SEVENTH FRAMEWORK PROGRAMME

SAFEGUARD

SAFETY EVALUATION OF ADVERSE REACTIONS IN DIABETES

Safety Evaluation of Adverse Reactions In Diabetes

HEALTH-282521

Common Study Protocol for Comparative Studies

WP4 – Observational Studies

Final
V 1.15

Lead beneficiary: EMC
Date: 13/06/2012
Nature: Report
Dissemination level: CO
(Confidential, only for members of the Consortium and Commission Services)

TABLE OF CONTENTS

Document Information

Grant Agreement Number	Health- 282521	**Acronym**	SAFEGUARD
Full title	Safety Evaluation of Adverse Reactions in Diabetes		
Project URL	www.safeguard-diabetes.org		
EU Project officer	Iiro Eerola (Iiro.EEROLA@ec.europa.eu)		

Deliverable	Common Study Protocols for Observational Studies – Comparative studies			
Work package	**Number**	4	**Title**	WP4: Observational Studies

Delivery date	**Contractual**	31/05/2012	**Actual**	
Status	Version 1.15		final ☐	
Nature	Report ☑ Prototype ☐ Other ☐			
Dissemination Level	Public ☐ Confidential ☑			

Authors (Partner)	Ingrid Leal (EMC), Silvana Romio (EMC), Miriam Sturkenboom (EMC)			
Responsible Author	Silvana Romio (EMC)		**Email**	s.romio@erasmusmc.nl
	Partner	EMC	**Phone**	+31-10-7044-128

Description of the deliverable	
Key words	

Document History

Name	Date	Version	Description
I Leal	06-01-2012	V1.1	First draft for internal revision.
SA Romio	09-01-2012	V1.2	Review of first draft
I Leal	02-05-2012	V1.3	Updated draft
SA Romio	21-05-2012	V1.4	Minor comments
JD Seeger, L Scotti	25-05-2012	V1.5	Review of confounding section
SA Romio	28-05-2012	V1.6	Integration/review all deliverable
C.de Vries,SA Romio	30-05-2012	V1.7	Integration of comments
M. Sturkenboom	01-06-2012	V1.8	Review
I Leal	05-06-2012	V1.9	Integration of comments
SA Romio	06-06-2012	V1.10	Integration/review all deliverable
M. Sturkenboom	07-06-2012	V1.11	Review
SA Romio, I Leal	08-06-2012	V1.12	Final review
SA Romio, I Leal	28-06-2012	V1.13	Integration of comments from reviewers
SA Romio, I Leal	13-07-2012	V1.14	Final integration
SA Romio, I Leal	17-08-2012	V1.15	Integration of seccion for quality assurance and study timelines

Definitions

- Partners (also named as beneficiaries) of the SAFEGUARD Consortium are referred to herein according to the following codes:

  EMC - Erasmus Universitair Medisch Centrum Rotterdam (Netherlands) - Coordinator
  SYNAPSE - Synapse Research Management Partners S.L. (Spain) - Beneficiary
  PHARMO - PHARMO Coöperatie UA (Netherlands) - Beneficiary
  F-SIMG - Fondazione Scientifica SIMG-ONLUS (Italy) - Beneficiary
  UBATH - University of Bath (UK) - Beneficiary
  AEMPS - Agencia Española de Medicamentos y Productos Sanitarios (Spain) - Beneficiary
  CMNS - Consorzio Mario Negri Sud (Italy) - Beneficiary
  DSRU - Drug Safety Research Trust (UK) - Beneficiary
  CUNI - Univerzita Karlova v Praze (Czech Republic) - Beneficiary
  VUA - Vereniging Voor Christelijk Hoger Onderwijs Wetenschappelijk Onderzoek en Patientenzorg (Netherlands) - Beneficiary
  BWH - The Brigham and Women's Hospital, Harvard Medical School (US) - Beneficiary
  UNIMIB - University of Milano-Bicocca (Italy) - Beneficiary
  Uni-HB - Universitaet Bremen (Germany) - Beneficiary
  RTI-HS - RTI Health Solutions (US) - Beneficiary

- **Grant Agreement**: The agreement signed between the beneficiaries and the European Commission for the undertaking of the SAFEGUARD project (HEALTH-282521).
- **Project**: The sum of all activities carried out in the framework of the Grant Agreement.
- **Work plan**: Schedule of tasks, deliverables, efforts, dates and responsibilities corresponding to the work to be carried out, as specified in Annex I to the Grant Agreement.
- **Consortium**: The SAFEGUARD Consortium, conformed by the above-mentioned legal entities.
- **Consortium Agreement**: Agreement concluded amongst SAFEGUARD participants for the implementation of the Grant Agreement. Such an agreement shall not affect the parties' obligations to the Community and/or to one another arising from the Grant Agreement.
- **Deliverable review**: An evaluation procedure by one or more reviewers, which precedes the distribution of a deliverable (as defined in the work plan) to the European Commission.
- **Quality assurance**: All the planned and systematic activities implemented to provide adequate confidence that an entity will fulfil requirements for quality.
- **Quality policy**: A set of principles on which quality assurance procedures are based.
- **Risk**: Uncertainty that may have a significant impact on the execution or outcome of the project, and which effect may be negative – a *threat* risk - or positive – an *opportunity* risk.
- **Foreground:** Means the results, including information, whether or not they can be protected, which are generated by activities in the Project. Such results include rights related to copyright; design rights; patent rights; plant variety rights or similar forms of protection.
- **Background:** Means information which is held by participants prior to their accession to the Grant Agreement, as well as copyrights or other intellectual property rights pertaining to such information, the application for which has been filed before their accession to the Grant Agreement, and which is needed for carrying out the indirect action or for using the results of the indirect action.

Abbreviations

The following abbreviations are used in this report:

- **ATC** – Anatomical therapeutic chemical classification system
- **BMI** – Body Mass Index
- **DDD** – Defined Daily Dose
- **DDP**-4 – Dipeptidyl peptidase 4
- **DM** – Diabetes Mellitus
- **EU –** European
- **GP** – General Practitioner
- **ICD-9-CM** – International Classification of Disease, 9th rev., Clinical Modification
- **ICD-10-GM** – International Classification of Disease, 10th rev., German Modification
- **ICPC** – International Classification of Primary Care
- **IV** - Intravenous
- **AMI** – Acute Myocardial Infarction
- **OR** – Odds ratio
- **OTC** – over-the-counter medication
- **RX** – prescription
- **SCD –** Sudden Cardiac Death
- **SUD** – Sudden Unexpected Death
- **T2DM** – Type 2 Diabetes Mellitus
- **TM –** Total Mortality
- **UK** – United Kingdom
- **USA –** United States of America
- **VA –** Ventricular Arrhythmia
- **VF** – Ventricular Fibrillation
- **VT –** Ventricular Tachycardia
- **WHO** – World Health Organization

1. Background

According to the International Diabetes Federation it is estimated that in 2010, 284 million people had diabetes mellitus (DM), of them, 55 million were in Europe[1]. By 2030, it is expected that there will be 438 million with DM, 66 million in Europe[2]. Type 2 Diabetes Mellitus (T2DM) accounts for 90-95% of DM.

T2DM is a chronic and progressive disease characterized by hyperglycemia due to a defect in insulin signalling (insulin resistance) and insulin deficiency. These features leads to hyperglycemia and dyslipidaemia, which can impair insulin secretion and action[3] and further cause microvascular (diabetic nephropathy, neuropathy and retinopathy) and macrovascular (coronary artery disease [CAD], peripheral arterial disease [PAD] and stroke) complications[4]. Cardiovascular and cerebrovascular diseases (mainly ischemic heart disease, heart failure [HF], stroke and sudden death [SD]) are the leading causes of death in patients with diabetes. [5] Several studies have demonstrated that the intensive control of blood glucose levels reduced the risk of microvascular complications; unfortunately, the reduction of macrovascular complications has not been proven[6].

The treatment of T2DM is based on lifestyle changes and pharmacological treatment. Its main goal is to prevent and control hyperglycemia and then reduce its complications. There are several treatment options to reduce blood glucose levels. For decades treatment options were based on lifestyle changes, metformin, sulfonylureas, and insulin. By the end of the 90's and during the last decade, new compounds with different mechanisms of action have been developed and authorized for marketing[7]: thiazolidinediones (TZD), meglitinides, Glucagon-like peptide 1 receptor (GLP-1R) agonists, Dipeptidyl-peptidase 4 (DPP-4) inhibitors and amilyn analogs. The progressive nature of T2DM and its associated glycemia, tends to lead to increases in the dose and the use of combinations of non-insulin blood glucose lowering drugs (NIBGLD) or the addition of insulin over time to meet the goals for glycemia control. According to the Position statement issued by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD), the management of hyperglycemia is based on subject characteristics (e.g. life expectancy, comorbidities, vascular complications, etc) and a proposed sequencing of anti-hyperglycaemic therapy. This recommendation starts with lifestyle changes (diet, weight control and physical activity) and then recommends an add-on therapy strategy starting from monotherapy to only insulin strategies if combinations of NIBGLD and basal insulin fail to achieve HbA1c target[8]. Since patients with T2DM usually have other cardiovascular risk factors such as dyslipidemia, hypertension, hypercoagulability and obesity, the treatment of these patients also comprises treatment for these comorbidities.

Although safety issues associated with blood glucose lowering drugs are not new, recently, the safety of these treatments has been questioned and highly publicized. It has been reported that some of them increase the risk or modify the prognosis of diseases such as cancer, cardiovascular (CVD) or pancreatic diseases.[5]

The first TZD in the market was troglitazone which was approved by fast track process by the end of 1996 in the United States of America (USA), and in October of 1997 in the United Kingdom (UK) where it was withdrawn just 2 months later due to liver toxicity[9]. In the USA it remained in the market until 2000 when new options (Rosiglitazone and Pioglitazone) with the same mechanism of action, but without the hepatic toxicity of troglitazone were available [10]. Rosiglitazone and Pioglitazone were approved by a fast-track process in 1999 in the USA and in 2000 in Europe. In 2007, Nissen *et al* published a meta-

analysis of randomized clinical trials, reporting an increased risk of myocardial infarction (OR=1.43, p-value=0.03) with rosiglitazone use[11]. On the same day (21 May 2007) the US Food and Drug Administration (FDA) issued a safety alert about this increased cardiovascular risk[12]. Similarly, on January 24th, 2008, the European Medicine Agency (EMA) issued a warning against the use of rosiglitazone in patients with ischemic heart disease and/or peripheral arterial disease[13]. The meta-analysis conducted by Nissen et al was criticized because of the variability of the study designs and outcome assessment among others.

Based on all the findings the treatment guidelines were updated and the use of rosiglitazone was no longer recommended [14]. In August 2007, FDA issued a new alert, but this time, the alert was for both available TZDs (rosiglitazone and pioglitazone) because of the potential of causing or exacerbating congestive heart failure in specific populations [15, 16]. Finally, on September 23rd, 2010 EMA recommended the suspension of rosiglitazone in the EU[17], while FDA restricted its use to patients already being successfully treated with this drug and whose glycemia cannot be controlled with other anti-diabetic drugs and whom after consulting healthcare professional, do not wish to switch to pioglitazone as monotherapy[18].

