



# Guidelines for training personnel involved in population based registries

WP8 - task 1 report

30.10.2017

Project Leader:  
Simona Giampaoli  
Istituto Superiore di Sanità  
Viale Regina Elena 299  
00161 Rome, Italy  
Tel +390649904231  
Fax +390649904227  
E-mail [simona.giampaoli@iss.it](mailto:simona.giampaoli@iss.it)



This project is funded  
by the Health  
Programme of the  
European Union

## CONTENTS

1. INTRODUCTION	3
2. POPULATION-BASED REGISTRIES: DEFINITION, STRENGTHS AND WEAKNESSES	8
Main steps for the implementation of a population-based registry	11
Build the team	13
3. IMPLEMENTATION	14
STEP 1: Define target population and data sources	15
Target population	15
Population size calculation	18
Geography	22
Duration	23
Setting	24
Data sources	24
Examples	27
STEP 2 - Carry out record-linkage of administrative data	28
Core data set	28
Procedure to detect mistakes in subject identification variables and to eliminate duplications in databases	31
STEP 3 - Perform a pilot study and validate routine data	33
Events	33
Identification of events	33
Event validation	35
Diagnostic criteria for Acute Myocardial Infarction/Acute Coronary Syndromes (AMI/ACS)	36
Diagnostic criteria for Stroke	47
Criteria for Type-1 Diabetes Mellitus cases identification	51
STEP 4 - Set up of a population-based registry	52
Description of the software for the implementation of the population-based registry of Coronary events (used in the EuroMed project for the coronary registry of Zagabria population)	54
Description of the software for the population-based registry of Cerebrovascular events	57
Description of the software 'RIDI-PROG' for the Italian RIDI - Type-1 Diabetes Mellitus registry	60
STEP 5: Analyses and dissemination of results	60
Incidence rate	61
Attack rate	61
Case fatality	62
Survival rate	62

Planning a fruitful dissemination of results	63
4. STEPWISE PROCEDURE: THE EXAMPLE OF CORONARY AND CEREBROVASCULAR REGISTRIES	63
5. QUALITY CONTROL	67
Data quality assurance	67

## 1. INTRODUCTION

These guidelines are intended for the training of the team that shall implement population-based registries. Purposes, goals, governance, costs, protocol elaboration, project planning and other preliminary stages of a registry are described in the WP8 'Platform for population-based registries Manual of operations - Task 1'. They are an integrated part of that Manual and are enriched and completed **with slide presentations available at the 'BRIDGE' page of the [www.cuore.iss.it](http://www.cuore.iss.it) website**. They are intended for investigators and personnel involved in data collection, record-linkage, event validation, data quality assessment, statistical analyses and dissemination of results at local level.

In training personnel, it is crucial to explain hypotheses, purposes and methods to pursuit good data quality, and achieve the possible results of the registry. The more the staff involved in the register knows these aspects, the better will be the quality of the collected and elaborated data of the registry.

This report aims at providing recommendations to train registry personnel, from general procedures and methods, to the step-wise procedure for the implementation of the population-based registry.

This report is complementary to the 'Platform for population-based registries Manual of operations - Task 1' and represents a tool to train personnel involved in the population-based registry who shall work to provide statistics on disease occurrence, incidence and survival rates, and the core indicators recommended by the short list of the ECHIM Project.

The staff involved in the activity of a Registry should be qualified through theoretical or practical training courses. Every training should be focused on how to implement the knowledge about different aspects of the registry. The training has to be planned in advance and tailored to its audience's needs.

In planning a training programme, the following questions should be taken into account:

- Who is the audience?
- What are the learning objectives?
- What are the best mechanisms to disseminate the information?
- What is the best approach to ensure that learning has occurred?

During the BRIDGE Health Project, a questionnaire was administrated to the people responsible of registries involved in the WP8-Task 1 for the collection of information about the methods to train personnel on main registry activities. Table 1 and Table 2 show the main results obtained from the questionnaire on items related to source of information, data collection, data input and control, statistical analysis, record-linkage, validation, standardization, and data quality.

All registries have trained their personnel on the following items: source of information used to identify events, data collection methods, data input and data checking, methods

to assess incidence, survival, and case fatality rates. Some registries have held periodical workshops to analyse and discuss the data collected, with the aim of improving efficiency and quality of data collection. Training on the standardisation of procedures and methods to collect, validate, analyse and assess indicators have been promoted by registries dealing with chronic diseases (Cancer, Diabetes, Coronary diseases, Stroke) and injuries. Most registries have considered validation procedures in their trainings and some also considered record-linkages. Training on data quality was an aspect faced by five registries (Cancer, Diabetes, Transplant, Coronary, Stroke).

In most cases, training courses are organised by National Public Health Institutions and Universities, and the participants (mainly Epidemiologists, Technicians, Medical Doctors) belong to Universities, Regions, and national and local Health Organisations. All the courses are theoretical and performed on site. Sometimes, practical sessions are also organised. Courses are mainly residential, only few registries provide on-line courses.

The various registries do not establish how often the courses must be held; when they do, the frequency is annual or biannual/triannual. Only few registries (Rare Diseases, Cancer, Coronary, Stroke, and Maternal Mortality) foresee that trainers verify the learning level of personnel in registry procedures by performing site visits or audits after the course.

