have access to practice or patient identifiers. However, the data are pseudo-anonymised in that THIN Additional Information Services (THIN AIS) can contact the general practitioners (GPs) so that the GP can provide additional information or contact patients. THIN Data have been used extensively in medical research since 2003 in the UK, Europe and the United States, with over 500 peer review publications utilising the THIN data source. The age and gender profile of the active patient population in THIN has been shown to be comparable to the UK population. Graphs comparing THIN with the Office for National Statistics UK population estimates for 2011 (latest available). Data within THIN are regionally representative as far as is possible within the distribution of the Vision practice software from which they are collected, representing more than 6% of the UK population. The regional representation of patients within THIN Data (September 2012 update)

THIN Data have also been shown to be generally representative of the UK in terms of Quality and Outcomes Framework chronic disease parameters. In addition, a study has been performed which compares THIN with data from practices using a different general practice software system (EMIS) and it was shown to match closely with these data, with the main exception that THIN patients are slightly more highly representative of the more affluent social class. As this socioeconomic information is available in THIN, researchers are able to adjust for it in analyses.

All studies using THIN, where the intention is to make public the study results, are subject to obtaining relevant prior ethical approval of the protocol.

Database updates: The database is updated 3 times per year.

Data subsets and variables: Demographics, Age, height, weight, BMI, smoking, social deprivation (patient postcode allocated IMD& Townsend), length of time with GP, transfer out date, death date, drug prescription (drug name, dose, frequency, pack size). This includes vaccinations and batch numbers. All Gps have electronic links to laboratories and lab tests requested by GPs are automatically uploaded. Appointments with GP and primary care based healthcare professionals

Limitations: THIN data has good information on therapy prescribed in general practice but this does not necessarily equate to therapy dispensed or take account of patient compliance with treatment. In addition it is not possible to include treatment bought by patients over the counter the study will therefore be restricted to GP prescriptions. Laboratory tests requested in secondary care are not available in the THIN database. Currently the THIN database has linked 60% of GP practices in England to the NHS secondary care “Hospital Episodes Statistics” database.

- Aarhus University Hospital Database (AUH)

Database description: The Aarhus University Hospital database is a system of linkage datasets in the area of the Central Denmark Region and the North Denmark Region. These are the two of five Danish Regions with a combined population of 1.8 million inhabitants and is representative of the population of Denmark. The population is entirely covered by a system of linkable registries and other administrative data sources. Since the healthcare is free and tax-supported in Denmark anyone will be recorded in these databases regardless of for instance age or income. The Civil Registration System holds key demographic data on all inhabitants in the population and maintains the civil registration code which is assigned to everyone at birth. The AUH system of databases includes data from in-patient, outpatient and emergency room visits from all somatic hospital in the two Regions. Surgical procedures and selected in-hospital

treatments are available since 1999. In addition, prescriptions dispensed at the pharmacies, laboratory measurement and causes of death are available.

Database updates and data time lag: The database is updated on a yearly basis.

Data subsets and variables: The database contains patient demographic data (CPR-number, birthdate, sex, residence, data on migration and death), prescriptions (ATC), hospital diagnoses (ICD-10), selected treatments and surgical procedures (NOMESCO), laboratory measurements (NPU) and Causes of Death (ICD-10)

Limitations of the database: Only diagnoses from hospital admissions and ambulatory care is included in the Hospital Discharge Registry and hence diagnoses from primary care is not available. The prescription database captures prescriptions dispensed in all pharmacies outside hospital and hence medication prescribed at the GP or in ambulatory care is recorded, but in-hospital medication is lacking. In addition, indication and instructions for use is lacking on all prescriptions in the prescription database.

Publications:

- *Nexø BA, Pedersen L, Sørensen HT, Koch-Henriksen N. Treatment of HIV and Risk of Multiple sclerosis*. Epidemiology. 2013;24:331-2.

- *Existing data sources for clinical epidemiology: The Danish National Database of Reimbursed Prescriptions.* Clin Epidemiol. 2012;4:303-13.