Drug-induced pancreatitis is considered a rare diagnosis and it is estimated that less than 2% of pancreatitis cases are induced by a drug[19]. According to epidemiological studies, subjects with T2DM have an increased risk of pancreatitis of 2.83 relative to subjects without diabetes[20]. Health authorities have received reports of acute pancreatitis in patients taking exenatide and sitagliptin. Exenatide belongs to one of the new classes of NIBGLD, andi its effects are mediated via its incretin actions.
Exenatide and liraglutide are agonists of the GLP-1R and mimic the effect of the Glucagon-like peptide 1 (GLP-1), improving the secretion of insulin that is dependent on oral ingestion of glucose, inhibiting glucagon release, and increasing the feeling of satiety. Exenatide was initially approved in the USA in 2005[21] and in 2006 by the EMA[22]. In October 2007 the FDA issued an alert for exenatide (Byetta) due to 30 reports of pancreatitis in patients taking this drug.[10] Liraglutide is the latest GLP-1R agonist approved by the FDA and EMA in 2010 and 2009 respectively. In clinical trials, more pancreatitis cases were reported in subjects taking liraglutide than those in other medications for T2DM [12].

Sitagliptin, vildagliptin and saxagliptin are DPP-4 inhibitors and they lower glucose based also on their incretin effects. The first DPP-4 inhibitor, sitagliptin, was approved by the FDA in October 2006[23]. In October 2009 the FDA issued an alert of acute pancreatitis after it was reported that 88 subjects taking sitagliptin or the combination sitagliptin/metformin had episodes of acute pancreatitis, two of those episodes were hemorrhagic or necrotizing pancreatitis.

Preclinical data showed attenuated expression of proinflamatory genes with the activation of GLP-1R[24] but until now, the mechanism by which exenatide could cause acute pancreatitis is unknown. Results of post marketing studies are controversial; a study conducted using the US FDA database of reported adverse events concluded that there is a 6 fold increase in the odds ratio (OR) for reported pancreatitis with the use of sitagliptin or exenatide [25]. On the other hand, two cohort studies performed in databases in the USA, did not find an association between exenatide and acute pancreatitis [26, 27].

It has been reported that GLP-1R agonists increase heart rate. This finding was observed in preclinical studies[28] and in patients treated with exenatide although not statistically significant. Gill and colleagues reported a non statistically significant increase in heart rate of 1.5 ± 1.8 beats/min (p=0.4) from baseline and a statistically significant increase of 3.4 ± 1.6 beats/min with dose escalation from 5 to 10 mcg[29]. GLP-1Rs have been detected in cardiomyocytes, muscle and vascular smooth muscle cells. Specifically,

the cardiovascular effect proposed is an increased inotropic action and improved myocardial glucose uptake [30]

Although the mechanism and the magnitude of the association has not been completely understood or described, several publications have reported an increased risk of sudden cardiac death (SCD) in subjects with T2DM. Among the risk factors that could contribute to this association are silent myocardial ischemia, autonomic nervous system dysfunction, abnormal cardiac repolarization, hypoglycaemia, hypercoagulable state, cardiomyopathy and impaired respiratory response to hypoxia and hypercapnea. [31] On the other hand it has been reported that some drugs could cause Qtc prolongation and the regulatory agencies have issued guidelines to assess the cardiovascular risk in drugs for T2DM before marketing[32]. These guidelines apply to recently developed drugs as a requirement for marketing authorization, for example FDA requested additional information on QTc prolongation for a GLP-1 agonist[33]. Therefore, it is important to assess the risk of cardiovascular events among drugs already in the market, including the recently launched medications.

DM has been reported as a risk factor for several types of cancer, including liver and pancreatic cancer each with a relative risk of 2 or higher [34]. Hyperinsulinemia (one of the characteristics in T2DM) and activation of the Insulin-like growth factor (IGF) receptor by insulin have been mentioned as the possible pathophysiologic causes with mitogenic effects[35]. Treatment options for DM have been reported as possible risk factors for cancer as well.

It has been reported that incretin based therapies could cause pancreatic cancer. One potential explanation is that the long-term activation of the GLP-1R could increase the risk of cancer in the pancreas [25], as a preclinical study reported increased β-cell proliferation with exenatide and liraglutide. Additionally, sitagliptin increased pancreatic ductal hyperplasia [35]. Cancer related to incretin-based therapies has not been reported in humans. An increased risk of pancreatic cancer with insulins (OR: 2.29,
95 % CI: 1.34 – 3.92) and sulfonylureas (OR: 1.90, 95 % CI: 1.32 – 2.74) and a risk reduction with the use of metformin (OR: 0.43, 95 % CI: 0.23 – 0.80) have been reported in observational studies.[36]

Several methodological issues could have played an important role in the observed association between NIBGLD and cancer such as, reverse causality, indication and detection bias. It has been reported that the risk of pancreatic cancer is higher during the first months after the diagnosis of diabetes and decreases over time[37]. Detection bias could be present since the number of visits to the physician increases when diabetes is diagnosed, and therefore the opportunity for cancer detection among subjects with diabetes is higher than in the general population. Reverse causality could be present since hyperglycemia might be a symptom of an undiagnosed tumour present in the pancreas. Finally, indication bias might be present because most of the studies used metformin monotherapy as a comparator, and this corresponds to first line treatment in T2DM[38], so patients using metformin are usually those with recent diabetes diagnosis and with fewer long-term T2DM complications. Additionally metformin has shown inhibition of cell proliferation, reduction of colony formation and cell cycle arrest in cancer cell lines[35], so that the associations found could reflect an increased risk of cancer with these drugs or a decreased risk with the use of metformin.

If these new therapies are associated with an increased risk of cancer, it would be important to distinguish if they induce or promote the malignancy. Specifically for pancreatic cancer, it was reported that the time for tumour induction and development since the first cell mutation is around 10 years[39]. Since these

new therapies have been in the market for less than 10 years, if the association is observed, maybe they promote but, not induce tumour progression.

In July, FDA[40] and EMA[41] issued alerts regarding the association of pioglitazone and bladder cancer. Lewis et al [42] studied the risk of bladder cancer in patients treated with pioglitazone compared to non-pioglitazone users. The results of an interim analysis did not show an increased risk, although they found a weakly increased risk after 2 years of exposure.

Until now, data regarding the safety of NIBGLD are inconclusive because of methodological issues due to the complexity of the disease, the treatment patterns of use (switching and combined therapies), comordidities and the comedication used for these comorbidities in subjects with T2DM. Additionally, diabetes shares many risk factors with different types of cancer, pancreatic and cerebrovascular diseases, and the inter-relatedness of the risk factors with the diseases have made it difficult to study these possible associations.

Since the number of people with diabetes is increasing and its treatment exposes large numbers of people to these drugs, their safety with respect to some types of cancer, CVD or pancreatic diseases represents a public health priority. The nature of treatment for T2DM, with chronic use of progressively higher doses of NIBGLD along with combinations including combinations with insulin, the issuance of several warnings regarding the safety of these medications, requires study using the most advanced methodological strategies available.

2. Study Objectives

The primary objective of these observational studies is to estimate the risk of myocardial infarction (MI), heart failure (HF), ventricular arrhythmia (VA)/sudden cardiac death (SCD), ischemic stroke (IS), hemorrhagic stroke (HS), acute pancreatitis (AP), pancreatic cancer (PC), bladder cancer (BC) and total mortality (TM) associated with the use of NIBGLD and insulins and insulin analogs in subjects with T2DM.compared with non users.

Secondary objectives are:

- To assess the background rates of the different events of interest in the population of subjects with T2DM

3. Methodology

3.1. Study design

The association of NIBGLD, insulins and insulins analogs in subjects with T2DM with each outcome of interest will be evaluated. Nested case control studies in a cohort of T2DM patients will be conducted to assess the association of NIBGLD, insulins, and insulin analogs with MI, HF, VA/SCD, IS, HS, AP, BC and PC.

A separate study will be conducted for each outcome. All analyses will be performed in each database separately and the heterogeneity between databases will be examined. A pooled estimate obteined using a meta-analytic approach will be calculated. The choice of the model (fixed or random effect) will depend on the results obtained from the test for heterogeneity.

A dynamic retrospective cohort study will be carried out to estimate the association between NIBGLD, insulins, insulin analogs and TM. For this outcome, the cohort study has some advantages on the case control design (e.g., time varying covariates can be taken into account, time to event can be studied). This is a better approach because information at baseline or at specific points in time is not enough to characterize the patient clinical history

To estimate the background rates of the outcomes of interest, a dynamic retrospective cohort study will also be used.

3.2. **Data Sources**

Data collected in 9 different electronic health databases (DBs) from 5 different European countries and from the USA will be used in the analysis. These DBs include different types of data sources (electronic medical records, administrative DBs and record linkage systems), different settings (out and in patient care, primary care), and the use of different terminology and coding systems for the registry of events and medications. In order to have homogeneous information for data extraction, events, including outcomes and covariates will be uniformly defined based on current literature (medical textbooks and guidelines issued by medical associations) and then all events of interest as well as drug codes will be mapped.

In order to benefit from previous collaborative work performed by some of the partners, information regarding mapping of some events will be obtained from other EC-funded drug safety projects such as ARITMO (Arrhythmogenic potential of drugs), SOS (The Safety of non-steroidal anti-inflammatory drugs) and EU-ADR (Exploring and understanding Adverse Drug Reactions by integrative mining of clinical records and biomedical knowledge).

The databases are described below:

3.2.1. IPCI

Database description

In 1992 the Integrated Primary Care Information Project (IPCI) 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 1.2 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.

3.2.2. PHARMO Database

Database description

The PHARMO medical record linkage system is a population-based patient-centric data tracking system that includes high quality and complete information of patient demographics, drug dispensings, hospital morbidity, clinical laboratory, and date of death of 3.2 million community-dwelling inhabitants of 65 municipal areas in the Netherlands. The drug dispensings originate from out-patient-pharmacies. This core dispensing database is linked on a patient level with different databases, among which the hospital morbidity data "The Dutch National Medical Register (LMR)". This register comprises all hospital admissions in the Netherlands, i.e., admissions for more than 24 hrs and admissions for less than 24 hours for which a bed is required. Only hospital admissions for the out-patient-pharmacy patients are collected in the PHARMO database. Clinical laboratory tests are available for a subset of the out-patient-pharmacy patients in a completely computerized format. Dates of death are available from the Central Bureau of Genealogy (CBG). The CBG is the Dutch information and documentation centre for genealogy, family history and related sciences. Data are collected since October 1994 and include mortality. The CBG returns date of death for the out-patient-pharmacy patients. The linkage method used for individuals of the separate databases is probabilistic.

Database updates and data time lag

The linked databases in the PHARMO database network are updated every year. Databases are linked when the hospital admission data of the preceding calendar year become available; the updated database becomes available in the second half of the year. In between the outpatient pharmacy data is updated every month. Dates of death returned from the CBG have a lag time of 2 years.

The PHARMO database network covers the period 1998-2010.

Data subsets and variables

The PHARMO databases contain the following information:

- **Socio-demographic data:**
  Unique anonymous person identification number
  Gender
  Birth year
  Last known ZIP-code
  Date first contact
  Date last contact
  Reason last contact
  Date of death

- **Outpatient dispensing drug data:**
  Unique person identification number
  Unique pharmacy identification number
  Type prescriber (GP, specialist)
  ATC
  Molecule name
  Dispensed quantity (number of units)
  Type of unit (fluid, tablets etc.)
  Dispensation date
  DDD (number of DDD in one unit)
  Duration of dispensing
  Number of units to take each day (free text in Dutch)
  Strength of one unit

- **Hospital data:**
  Unique person identification number
  Unique hospital identification number
  Main diagnoses are coded in ICD9-CM
  Main diagnostic/surgical procedure
  Side diagnoses
  Dates of hospital admission and discharge
  Type of care (day/clinical)

Limitations of the database

- Date of first entry, last entry in the population might be subject to misclassification.
- Linkage is highly sensitive and specific but does not exclude a small percentage of linkages as misclassified

- CBG data have a lag time of 2 years compared to 1 year or less for the other sources of data. Clinical lab tests are only available for a subset of the PHARMO database

3.2.3. Health Search Database/CSD Longitudinal Patient (HSD)

Database description

The Health Search/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, 700 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.