The most important aspects usually evaluated in a training course are:

#### *Sources of information*

A wide variety of sources of information are available for Registries, depending on the type of disease or condition under surveillance and their related events. For example, to monitor Non Communicable Diseases (NCDs) in the general population, at least the following sources of information should be available: mortality records with death certificates; HDRs with clinical records, autopsy registry, nursing home and clinic records, emergency and ambulance services, GP clinical records, drug dispensing registry, exemptions.

#### *Data collection*

The methods adopted to define the event and to identify suspected events are important to obtain a good final result. They should be standardized, follow quality control criteria, and comply with the aim of the registry. Population-based registries are usually based on integration or linkage of several data sources, e.g. hospital discharges, death certificates, drug prescriptions.

#### *Data input and check*

Data inputs in the registry database need to be checked for possible errors. Errors can concern wrong coding or double entries, and involve quantitative (format, plausibility of values, anomalous data, range, distribution, number of missing values, and consistency with other related variables, e.g. format, value, distribution) or qualitative variables (format, allowed codes, distribution, missing values, and consistency with other related variables). Reduction of errors can be obtained by a periodic and appropriate training of the staff.

Data input is performed by using specific software that already includes check procedures.

### *Data quality assessment*

Overall, a good quality of data in a registry is ensured by checks on the different dimensions of quality: completeness, validity, and timeliness. The relevance of a single dimension depends on the registry's type and objectives, as well as its scope and methodology. Data quality dimensions are components that allow the user to quickly identify specific problematic aspects of data, interpret statistics in the right context, and compare results with other statistics.

### *Completeness*

It can be referred to the number of registered events, or to the recorded information. The former is assessed by the 'coverage rate' and describes the extent to which all the expected events are registered; the latter refers to the availability of all information about the event. Missing data indicate the presence of problems in data collection validity. A variable can be controlled by the point of view of the format, the admitted range of values, dates, classification, etc.). Errors in a variable can occur during event identification, data entry, coding and editing; the best way to discover these errors is to re-identify, re-code, and re-enter data.

### *Timeliness*

In population-based registries, timeliness refers to the length of time between the occurrence of the event and the dissemination of the results. Registries may tend to delay the dissemination of their results in order to achieve better completeness.

To obtain a good quality data collection, it is essential to prepare the manual of operations. This is a fundamental tool to plan a good training.

**Table 1 - Purpose of training**

Registry	Source of information	Data collection	Data input & check	Statistical analysis	Record linkage	Validation	Standardisation	Data quality	Other
Rare disease	YES	YES	YES	na	na	YES	YES	na	
Cancer	YES	YES	YES	YES	na	na	YES	YES	
Cardiovascular diseases	YES	YES	YES	YES	YES	YES	YES	YES	
Stroke	YES	YES	YES	YES	YES	YES	YES	YES	
Diabetes	YES	YES	YES	YES	YES	YES	YES	YES	
Arthroplasty	YES	YES	YES	na	na	na	na	na	
Injuries	YES	YES	YES	na	YES	YES	YES	na	
Congenital hypothyroidism in infancy	YES	na	na	na	na	na	na	na	
Transplants	YES	YES	YES	na	na	na	na	YES	claim of declaration of will
Maternal mortality	YES	YES	YES	na	YES	YES	YES	na	

na = not available

**Table 2 - Training structure**

Registry	Organiser	Membership/ Participants	Course type <sup>^</sup>	Course mode <sup>^</sup>	Periodicity	Participant qualification*	Testing by trainer
Rare diseases	NIH	Region, HLO	T,P	R	every year	EP,MD,TE,NU	YES
Cancer	University, AIRTUM <sup>^^</sup>	University, AIRTUM	T,P	R	every year	RE,EP,ST,MD,TE	YES
Cardiovascular diseases	NIH	University, Region, HLO	T,P	R	na	EP,MD	YES
Stroke	NIH	University, Region, HLO	T,P	R	na	EP,MD	YES
Diabetes	University, National Coord., Local Registries Coord.	University, HLO, Hospital	T	R	na	EP,MD,TE	NO
Arthroplasty	NIH, Region	Region, HLO, Hospital	T	R	na	EP,MD,T	NO
Injuries	NIH, Region, Hospital	Region, HLO, Hospital, Prevention Dept.	T,P	R	na	EP,ST,MD,TE,NU	YES
Infant congenital hypothyroidism	NIH		na	na	every 2 years	na	na
Transplants	Nat. Transplant Center	HLO	T,P	na		na	na
	Reg. Transplant Center	Municipality		R		MD,TE,NU,staff of Municipality	NO
Maternal mortality	NIH, Region	Region, Hospital	T,P	R, On-line	every 3 years	MD,EP,OB,ST	YES

\* EP=Epidemiologist, MD=Medical Doctor, TE=Technician, NU=Nurse, ST=Statistician, RE=Researcher

<sup>^</sup> T = Theoretical, P = Practical, R = Residential

<sup>^^</sup>AIRTUM = Italian Association of Cancer Registries

NIH = National Institute of Health; HLO = Health Local Organization; na = not available



## 2. POPULATION-BASED REGISTRIES: DEFINITION, STRENGTHS AND WEAKNESSES

A population-based registry is an organized system that uses observational study methods to collect all new cases of a disease in a defined population (most frequently a geographical area); data serve for one or more predetermined scientific, clinical and health policy purposes.

The “**core**” activity is to provide information on **incidence and survival**; a different number of variables can be collected, allowing to study the effects of various prevention aspects, treatments and caring services.