- *Statin Prescriptions and Breast Cancer Recurrence Risk: A Danish Nationwide Prospective Cohort Study*. J Natl Cancer Inst. 2011;103:1461-8.

Database description: The Aarhus University Hospital database is a system of linkage datasets in the area of the Central Denmark Region and the North Denmark Region. These are the two of five Danish Regions with a combined population of 1.8 million inhabitants and is representative of the population of Denmark. The population is entirely covered by a system of linkable registries and other administrative data sources. Since the healthcare is free and tax-supported in Denmark anyone will be recorded in these databases regardless of for instance age or income. The Civil Registration System holds key demographic data on all inhabitants in the population and maintains the civil registration code which is assigned to everyone at birth. The AUH system of databases includes data from in-patient, outpatient and emergency room visits from all somatic hospital in the two Regions. Surgical procedures and selected in-hospital treatments are available since 1999. In addition, prescriptions dispensed at the pharmacies, laboratory measurement and causes of death are available.

Database updates and data time lag: The database is updated on a yearly basis.

Data subsets and variables: The database contains patient demographic data (CPR-number, birthdate, sex, residence, data on migration and death), prescriptions (ATC), hospital diagnoses (ICD-10), selected treatments and surgical procedures (NOMESCO), laboratory measurements (NPU) and Causes of Death (ICD-10)

Limitations of the database: Only diagnoses from hospital admissions and ambulatory care is included in the Hospital Discharge Registry and hence diagnoses from primary care is not available. The prescription database captures prescriptions dispensed in all pharmacies outside hospital and hence medication prescribed at the GP or in ambulatory care is recorded, but in-hospital medication is lacking. In addition, indication and instructions for use is lacking on all prescriptions in the prescription database.

Publications:

- *Nexø BA, Pedersen L, Sørensen HT, Koch-Henriksen N. Treatment of HIV and Risk of Multiple sclerosis*. Epidemiology. 2013;24:331-2.

- *Existing data sources for clinical epidemiology: The Danish National Database of Reimbursed Prescriptions.* Clin Epidemiol. 2012;4:303-13.

- *Statin Prescriptions and Breast Cancer Recurrence Risk: A Danish Nationwide Prospective Cohort Study*. J Natl Cancer Inst. 2011;103:1461-8.

- PHARMO Network Database (PHARMO)

Database description: The PHARMO Database Network is a population-based network of healthcare databases and combines data from different healthcare settings in the Netherlands[10]. These different data sources, including in- and out-patient pharmacy, clinical laboratory, and hospitals are linked on a patient level through validated algorithms. The longitudinal nature of the PHARMO Database Network system enables to follow-up more than 4 million (25%) residents of a well-defined population in the Netherlands for an average of ten years.

Database updates and data time lag: The data sources are linked on an annual basis, meaning that the average lag time of the data is one year. The updated database becomes available in the second half of the year.

The PHARMO database network currently covers the period 1998-2013 (an update covering the period 1998-2014 is forthcoming).

Data subsets and variables: All electronic patient records in the PHARMO Database Network include information on age, sex, socioeconomic status and mortality.

The Out-patient Pharmacy Database comprises GP or specialist prescribed healthcare products dispensed by the out-patient pharmacy. The dispensing records include information on type of product, date, strength, dosage regimen, quantity, route of administration, prescriber specialty and costs.

The In-patient Pharmacy Database comprises drug dispensings from the hospital pharmacy, given during a hospitalisation. The dispensing records include information on type of drug, start and end date of use, strength, dosage regimen and route of administration.

The Clinical Laboratory Database comprises results of tests performed on clinical specimens. These laboratory tests are requested by GPs and medical specialists in order to get information concerning diagnosis, treatment, and prevention of disease. The electronic records include information on date and time of testing, test result, unit of measurement and type of clinical specimen.

The Hospitalisation Database comprises hospital admissions from the Dutch Hospital Data for more than 24 hours and admissions for less than 24 hours for which a bed is required. The records include information on discharge diagnoses, procedures, and hospital admission and discharge dates.

Limitations of the database: Data collection period, catchment area and overlap between data sources differ. Therefore, the final cohort size for any study will depend on the data sources included and the study design.