3.2.4. Regional Database Puglia

Database description

The regional databases of Puglia include hospital discharge records, prescription databases, and the civil registry, for the period 2002-2009. Shortly, data on 2010 will be available. In addition the archive of physicians for 2005 is available.

Data subsets and variables

Prescription databases (last update: January 2009)

Prescription databases provide data on the community prescriptions reimbursed by the NHS with information on type and quantity of dispensed drug (generic and brand names) and dispensing date with drugs coded according to ATC classification system. This database provides the following information for each reimbursed prescription:

- Number of prescription
- Date of prescription
- Date of dispensing
- Identification number of the dispensed product (MINSAN)
- ATC code
- Number of packages
- Cost of prescription
- Pharmacy code
- City code
- First name and surname of the patient
- Date of birth of the patient
- Gender of the patient
- Fiscal or sanitary code (Tax code [alphanumerical code for personal identification, SSN cultural equivalent])
- Prescriber identification number

Hospital discharge records (last update: January 2009)

Hospital discharge records include information on primary diagnoses and up to five co-existing conditions, performed procedures (diagnostic and therapeutic interventions), date of hospital admission and discharge, and in-hospital death. All diagnoses are coded according to the ICD-9 CM.

In particular, this database provides the following information for each hospital admission:

- Hospital
- Hospital code
- Hospital discharge record
- Family name
- First name
- Date of birth
- Gender
- Place of birth (code)
- Citizenship
- Residence Region (code)
- Residence Local Health Authority (LHA code)
- Residence city (code)
- Fiscal or sanitary code (Tax code (alphanumerical code for personal identification, SSN cultural equivalent))
- Marital Status (Single; married; legally separated; divorced; widowed; not declared)
- School education (Primary/None; Middle school; Secondary school; University degree)
- Type of admission (Ordinary Hospitalization; Day hospital)
- Type of ordinary hospitalization (Scheduled not urgent; Urgent; Involuntary psychiatric treatment; Scheduled with pre-hospitalization)
- Birth weight

- Reason for day hospital (Diagnostic procedure; Day Surgery; Therapeutic procedure; Rehabilitation)
- Number of days in day-hospital
- Admission date
- Admission Unit
- Origin of the patient (GP; direct access; other region; other private institute; other public institute…)
- Traumatisms or poisoning (Industrial accident; domestic injury; car accident; suffer acts of violence; self-injury; other)
- Discharge date
- Discharge Unit
- Discharge modality (Death; ordinary; voluntary; transfer to other structure…)
- Transfer Unit and date
- ICD-9 underlying (main) discharge diagnosis code
- Other ICD-9 underlying discharge diagnosis codes
- Principal Procedure Code and Date
- Other Procedure Codes and Dates
- DRG

Civil registry (last update: January 2011)
Population registry with patients' demographics information, as gender, date of birth, fiscal code, as subjects' identifiers of all patients of LHA.

Limitations:

- No information on: Race/Ethnicity; Laboratory values; Dosing regimen; Symptoms
- It is not possible to distinguish between type 1 and type 2 diabetes mellitus
- It is not possible to assess the burden of diabetes in terms of mortality, since regional death registries are not uniformly available in all areas and are updated with a substantial delay.

3.2.5. Regional Database Lombardy (SISR)

Database description
In the Sistema Informativo Sanitario Regionale (SISR) database, data are obtained from the electronic healthcare databases of the Lombardia region. Lombardia is the largest Italian region with about nine million inhabitants, about 16% of the population of Italy. This population is entirely covered by a system of electronically linkable databases containing information on health services reimbursable by the National Health Service, including hospital admission and outpatient prescriptions of drugs free of charge.
The SISR database has a full population coverage (i.e. the population covered is not selected by any criteria) and the available information is related to drug prescriptions and to hospital admissions for the period 2000 -2010.

Database updates and data time lag
The SISR database is updated yearly. However, the possibility of access to new data is not guaranteed.

Data subsets and variables
The SISR database contains the following information:

Patient register:
- Patient ID: unique person identification number used for record linkage
- Sex
- Birth date
- Date of transferring out: the date on which a person leaves the database
- Cause of transferring out : the cause of exit from the database (death or migration)

Prescription**:** Contains all outpatients prescriptions of drugs reimbursable by the NHS
- Patient ID: unique person identification number used for record linkage
- ATC code of the drug
- AIC (Marketing Authorization): Unique code, released by AIFA (Italian Drug Agency), used to identify each box of each drug in commerce
- Prescription Date
- Quantity: number of prescribed boxes
- Using the AIC code it is possible to link drug prescriptions to a drug register which contains information on the commercial name of the drug, the quantity of active principle of the drug contained in one box, defined daily doses (DDDs) of the active principle, and the estimated coverage of one box.

Hospitalization: Contains all hospitalisations occurring in the public and private hospitals in Lombardy
- Patient ID: unique person identification number used for record linkage
- ICD-9-CM codes for diagnoses: there are 6 fields (one for the main diagnosis and 5 for the secondary diagnoses) containing ICD-9 codes
- Diagnostic procedures/surgery code: there are 6 fields (each field corresponds to a different procedure)
- Hospitalization Date: date of hospital admission
- Discharge Date: date of discharge from the hospital
- Procedures Date: there are 6 fields containing the date of the associated procedures

Limitations of the database
- The DB does not contain information on over-the-counter (OTC) medication.
- The DB does not contain information on outpatient care.
- The DB does not contain information on anthropometric measures and lifestyle (e.g. weight, being a smoker).
- DDDs and the amount of drug prescribed are available in the database, but there is no information on the prescribed dose.
- It is not possible to inspect hospital medical charts for validation through the regional database.

3.2.6. CPRD

The Clinical Practice Research Datalink (CPRD) contains anonymised data from general practice for between 4 and 7% (depending on calendar year) of the UK population. The calendar period covered is 1987 – current although data from before ~2000 can only be reliably used for purposes of incidence & prevalence calculation if they are combined with the locked CPRD data (data lock April 2002). This is a

consequence of the changeover of the GP practice software from a DOS-based (VM6) to a Windows-based (Vision) program and the opportunity at the time to erase historic data for patients no longer registered with the practice. As a result, the numerator & denominator data for the 1990s are unreliable for the calculation of incidence and prevalence rates.

Data are updated approximately bimonthly although this varies by practice. There are no data subsets as such. All anonymised data recorded by the GP deemed relevant for clinical management of the patient are available to researchers and comprises of demographic data including the dates the patient was registered with the practice, diagnoses and symptoms, prescriptions, hospital referral and discharge data as well as some information on major procedures carried out in hospital. For outpatient visits, hospital consultants are not allowed to issue prescriptions and hence an important proportion of prescribing initiated in hospital is captured in the CPRD (because hospital consultants will ask the GPs to prescribe the medication needed). Covariate information is available on body mass index, height, weight, alcohol intake, smoking status, age, sex, socio-economic status, and comorbidity. No or extremely limited information is available on ethnicity, diet, physical activity level.

Limitations of the database include incompleteness of information on the indication for prescribing, on lab test results, and family history, as well as virtually complete lack of data on ethnicity, diet, physical activity levels, environmental and occupational exposures, medicines received and procedures undergone in hospital, and non-compliance. In addition, for data confidentiality reasons researchers do not have routine access to the free text fields; this is available at an additional cost of £0.05 per word plus £1000 administrative fee.

3.2.7. BIFAP (AEMPS)

Database description

BIFAP (Base de Datos para la Investigación Farmacoepidemiologica en Atención Primaria) database is a longitudinal population-based database of anonymzed computer based medical records of general practitioners (GPs) throughout Spain (Salvador-Rosa A, 2002). BIFAP is a non-profit research project, kept by the Spanish Medicines Agency (AEMPS), a public agency belonging to the Spanish Department of Health. The project started in 2003, including anonymized information from 2001 onwards, and the database covers data from approximately 1,260 GPs from 9 different autonomous communities in Spain. From those, 1,045 are GPs and 215 pediatricians. The database captures data on 3,948,464 patients corresponding to 17,735,987 person-years of follow-up.

In the Spanish health care system, patients are registered with a single GP who acts as a gatekeeper for and receiver of information from primary and secondary care.

The dataset is comparable with the Spanish population with respect to its age and sex distribution. Downloads are made periodically and the information is sent to the gatekeeper who de-identifies all information before further access is provided.

The GPs' electronic medical records contain coded and anonymous data on patient demographics, prescription details, clinical events, specialist referrals, laboratory test results. The International Classification of Primary Care (ICPC-1) is the coding system for patient complaints and diagnoses, although this information can also be entered as free text[43]. Prescription data information in BIFAP includes product name, quantity dispensed, dosage regimens, strength and indication[44]. Prescriptions are coded according to the Anatomical Therapeutic Chemical (ATC) classification scheme recommended

by the WHO[45]. The system complies with European Union guidelines on the use of medical research and has been proven valid for pharmaco-epidemiological research[46].

Database updates and data time lag

The out-patient-pharmacy database is updated every year with a time lag of approximately 3 months and covers the period 2001-2009.

From 2012 (for SAFEGUARD), BIFAP database will contribute with data up to the end of 2011.

Data subsets and variables

The BIFAP database contains the following information:

- **Socio-demographic data:**
  Unique anonymous person identification number
  Date of birth (dd/mm/yyyy)
  Gender
  Geographic Region (include information of GPs from 9 out of 17 Regions in Spain)
  Weight
  Height
  Smoking status
  Prescribing physician location/practice
  Prescribing physician code
  Registration status
  Start data in database system
  End date in database system
  Reason for end date recorded
  Date of death; cause of death not registered
  Start date with a specific practice
  Transfer/end date with a specific practice
  Start date for practice in system
  End date for practice in system

- **Outpatient dispensing drug data:**
  Unique anonymous person identification number
  Prescriber (GP, paediatrician)
  Drug prescribed (unique product name)
  Drug coding system (ATC codes)
  Therapeutic class
  Prescription date
  Prescribed dosage and quantity (number of units)
  Type of unit (fluid, tablets etc.)
  Strength of one unit
  Duration of prescription

Number of units to take each day

- **Outpatient visits (GP):**
  Unique anonymous person identification number
  Diagnosis coded using (ICPC-2); unlimited number of diagnosis allowed
  Free text comments attached to the diagnosis.
  Referral to specialist and reason
  Tests ordered by GPs
  Name and date of tests ordered

Limitations of the database

- The dispensing of drugs is based on the prescriptions registered by the GPs, dispensings without prescription or prescriptions of other physicians are not included. In addition, it is not possible to know if a patient did actually take a dispensed drug or not.
- There is limited information available on laboratory test results, on diagnostic test results such as X-ray, MRI, etc., on results of referral visits, on hospital admission and discharge diagnosis, and on cause of death. This information is only available if the GPs include it in the electronic medical records, either in a structured way or as free text.

Information in the BIFAP database is anonimized and no personal identification is included. Consequently, in BIFAP it is not possible to access detailed information stored in other levels of primary and secondary care, and besides this, BIFAP database can not be linked to other data sources (such as hospitalary or mortality registries, etc).

3.2.8. GePaRD

Database description

The German Pharmacological Research Database (GePaRD) consists of claims data from four German statutory health insurance (SHI) providers. It covers about 14 million insurants throughout Germany who have at any time since 2004 been enrolled in one of the four SHIs. The database population represents approximately 17% of the German population of about 82 million inhabitants.

