Identification of type 2 diabetes cases in a set of databases participating to the EMIF project

Protocol version: 1.5

Title	Identification of T2DM in a set of databases participating in the EMIF project.
Medicinal product(s) / Device(s)	Not applicable
Event(s) of interest	Type 2 diabetes
Research question and objectives	To describe the impact of different identification strategies of type 2 diabetes (T2DM) cases in heterogeneous sources of EHR participating in the EMIF project
Country(ies) of study	Denmark, Estonia, Italy, The United Kingdom, The Netherlands, Spain
Protocol author(s)	Giuseppe Roberto (ARS), Ingrid Leal (EMC), Rosa Gini (ARS)

TABLE OF CONTENTS

LIST OF ABREVIATIONS:

AUH	Aarhus University Hospital
ARS	Agenzia Regionale di Sanità della Toscana
DB	Database
DC	Data custodian
DM	Diabetes Mellitus
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
T1DM	Type 1 diabetes mellitus
T2DM	Type 2 diabetes mellitus
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
Ingrid Leal	EMC	Data custodian / Researcher
Naveed Sattar	British Heart Foundation Glasgow Cardiovascular Research Centre	Researcher
Katrina Loomis	Pfizer Worldwide Research and Development	Researcher
Paul Avillach	EMC	Researcher
Peter Egger	GSK	Researcher
Rients van Wijngaarden	PHARMO	Data custodian / Researcher
David Ansell	THIN	Data custodian / Researcher
Sulev Reisberg	University of Tartu	Data custodian / Researcher
Alessandro Pasqua	GENOMEDICS – Health SearchHSD	Data custodian / Researcher
Lars Pedersen	AUH	Data custodian / Researcher
James Cunningham	UNIMAN	Data custodian / Researcher
Lara Tramontan	Pedianet	Data custodian / Researcher
Miguel Angel Mayer	IMASIS	Data custodian / Researcher
Ron Herings	PHARMO	Data custodian / Researcher
Preciosa Coloma	EMC	Data custodian / Researcher
Francesco Lapi	ARS	Researcher
Miriam Sturkenboom	EMC	Data custodian / Researcher
Johan van der Lei	EMC	Platform leader
Martijn Schuemie	Janssen Research & Development	Work package leader
Peter Rijnbeek	EMC	Work package leader

DOCUMENT HISTORY

Name	Date	Version	Description
I. Leal	8-07-2015	1.0	First Draft
G. Roberto, R. Gini	30-07-2015	1.1	Text Revision
G. Roberto and R. Gini (ARS), I. Leal (IPCI), R. van Wijngaarden (PHARMO), D. Ansell (THIN), S. Reisberg (UTARTU), A. Pasqua (GENOMEDICS), L. Pedersen (AUH), L. Tramontan (PEDIANET), MA. Mayer (UPF)	15-09-2015	1.2	Revision of the document contents. Description of the participating data sources added by data custodians
I. Leal, G. Roberto, R. Gini	25-09-2015	1.3	Pre-final version
N. Sattar (BHF), K. Loomis (PFIZER), P.Avillach (EMC), P. Egger (GSK), R. van Wijngaarden (PHARMO), D. Ansell (THIN), S. Reisberg (UTARTU), M. Tammesoo (UTARTU) H. Alavere (UTARTU), A. Pasqua (GENOMEDICS), L. Pedersen (AUH), J. Cunningham (UNIMAN), L. Tramontan (PEDIANET), MA. Mayer (UPF), R. Herings (PHARMO), P. Coloma (EMC), F. Lapi (ARS), M. Sturkenboom (EMC), J. van der Lei (EMC), M. Schuemie (JANSSEN), P. Rijnbeek(EMC)	30-09-2015	1.4	Protocol revision and approval from data custodians and all study participants/contributors
G. Roberto, I. Leal, R. Gini	01-10-2015	1.5	Final Version

AMENDMENTS AND UPDATES

Version	Description of changes	Study protocol section	Date of effectiveness

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-data base observational studies and generate high quality. For this purpose, a template data derivation process was specifically developed and the identification of Type 2 diabetes mellitus (T2DM) was used as a test case.