- IMASIS (IMASIS)

Database description: The IMASIS information system is the Electronic Health Records (EHRs) system of the Parc Salut Mar Barcelona Consortium that is a complete healthcare services organization[11]. Currently, this information

system includes the clinical information of two general hospitals, one mental health care centre, one social-healthcare centre and five emergency room settings in the Barcelona city area in Spain. In the future the system will include information of the EHRs of thirteen primary care teams. The Hospital del Mar is the principal public health facility, while social-public health services are concentrated at the Esperança Hospital and the Forum Centre. It also provides services for mental health and addiction for adults, children and youths at the Dr Emili Mira Centre. The first version of IMASIS information system was designed in 1984 and afterwards it was completely implemented and extended in several phases. Currently IMASIS includes administrative and clinical information of patients who have used the services of this healthcare system since 1990 and from different settings such as admissions, outpatients, emergency room and major ambulatory surgery. The database contains information on approximately 1.4 million patients and half of them have at least one diagnosis coded using "The International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM)".

Data subset and variables: IMASIS-2 is the relational model database containing anonymized patient information from IMASIS. It includes socio-demographic information such as date of birth, gender or dates of visits and admissions, and clinical information such as main and secondary diagnosis and procedures and, for a subset of hospital admissions, drug prescriptions and laboratory tests as well. All this data can be linked using a unique anonymous person identification number. Diagnoses and procedures are coded in ICD-9-CM. Drug names are coded with a local terminology and with the drug national code of the Spanish Medicines Agency.

Data updates and data time lag: Currently IMASIS-2 database is updated every 6 months. The last update was performed in March 2015. This update included additional information about admissions, outpatients and emergency room as well as drugs and laboratory results of the different hospitals of the system, which were not present in previous versions.

Limitations of the database: One of the limitations of the database is in relation to the fact that there is additional clinical information that could be useful but which is in a free text format. This type of information would require the use of specific text mining techniques to exploit it. It should also be borne in mind that the clinical information comes from several settings that were included in different moments and steps in the implementation process of the information system, and for that reason it is possible to find records at varying stages of completeness.

- The Estonian Genome Center of University of Tartu (EGCUT)

Database description: EGCUT is a biobank, which collects, stores and uses biological samples and phenotype information for about 52000 volunteer-based adult donors (age >= 18 years). The cohort covers 5% of the Estonian population and closely reflects the age, sex and geographical distribution in the population. EGCUT is established and maintained according to Human Genes Research Act (https://www.riigiteataja.ee/en/eli/531102013003/consolide). All participants have signed a broad informed consent, which allows the continuous update of epidemiological data through periodic linking to national electronic databases and registries.

Database updates and data time lag: Full health profile is described for each participant when person becomes a gene donor. This is conducted via Computer Assisted Personal Interview (CAPI) within 1-2 hours appointment at a doctor's office. For some donors the full health profile is described again after few years during follow-up. In addition, new information about the donors is gathered 1-2 times a year from national electronic databases and registries. However,

harmonizing and connecting the retrieved data to general database is not performed on regular basis yet and is expected to start to do so in the upcoming years.

Data subsets and variables: EGCUT has two types of data:

- Genetic data (derived from tissue samples):
  - whole-genome sequences (2300)
  - exome sequences
  - genotypes (more than 21500 donors, HumanOmniExpress, HumanCoreExome, , Exome, PsychArray, HumanCNV370, Cardio-Metabo, ImmunoChip – Illumina platforms)
  - imputed data
  - biochemistry: sugar, lipids, cholesterol
  - other measurements derived from tissue samples
- Phenotype data
  - diseases (ICD-10)
  - prescriptions (ATC)
  - personal data (place of birth, place(s) of living, nationality, education etc.)
  - genealogical data (family history of medical conditions spanning four generations)
  - lifestyle data (physical activity, dietary habits - FFQ, smoking, alcohol consumption, women's health, quality of life)
  - objective measurements (height, weight, BMI, heart rate etc.)
- Additional modules (e.g. disease specific data collections) are and can be added flexibly to the system. There are 40,000 participants with MCTQ (chronotype) data, and 15,000 with both MSTQ and genome-wide microarray (GWAS) data, 3,000 participants have filled the NEO-PI-R questionnaire, including GWAS data available on 2,700 participants.