For some NCDs, population-based registries are the best data source for incidence and survival rates, in particular for those diseases that have an acute onset, such as coronary and stroke events, and injuries. Registries consider both fatal and non-fatal events occurring in hospital and out-of-hospital, all new cases and recurrent events in a defined general population, whether treated at home or in hospital, in whichever season of the year or time of the day they may occur, and would also include sudden fatal cases unable to reach the medical service, thus providing estimates of key indicators such as incidence and case fatality rates and survival rates. For other NCDs such as cancer and type 1 diabetes, the definition of onset is an arbitrary concept since it is a continuum of the disease’s natural history; in that case, incidence (new case of diseases) corresponds to the time of the clinical diagnosis, after a patient has been presented to medical attention.

**The burden of disease** and its probable future evolution can be evaluated in terms of incidence and mortality, but other dimensions can be considered, such as prevalence, years of life lost, quality or disability-adjusted life years: a deep knowledge of the history of a disease may help to project trends into the future and assess probable effects in changing risk factors.

Focusing on general population, population-based registries may provide a **comprehensive picture of a disease in the community**, highlight problem areas and suggest where treatment facilities are most in need of improvement; they may also provide information systems needed to plan healthcare services, and develop and test which methods are most useful as a basis for preventive actions. This is crucial in order **to plan research purposes**, identify causes and monitor progress in prevention, produce annual reports, **orientate preventive actions**, make comparisons among countries in order to achieve better knowledge and more effective interventions, and **support decision making**.

Information from multiple sources contributes to the population-based registry database; therefore it is important to link all the records pertaining to an individual to **avoid duplicate registration**. A **Personal Identification Number (PIN)** is ideal for this purpose; however only Nordic Countries have it. The PIN for each subject is a strong tool in linkage procedures between different sources of data, such as hospital discharge diagnoses, GP records, and death certificates; alternatively, multiple variables (e.g. date and place of birth, sex, residence) may be used for record-linkage. This is not

possible in most European countries due to privacy and ethical norms and laws, despite the new European law on this matter.

Potentially, an individual can have more than one cancer or more than one coronary event; with improved survival, this is becoming more frequent. Incidence and survival rates relate to a specific event, so the **new case must be distinguished from the recurrent** (attack rate). The definition of the event should take into account both the ICD codes reported in the hospital discharge diagnoses (main or secondary) or in the causes of death (underlying or secondary) and the duration of the event. In the case of coronary and stroke events, hospital admissions and deaths occurring within 28 days (onset is day 1) are considered to reflect the same event).

**The quality of registry data is evaluated by its completeness, validity and timeliness.** Completeness of case ascertainment should be as close as 100%; validity, (the accuracy of recorded data) can be increased by checks on recorded data. The rate accuracy is related to the completeness and quality control of data collected for numerator (all events from death certificates, hospital discharge registry, GPs, ...) and denominator (census or population under surveillance). Completeness also depends on tracing pauci or asymptomatic subjects treated **outside hospitals** (nursing homes, clinics, GPs). A valid population-based registry should also collect non-fatal events in the target population occurring **outside the area** of surveillance.

Identification of events can be obtained by “hot pursuit” or “cold pursuit”. **Hot pursuit** means identifying case admissions to hospital usually within one or two days from event onset and acquiring relevant information by visiting the ward or interviewing the patient. Information bias is minimised by the hot pursuit approach as information is collected immediately after the event. The process is very expensive for the numerous suspected events collected for validation.

**Cold pursuit** implies the use of routine and delayed procedures, by means of hospital discharge, and review of clinical and death records. The process is easier and less expensive than hot pursuit; the number of cases studied is typically smaller because discharge diagnoses are more precise and specific than those on admission, but there is a possibility of missing important information when these diagnoses are not recorded in the hospital discharge registry. Both methods, hot and cold pursuit, are used to identify suspected events, which are subsequently validated using standardised methods to define diagnostic criteria.

Population-based registry **must be validated**. Validation provides the means to take into account bias from diagnostic practices and changes in coding systems; it traces the impact of new diagnostic tools and re-definition of events; ensures data comparability within the registry (i.e. different sub-populations, different time points, etc.); ensures data comparability with other registries within and between countries. The strength of population-based registries lies in the possibility of **validating each single event according to standardised diagnostic criteria** and collecting disease-specific clinical and paraclinical data. This process implies a **great effort in training personnel**, in implementing quality controls for reading and collecting information, for the

classification of events according to standardised diagnostic criteria, and for local site visits to assure that standard level are respected and maintained.

Incidence and survival rates from population-based registries have been criticized for the validity of information they provide, in particular when rates are compared in different populations or in different periods of time; a problem common to all comparative studies is the effect of changes in disease classification and coding over time.

Survival, in terms of average number of years lived after a disease, is easy to measure, but it is a “crude indicator” of outcome; years of life are of little value if they are accompanied by disability. **The measurement of health-related quality of life should be part of population based information.**

In some countries, population-based registries have a legal basis (cancer registries); other registries operate on a voluntary basis (coronary and cerebrovascular events), covering populations not representative of the entire country. This means that registry associations have great importance in recommending **common definitions, coding, quality control methods and team training**. A limited geographic coverage is adequate for many descriptive activities, although national data are important to avoid losing migrating subjects or events treated out of the surveillance area. The advantage of using population-based data is that **they relate to the whole community rather than to a single institution or self-selected and atypical subgroups of patients** (those reaching or recovered in a specific centre).