Membership in an SHI is compulsory in Germany for employees with an annual income up to approximately 47.000 €. Subjects with higher incomes can choose private health insurance providers instead of an SHI and are probably underrepresented in SHIs. However, some of these higher-income subjects are voluntary members of SHIs, most often because SHIs provide free health insurance for unemployed family members (children and spouse) whereas in private health insurance plans all family members have to be paid for. About 70 million people (85% of the German population) are SHI members, including children and insurants who are retired or unemployed and about five million voluntary members.

Three of the four SHIs contributing to the database are so called ‘Ersatzkassen’ which are more likely to insure people of middle to higher socio-economic status. The database also includes data from one ‘Allgemeine Ortskrankenkasse’, an SHI which has traditionally insurants of lower socio-economic status. Two large SHIs contributing to the database together insure more than 13 million subjects all over

Germany. We therefore expect the data to be adequately representative with respect to age, sex, and region of residence.

Since German SHIs pay the costs for ambulatory physician visits, hospital stays and prescription drugs for their enrolled members, information on these health services are contained in the database.

An advantage of data from German SHIs is the stability of their membership which makes long term follow-up studies feasible. In the BIPS database membership is stable in about 75% of all subjects from 2004 to 2006. However, insurants leaving a specific SHI and entering one of the other three participating SHIs cannot be identified as the same individual (synonym error).

Database updates and data time lag

At the moment the database includes about 14.3 million subjects covering the years 2004 until 2009. Usually, the database is updated annually and data from the most recent year should be available in the late autumn of the following year. After data delivery another three months for in house preparation and validation are needed before updates of the database are finalised.

Data subsets and variables

The GePaRD contains the following information:

- **Socio-demographic data:**
  Unique person identification number: allows longitudinal analysis and linkage between the subsets
  Family identification number: identifies members of a family who are insured together
  Year of birth
  Sex
  Region of residence
  Nationality (German/other)
  Indicators for social status
  Dates of insurance coverage (entry and exit)
  Reasons for end of coverage (including death)

- **Hospital data:**
  Unique person identification number
  Unique hospital identification number
  Hospital diagnoses are coded in ICD-10-GM (at least 4 digits). Diagnosis at admission, main diagnosis at discharge, and a variable number of accessory diagnoses are available
  Dates of hospital admission and discharge
  Reason for admission
  Reason for discharge (including death)
  Diagnostic and surgical procedures (OPS Codes)

- **Outpatient prescription drug data:**
  Unique person identification number

Unique pharmacy identification number
Unique physician identification number: allows identification of speciality of prescribing physician
PZN (Pharmazentralnummer): a pharmaceutical reference identification number
Prescribed quantity (number of packages)
Prescription date
Dispensation date

A central pharmaceutical reference database with all PZN on the German market has been built up by BIPS. It contains information on generic name, brand, manufacturer, packaging size, strength, defined daily dose (DDD), pharmaceutical formulation, and ATC code. Information from the central pharmaceutical reference database is linked to the SHI database via the PZN.

- **Outpatient medical treatment data:**
  Unique person identification number
  Unique physician identification number: allows identification of specialty of consulted physician
  Ambulatory diagnoses are coded in ICD-10-GM (at least 4 digits). These diagnoses are not linked to a definite date, but refer to a quarter, as physicians' claims are collected quarterly.
  Diagnostic certainty: coded as certain, suspected, excluded, status post
  Dates of treatment / visits
  Types of treatment / diagnostic procedures with exact date (EBM codes, developed for payment of physicians for the outpatient treatment of German SHI patients)

Limitations of the database

- Exact date of birth is not known, only birth year available.
- Database contains no information on hospital or OTC medication.
- Only prescribed quantity, not prescribed dose available for medication data.
- Exact date for outpatient diagnoses is not known, only quarter available, however ambulatory diagnostic or therapeutic procedures (EBM codes) come with exact date.
- No laboratory values are contained in the database, but ordering of lab values is contained with exact date.
- The diagnostic certainty is missing for some ambulatory diagnoses, mostly in 2004.
- No information on diagnoses, treatments, and prescriptions for occupational accidents and during rehabilitation is available as they are insured by a different carrier.

3.2.9. Medicare

General description

The Caremark-Medicare linked dataset provides healthcare transaction data on community-dwelling patients 65 years and older who receive their health insurance through Medicare and have prescription drug coverage through Caremark (a pharmacy benefits management company). The Caremark portion of the dataset includes pharmacy claims for medications dispensed to people who have

this form of coverage, while the Medicare portion includes data on inpatient and outpatient services (Medicare Part A and B), as well as enrollment and demographic data. For research purposes, we link the data between these data sources using multiple identifiers. The date range is from 2005 through 2008.

Type of database (GP/claims)

Administrative claims from pharmacies, inpatient facilities, and outpatient offices/clinics/ancillary medical services.

Database updates and data time lag

The Caremark data is updated frequently with an approximate 2-day delay from the date of prescription dispensing until its incorporation in the data source. The Medicare data is updated less frequently with a lag between service and availability in the data source for medical services of between 1-1.5 years.

To be usable for research at BWH, there is a lag of 1.5 – 2 years to obtain the most recent data and conduct the linkage between Medicare and Caremark. Thus, the year 2009 data is anticipated to be available for research in mid-2012.

Data subsets and variables

- Prescriptions: drug name (brand and generic entity), dose, formulation, days supply, number of units dispensed, dispensing date
- Inpatient and outpatient services: medical procedures (ICD-9 procedure codes), outpatient procedures (CPT-4 codes), inpatient and outpatient diagnoses (ICD-9 diagnosis codes), acute inpatient hospitalizations, emergency room visits, skilled nursing facility stays, hospice data, durable medical equipment
- Demographics, including race and ethnicity, as well as vital status
- Payments: reimbursement for inpatient and nursing home, copayments for prescriptions

Limitations of the database

Variables such as vital signs, lab test results, body mass index, and smoking are not captured; there is no cause of death recorded in the data source; diagnosis code correspond to the reason for the service and not the findings of the service; diagnosis codes could be misapplied

An overview of the databases is presented in the Annex II. “Characteristics of the healthcare databases”.

3.3. **Study period**

The study period will encompass from January 1st, 1998 (the earliest available data) to the last data drawn down in each DB (2010-2011). The time period of data availability for each database is shown in table 1.

Table 1. Time period for of data availability of healthcare databases

	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
IPCI*														
PHARMO														
HSD														

R. LOMBARDY														
R. PUGLIA														
GePaRD														
CPRD														
BIFAP														
MEDICARE														

* Due to a system conversion, the new version "IPCI II" is available from 2006. All previous data can be used for historic data.

3.4. **Study population and eligibility periods**

The study population will include all cohort members as defined below and with at least 365 days of enrolment in the respective database.

3.4.1. Cohort Definition

3.4.1.1 Inclusion Criteria

Cohort members have to fulfill all of the following eligibility criteria:

- At least 365 days of continuous enrolment in the database before cohort entry.
- At least one prescription/dispensing of a NIBGLD (ATC: A10B – Annex 1) or insulin or insulin analogues (ATC: A10A – Annex 1) in the study period. In order to be sure that only subjects with T2DM are included in the cohort, subjects that are included with only prescription/dispensing of insulin or insulin analogs, they should have a prior diagonosis of T2DM or a prescription of NIBGLD in the 365 days period prior to the cohot entry.
- No a prior prescription/dispensing of a NIBGLD or insulin or insulin analog in the 365 days before cohort entry.
- No previous malignant cancer at any time before study entry

3.4.1.2 Cohort entry

Cohort entry is defined as the date of the first study drug prescription/dispensing during the study period that fulfils the cohort entry criteria.

3.4.1.3 Cohort Exit

Cohort exit is defined as the earliest of the following dates:

- Last data drawn down (database-specific but generally 2010-2011).
- Occurrence of the study outcome (specific for each outcome of interest)

- Transfer out of database / end of registration / end of membership / institutionalization (depending on type of database).
- Death.

3.5. **Events of interest**

The events of interest that will be evaluated and their clinical definition are the following:

Myocardial Infarction (MI)

Cell death of cardiac myocytes caused by ischemia, which is the result of a perfusion imbalance between supply and demand.[47].

Heart Failure (HF)

Complex clinical syndrome characterized by systemic perfusion inadequate to meet the body's metabolic demands as a result of impaired cardiac pump function. It is characterized by specific symptoms (dyspnea and fatigue) in the medical history and signs (edema, rales) on the physical examination. [48]

Ventricular arrhythmia (VA) and Sudden Cardiac Death (SCD) [49]

- Ventricular arrhythmia:
    - Ventricular arrhythmia includes both ventricular tachycardias (VT) and ventricular fibrillation or flutter (VF):
        - Ventricular Tachycardia: cardiac arrhythmia of three or more consecutive complexes in duration emanating from the ventricles at a rate of greater than 100 bpm (cycle length less than 600 ms).
            - Non-sustained VT: Three or more beats in duration, terminating spontaneously in less than 30 s.
            - Sustained VT VT greater than 30 s in duration and/or requiring termination due to hemodynamic compromise in less than 30 s .
        - Ventricular fibrillation: Rapid, usually more than 300 bpm/200 ms (cycle length 180 ms or less), grossly irregular ventricular rhythm with marked variability in QRS cycle length, morphology, and amplitude.
        - Ventricular flutter: A regular (cycle length variability 30 ms or less) ventricular arrhythmia approximately 300 bpm (cycle length—200 ms) with a monomorphic appearance; no isoelectric interval between successive QRS complexes.
        - Torsades de pointes: VT associated with a long QT or QTc, and electrocardiographically characterized by twisting of the peaks of the QRS complexes around the isoelectric line during the arrhythmia: a) “Typical,”

initiated following "short-long-short" coupling intervals. b) Short coupled variant initiated by normal-short coupling.

- Cardiac arrest: An abrupt loss of effective blood flow, sufficient to cause immediate loss of consciousness, leading immediately to death if untreated. The most common electrical mechanisms for cardiac arrest are VF and pulseless VT, but substantial numbers of cardiac arrests begin as severe bradyarrhythmias, asystole, or pulseless electrical activity.

- Sudden cardiac death
  - Sudden cardiac death (SCD): Death from an unexpected circulatory arrest, usually due to a cardiac arrhythmia occurring within an hour of the onset of symptoms.
- Sudden cardiac arrest: Death from an unexpected circulatory arrest, usually due to a cardiac arrhythmia occurring within an hour of the onset of symptoms.

Hemorrhagic stroke (HS)

A disease of abrupt onset with neurologic damage due to hemorrhage into brain tissue (parenchymatous hemorrhage) or hemorrhage into the spaces surrounding the brain, most frequently the subarachnoid space. Subdural and epidural hemorrhage will not be considered in the definition. [50]

Ischemic stroke (IS)

Infarction of central nervous system tissue, these may be either symptomatic or silent. Symptomatic ischemic strokes are manifest by clinical signs of focal or global cerebral, spinal, or retinal dysfunction caused by central nervous system infarction. A silent stroke is a documented central nervous system infarction that was asymptomatic.[51]

Acute Pancreatitis (AP)

An inflammatory disease of the pancreas, characterized by a discrete episode of abdominal pain and elevated serum amylase and lipase levels. Distinguished from chronic pancreatitis by complete restitution of the pancreas both morphologically and functionally after the derangements that precipitated the attack have been corrected.[52]

Pancreatic Cancer (PC)

A disease in which malignant (cancer) cells are found in the tissues of the pancreas.[53]

Bladder Cancer (BC)

Cancer that forms in tissues of the bladder. Most bladder cancers are transitional cell carcinomas (cancer that begins in cells that normally make up the inner lining of the bladder). Other types include squamous cell carcinoma (cancer that begins in thin, flat cells) and adenocarcinoma (cancer that begins in cells that make and release mucus and other fluids). The cells that form squamous cell carcinoma and adenocarcinoma develop in the inner lining of the bladder as a result of chronic irritation and inflammation.[53]

Total Mortality (TM)

Death from all causes.