The objectives of this study are: a) to establish a set of standard algorithms useful to identify patients with type 2 diabetes (T2DM) across heterogeneous sources of health data, b) to describe the data source-tailored combinations of standard algorithms recommended by the relevant local data base experts, c) to assess the impact of individual standard algorithms on the population of cases identified across different data sources.

BACKGROUND

In recent years, an increasing number of projects have been focusing on re-using existing electronic health records (EHR) for clinical research.[1] In particular, huge efforts have been made to combine heterogeneous health data from isolated environments and perform valid multi-data source observational studies.[2,3] In fact, leveraging data from different health care settings has a tremendous potential for filling important gaps of knowledge in health sciences allowing for the identification and study of specific populations or events of interest using otherwise unconceivable sample sizes.

In this context, at the end of 2012, the European Medical Information Framework (EMIF) project was launched with the main objective of building an infrastructure for the efficient re-use of existing health care data for epidemiological research (http://www.emif.eu/). Within the project, the EMIF-Platform represents a federation of heterogeneous sources (e.g. administrative, hospital or primary care databases, disease registries, biobanks) that currently collects real world data on around 40 million European citizens. Such data sources may differ in terms of database structure, content, reasons for recording, language, coding terminologies and healthcare system organization. Therefore, one of the major challenges for the EMIF project is to deal with their heterogeneity. For this purpose, a template data derivation process was specifically developed and the identification of Type 2 diabetes mellitus (T2DM) was used as a test case.

T2DM is a chronic clinical condition characterized by hyperglycemia due to insulin resistance and a progressive deficiency in insulin production. It is diagnosed and followed-up using laboratory tests for blood glucose measurements while treatment comprises interventions in the life style (i.e. diet and exercise) and use of medications. T2DM comprises about 90% of all diabetes cases worldwide.[4]

In order to identify subjects with this condition in EHR, diagnosis records can be used, as well as records from other data domains that collect information on activities routinely conducted as part of the patients' clinical care and follow-up (e.g drug prescriptions, laboratory tests and results).[5-7]Therefore, on the basis of the specific research question and data source characteristics, different algorithms can be adopted for T2DM identification. Since different algorithms may differ in terms of sensitivity and positive predictive value, the choice of a specific case identification strategy can have an important impact on the size and type of the population of patients retrieved[5] and should be taken into account when interpretating study results.

STUDY OBJECTIVE

a) To establish a set of standard algorithms useful to identify patients with type 2 diabetes (T2DM) across heterogeneous sources of health data,

b) to describe the data source-tailored combinations of standard algorithms recommended by local data base experts to identify T2DM,

c) to assess the impact of individual standard algorithms on the population of cases identified across different data sources.

MATERIALS AND METHODS

Data sources

Nine data sources from six different European countries (Italy, Denmark, The United Kingdom, Estonia, The Netherlands and Spain) will participate to the study. A brief description of the participating data sources provided by the relevant data base 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[8 9]. 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.

Integrated Primary Care Database (IPCI)

Database description: In 1992 the Integrated Primary Care Information Project (IPCI)[10] 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 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/

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.

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. 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. 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

Pedianet

Database description: Pedianet is a longitudinal observational database that collects epidemiological clinical data for clinical research from family paediatricians involved in the Pedianet network in Italy. This system is based on the transmission of specific data from computerised clinical files, which the paediatricians in the network fill out during their daily professional activities. Informed consent is required from the parents. Such data is collected anonymously by a central server in Padua, where it is validated and elaborated.

Pedianet is an independent network. The coordination of the projects and data analysis is carried out by a scientific committee that include internationally renowned paediatricians, epidemiologists and researchers. Approximately 400 paediatricians throughout the country have taken part in Pedianet projects.

The paediatric population involves infants, toddlers, children and adolescents up to 14 years. The database contains information on about 400000 patients. This is the cumulative amount of patients who have ever been part of the dynamic cohort of patients who have been registered.