Limitations of the database: The anonymous data of the gene donors are available for research projects. Access to the data is described in http://www.geenivaramu.ee/en/access-biopank/data-access

- SIDIAP

Database description: The Information System for the Development of Research in Primary Care (SIDIAP: www.sidiap.org) is a primary care anonymized database with continuous data collection since 2006 on a total of almost 6.5 million individuals (80% of the Catalan population and 10.2% of the total population of Spain), of which approximately 5.5 million are currently active. SIDIAP includes the population registered in 274 primary care practices throughout Catalonia, with a total of 3,414 general practitioners (GP) and is highly representative of both urban and rural areas. The average follow-up for individuals is 7.4 years. The information contained in SIDIAP is collected by health professionals during routine visits over time, and then it provides a good source of longitudinal population-based data. The information recorded includes demographic and lifestyle factors relevant to primary care settings (body mass index, smoking status, alcohol use, etc); clinical diagnoses, outcomes, and events (coded according to the International

Classification of Diseases, 10th revision [ICD-10]); referrals to specialists and laboratory tests; and prescribed and dispensed drugs by community pharmacies. The quality of SIDIAP data has been previously documented, and the database has been widely used to study the epidemiology of a number of health outcomes.

Database updates and data time lag: The database

e is updated on a yearly basis.

Data subsets and variables:

For all patients of the primary care teams (PCT) of the Catalan Institute of Health (CIH), SIDIAP has available the following information which is linked to a unique and anonymous identifier:

1. Information originated from the electronic healthcare records ECAP:

- Demographic data: Date of birth, gender, nationality, allocated PCT and professionals, MEDEA index
- Visits to Primary Care Service: Date, type and professional in charge
- Health Problems: ICD-10 code, date of diagnosis and expiry date. Available for acute and chronic health problems
- Clinical variables: Date and measurement. Clinical variables available are, for instance, blood pressure, smoking habit and BMI
- Immunisations: Vaccine administered and date
- Referrals: Date and department of referral
- Death: Date of death
- Prescriptions
- Sick leave: Date and ICD-10 diagnosis

2. Information on laboratory results: Since 2006, information is available on the results of laboratory analysis requested by PCT of the CIH. This information is obtained directly from the databases of the laboratories and therefore does not depend on the manual register of professionals in the PCT. Validation and standardization of protocols to guarantee data quality.

3. Information on medication dispensed in pharmacies: Since 2005, information is available on all pharmaceutical products with a prescription of the national health system signed by a professional of the Catalan Institute of Health and dispensed in pharmacies. This information is directly obtained from the joint database of CatSalut with Catalan pharmacies.

Limitations of the database: In SIDIAP, although the date of death is available because it is recorded in the electronic clinical history by GPs, the exact cause of death is unknown. Inpatient diagnoses are only available for a subset of the SIDIAP population from the hospital discharge registry of the CHI hospitals. Diagnosis and medications used or prescribed in inpatient setting are not available.

Publications: http://www.sidiap.org/index.php/dissemination/articles

Setting

Source population

All subjects in the corresponding catchment area of the participating data sources.

Study population

The study population in each participating data source will include all active subjects at 1st January 2013 (cohort entry) with at least 365 days of look-back (special criteria will be applied to EGCUT data, see Annex 2).

Study design

Descriptive, retrospective multi-national, multi-database cohort study.
Prevalence and incidence of AMI observed between 1st January and 31st December 2013 will be calculated in each participating data sources according to different case definitions.