3.5.1. Event identification

Events will be identified on the basis of diagnosis codes/text in the claims or medical records. In claims records we will only consider the primary hospital discharge diagnosis code, except for cancer that can be identified also on the basis of secondary diagnoses codes. In Medical records diagnoses codes and text may be utilized.based on the characteristics of each database, cases might be operationally identified by algorithms including codes, text files, diagnosis, procedures, medications, etc. Required codes and free text phrases will be retrieved from mapping of concepts according to the Unified Medical Language System (UMLS), according to a method that was defined and described for the EU-ADR project [54].
For TM, all deaths recorded in each database during the study period will be identified.

Details of the codes and algorithms that will be used in each database for the identification of the events of interest will be provided in the deliverable 4.2 "Report on Terminology Mapping"

3.5.2. Case validation

Validation of events will be done in selected databases where review of clinical charts and/or letter is possible i.e., IPCI, CPRD, BIFAP, SIMG, PHARMO.

Common questionnaires will be developed to gather the information from the chart reviews and specialist letters from the selected databases. From each of these databases, a random sample of 50 cases per outcome will be selected. Cases will be classified as "definite case", "Non case" and "Non assessable case", then, the positive predictive value (PPV) will be calculated.

This process will provide an insight regarding the extent of misclassification and will allow for analytic approaches to address it (regression calibration).

3.5.3. Case assignment

For each of the following cardiovascular outcomes, MI, VA/SCD, HS, and IS, the first event occurring during the follow up after cohort entry will be identified.

Cases that had an event prior to cohort entry they will be classified as having a prior history of that specific event.

For HF, AP, PC, BC and TM will consider as cases only the persons experiencing this event for the first time after cohort entry (first ever event). For each of these specific study outcomes, subjects will not be eligible for cohort entry if they have this event in the year prior to cohort entry. This is to avoid

confounding by contra-indication (HF), misclassification due to chronic pancreatitis (AP) and nature of the event (PC, BC).

The index date will be defined, for the cases, as the the date of the first occurrence of an event of interest during the study period after the cohort entry.

3.6. **Selection of Controls**

For the nested case-control, studies we will randomly select up to 10 controls per each case from the study cohort using risk set sampling. If necessary, for specific subgroup analysis, the number of controls can increase.

Controls will be matched to a case by:

- Database
- Sex
- Year of birth ± 1 (The age of the case will determine the category for the analysis)
- Year of cohort entry

For controls, the index date will be the defined as the date that results in the same time of follow-up as for the respective case.

3.7. **Exposure**

Individual drugs from each of the NIBGLD classes, insulins and insulin analogs, will be considered for evaluation in this study and are in the following list (Annex I):

- Biguanides
- Sulfonylureas
- Thiazolidinediones
- Meglitinides
- ∝-glucosidase inhibitors
- GLP-1 analogs
- DPP-4 inhibitors
- Amilyn analog
- Insulin and insulin analogs
- Others (Benfluorex)

A new user will be defined as a subject that does not have a NIBGLD or insulin or insulin analogue prescription/dispensing of the same class in the previous 365 days.

3.7.1. Timing of exposure

The effect of timing of exposure (hazard function)[55] will be assessed in different ways to evaluate the effect of the drugs in terms of cumulative dose and duration, and the shape of the hazard function during use and after stopping use.

To allow for flexible modelling of the hazard function we will use short duration categories which can be aggregated if the risk is homogeneous in those categories.

Case control

Timing of exposure will be assessed with reference to the index date for the case control study and categorized as follows:

- Current use: if the drug prescription duration covered the index date or ended at most 30 days before (i.e. carry-over effect).
- Recent use: most recent prescription for study drug ended between 31 and 62 days before index date
- Past use: most recent prescription for study drug ended between 63 and -90 days before index date.
- Distant past: most recent prescription for study drug ended more than 90 days before
- No use: no prescription for this drug prior to the index date

In addition, to study the chronic effect of drugs of interest, the proportion of days covered (PDC) i.e., the ratio between cumulative days of use and total number of days of use, can be used to study the effect of these drugs. Duration can be also used in alternative to this index.

These definitions will be refined based on the results of the drug utilization study.

If analyses are conducted within class, the drug with most recent exposure will be selected over others that are more distant from the index date.

The chronic an progressive nature of T2DM leads to dynamic treatment patterns, with treatment usually started as monotherapy but later 2 or more drugs with different mechanism of action are combianed in order to reach and maintain the glycemic goals. The effect of diabetes therapy regimen will be assessed in current users based on the number of drugs for T2DM (AnnexI) in the following categories:

- Monotherapy of NIBGLD: Only one NIBGLD
- Dual therapy without insulin: two NIBGLD
- Three or more NIBGLD without insulin: three or more NIBGLD
- Insulin therapy: Only insulins or insulin analogues.
- Combined therapy including insulin: Regimen of 2 or more drugs including NIBGLD and insulin or insulin analogs.

A subject could be treated with one or more drugs belonging to different classes or change from one therapy to another. In order to identify these changes, a discontinuation, concomitance and switching will be defined as follows[56]:

- A discontinuation of drug A will be defined when no new prescription/dispensing of A is registered in a period of 1.5 times the length of the last prescription of A after the end of the last prescription/dispensing of A.
- Concomitance of drug A and B or addition of B to A, will be defined as starting drug B while obtaining a refill for drug A, within a window of 2.5 times the length of the last prescription of A from the start date of the prescription of A.
- Switching from drug A to drug B is defined as starting B during or after the last prescription of A while not obtaining a refill for drug A within 2.5 times the length of the last prescription/dispensing of drug A.

Cohort
For the dynamic cohort study, each day of follow-up after cohort entry for each cohort member will be classified according to type of exposure to each drug of interest, using a similar classification as defined above. Exposure classification for the type of drug will be mutually exclusive i.e., days will not be double counted if multiple drugs are used but rather categorized by the different drugs at that day, current use will overrule more distant use.

3.7.2. Duration of a prescription

In order to assess timing and duration effects we need to use information on the duration of prescriptions. The defined daily dose (DDD) will be used to estimate the duration as (strength*package size)/DDD.

We will estimate the effect of duration within current users. Duration of use of the same drug or drugs within the same class will be categorized as:

- < 1 week
- 8-14 days
- 15-30 days
- 31—60 days
- 61-90 days
- 91-120 days
- 121-150 days
- 151-180 days
- 181-210 days
- 211-240 days
- 241-270 days

- 271-300 days
- 301-330 days
- 331-365 days
- 366– 730 days
- 731 – 1095 days
- > 1096 days

If necessary, categories will be aggregated.

3.7.3. Dose

The prescribed daily dose (PDD) will be used if recorded in the database (i.e., IPCI, PHARMO, THIN and BIFAP).

For all other databases the average daily dose of consecutive prescriptions will be estimated by strength, package size, and duration of prescription interval (i.e., time between the penultimate and ultimate prescription before index date) as (strength x package size) / duration of prescription interval. If prescriptions are not consecutive or in case of single prescriptions, the defined daily dose (DDD) will be used.

Dose will be categorized as follows in current users of the study drug:

- low: < 0.5 DDD
- normal: 0.5 – 1.0 DDD
- high: 1.0 – 1.5 DDD
- very high: ≥ 1.5 DDD

Classes of dosage are closed on the left; reference for dose comparisons is normal dosage.

3.8. **Covariates**

Demographic and lifestyle information, co-morbidities, and drug use as listed below will be considered as potential confounders/ risk factors. For each analysis, the specific list of confounders will be specified in the statistical analysis plan (SAP).

As databases contain different types of information and level of detail, individual strategies will be applied to gather the best information possible for each database. Separate variables will be used to distinguish ascertainment of a condition via diagnosis codes, drug use as proxy, procedure codes, or laboratory result.

Confounders will be assessed at cohort entry or in a prespecified window at index date (this window will vary from one to another covariate and will be specified in the SAP); baseline assessment period is defined as the 12 months time-period before cohort entry unless otherwise specified.

Table 2: Covariates that will be considered as confounders/risk factors for different events

COVARIATES	MI	HF	IS	HS	VA /SCD	AP	PC	BC	TM
Demographics/lifestyle									
Age [a]	x	x	x	x	x	x	x	x	x
Sex [a]	x	x	x	x	x	x	x	x	x
Race/ethnicity [a]	x	x	x	x	x	x	x	x	x
Country of origin /DB [a]	x	x	x	x	x	x	x	x	x
Smoking status [a]	x	x	x	x	x	x	x	x	x
Drug abuse [a]	x			x	x				
Obesity/BM [a] I	x	x	x	x	x	x	x	x	x
Weight loss [b]						x			
Alcohol abuse/dependence/alcohol intake [b]	x	x	x	x	x	x	x	x	x

[a] Assessed at cohort entry [b] Assessed at index date

COVARIATES	MI	HF	IS	HS	VA /SCD	AP	PC	BC	TM
Health care utilization [a]									
Number of physician visits in year prior [a]	x	x	x	x	x	x	x	x	x
Number of different drugs utilized in year prior (ATC-7 level) [a]	x	x	x	x	x	x	x	x	x
Number of Hospitalizations in year prior [a]	x	x	x	x	x	x	x	x	x

[a] Assessed at cohort entry

COVARIATES	MI	HF	IS	HS	VA /SCD	AP	PC	BC	TM
DM related co-variates									
Year of cohort entry [a]	x	x	x	x	x	x	x	x	x
Duration of T2DM (from 1st Dx) [b]	x	x	x	x	x	x	x	x	x
Hypoglycemic events [b]	x		x		x				
HbA1c levels [c]	x	x	x	x	x	x	x	x	x

[a] Assessed at cohort entry; [b]Assessed at index date; [c] Assessed in the period 2 month prior to cohort entry till one month after cohort entry

COVARIATES	MI	HF	IS	HS	VA /SCD	AP	PC	BC	TM
Co-morbidity									
Myocardial Infarction (MI) [a]	x	x	x		x				x
Cardiac conduction disorders (other than ventricular arrhythmia /AF) [a]					x				x
Atrial fibrilation / flutter [a]	x	x	x		x				
Ventricular arrhythmia [a]					x				x

Pericardial diseases [a]					x				
Pulmonary hypertension [a]				x	x				
Cardiomyopathies [a]	x	x	x		x				x
Genetic arrhythmia syndromes (Long QT syndrome, Short QT syndrome,Brugada syndrome,Catecholaminergic VT) [a]					x				x
Congenital heart disease [a]	x	x			x				x
Valve disorders [a]	x	x	x		x				x
Ischaemic heart disease/ coronary heart disease [a]	x	x	x		x				x
Heart failure [a]	x		x	x	x				x
Peripheral artery disease [a]	x	x	x	x	x				x
Hypertension [a]	x	x	x	x	x				x
Thrombosis/embolism [a]	x		x	x					x

COVARIATES	**MI**	**HF**	**IS**	**HS**	**VA /SCD**	**AP**	**PC**	**BC**	**TM**
Coagulopaties [a]	x		x	x					x
Hypokalemia [b]					x				x
Hypomagnesemia [b]					x				
Hypercalcaemia [b]					x	x			x
Hyperlipidemia [a]	x		x	x		x			x
Chronic kidney disease [b]	x	x		x	x				x
Chronic liver disease [b]	x	x	x	x	x	x	x	x	x
Cancer (only malignant) [b]	x	x	x	x	x	x	x	x	x
Severe COPD [a]	x	x			x				x
Stroke [a]	x	x	x	x					x
TIA [a]			x	x					x
Cerebral aneurysm [a]				x					x
Endoscopic retrograde cholangiopancreatography (ERCP) [b]						x			
Gallstones [b]						x			
Kidney stones [a]								x	
Bladder stones [a]								x	
History of pancreatitis [a]						x	x		
Metabolic and inflammatory conditions (Myocarditis,Rheumatic diseases,Endocarditis,Sarcoidosis,Amylodosis,Fabry disease,Hemocromatosis, Endocrine disorders and diabetes,End-stage renal failure,Obesity, dieting and anorexia) [a]					x				
Recurrent urinary tract infection [a]								x	

history of HPV infection/genital warts [a]								x	
Paraplegia [a]								x	
Gastric ulcer [a]								x	
Pelvic radiation exposure [a]								x	
Head trauma [a]				x					
Hepatitis C							x		