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

Data subsets and variables: The database includes identification information (e.g. age, sex, patient identification, GP registration information), information about visits to the paediatrician (reason for the visit and diagnosis), information about drug, instrumental exams and specialist visits prescriptions, specialist visits and instrumental exams results, free text patient's diary, hospitalizations and emergency room access, patient's measurements (e.g. weight, height, head circumference), and death. Diagnoses are coded according to the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) or are indicated as free text. Drug names are coded according to the ATC classification system recommended by the WHO.

Limitations of the database: A limitation of the database is that some information is available in narratives, as free text, especially information about diagnosis and symptoms. However, in recent data and in the future almost all diagnosis

information is being coded. Moreover, we are developing text-mining algorithms to automatically extract information form free text fields. The recordings of diagnosis related to specialist visits or hospitalisation/emergency room might not be completed by the paediatrician or might be delayed. It is known, however, that this proportion is minor.

Publications:

- *Use of moxicillin, amoxicillin/clavulanate and cefaclor in the Italian pediatric population.* Journal of Pediatric Infectious Diseases 9 (2014) 1–9

- *Assessment of pediatric asthma drug use in three European countries; a TEDDY study*. Eur J Pediatr. 2011 Jan;170(1):81-92. doi: 10.1007/s00431-010-1275-7. Epub 2010 Sep 2.

- *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. doi: 10.1002/pds.2053. Epub 2010 Nov 8.

- *In-label and off-label use of respiratory drugs in the Italian paediatric population.* Acta Paediatr. 2010 Apr;99(4):544-9. doi: 10.1111/j.1651-2227.2009.01668.x. Epub 2010 Jan 27.

Table 1. Database characteristics

Name	**Type of data source**	**Catchment area**	**Source population size**	**Diagnoses** (setting, coding system)	**Medication** (coding system)	**Diagnostic procedures** (coding system)	**Laboratory results** (coding terminology for measurments)
ARS	Record linkage system	Tuscany (Italy)	3.7 millions	Inpatient, ICD9CM	ATC	ICD9CM or local terminology	-
AUH	Record linkage system	The northern and central region of Jutland. (Denmark)	1.8 million	Inpatient, secondary care ICD10	ATC	NOMESCO	-
PHARMO	Record linkage system	Netherlands (Certain regions, mainly South East and North-West)	4 million	Inpatient, ICD9CM	ATC	Local terminology	Local terminology
HSD	Primary care	Italy	2.5 millions	Primary care, ICD9CM	ATC	Local terminology	Local terminology
THIN	Primary care	United Kingdom	3.5 millions	Primary care, RCD	ATC	Local terminology	Local terminology
IPCI	Primary care	Netherlands	1 million	Primary care, ICPC/free text	ATC	Local terminology	Local terminology
IMASIS	Hospital	Barcelona (three city districts)	1.4 millions	Admissions, outpatients, major ambulatory surgery and emergency room ICD9CM	Local terminology & the Spanish Medicines Agency codes	ICD9CM	Local terminology

EGCUT	Biobank	Estonia	52000	Primary care/Self reported, ICD10	ATC	Local terminology	Local terminology
Pedianet	Primary care	Italy	350000	Primary care, ICD9CM, free text	ATC	Local terminology	Local terminology

Study design

Descriptive, cross-sectional, retrospective multi-database study.

Setting

Source population

All subjects recorded in the participating databases.

Study population

The study population will include all active subjects in the participating databases at two distinct index dates: 1st January 2009 and 1st January 2012, respectively.

Study period

The study period considered for the analysis will be from the start of data availability to the 1st January 2012 (except for EGCUT, see Annex 2 – Rationale for ad hoc patients file preparation in EGCUT).

Event definition

T2DM is a chronic disease characterized by hyperglycaemia due to insulin resistance and pancreatic beta-cell failure.[11]

According to the ESC/EASD guideline, DM is diagnosed as follows:

Diagnose/ measurement	WHO 2006[7]/2011[7]	ADA 2003 and 2012[5,6]
Diabetes		
HbA_{1c}	Can be used If measured ≥6.5% (48 mmol/mol)	Recommended ≥6.5% (48 mmol/mol)
	Recommended	
FPG	≥7.0 mmol/L (≥126 mg/dL)	≥7.0 mmol/L (≥126 mg/dL)
	or	or
2hPG	≥11.1 mmol/L (≥200 mg/dL)	≥11.1 mmol/L (≥200 mg/dL)

This table was adapted from: European heart journal. 2013;34(39):3035-87.[11]

From the cases fulfilling the above mentioned criteria, non-T2DM types (i.e. gestational diabetes, T1DM) should be excluded.