Event definition

For the purposes of this study, AMI was defined in accordance with the third universal definition of myocardial infarction [12] as a condition of prolonged myocardial ischaemia leading to cell death as identified by any evidence of myocardial necrosis recorded in a clinical setting which is consistent with acute myocardial ischaemia.
Chest pain represents the main symptom associated to AMI. It may move to the shoulder, arm, back, neck, or jaw. It usually lasts more than 15-20 minutes. The discomfort caused by AMI may also seems a heartburn. Other symptoms associated to the onset of AMI are: dyspnoea, nausea, weakness, cold sweat, or feeling tired. Symptoms typically occur few minutes before the occurrence of the event. Diagnosis of AMI is based on anamnesis, electrocardiogram (ECG) showing new or presumed new significant ST-segment–T wave (ST–T) changes or new left bundle branch block (LBBB). Development of pathological Q waves in the ECG, cardiac biomarkers (troponins), imaging evidence of new loss of viable myocardium or new regional wall motion abnormality obtained through cardiac magnetic resonance imaging (MRI), and myocardial perfusion single-photon emission computerized tomography (SPECT).

Event Operationalization

In order to identify patients with AMI from the selected data sources, a set of standard algorithms, referred to as "component algorithms"[8], will be created. Each component algorithm will be used to identify those subjects who have a specific pattern of records from a single data domain, e.g. diagnoses, procedures or laboratory/test results(1). Component algorithms will be subclassified according to the settings of data collection in case this is expected to be associated to significant difference in terms of sensitivity and PPV with respect to the identification of AMI cases (e.g. diagnoses of AMI from inpatient, primary or secondary care setting, death registry). We will investigate the extent to which different components overlap in identifying the same subjects, and what is the unique contribution of each component, across different data sources.

To create the list of component algorithms, two sources of knowledge will be leveraged and integrated: a central expert-based clinical and operational definition of AMI (top-down approach) and the existing local expertise on AMI identification (bottom-up approach) provided by local data source experts. The final list of components will be homogeneous across data sources, and will be obtained through an iterative process.

Example of component algorithms for AMI identification

Algorithm acronym	Algorithm name	Description	Selection rules	Rules to identify subjects	Rules to identify date
DIAG_AMI_PC	Diagnosis in primary care	Patients who have at least one diagnosis recorded in a primary care setting	(AMI) occurs in [diagnosis fields] of [records collected during primary care]	all subjects such that selection rule holds once or more	date of first record
DIAG_AMI_INP	Diagnosis in inpatient care	Patients who have at least one diagnosis recorded during a hospital admission	(AMI) occurs in [diagnosis fields] of [records collected during inpatient care]	all subjects such that selection rule holds once or more	date of first record
DIAG_AMI_ER	Diagnosis in emergency admission	Patients who have at least one diagnosis recorded during a visit in the emergency room	(AMI) occurs in [diagnosis fields] of [records collected during emergency care]	all subjects such that selection rule holds once or more	date of first record
DIAG_AMI_SC	Diagnosis in inpatient care	Patients who have at least one diagnosis recorded during a specialist encounter	(AMI) occurs in [diagnosis fields] of [records collected during secondary care]	all subjects such that selection rule holds once or more	date of first record
DIAG_AMI_DEATH	Diagnosis in inpatient care	Patients who have at least one diagnosis recorded as cause of death	(AMI) occurs in [diagnosis fields] of [records collected at death]	all subjects such that selection rule holds once or more	date of first record
DIAGTEST_ECG_AMI	ECG results positive for AMI	Patients who have at least one ECG results positive for AMI	(ECG) occurs in [code of test field] of [records collecting laboratory test results] AND [result field] of the same record suggests new significant ST-segment–T wave changes or new left bundle branch block	all subjects such that selection rule holds once or more	date of first record

Algorithm acronym	Algorithm name	Description	Selection rules	Rules to identify subjects	Rules to identify date
PROCEDURE_AMI_THROMB	Procedure for admistration of thrombolytic therapy	Patients who have at least one record of thrombolytic therapy administration procedure	(Thrombolitic therapy) occurs in [code of procedure field] of [records collecting procedures in inpatient care]	all subjects such that selection rule holds once or more in a year	date of first record
PROCEDURE_AMI_PCI	Surgical procedure for Percutaneous Coronary Intervention (PCI)	Patients who have at least one record of surgical procedure for PCI	(Percutaneous coronary intervention) occurs in [code of procedure field] of [records collecting procedures in inpatient care]	all subjects such that selection rule holds once or more in a year	date of first record
LABVAL_AMI_CK	Blood creatin kinase MB levels higher than threshold	Patients who have at least one record of creatin kinase MB levels with values higher than threshold	(blood creatin kinase MB measurement) occurs in [code of test field] of [records collecting laboratory test results] AND [result field] of the same record are higher than threshold	all subjects such that selection rule holds once or more in a year	date of first record
LABVAL_AMI_TROPONIN	Blood troponin levels higher than threshold	Patients who have at least one record of blood troponin levels with values higher than threshold	(blood troponin measurement) occurs in [code of test field] of [records collecting laboratory test results] AND [result field] of the same record are higher than threshold	all subjects such that selection rule holds once or more in a year	date of first record