[a] Assessed at cohort entry

COVARIATES	MI	HF	IS	HS	VA /SCD	AP	PC	BC	TM
Drugs									
Phenacetin use [b]								x	
Drugs that prolong QTc (AZCERT list) [b]					x				x
Drugs that cause pancreatitis [b]						x			
Use of anticoagulants (B01A) [a,b]	x	x	x	x	x				x
Use of aspirin and other antiplatelets (e.g. clopidogrel, ticlopidine, prasugrel) [a] [a,b]	x	x	x	x			x		x
Antiarrythmic drugs [b]	x				x				x
Number of cardiovascular drugs (ATC=C) [a]	x	x	x	x	x	x	x	x	x
lipid lowering drugs (C10) [a,b]	x	x	x	x	x	x	x	x	x
Glucocorticoids (H02AB) [a,b]	x	x	x		x	x			x
ACE inhibitors (C09A, C09B) [a,b]	x	x	x	x		x			x
AT II antagonists (C09C, C09D) [a,b]	x	x	x	x	x	x			x
Diuretics (C02C, C03A, C03B, C03C) [a,b]	x	x	x	x	x	x			x
Calcium antagonists (C08) [a,b]	x	x	x	x	x	x			x
Cardiac glycosides (C01A) [a,b]	x	x	x		x				x
Vasodilators (C01D) [a,b]	x	x	x		x				x
Respiratory drugs (R03) [a]	x	x							x
Opioids (N02A) [a,b]					x				x

[a] Assessed at cohort entry; [b]: assessed at index date

Not all covariates (e.g., regarding life style) are contained in all databases and not all variables contain the information in desired detail. Considering the fact that the type of information varies from database to database, specific strategies will be applied to gather the best and homogeneous information from each database. Mapping of disease concepts will be based on the UMLS, when available mappings will be obtained from the EU-ADR, SOS, and ARITMO projects. The protocol will be updated with the codes and strategies once this information is available.

4. Statistical analysis

4.1. Descriptive analyses

4.1.1. Harmonization

Incidence rates (IR) and direct standardized incidence rates (SIRs) of each outcome of interest in each database will be estimated at the population level for harmonization purposes. IR and SIRs will be calculated stratifying by age categories (categories of 5 years) and sex. Crude incidence rates with the 95% confidence interval (95% CI) will be calculated dividing the number of events by the total number of person-years (py). Standardization will be done using the WHO population as reference[57].

4.1.2. Rates during drug use

The incidence rate for each outcome of interest by each of the individual study drug use will be calculated using exposure time days (for current/recent/past/non-use) as denominator and the number of events during current, recent, past and non use as numerator. The crude and age adjusted incidence rates with the 95% Confidence Interval (CI) will be calculated dividing the number of events occurring during the exposure to each of the study drugs by the total number of person-years (py) of exposure using Poisson regression. In order to estimate the hazard function, the rates for current and recent/past users will be further split by duration of use and time since last use.

4.1.3. Co-variates

The study cohort will be described according the distribution of covariates. Potential channelling will be addressed in the drug utilization study.

4.2. Main analysis

4.2.1. Case control analysis

The main objective of the nested case control studies is to assess the risk of the outcomes of interest (MI, HF, VA/SCD, HS, IS, AP, PC and BC), associated with the use of NIBGLD and insulins and insulin analogues. For the estimation of the risk, cases of MI, HF, VA/SCD, HS, IS, AP, PC and BC will be compared with matched controls and adjusted for potential confounders. Conditional logistic regression will be used to calculate the unadjusted and adjusted odds ratios (ORs) with their 95% confidence intervals (CIs) with reference to another active compound which will be selected based on drug utilization studies.

All analyses will at first be performed for each database separately and the heterogeneity among databases will be examined through heterogeneity indexes (e.g., I^2).

4.2.2. Cohort analysis

For the cohort studies, hazard ratios and incidence rates (IR) as well as the relative risk for TM with their 95% CIs will be estimated using Cox-regression analysis. First drug exposure will be assigned as index drug and compared to a reference category that will be decided based on the results of the drug utilization study. Time varying analyses will be conducted for estimation of the effect of duration of treatment. Indeed cumulative exposure can be considered in the model as a time varying covariate instead of exposure status.

4.3. **Sub-analysis**

4.3.1. Examination of Age and Sex Effects

Because matching was done on database, sex, year of birth (±1) and year of cohort entry, age and sex effect cannot be examined. In the cohort study, to determine whether there is an interaction between age and sex with the use of NIBGLD and insulins and insulin analogues and the risk of the outcomes of interest, an interaction term between each study drug and age and sex can be included in the model.

4.3.2. Class effect

To determine whether increased risks of each outcome of interest are homogeneous within each drug class or whether there is heterogeneity within each class. Drugs will be grouped as mentioned in Section 3.7 "Exposure definition":

4.3.3. Dose effect

The risk associated with different daily and cumulative dosages of the same drug or drug group will be compared with the reference within each dose category for current users.

4.3.4. Duration effect

To determine the effect of the duration of treatment with the study drugs on the risk of each outcome, the risk associated with different cumulative durations of use of the same drug or drug group will be compared.
Sensitivity analyses will be conducted based on different methods of duration of use estimation.

4.3.5. Co-medication effect

Since most of the patients with T2DM frequently have other comorbities and risk factors such as lipid metabolism disorders, hypertension, hypercoagulability disorders among others, they use other types of drugs for these co morbidities (i.e., ARB, ACE inhibitors, statins, etc). The odds ratio (OR) of the co-medication as well as the OR of the interaction between co-medication and individual study drugs will be examined. Additionally, a stratified analysis will be performed if data allows for it.

4.3.6. Co-morbidities

Subjects with T2DM frequently have other comorbidities such as prior history of cardiovascular diseases. Therefore they could have a higher risk of a new episode of a cardiovascular event for example. The objective of this subanalysis is to determine the modifying effect specific comorbidities on the risk of the events of interest.

Specific topics: Confounding

Observational studies must inevitably address the potential for confounding to explain part or all of an observed association between an exposure and an outcome. A number of different methodologies have been developed to address confounding either based on the study design or analysis of an observational study, with the choice among the methods depending on aspects of the study and potential confounder, including the association to be analyzed, the design of the study, the associated model, features of the information contained in the data and the availability of information. In Annex III we summarize approaches to address confounding according to the stage of research on which they primarily focus with the main advantages and disadvantages (pros and cons).

Methods will be selected after conducting the drug utilization analysis and all details will be specified in the SAP associated to each analysis. Model for the Propensity Score (PS), quantiles, as well as other details of the analysis will be specified in the Statistical Analysis Plan (SAP) associated to the study protocol. Methods such as Inverse Probability of Treatment Weighted (IPTW) and Instrumental Variable (IV) techniques will be applied in another stage of the project. Of course, these methods can be combined where appropriate.

Propensity score (PS), Disease score risk (DSR), High dimensional propensity score
In performing a cohort study, stratification can be done by propensity score (PS) or disease risk score (DRS). In case control studies, PS and DRS can be used. In case control studies, stratification can be done using stratification scores [58] or DRSs that are analogous to the PS for cohort studies. Nested case control studies will be used to study the association between NIBGLD, insulin and insulin analogs and MI, HF, VA/SCD, IS, HS, AP and cancer. For all cause mortality, dynamic cohort studies will be performed. For these models, PS will be applied. Indeed, due to the large set of confounders, PS and DRS approaches are good methods to reduce the number of variables for adjustment. The use of PS has also a further advantage: the same PS used in the whole cohort can be used in subgroup studies, e.g., patients at high risk [59]. High dimensional PS will be explored as well for this type of study.

Multivariable adjustment
For all outcomes, classical multivariable adjustment approache will be applied [60] For cancer and mortality, dynamic cohort studies will be performed. Also for these models, multivariable adjustment will be applied. Selection of confounders will be specified in the SAP. Different methods will be considered e.g., change-in-estimate method [61] and backward selection approach.

Monte Carlo Sensitivity Analysis
Because of the different characteristics of the DBs considered in the analysis, information on some covariates may not be available in all DBs. Therefore, the set of covariates to use in both adjustment and PS approaches could be different among DBs. To deal with this heterogeneity in the information, two different approaches will be applied: the first one will consider a common set of covariates for which information is available for all DBs (i.e., a minimal covariate set) while the second one will consider a set of covariates accordingly with the specific information (and on more covariates) available in each DB. The comparison of the estimates obtained from the two approaches will help in deciding if the application of an external adjustment method as Monte Carlo Sensitivity Analysis is actually needed.

Instrumental variable
An instrumental variable analysis can be used to control for unmeasured confounding, using physicians prescribing preference as instrument[62]. Details will be specified in the SAP for this sub study.

4.4. **Sensitivity analysis**

Different sensitivity analyses will be performed to evaluate the robustness of the estimates with respect to: a) exclusion of subjects affected by the outcome of interest before cohort entry; b) inclusion of secondary diagnosis in the outcomes' definition; c) exclusion of subjects with less that one year of follow up.

5. Quality assurance, timelines and reporting

The studies will be conduced according to the Good pharmacoepidemiology Guidelines (GPP) issued by the International Society of Pharmacoepidemiology [46] following the Guide on Methodological Standards in Pharmacoepidemiology from The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP)[63].

5.1. **Validation of study outcomes:**

The validation is one of the activities conducted as part of the quality control. For this the positive predictive value (PPV) of the case identification process will be calculated for all outcomes. This activity will be carried out only in DBs where additional information or specialist notes can be obtained. Charts and letters will be obtained for a random sample of 50 cases per outcome per DB. The validation will be done through a common questionnaire that will be developed. The Databases that will participate in the validation are: IPCI, CPRD, AEMPS, SIMG and PHARMO.

5.2. **Remote Research Environment**

A study-specific central remote research environment (RRE) for secure access by consortium members will be used. Due to data protection and ethical considerations, each partner will work with local data to create output files using Jerboa © v2.9.21, a custom-built software written in Java ™ [64]. These output files will contain only anonymised de-identifiable data that will be shared in the RRE where consortium members will have a secure and restricted access and where data will be analysed. Details of the RRE will be given in D4.2.

For all analysis SAS ® version 9.2 will be used.

5.3. **Scientific Advisory Board**

The Scientific Advisory Board (SAB) with consultative function, will be formed by independent experts external to the project, so that the expertise and knowledge necessary to assist the Steering Committee on scientific and technical grounds are gathered. There will be 2 fixed SAB members, who will be selectively complemented by other experts during the project's development depending on the issues to be discussed. The SAB will usually meet once per year.

5.4. **Reporting and dissemination of results.**

The study will be implemented according to the following timelines:

Activity	Date
Submission of Study protocol to the Commision	13-July-2012
Start of the study	20-Sep-2012
Interim Report on Terminology mapping to the Commission	31-Jan-2013
Final Report to the Commission	30-Apr-2014

The work plan structure in SAFEGUARD has been carefully designed to cover all aspects requiring specific effort towards a successful completion, and divides activities into eight work packages (WP). Among them WP8 is in charge of the dissemination and communication of results.