Event Operationalization

In order to identify patients with T2DM from the selected data sources, a set of standard algorithms, from now on referred to as "component algorithms", will be created. A component algorithm will be the identification of those subjects who have a specific pattern of records from a unique data domain (eg diagnoses, drug prescription or laboratory results), for instance:

1- Select subjects with at least two records of non-insulin antidiabetics prescription in one year;

2- Select patients with at least one record of insulin prescription in one year;

3- Select patients with at least one record of T2DM diagnosis recorded in a defined healthcare setting;

4- Select patients with at least one record of glucose measurement result above a defined threshold.

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 T2DM (top-down approach) and the existing local expertise on T2DM derivation (bottom-up approach) provided by local database experts.

Example of component algorithms for T2DM identification

Algorithm acronym	Algorithm name	Description	Selection rules	Rules to identify subjects	Rules to identify date
T2DM_ALG_A	Diagnosis in primary care	Patients who have at least one diagnosis recorded in a primary care setting	(Diabetes type 2) occurs in [diagnosis fields] of [tables collected during primary care]	all subjects such that selection rule holds once or more	date of first record
T2DM_ALG_B	Diagnosis in inpatient care	Patients who have at least one diagnosis recorded during a hospital admission	(Diabetes type 2) occurs in [diagnosis fields] of [tables collected during inpatient care]	all subjects such that selection rule holds once or more	date of first record
T2DM_ALG_C	Glycated hemoglobin values higher than threshold	Patients who have at least one result recorded from a glycated hemoglobin test which is higher than 6.5% (48 mmol/mol)	(Glycated Haemoglobin) occurs in [code of test field] of [tables collecting laboratory test results] AND [result field] of the same record is higher than 6.5% (or 48 mmol/mol, according to unit of measurement adopted in the table)	all subjects such that selection rule holds once or more	date of first record
T2DM_ALG_D	Non-insulin antidiabetic utilization	Patients who have at least two prescriptions of non-insulin antidiabetics in a calendar year	(Drugs used in diabetes, excl insulin) occurs in [ATC field] of [drugs tables]	all subjects such that selection rule holds twice or more in a year	date of second record

T2DM_ALG_E	Utilization of tests indicated for diabetes	Patients who have at least three glycated hemoglobin tests in two calendar years, or two per year in five consecutive years	[1] (Glycated Haemoglobin) occurs in [code of test field] of [tables collecting laboratory test results or dispensing] OR [2] (Blood glucose measurement) occurs in [code of test field] of [tables collecting laboratory test results or dispensing]	all subjects such that selection rule [1] holds three times or more in two years OR all subjects such that selection rule [2] holds twice per year or more in five consecutive years	earlier between date of third record in selection rule [1] and date of the second record in 5th year in selection rule [2]

The Unified Medical Language System (UMLS) will be used to build a shared semantic foundation across the different coding systems in use in each data base:[3] medical concepts concerning diagnoses, pharmacological treatments and diagnostic procedures pertinent to T2DM will be identified and projected to local terminologies.

Table X. Example of terminology mapping output

DIABETES TYPE 1

CUI	ICD9CM	ICD10	ICPC	READ	ATC
C0375131	250.41				
C0375118	250.11				
C0375148	250.83				
C0375116	250.03				
C0375127	250.31				
C0011854		E10	T89001	X40J4	
			T89002		
			T89003		
			T90004		
			T90006		
			T90008		
C0375123	250.21				
C0375146	250.81				
C0342302				66AJ1	
C0375150	250.91				
C0375135	250.51				
C0375125	250.23				
C0375133	250.43				
C0375129	250.33				
C0375152	250.93				
C0375114	250.01				
C0375136	250.53				
C0375120	250.13				