Due to their accuracy in identifying and validating events, results from population-based registries are available usually with a **delay of 3-5 years** in comparison to current administrative data and statistics. All these issues make population-based registries **very expensive**, therefore **this kind of registries can be usually maintained only for a limited period, in a defined population of a reasonable size, to answer the specific questions for which they were instituted**; local or regional registries may not be representative of the whole country; these are the major limits for the implementation of a population-based registry.

With population-based databases, **descriptive studies** may be conducted, examining the differences in incidence and lifestyles; usually these descriptive studies are important to generate hypothesis on risk factors; for example, the role of salt intake in gastric cancer was suggested looking at the difference in different countries, studying migrant and time trends.

**HES** and health information surveys can further supplement the information collected from population-based registries with additional details on socio-demographic characteristics, lifestyles, risk factors, and physical/biological measurements.

A population-based registry is intended for health professionals, researchers and policy makers.

Training of personnel involved in the different steps of a registry implementation is an important aspect. Even though informatic supports are commonly available and used, all the operations to implement a population-based registry have to be conducted on the basis of standardized procedures and methods. This demands specific training for all type of personnel involved in the different steps. Formal training courses, audits and site visits are recommended in order to apply and maintain standardized procedures; it is important to avoid the establishment of individualized practices. Provisions should be made for training courses based on the specific manuals of operations of registries.

The following paragraphs present many issues relating to practical design and operational activities, evaluation principles, and good practices that are included in specific training courses for population-based registry personnel.

### *Main steps for the implementation of a population-based registry*

There is variability in size, scope, and resource requirements for registries. Population-based registries may cover a population large enough to produce stable and robust disease rates. They may target rare or common conditions. They may require the collection of limited amounts of variables, operate for short or long periods of time, and should be financially sustainable. The scope of a registry may be adapted over time to reach broader or different populations. Registries require good planning in order to be successful.

When planning a registry, these initial steps are desirable: (1) formulating the purpose of the registry; (2) determining if a registry is the appropriate tool to achieve the purpose; (3) identifying key stakeholders interested in the disease registry; and (4) assessing the feasibility of a registry. Once a decision is made to proceed, the next considerations are strictly related to the methods to be adopted, specifically in relation to (5) team building; (6) establishing the governance; (7) defining the scope and rigour needed; (8) defining the data set, outcomes, and target population; (9) developing a protocol; and (10) developing a project plan. These steps are described in details in the 'Platform for population-based registries Manual of operations - Task 1' and are the task of governance members, decision makers, and people responsible for the registry.

Periodic evaluations of data collected in the registry ensure that the objectives are met. This is particularly important for registries that collect data over many years. When objectives are no longer met, or when the diagnostic criteria of a disease are changed, the registry needs to be adapted, or collection of new data should be stopped.

The first step in planning a registry is the formulation of a clear, well defined and rational purpose. A defined purpose helps to clarify the quality of data to be collected.

For example, the *WHO MONICA Project* (MONItoring Trends and Determinants in CARdiovascular Disease) was designed to answer key questions arising from the 1978 Bethesda Conference on the Decline in Coronary Heart Disease Mortality, which were: *“Are the reported declines in coronary heart disease mortality genuine? If so, how much this can be attributed to improved survival rather than to a decline in coronary event*

*rates? Are these trends related to changes in risk factors and health care?"*. MONICA was a very wide project conducted between mid-'80s and mid-'90s overall in the world; it allowed for the first time (a) to collect and register 166,000 events in men and women aged 35-64 years, during a 10-year surveillance of 37 populations in 21 countries; (b) to classify all suspected events in fatal and non-fatal definite events, possible, ischemic cardiac arrest with successful resuscitation as well as insufficient data, following the same standardized diagnostic criteria (site and duration of chest pain, evolution of ECG findings -all coded according to the Minnesota Code-, variation of cardiac enzyme values, history of Ischaemic Heart Disease-IHD, and, if performed, necropsy). The 10-year duration of the WHO MONICA Registry was sufficient to demonstrate that contributions to IHD mortality change varied, but, in populations in which mortality decreased, coronary event rates contributed for two thirds and case fatality for one-third. The extent of these trends was related to changes in known risk factors (systolic blood pressure, total cholesterol, smoking habit and body mass index) daily living habits, health care and major socio-economic features measured at the same time in defined communities in different countries.

The questions to consider are whether a registry is needed to address the purpose and which data collection is appropriate, between prospective or retrospective collection. In particular, if data already exist, it must be assessed whether they have the necessary quality to answer the question, whether they are accessible, or whether a new data collection is needed. For example, could the necessary data be extracted from electronic medical records? May the registry avoid re-collecting data? Are data accessible? Is it possible to link administrative data to other relevant data sources? Answers to all these questions and issues should be addressed by the people responsible for the registry before starting its implementation.

After the MONICA experience, some registries implemented CVD surveillance by adopting a simplified methodology. Starting from electronic databases of mortality and Hospital Discharge Records, occurred coronary and cerebrovascular events in the resident adult population were estimated, thus avoiding to re-collect data and to double count fatal events; it was possible to implement the simplified methodology in limited areas of different countries (regions, provinces, geographical districts, towns) where record-linkage between mortality and HDR was possible.