The Unified Medical Language System (UMLS) will be used to build a shared semantic foundation across the different coding systems in use to record information in each data base[13] medical concepts concerning diagnoses, pharmacological treatments, procedures or test results pertinent to AMI will be identified and projected to local terminologies.

Table X. Example of terminology mapping output for different terminologies in use to code diagnoses.

Medical concept: AMI

CUI	ICD9CM	ICD10	ICPC	READ
12345678	xxxx	xxxx	xxx	Xxx
39760372	yyyyy	yyyyy	yyyy	Yyyyy

CUI: Concept Unique Identifier from the Unified Medical Language System

Local experts will be requested to extract all the possible components from their data source (for instance if inpatient care is not recorded in a data source, they will not extract the corresponding components).
Each extracted component algorithm will be intended as a building-block to create more complex case-identification strategies: using a custom-built analysis tool (a Microsoft Access interface for Stata and LaTeX softwares, see annex 3), local experts and central investigators will be able to test the extracted components in different logical combinations by using Boolean operators AND, OR, AND NOT). Such logical combination of component algorithms will be referred to as "composite algorithms". This approach will allow the use of each extracted component as inclusion, exclusion or refinement criterion for the identification of AMI (e.g. "Select patients with at least a ECG results positive for AMI" AND "Patients with at least a record of PCI").

Each local expert will recommend a data source-tailored composite algorithm for the identification of AMI cases as the preferred case-finding algorithm to be applied to that data source. The local expert will also provide a comment on the reasons behind the choice of that specific composite algorithm. The local experts and central investigators will agree on an estimate of the algorithm's sensitivity and PPV in detecting prevalent and incident cases, using evidence available from literature, results from the analysis of the component algorithms in the same and in the other data sources, as well as mathematical equations linking validity indices with observed and actual prevalence and incidence[14;15].

Data analysis

Statistical hypothesis

Not applicable. This is a hypothesis-free descriptive study.

Statistical methods

Prevalence of patients with AMI will be computed as the ratio of the total number of patients with AMI observed during 2013 on the total number of subjects active at 1st January 2013. Incidence of AMI cases occurred in 2013 will be computed as the ratio of the total number of incident AMI cases observed (patients with AMI during 2013 and no AMI at any time before cohort entry) on the total number of subjects at risk for the event (i.e. patients with no AMI at any time before cohort entry).

The impact of each extracted component algorithm to the population of both prevalent and incident AMI cases identified with the recommended composite algorithm will be assessed per data source and presented as reported in the following example table.

TableX. Impact of individual component algorithms on the population of cases retrieved in each participating data source through the recommended composite algorithms.

		Recommended composite algorithm (A)		
		Data Source X	**Data Source Y**	**Data Source Z**
Component Algorithm (B)	Tot. data base population N in A % of A in N			
Algorithm 1	N in B % of B in A PR if B is added to A			
Algorithm 2	N in B % B in A PR if B is added to A			
Algorithm 3	N in B % B in A PR if B is added to A			

N- Number of subjects

A- Recommended composite algorithm

B- component algorithm

PR- Percentage Ratio of patients in A or in B with respect to the percentage of A in the data base population

The age band specific prevalence and incidence of AMI cases obtained from the application of the recommended composite algorithms, as well as the individual extracted components, will be compared across data sources considering the following age categories:

- 45 to 64 years
- 65 to 84 years
- 85+ years

Trends of prevalence and incidence will be estimated using the preferred algorithm, both crude and age adjusted[16] with ad hoc statistical models taking into account estimated PPV and sensitivity[15].