A "Communication Plan" has been set up for the dissemination of results of the studies conducted by the SAFEGUARD Consortium.

Speciffically for the comparative studies, the dissemination undertakings will entail primarily, though not exclusively, scientifically-driven interactions that will include, at least:

- Publication of scientific papers in peer reviewed journals.
- Presentations at relevant events (congresses, meetings, workshops, etc)
- Individual presentations and meetings with key stakeholders.

5.5. **Protocol amendments**

Amendments to the study protocol will be generated as needed during the conduction of the study and will be properly documented in a new version of this document. The rationale of the amendment will be documented as well.

6. Strengths and Limitations

One of the strengths of these strudy is that 9 databases from EU and USA will contribute data on users of drugs for T2DM. The sizes of the data will permit us to study less often prescribed medications, rare co-medications and rare co-morbidities. The heterogeneity of the DBs will allow us to study drugs that are available in the market in EU but also in the USA and will allow examining the reasons for varying risk estimates for each endpoint and individual compound across different characteristics (e.g. type of database, country, etc.).

The common study protocol and the use of common software (i.e Jerboa) makes it possible to overcome some of the common problems with multiple database studies, especially, whether differences in results arise from differences in the databases or differences in programing or study definitions. Under a common study design and standardized protocol the differences due to variations in study design are mitigated.

Another important advantage of the use of existing data is the number of co-variables contained in the databases and that the data are available with relatively low cost and time requirements compared to other types of studies. It also helps on avoiding certain type of bias such as selection or recall bias.

The limitations of this study will be mainly due to the availability and level of detail of data. Not all potential confounders (e.g., regarding lifestyle) are contained in all databases and not all variables contain the information in desired detail. Particularly, information on the prescribed dose and duration of a prescription is not contained in all databases and has to be estimated, which might lead to misclassification of exposure. However, sensitivity analyses will be performed to assess the effect of different methods to estimate the prescribed dose.

Some misclassification of endpoints as well as confounders is possible. Deaths, particularly before arrival to the hospital, are one example of this issue. Validation studies have shown that coding is reliable in many of the databases. However, a validation study is planned which makes it possible to assess the rate of misclassification and correct the estimates accordingly.

7. Ethical, Data Privacy, and Legal Issues

The study protocol will be submitted to local Scientific and Ethical Advisory Boards according to local requirements. Each participant database will process personal data collected in national/regional electronic health record databases. A Remote Research Environment (RRE) will be crated as a repository and secure environment for all information extracted from the participant databases. In this RRE only aggregated unanonymized data will be shared, information at individual level will be managed at local level only.

According to the local legal/ethical requirements of each country, the study protocol will be submitted to the Review Boards / Ethics Committees if needed.

ANNEX I. Non-insulin blood glucose lowering drugs (NIBGLD) and insulins

NIBGLD (A10B)	
Substance	**ATC code**
Biguanides (A10BA)	
Metformin	A10BA02
Sulfonamides, urea derivates (A10BB)	
Glibenclamide (Glyburide - USA)	A10BB01
Chlorpropamide	A10BB02
Tolbutamide	A10BB03
Tolazamide	A10BB05
Carbutamide	A10BB06
Gliclazide (not marketed in the USA)	A10BB09
Glimepiride	A10BB12
Glipizide	A10BB07
Acetohexamide	A10BB31
Combinations of oral blood glucose lowering drugs (A10BD)	
Metformin/glibenclamide	A10BD02
Metformin/rosiglitazone	A10BD03
Rosiglitazone/glimepiride	A10BD04
Pioglitazone/metformin	A10BD05
Pioglitazone/glimepiride	A10BD06
Sitagliptin/metformin	A10BD07
Vildagliptin/metformin	A10BD08
Metformin/Glipizide	A10BD02
Pioglitazone / Alogliptin	A10BD09
Metformin / Saxagliptin	A10BD10
Metformin/Linagliptin	A10BD11
Metformin / Repaglinide	
Sitagliptin / Simvastatin	
Alpha-glucosidase inhibitors (A10BF)	
Acarbose	A10BF01
Miglitol	A10BF02
Voglibose	A10BF03
Thizolidinediones (A10BG)	
Rosiglitazone	A10BG02
Pioglitazone	A10BG03
Dipeptidyl peptidase 4 inhibitors (A10BH)	

Sitagliptin	A10BH01
Vildagliptin (Not in USA)	A10BH02
Saxagliptin	A10BH03
Alogliptin	A10BH04
Linagliptin	A10BH05
Meglitinides (A10BX)	
Repaglinide	A10BX02
Nateglinide	A10BX03
GLP1 analog (A10BX)	
Exenatide	A10BX04
Liraglutide	A10BX07
Amilyn analog (A10BX)	
Pramlintide (only USA)	A10BX05
Other (A10BX)	
Benfluorex (Not in USA)	A10BX06

INSULINS AND ANALOGUES (A10A)	
Substance	*ATC Code*
Insulins and analogues for injection, fast-acting (A10AB)	
insulin (human)	A10AB01
insulin (beef)	A10AB02
insulin (pork)	A10AB03
insulin lispro	A10AB04
insulin aspart	A10AB05
insulin glulisine	A10AB06
combinations	A10AB30
Insulins and analogues for injection, intermediate-acting (A10AC)	
insulin (human)	A10AC01
insulin (beef)	A10AC02
insulin (pork)	A10AC03
insulin lispro	A10AC04
insulin aspart	A10AC30
Insulins and analogues for injection, intermediate-acting combined with fast-acting (A10AD)	
insulin (human)	A10AD01
insulin (beef)	A10AD02
insulin (pork)	A10AD03
Insulin lispro	A10AD04
insulin aspart	A10AD05
combinations	A10AD30
Insulins and analogues for injection, long-acting (A10AE)	
insulin (human)	A10AE01
insulin (beef)	A10AE02
insulin (pork)	A10AE03
insulin glargine	A10AE04
insulin detemir	A10AE05
combinations	A10AE30
Insulins and analogues for inhalation (A10AF)	
insulin (human)	A10AF01

ANNEX II. Characteristics of the healthcare databases

Country	**Italy**			**NL**		**UK**	**Spain**	**Germany**	**USA**
Name of data source	**Health Search**	**Regional database Puglia**	**Regional database Lombardy**	**IPCI**	**PHARMO**	**CPRD**	**BIFAP**	**GePaRD**	**Medicare**
Type of data source	Electronic medical record	Administrative database	Administrative database	Electronic medical record	Record linkage system	Electronic medical record	Electronic medical record	Administrative database	Administrative database (65 yrs +)
Period covered now	2000-2010	2002-2009	2002-2010	2007-now	1998-2010	2000-now	2001-2009	2004-2008	2005-2008
No. source population	1.4 million	5 million	9 million	1.1 million	4 million	8 million	3.2 million	> 14 million	>4 million
Setting	Primary care	Outpatient care	Outpatient care	Primary care	Out and in patient care	Primary care	Primary care	In- and outpatient care	In- and outpatient care
Diagnosis terminology*	ICD-9 and free text	ICD-9	ICD-9	ICPC and free text	ICD-9	READ codes	ICPC and free text	ICD-10-GM codes	ICD-9
Drugs	Prescriptions	Dispensings	Dispensings	Prescriptions	Dispensings	Prescriptions	Prescriptions	Dispensings	Dispensings
Drug coding system **	ATC	ATC	ATC	ATC	ATC	Multilex	ATC	ATC	NDC System
Laboratory values	Yes	No	No	Yes	Yes (subset)	Yes	Yes	No	No

* ICD-9= International Classification of Diseases, 9th edition; ICD-10= International Classification of Diseases, 10th edition; ICD-10-GM= International Classification of Diseases, 10th edition German Modification; ICPC= International Cassification of Primary Care; READ: It is the clinical terminology system used in General Practice in the United Kingdom.

** ATC= Anatomical Therapeutic Chemical Classification System; NDC System=National Drug Code System; Multilex BNF=British National Formulary

Annex III. Methods for confounding control

Method	Pros	Cons
Design based		
Restriction	Easily understood and effective at eliminating confounding.	Can lead to substantial loss of power as well as difficult generalization of results, especially if restriction is applied to more than one characteristic. Only works for measured confounders.
Stratification	Intuitive approach to address confounding that can be presented in a transparent manner. Provides insight into effect-measure modification. Can provide both stratum-specific and pooled estimates.	There will be many strata if several variables involved, and this will complicate presentation of results (many tables). Sparse data can jeopardize the analysis. Only works for measured confounders.
Matching	Intuitive approach to address confounding that can be presented in a transparent manner. Flexible since different types of matching can be applied to different variables such as individual variables or summaries of variables (such as a propensity score).	Loss of power if fraction of exposed patients matching is low or if the number of events lost due to lack of matches is high. Residual imbalance may still exist. Only works for measured confounders. Does not permit proper evaluation of the effect of matching variable.
Within-pt	Time-independent confounders (known and unknown) are accounted for.	Well defined exposure period and event date is essential. Best for transient exposures and acute outcomes. Limited utility for chronic exposures/outcomes.
Active Comparator Group	Addresses confounding (measured and unmeasured) associated with receipt of treatment (relative to an untreated comparison group) Addresses confounding by indication (if active comparator has similar indication)	Active comparator with similar indication may not exist. Channeling bias can be an issue The effect of the active comparator on the risk of the outcome must be established
Analysis based		
Adjustment	Broadly understood and straight-forward to apply.	Model building and variable selection can be complicated. Assumptions may not be recognized so violations may be

Method	**Pros**	**Cons**
		ignored. Only works for measured confounders.
PS	Particularly useful with many covariates and infrequent outcome	Model selection for PS may be complicated Only works for measured confounders. Separate scores should be calculated for users of various comparators (reference groups)
HDPS	Particularly useful with many covariates Can identify confounders that are not expected	Only works for measured confounders Separate scores should be calculated for users of various comparators (reference groups)
IPTW	All patients are used in analysis.	Model selection for weights may be complicated. Only works for measured confounders.
DRS	Useful with many covariates and infrequent exposure A single DRS can be applied for all exposures of interest The interpreatation of DRS is clearer than the PS and DSR distributions across databases will help to understand confounding	Only works for measured confounders
IV	Good approach to address unmeasured confounders	Choice of instrumental variable may be difficult. Assumptions may not be recognized or untestable so violations may go unnoticed.
External adjustment	Take information from other sources so that it is not necessary to have the information within the data source.	Representativeness of the external data source