After deciding that a registry is the appropriate method for data collection, it is important to consider the state of knowledge. Other factors that may influence this decision include the size of the population of interest, how to identify the events, the length of the observational period needed to achieve the objective, the amount of funding available. Given the partial or preliminary information on disease occurrence and purposes, it is possible to assess the size of the population of interest and the duration of the observational period needed to estimate stable and reliable indicators such as incidence/attack rate, case fatality, and survival rate. Once decided the epidemiological definition and the adopted diagnostic criteria for event identification,

and consequently the basic information to be collected, it will be possible to estimate the funding needed and verify registry feasibility. Registries may be the most appropriate choice for some research questions. For example, population-based registries are particularly useful in situations where it is necessary to estimate the occurrence of a chronic disease in the population and hospital-based registries are insufficient to address this objective; for example, sudden coronary and cerebrovascular fatal events that do not reach the hospital still represent a significant proportion of overall fatal events (around 30%); these fatal events would be lost (not registered) if only hospital registries were used. A correct estimation of the proportion of fatal events occurred suddenly out of hospital without any medical assistance is of primary importance to plan preventive actions in the general population, since such interventions represent the only way we have to reduce or postpone those fatal events that do not reach the hospital.

The key element in the feasibility of a new registry is related to data accessibility and funds. The expenses depend on the scope of the registry, the data collection method, the objectives (research or surveillance), and the data validation process. For population-based registries, systematic direct collection methods are more expensive and time consuming to implement, compared to a simplified method based on the validation of a random sample of events, or on administrative databases and health data routinely collected by national, regional and local health authorities.

### *Build the team*

The team should understand the objectives of the registry, its data sources, and the importance of data validation; this is the only way in which the registry team will be able to collect and use data in the most appropriate way and for the most appropriate interpretation.

*Fieldwork coordinators:* they coordinate the work and the personnel involved in data collection and data input, prepare the annual report (including results and quality reports), manage timelines and deliverables, and ensure communication with stakeholders.

*Clinicians and expert epidemiologists of the disease/condition under surveillance:* they are usually included as members of the Registry Steering Committee and participate both in the early phase of the registry planning, and in the delicate activity of applying diagnostic criteria to validate suspected events.

*Data collection and database:* These experts may need to input and clean the data, to check their quality, to protect, group and store data, according to their level of security of sensitiveness.

*Statistical analysis:* Epidemiologists and biostatisticians with experience in registry data analysis in the specific disease field are necessary to analyse the collected data. A

software to digitalise most operational activities of the registry is needed for the following steps: record-linkage, identification of suspected events, selection of a sample of events to be validated, collection of overall clinical data of the single event to be validated, algorithm for event validation diagnostic criteria, estimation of PPVs, elaboration of main registry indicators, quality control procedures.

*Ethical Committee at local level:* it is needed to protect the patients and database owners by ensuring that the registry complies with all ethic and privacy rules and with patient information.

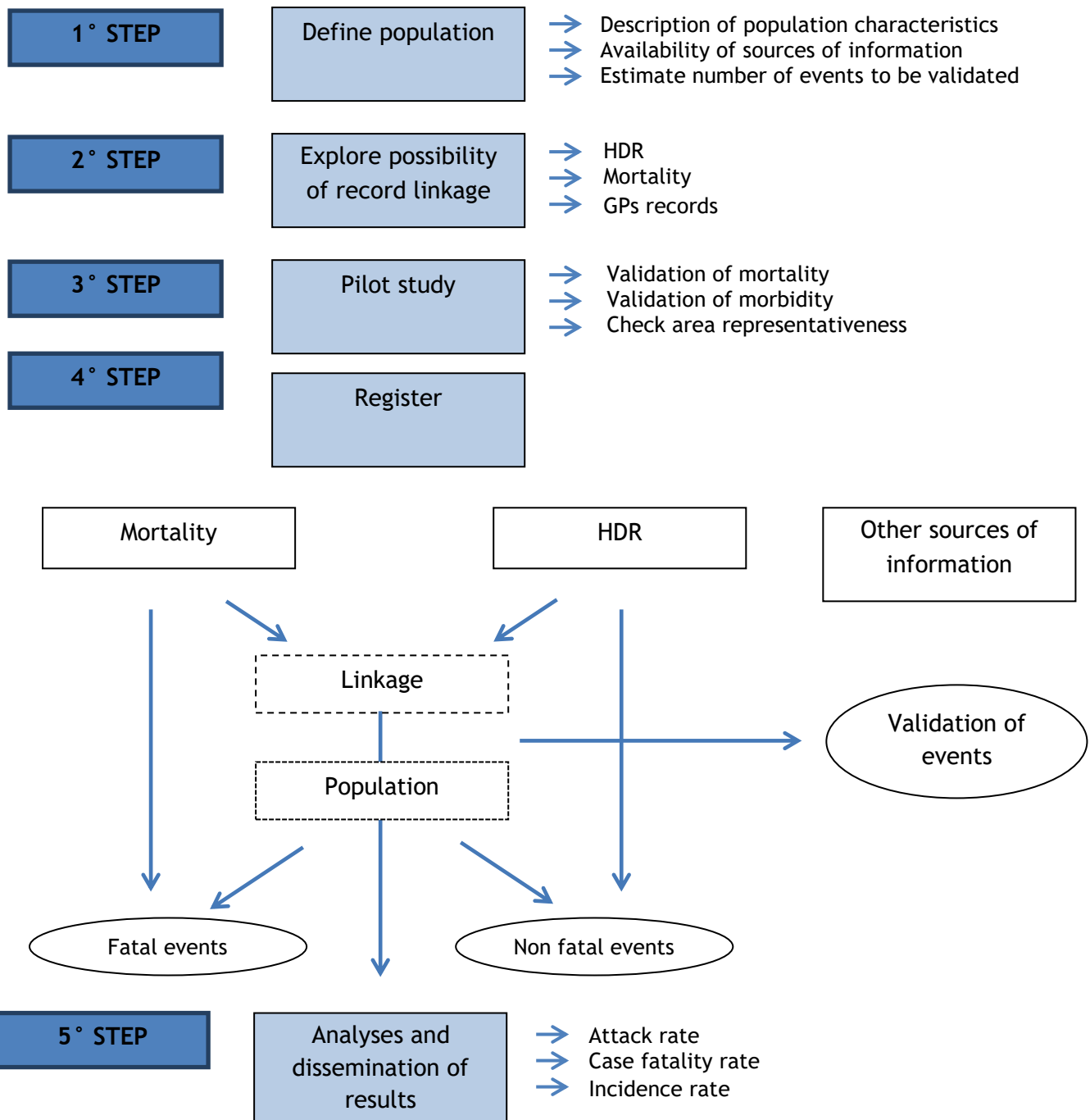
*Quality assurance:* This is another important component for the success of a registry. Expertise in quality assurance will help in planning a good registry. Goals for quality assurance should be established for each registry, and the efforts made and the results achieved should be described.