Data management and processing

A distributed network approach has been adopted in EMIF to allow partners for maintaining control of their data and to benefit from local data base expert consultation on the appropriate use of data and interpretation of results. Therefore, anonymized row patient-level data will be extracted and managed locally. A custom-built Java based software called Jerboa Reloaded (http://www.emif.eu/emif/scientific-publications/deliverables/data-extraction-software-v1) representing an updated version of Jerboa, which was used in previous multi-data source research projects[17] (see Annex 1). Jerboa Reloaded will be run by local data base experts allowing the standardization of the data analysis process. After providing formal written approval, DCs will upload the output file with the analytical aggregated dataset produced by the software to the private remote research environment (PRRE) for analysis.

Data will be analyzed in the PRRE using the custom-built analysis tool that will estimate the contribution of each component in the identification of subjects with AMI in each DB (See “Statistical methods” under the section “Data Analysis”).

Once the preferred algorithm is chosen, each data custodian will run a second script of Jerboa Reloaded on the same input files. The second script will create a dataset of counts of the preferred algorithm only, across age bands and years. This second file will be shared in the PRRE and will undergo further statistical analysis, to estimate observed and adjusted incidence and prevalence.

The input files used to run the Jerboa Reloaded, as well as the queries used for extraction, data management and input file creation, will be maintained in each participating institution together with relevant Jerboa Reloaded output files which will be also uploaded and archived in the PRRE together with analysis results.

Software and hardware

The following software will be used:

- Jerboa reloaded module algorithm comparison, version: v2.5.0.3
- Analysis tool, version 2.0 based on a Microsoft Access interface for Stata 13.0 and LaTeX
- Stata 13.0

QUALITY ASSURANCE

Jerboa software module has been double coded in SAS. The analytical datasets are homogeneously produced using Jerboa in all databases locally. All analyses will be conducted using the same analytical tool (“Analysis tool”) developed for this study.

A study-specific PRRE for secure access will be used. Due to data protection and ethical considerations, each partner will work with local data to create output files that will contain only aggregated anonymized data that will be shared in the PRRE where only the use case participants (data custodians) will have a secure and restricted access and where data will be analyzed.

LIMITATIONS OF STUDY METHODS

A validation of the records extracted in the databases will not be performed, since this is out of the scope of this study.

ETHICAL CONSIDERATIONS

All databases will submit this protocol in order to fulfill their local ethical guidelines and procedures.

DISSEMINATIONS AND COMMUNICATION STRATEGY

Data generated through this research will be shared among data partners before October 2016. Every data source will be free to reuse data generated from treatment of their own data.

A study report summarizing all main results will be produced and shared with data partners before November 2016.

The findings from this study will be submitted to a peer-review international journal by December 2016.

ANNEXES

Annex 1. Jerboa instructions

Available upon request by contacting the principal investigator.

Annex 2. Rationale for ad-hoc Patient file preparation in EGCUT

EGCUT is a data source that collects information from interviews of donors of biological samples.

Due to the cross-sectional nature of this data source, participants' (i.e. donors) observation period in EGCUT starts and ends on the same day (i.e. on the date of the interview), except for a sample of patients corresponding to about 4% of the total population (around 2000 patients) for which at least 2 interviews are available.

Therefore, the definition of "start date" and "end date" adopted for Patients file preparation in the longitudinal data sources of the EMIF-Platform cannot be applied to EGCUT, otherwise the ordinary designs implemented in Jerboa Reloaded for Primary Data Extraction would not detect any active subjects on a given date.

Moreover, in EGCUT information on participant's medical history, such as disease diagnoses and drug use, are collected on the day of the interview and recorded with "special" criteria:

i) Diseases diagnosed at any time before the interview are recorded. Onset date is also recorded either if the participant's GP has this information or if medical documentation is provided by the participant or it is just mentioned by the participant as he remembers (the participant is also asked whether he/she is suffering from a disease at present or not). Usually at least approximate year of the diagnosis is recorded. If it is not known, the field is left empty. As Jerboa does not allow empty dates, the date of the interview is given as diagnosis date on such cases in Event file.

ii) drugs used in the last two months preceding the interview to treat one of the mentioned diseases diagnosed are also recorded. No additional date is collected for drug exposure.