References

1. IDF, *Diabetes Atlas. Diabetes Estimates 2010, 7 regions*, 2010.
2. IDF, *Diabetes Atlas. Diabetes Estimates 2030, 7 regions*, 2010.
3. Muoio, D.M., *Molecular and metabolic mechanisms of insulin resistance and β- cell failure in type 2 diabetes.* Nature Reviews Molecular Cell Biology, 2008. **9**(March): p. 193-205.
4. Fowler, M.J., *Microvascular and Macrovascular Complications of Diabetes.* Clinical Diabetes, 2008. **26**(2): p. 77-82.
5. Bergenstal, R., *Type 2 Diabetes: Assessing the Relative Risk and Benefits of Glucose-lowering Medications.* The American Journal of Medicine, 2010. **123**(4): p. 374.e9-374e.18.
6. UKPDS, *Intensive blood-glucose control with sulphonylureas or insulin compared with type 2 diabetes (UKPDS 33).* Lancet, 1998. **352**: p. 837-853.
7. EMA, *European public assessment reports. Diabetes Mellitus Type 2.*, 2011.
8. Inzucchi, S.E., et al., *Management of hyperglycaemia in type 2 diabetes: a patient-centered approach. Possition statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD).* Diabetologia, 2012. **55**: p. 1577-1596.
9. Gale, E.A.M., *Troglitazone: the lesson that nobody learned?* Diabetologia, 2006. **49**(1): p. 1-6.
10. (FDA), F.a.D.A. *News Release: Troglitazone (Rezulin).* 2000 24/Nov/2011]; Available from: http://www.fda.gov/Safety/MedWatch/SafetyInformation/SafetyAlertsforHumanMedicalProducts/ucm173081.htm.
11. Nissen, S.E. and K. Wolski, *Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes.* New England Journal of Medicine, 2007. **356**(24): p. 2457-2471.
12. FDA. ***FDA ALERT #1: 5/21/2007, updated 11/19/2007****: FDA adds a new warning about the risk of myocardial ischemia to the boxed warning for Avandia, for treatment of Type 2 Diabetes*. 2007 21/Nov/2011]; Available from: http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm143406.htm#2007_5.
13. EMEA, *Press Release: EMEA recommends new warnings and contraindications for rosiglitazone*, 2008.
14. Nathan, D., *Management of hyperglycaemia in the type 2 diabetes mellitus: a consensus algorithm for the initiation and adjustment o therapy. Update regarding the thiazolidinediones.* Diabetologia, 2008. **51**: p. 8-11.
15. FDA. ***FDA ALERT #2: [8/14/2007]. New boxed warning, and additional updated warnings, precautions and contraindications to emphasize that rosiglitazone may cause or exacerbate heart failure, particularly in certain patient populations***. 21/Nov/2011]; Available from: http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm143406.htm#2007_7.
16. FDA. ***FDA ALERT [8/2007] New boxed warning and additional updated warnings, precautions and contraindications to enphasize that pioglitazone may cause or exacerbate heart failure, in certain populations.*** 2007 21/Nov/2011]; Available from: http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm109136.htm.

17. EMA. *Press Release: European Medicines Agency recommends suspension of Avandia, Avandamet and Avaglim. Anti-diabetes medication to be taken off the market.* 2011 21/Nov/2011]; Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Press_release/2010/09/WC500096996.pdf.
18. FDA. *FDA Drug Safety Communication (2-03-2011): Avandia (rosiglitazone) labels now contain updated information about cardiovascular risk and use in certain patients.* 2011; Available from: http://www.fda.gov/Drugs/DrugSafety/ucm241411.htm#sa.
19. Nitsche, C.J., *Drug induced pancreatitis.* Best practice & Research Clinical Gastroenterology, 2010. **24**: p. 143-155.
20. Noel, R.A., *Increased risk of acute pancreattis and biliary disease observed in patients with type 2 diabetes. A retrospective cohort study.* Diabetes Care, 2009. **32**(5): p. 834-838.
21. FDA. *Drugs @ FDA. FDA Approved Drug Products. Exenatide*. 14/Nov/2011]; Available from: http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm?fuseaction=Search.Overview&DrugName=BYETTA&CFID=62765693&CFTOKEN=c3969a2c9eb5ee18-A266E25C-975E-5AA5-AD2D68D5BE17EE97.
22. EMA. *Human medicines. Product details. Exenatide - Byetta*. 14/Nov/2011]; Available from: http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/medicines/000698/human_med_000682.jsp&mid=WC0b01ac058001d124.
23. FDA. *Drugs @ FDA. FDA Approved Drug Products. Sitagliptin.* 14/Nov/2011]; Available from: http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm?fuseaction=Search.DrugDetails.
24. Drucker, D.J., *The Safety of Incretin-Based Therapies - Review of the scientific evidence.* J Clin Endocrinol Metab, 2011. **96**(7): p. 2027-2031.
25. Elashoff, M., *Pancreatitis, Pancreatic and Thyroid cancer with Glucagon-Like Peptide-1-Based Therapies.* Gastroenterology, 2011. **141**: p. 150-156.
26. Dore, D.D., *A cohort study of acute pancreatitis in relation to exenatide use.* Diabetes, Obesity and Metabolism, 2011. **13**: p. 559-566.
27. Dore, D.D., *Use of a claims-based active drug safety surveillance system to assess the risk of acute pancreatitis with exenatide or sitagliptin compared to metformin or glyburide.* Current Medical Research and Opinion, 2009. **25**(4): p. 1019-1027.
28. Yamamoto, H., *Glucagon-like peptide-1 receptor stimulation increases blood pressure and heart rate and activates autonomic regulatory neurons.* Journal of Clinical Investigations, 2002. **110**: p. 43-52.
29. Gill, A., *Effect of exenatide on heart rate and blood pressure in subjects with type 2 diabetes mellitus: a double-blind, placebo-controlled, randomized pilot study.* Cardiovascular Diabetology, 2010. **9**: p. 6.
30. Davidson, M.H., *Cardiovascular Effects of Glucagonlike peptide-1 Agonists.* The American Journal of Cardiology, 2011. **108**(3 Suppl): p. 33B-41B.
31. Bergner, D. and J. Goldberger, *Diabetes mellitus and sudden cardiac death: what are the data?* Cardiol J., 2010. **17**(2): p. 117-29.

32. FDA, *Guidance for Industry, Diabetes Mellitus — Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes*, F.a.D.A. U.S. Department of Health and Human Services, Center for Drug Evaluation and Research (CDER), Editor 2008: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071627.pdf.
33. Watters, K., N. Munro, and M. Feher, *QTc prolongation and diabetes therapies.* Diabet. Med. , 2012. **29**: p. 290-292.
34. Vigneri, P., *Diabetes and Cancer (Review).* Endocrine-Related Cancer, 2009. **16**: p. 1103-1123.
35. Giovannucci, E., *Diabetes and Cancer: A consensus Report.* Diabetes Care, 2010. **33**(7): p. 1674-1685.
36. Bodmer, M., et al., *Use of Antidiabetic Agents and the Risk of Pancreatic Cancer: A Case – Control Analysis.* Am J Gastroenterol, 2012. **107**: p. 620-626.
37. Johnson, J.A., *Time-varying incidence of cancer after the onset of type 2 diabetes: evidence of potential detection bias.* Diabetologia, 2011. **54**: p. 2263-2271.
38. Nathan, D.M., et al., *Medical management of hyperglycemia in type 2 Diabetes: A consensus Algorithm for the Initiation and Adjustment of Therapy. A consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes.* Diabetes Care, 2009. **32**(1): p. 193-203.
39. Yachida, S., *Distant metastasis ocurrs late during the genetic evolution of pancreatic cancer.* Nature, 2010. **467**: p. 1114-1117.
40. FDA. *FDA Drug Safety Communication: Update to ongoing safety review of Actos (pioglitazone) and increased risk of bladder cancer*. 2011 17/Nov/2011]; Available from: http://www.fda.gov/Drugs/DrugSafety/ucm259150.htm#sa.
41. EMA. *Press releaase: European Medicines Agency recommends new contraindications and warning for pioglitazone to reduce small increased risk of bladder cancer*. 2011 17/Nov/2011]; Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Press_release/2011/07/WC500109176.pdf.
42. Lewis, J.D., *Risk of bladder cancer among diabetic patiens treated with pioglitazone.* Diabetes Care, 2011. **34**: p. 916-922.
43. Lamberts, H., M. Wood, and I.M. Hofmans-Okkes, *International primary care classifications: the effect of fifteen years of evolution.* Fam Pract, 1992. **9**(3): p. 330-9.
44. Salvador Rosa, A., et al., *[The BIFAP project: database for pharmaco-epidemiological research in primary care]*. Aten Primaria, 2002. **30**(10): p. 655-61.
45. World Health Organization (WHO), *Anatomical Therapeutical Chemical Index With Defined Daily Dosages*2002, Oslo, Norway: World Health Organization Collaborating Center for Drug Statistics Methodology.
46. Epstein, M., *Guidelines for good pharmacoepidemiology practices (GPP).* Pharmacoepidemiol Drug Saf, 2005. **14**(8): p. 589-95.
47. Thygesen, K., *Universal Definition of Myocardial Infarction.* Circulation, 2007(116): p. 2634-2653.
48. Jessup M, et al., *Focused Update: ACCF/AHA Guidelines for the Diagnosis and Management of Heart Failure in Adults: A Report of the American College of Cardiology Developed in*

Collaboration with the International Society for Heart and Lung Transplantation. J. Am. Coll. Cardiol., 2009. **53**: p. 1343-1382.

49. Zipes, D., *ACC/AHA/ESC 2006 Guidelines for Management of Patients With Ventricular Arrhythmias and the Prevention of Sudden Cardiac Death—Executive Summary. A Report of the American College of Cardiology/American Heart Association Task Force and the European Society of Cardiology Committee for Practice Guidelines (Writing Committee to Develop Guidelines for Management of Patients With Ventricular Arrhythmias and the Prevention of Sudden Cardiac Death).* J. Am. Coll. Cardiol., 2006. **48**: p. 1064-1108.
50. Goldman, L. and D.A. Ausiello, *Chapter 430: Approach to Cerebrovascular diseases.*, in *Cecil Medicine*2008, Sauders Elsevier.
51. Easton, J.D. and J.L. Saver, *Definition and evaluation of transient ischemic attack: a scientific statement for healthcare professionals from the American Heart Association/American Stroke Association Stroke Council; Council on Cardiovascular Surgery and Anesthesia; Council on Cardiovascular Radiology and Intervention; Council on Cardiovascular Nursing; and the Interdisciplinary Council on Peripheral Vascular Disease. The American Academy of Neurology affirms the value of this statement as an educational tool for neurologists.* Stroke, 2009. **40**(6): p. 2276-2293.
52. Goldman, L. and D.A. Ausiello, *Chapter 147: Pancreatitis.*, in *Cecil Medicine*2008, Sauders Elsevier.
53. NCI. *Dictionary of Cancer* 21Dec2011].
54. Avillach, P., et al., *Design and evaluation of a semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European EU-ADR project.* Stud Health Technol Inform., 2010. **160**(Pt 2): p. 1085 - 1089.
55. Guess, *Exposure-time-varying hazard function ratios in case-control studies of drug effects.* PDS, 2006. **15**: p. 81-92.
56. Boccuzzi, S., et al., *Utilization of oral hypoglycemic agents in a drug-insured U.S. population.* Diabetes Care, 2001. **28**(8): p. 1411-1415.
57. WHO, *Age standardization of rates: a new WHO standard*, in *GPE Discussion Paper Series*, EIP/GPE/EBD, Editor 2001. p. 1-14.
58. A Allen, G.S., *Control for Confounding in Case-Control Studies Using the Stratification Score, a Retrospective Balancing Score.* Am.J. of Epidemiology, 2011. **173**: p. 752-760.
59. JA Rassen, A.S., J Myers, RJ Glynn1, KJ Rothman, S Schneeweiss, *One-to-many propensity score matching in cohort studies.* PDS, 2012. **21**: p. 69-80.
60. B Fireman, S.T., MG Butler, AS Go, HV. Joffe, DJ Graham, JC. Nelson, GW. Daniel, JV. Selby, *A protocol for active surveillance of acute myocardial infarction in association with the use of a new antidiabetic pharmaceutical agent.* PDS, 2012. **21**: p. 282-290.
61. Greenland, S., *Modeling and variable selection in epidemioloc analysis.* Am J. of Public Health 1989 **79**: p. 340-349.
62. JA Rassen, S.S., RJ Glynn, MA Mittleman, MA Brookhart *Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes.* Am J of Epidemiology, 2009. **169**: p. 273-284.
63. ENCePP, *Guide on Methodological Standards in Pharmacoepidemiology.* 2011.

64. Coloma, P., *Combining electronic health care databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project.* Pharmacoepidemiol Drug Saf, 2011. **20**(1): p. 1-11.