### 3. IMPLEMENTATION

This section describes the stepwise procedures required to implement a population-based registry. These procedures comply with the above reported recommendations.

The flow chart summarises these procedures (Figure 1).

Figure 1 - Description of stepwise procedures



**STEP 1: Define target population and data sources**

**Target population** A population-based registry may cover the entire country; where this is not feasible, the population under surveillance shall include the residents of a well-defined geographical and administrative area or region for which population data and vital statistics are routinely collected and easily available each year. Both urban and



rural areas should be under surveillance because differences often exist with regard to exposure to risk factors, treatment of diseases and access to facilities.

It is important to record all cases among those residing in the area, even if the event occurs outside the area (*completeness*). Likewise, all events treated at hospitals within the area but with residence outside the area must be excluded. If this is not possible, it is important to give an estimate of the magnitude of the lost events and establish whether this changes or interferes with the validity of the observed rate trends over a period of years.

It is also important to consider to what extent an area is representative of the whole country (*representativeness*): it should be representative according to the mortality rate, distribution of risk factors (socioeconomic status and health behaviour) and distribution of health services (specialised hospitals, GPs).

The population under surveillance should be selected to produce estimates of disease rates sufficiently robust from a statistical point of view, so that trends can be established and data comparability ensured. If the population-based registry is not a national registry, it is recommended to select more than one area in order to have a comprehensive picture for the whole country; coordination between the areas is recommended to ensure comparability. One of the goals for registry data may be the generalization of conclusions drawn for defined populations to be applied to broader ones. This implies that registries must use relatively broad inclusion criteria. As an example, results could be generalised to the overall population if the registry includes the resident population of several geographic areas representative of the country.

The definition of the target population will depend on many factors (e.g. frequency and occurrence of the disease in the population, data accessibility, scope, and cost).

In establishing the target population, attention should be paid to the access to that population data (e.g. mortality records, HDRs, clinical records). It is important to distinguish the ideal situation from the real one. In this regard, at least the following questions should be considered:

- How common is the disease of interest?
- Can eligible people be readily identified?
- Are other sources competing for data on the same persons?
- Is care centralized or dispersed (this is of fundamental importance to validate events for which clinical data are needed; if clinical records are spread among a large number of hospitals, collection of information for event validation is more expensive and time consuming than when events are concentrated in one or few hospitals)?
- How mobile is the target population (especially when people move out of the area covered by the registry; all hospitalizations, in and out of the registry area, should be identified for the resident population included in the registry and registered during the surveillance period)?

- Are data sources under quality control? Do they assure time continuity of data collection? Are they geographically homogeneous in the area under surveillance of the population-based registry?

An increased accessibility to the target population, and the completeness of the information needed for the registry, guarantee benefits in terms of enhanced representativeness and statistical power.

The target population should be selected taking the following parameters into account:

- *Age*

The age range depends on the registry aim;

- for registries covering coronary and cerebrovascular events, such as MONICA, the age range chosen was 35 to 64 years. The EUROCISS Project suggested a wider age-range, 35 to 74 years, or even up to 84 years, considering that more than half of the events occur in patients aged 65 or more; it is recommended to present morbidity and mortality data divided in decennial groups, in particular the age groups 35 to 44, 45 to 54, 55 to 64, and 65 to 74.
- for registries covering cerebrovascular events, such as MONICA, the age range chosen was 25 to 64 years. The EUROCISS Project suggested a wider age-range, 35 to 74 years, or even up to 84 years, considering that more than half of the events occur in patients aged 65 or more; it is recommended to present morbidity and mortality data divided in decennial groups, in particular the age groups 35 to 44, 45 to 54, 55 to 64, 65 to 74, and, if possible, 75 to 84;
- for registries covering type-1 diabetes, such as the RIDI (Registro Italiano Diabete mellito Insulino-dipendente), the age range chosen was 0-29 years, but the difficulty in retrieving data may justify a shorter interval (0-14 years). The shorter interval and age range up to 14 years is suggested by the fact that IDDM (Insulin-Dependent Diabetes Mellitus) is most common in the years close to puberty and the incidence can remain high in later years. The recommended age groups to present results are quinquennial; in particular the age ranges 0 to 4, 5 to 9, and 10 to 14.