C0375138 C0011855	250.61				
C0375142	250.71				
C0375144	250.73				
C0375140	250.63				

DIABETES TYPE 1

CUI	ICD9CM	ICD10	ICPC	READ	ATC
C0375151 C0011860	250.92	E11	T90005 T90007 T90009	X40J5	
C0375143	250.72				
C0375117	250.10				
C0375126	250.30				
C0375149	250.90				
C0375147	250.82				
C0375141	250.70				
C0375119	250.12				
C0375122	250.20				
C0375132	250.42				
C0375130	250.40				
C0375134	250.50				
C0375145	250.80				
C0375115	250.02				
C0375113	250.00				
C0376128	250.52				
C0375137	250.60				
C0375124	250.22				
C0375128	250.32				
C0375139	250.62				

CUI: Concept Unique Identifier from the Unified Medical Language System

Each local data base expert will extract only those components that apply to the data collected in the respective data source.

Using a custom-built analysis tool (a Microsoft Access interface for Stata and LaTeX softwares, see annex 3), local data base experts will be able to test the extracted components in different logical combinations by using Boolean operators AND, OR, AND NOT). From now on, logical combination of components 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 T2DM (e.g. “Select patients with at least a diagnosis of T2DM” AND NOT “Patients with at least a diagnosis of type 1 diabetes”).

Data analysis

Statistical hypothesis

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

Statistical methods

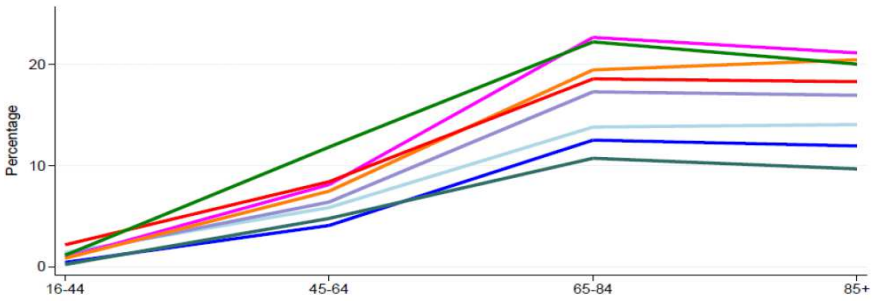
Results from individual component algorithms, as well as from each data source-specific composite algorithm, will be computed and presented as age band distribution of the percentage of the data base population identified.

The age categories will be:

- 0 to 15 years
- 16 to 44 years
- 45 to 64 years
- 65 to 84 years
- 85+ years

Example:

Figure x. Age band distribution of the percentage of data base population identified by component algorithm A in different data source



The contribution of each extracted individual component algorithm to the population of cases identified with the recommended composite algorithm will be assessed per data source and presented as reported in the following example:

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

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[2] (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 T2DM in each DB (See “Statistical methods” under the section “Data Analysis”).

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.

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. Results will not be intended as disease frequency estimates.

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 December 2015. 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 December 2015.

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

References

1. Richesson RL, Horvath MM, Rusincovitch SA. Clinical research informatics and electronic health record data. *Yearb Med Inform* 2014;9(1):215-23.

2. 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.

3. 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.

4. World Health Organization. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus. Geneva; 1999. Report No.: WHO/NCD/NCS/99.2.

5. Richesson RL, Rusincovitch SA, Wixted D, et al. A comparison of phenotype definitions for diabetes mellitus. *J Am Med Inform Assoc* 2013 Dec;20(e2):e319-e326.

6. Carstensen B, Kristensen JK, Ottosen P, et al. The Danish National Diabetes Register: trends in incidence, prevalence and mortality. *Diabetologia* 2008 Dec;51(12):2187-96.

7. Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc* 2012 Mar;19(2):212-8.

8. 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.

9. 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.

10. 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.

11. Ryden L, Grant PJ, Anker SD, et al. ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD: the Task Force on diabetes, pre-diabetes, and cardiovascular diseases of the European Society of Cardiology (ESC) and developed in collaboration with the European Association for the Study of Diabetes (EASD). *Eur Heart J* 2013 Oct;34(39):3035-87.

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 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 (we also ask if the participant 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.