Based on the information reported above, study-specific criteria must be applied for ad-hoc Patient file preparation in EGCUT, in order to allow jerboa to produce reasonable results.

Studies of prevalence of a condition (UC6)

-Start date = Birth date

-End date = 1st January of the year following that of the interview (if more than one interview is available for the same patient, the last recorded interview is considered). If death year also available - for those donors, if their death is before the 1st January, "deathyear-12-31" is applied as the end date.

This way at any given date the estimated prevalence of the condition in the population interviewed after that date is available. This information can be interpreted as the prevalence of the condition in a healthy sample of the population of the catchment area.

When collecting conditions from diagnostic codes, two algorithms may be recommended

a) <COND>_DIAG_PC for conditions where the date is estimated by the GP or from medical documentation

b) <COND>_DIAG_OTH when the date is the date of the interview

Studies of prevalence of drug utilization

- Start date = 31st December of the year preceding the interview

- End date = 1st January of the year following that of the interview (if more than one interview is available for the same patient, the last recorded interview is considered)

In this case the date of the interview is always assumed as the date of the exposure.

Notably, in EGCUT indication of use is always available.

Other studies

For other studies the case of EGCUT must be treated separately.

Annex 3. Analysis tool user's manual

Available upon request by contacting the principal investigator.

References

(1) Gini R, Francesconi P, Mazzaglia G, et al. Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey. BMC Public Health 2013;13:15.

(2) Morley KI, Wallace J, Denaxas SC, et al. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. PLoS One 2014;9(11):e110900.

(3) Valkhoff VE, Coloma PM, Masclee GM, et al. Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. J Clin Epidemiol 2014 Aug;67(8):921-31.

(4) Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc 2013 Dec;20(e2):e206-e211.

(5) Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013 Jan 1;20(1):117-21.

(6) Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Med 2015 Oct;12(10):e1001885.

(7) Gini R, Schuemie M, Brown J, Ryan P. Data Extraction And Management In Networks Of Observational Health Care Databases For Scientific Research: A Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies. eGEMs 2016;4(1, Article 2).

(8) Roberto G, Leal I, Sattar N, et al. Pharmacoepidemiol Drug Saf Sep 2015 24: 1 587 Abstracts of the 31st International Conference on Pharmacoepidemiology and Therapeutic Risk Management, August 22-26, 2015, Boston, Massachusetts, USA 2015:page535-6 2015;535-6.

(9) Vlug AE, van der LJ, Mosseveld BM, et al. Postmarketing surveillance based on electronic patient records: the IPCI project. Methods Inf Med 1999 Dec;38(4-5):339-44.

(10) Herk-Sukel MP, Lemmens VE, Poll-Franse LV, Herings RM, Coebergh JW. Record linkage for pharmacoepidemiological studies in cancer patients. Pharmacoepidemiol Drug Saf 2012 Jan;21(1):94-103.

(11) Sancho JJ, Planas I, Domenech D, Martin-Baranera M, Palau J, Sanz F. IMASIS. A multicenter hospital information system--experience in Barcelona. Stud Health Technol Inform 1998;56:35-42.

(12) Thygesen K, Alpert JS, Jaffe AS, Simoons ML, Chaitman BR, White HD. Third universal definition of myocardial infarction. Glob Heart 2012 Dec;7(4):275-95.

(13) Avillach P, Coloma PM, Gini R, et al. Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. J Am Med Inform Assoc 2013 Jan 1;20(1):184-92.

(14) Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ 1994 Jul 9;309(6947):102.

(15) Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. Am J Epidemiol 1978 Jan;107(1):71-6.

(16) Eurostat. Revision of the European Standard Population. 2013.

(17) Coloma PM, Schuemie MJ, Trifiro G, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol Drug Saf 2011 Jan;20(1):1-11.