Age-standardised rates are recommended and the European standard population should be used as reference.

- *Sex*

Differences in incidence and mortality between men and women are well documented in literature for most diseases. Therefore, it is important that the same high quality data collection methods are applied to both women and men.

- *Population size*

The size of the population under surveillance is determined by the number of events. The number of events is determined by the definition of the event and the event rate in the age groups included. In most cases, the population size has to be determined on the basis of mortality statistics. For example, the age-specific mortality rate of ischemic heart diseases is greater for men than for women. This means that, in order to estimate attack rates in middle-aged subjects with the same degree of precision, the population observed should be larger for women than for men. Usually, to be eligible to participate in an Acute Myocardial Infarction/Acute Coronary Syndrome (AMI/ACS) population-based register, a minimum amount of 300 coronary fatal events (men and women together) per year is necessary in the population groups aged 45 to 74 years. This minimum amount has been established to detect a 2% mortality trend decrease in event rates per year, taking into account that the population range under surveillance could vary from approximately 1,800,000 (all ages) in a low incidence country such as Italy and 200,000 (all ages) in a high incidence country such as Finland. This calculation is based on female attack rates, which are usually lower than male attack rates. If more than one area is enrolled, the same number of 300 total events should be considered for each single enrolled area.

### Population size calculation

See the power point presentation ‘Step 1 - Population size calculation’ available on the web site [www.cuore.iss.it](http://www.cuore.iss.it); website address: <http://www.cuore.iss.it/eng/bridge/WP8/TrainingPopulation-basedRegistries.asp>

In the planning of a surveillance programme, it is important to consider the required population size that would allow to obtain reasonably precise estimates of incidence rates, for instance. In this context, it will be necessary to take into account the most basic rate comparisons in order to meet the aims of the programme. In general, this would concern evaluations of changes in rates over time and population differences in rates. Here below, we present two different approaches that can be used to determine the required population size: A) Hypothesis testing approach; and B) Confidence interval approach. The calculations are illustrated by a worked example.

#### A) Hypothesis testing approach

Under the hypothesis of a given annual percent change in the attack rate, this approach allows to calculate the necessary population size based on a Poisson probability function where the minimal number of events to be registered per year is given by the following relation:

$$\text{Number of events per year} = X / k = 2 / k^3 * [(\Phi^{-1} (1-\alpha/2) + \Phi^{-1} (1-\beta)) / (t / 100)]^2$$

where

X = indicates the number of events over k years;

$\alpha$  = significance level;  $1-\beta$  = statistical power;

t = indicates the attack rate percent change per year;

$\Phi^{-1}$  = is the inverse of the Poisson probability distribution.

For example, for an 80% probability ( $1-\beta$ ) of detecting a 2% change in event rate per year over 5 years statistically significant at the 5% level ( $\alpha$ , two tailed test), the annual number of events needed is approximately 300:

$$\text{Number of events per year} = X / k = 2 / 5^3 * [(1.96 + 0.84) / (2 / 100)]^2 = 314$$

To give an example, Table 3 shows the numbers of events to be collected per year to have 80% probability of detecting a 2% or 1% change in attack rate per year over 10 years, significant at the 5% level (two tailed test), for men and women aged 45-74, for Coronary and Cerebrovascular events separately. In the table, estimated population sizes are given for a low CVD incidence country (Italy) and a high CVD incidence country (Finland); Coronary and Cerebrovascular attack rates used for the calculations derive from the Italian Progetto CUORE ([www.cuore.iss.it](http://www.cuore.iss.it)), and the Finnish National Cardiovascular Disease Register.

In Table 3, the column of the 'Events' shows the number of events to be collected per year to satisfy the chosen parameters; the following two columns indicate the country specific crude attack rates used to estimate the minimal numbers; then the number of men and women to be taken under surveillance in the country-specific population, calculated on the basis of events to be collected and country-specific attack rates; the following column shows the required total population size based on the number of men and women respectively, using the European standard population structure; finally, the last column shows the correspondent total population size to monitor after 10 years, under the assumption of a constant decrease, in order to maintain statistical power.

### *B) Confidence interval approach*

An alternative approach to test hypothesis and to estimate the population size to be monitored is based on the width of confidence intervals: a not too wide confidence interval could be requested. Given the hypothesis that the purpose of the registry is to estimate attack rates and change in attack rates over time rather than testing a predefined hypothesis, this approach might be appealing. This approach is mainly based on the balance between two competing parameters: the confidence level and the interval width. An increased confidence level will correspond to an increased interval width, and this means less information about the true rate. Given the confidence level and the interval width, it is possible to determine the related minimal population size. In a large population, or for incidence rates not too small, the Poisson probability distribution can be approximated by the Normal distribution; in this case, the estimation of the minimal population size (N) can be calculated using the following relation:

$$N \geq (2Z_{\alpha/2})^2 p(1-p) / w^2$$

where

$p$  = attack rate estimate;

$p(1-p)$  =  $\sigma$  = standard deviation estimate;

$\alpha$  = significance level; in this context a factor specified by the confidence level, e.g.  $\alpha=0.05/2$  would correspond to a 95% confidence interval;

$z$  = refers to the use of the standard Normal distribution to derive probabilities;

$w$  = the chosen absolute interval width.

For example, in a large population with an attack rate of 44.1 / 10,000, given the significance level of 5% ( $\alpha$ , two tailed test), and an absolute interval width of 20% of the attack rate, the minimal population size needed is approximately 87,000:

$$N \geq (2 * 1.96)^2 * 0.00441 * (1 - 0.00441) / (0.00441 * 20 / 100)^2 \geq 86,727$$

The estimation of the required population size to monitor time trends in event rates is important, and the results may limit the number of possible areas able to produce stable trend estimates. What matters is the annual number of events, and not the population size; in high attack rate countries, smaller populations can be studied, whereas larger populations would be needed in low attack rate countries. The limitations derived by the use of less than ideal population sizes could be reduced by:

- i) accepting a threshold for the annual rate of change higher than that used in the example considering a 2% per year. This would be relevant to areas with low but rapid rates;
- ii) increasing alpha and beta to lower the sample size. This would lower the power below 80% and/or increase  $\alpha$ , the significance level, from 5% to 10%;
- iii) pooling:
  - (a) results from age groups under the age of to 25 (small effect on numbers);
  - (b) results from age groups above the age of 74 (large effect);
  - (c) combining data from both sexes (moderate effect);
  - (d) combining data from two or more geographically separated areas within the same country to establish trends, and study them separately for other purposes;
  - (e) combining data within collaborative projects for centres in different countries, matched for certain characteristics such as initial event rates, risk factor trends, socio-economic characteristics, or health services.

While data pooling will increase numbers, it may conceal some important information.

It is recommended that the minimum period of observation is one complete calendar year, because of possible seasonal variations.

TABLE 3 - Minimal size of low and high risk population under surveillance required for fatal and nonfatal coronary and stroke events, ages 45-74 years

	<i>Attack Rate percent variation (t %)</i>	<i>Events</i>	<i>Attack rate (x 10,000)</i>		<i>Male and Female population required according to gender specific attack rates</i>		<i>Total population required using EU standard population structure</i>		<i>Total population required after 10 years under the assumption of continuous attack rate decrease</i>	
			<i>Men</i>	<i>Women</i>	<i>Male population</i>	<i>Female population</i>	<i>Total pop based on MEN</i>	<i>Total pop based on WOMEN</i>	<i>Total pop based on MEN</i>	<i>Total pop based on WOMEN</i>
<b>2%</b>										
<i>Total Coronary Events Attack rates</i>										
	<b>Italy</b>	314	44.1	12.8	71,192	245,277	444,948	1,532,984	544,563	1,876,191
	<b>Finland</b>	314	272.7	116.9	11,512	26,846	71,948	167,789	88,056	205,354
<i>Total Cerebrovascular Accidents Attack rates</i>										
	<b>Italy</b>	314	33.5	20.3	93,718	154,658	585,737	966,611	716,873	1,183,017
	<b>Finland</b>	314	112.0	61.2	28,044	51,317	175,276	320,730	214,517	392,536
<b>1%</b>										
<i>Total Coronary Events Attack rates</i>										
	<b>Italy</b>	1256	44.1	12.8	284,767	981,110	1,779,791	6,131,937	1,967,964	6,780,251
	<b>Finland</b>	1256	272.7	116.9	46,047	107,385	287,794	671,157	318,222	742,116
<i>Total Cerebrovascular Accidents Attack rates</i>										
	<b>Italy</b>	1256	33.5	20.3	374,872	618,631	2,342,949	3,866,443	2,590,663	4,275,232
	<b>Finland</b>	1256	112.0	61.2	112,177	205,267	701,104	1,282,921	775,229	1,418,560

- *Patient eligibility*

A patient is considered eligible for inclusion in a population-based registry only if he/she is resident in the area under surveillance, meets the selected criteria and has had an event within the defined surveillance time period.

In the Italian pilot Registry of Coronary events, a subject is considered eligible for inclusion in the population-based AMI/ACS register only if he/she is resident in the areas under surveillance (7 well-defined geographical areas of the country), meets the selected age criteria (35-74 years), and had an AMI/ACS event within the defined time period (2 calendar years).

In the Italian pilot Registry of Cerebrovascular events, individuals are considered eligible for inclusion in the stroke population-based register only if they are resident in the areas under surveillance (8 well-defined geographical areas of the country), meet the selected age criteria (35-74 years), and had a stroke event within the defined time period (2 calendar years).

In the RIDI type-1 diabetes register, a subject is considered eligible for inclusion if he/she is resident in the area under surveillance (15 well-defined areas of the country), meets the selected age criteria (0-14 years) and had a diagnosis of insulin-dependent diabetes mellitus according to WHO classification, within the defined time period of registry implementation.

**Geography** It is the selection of a geographical administrative area with a population big enough to provide stable estimates. This means, as reported in 'Population size', that a stable population in a representative area of the country with 300 fatal events in the age range 45 to 74 should be chosen in the case of coronary events under the hypothesis of a 2% annual reduction of attack rates in 10 years. If more areas are under surveillance, 300 fatal events should be considered for each separate area.

Population-based registries are established to identify regional differences in incidence rates to better understand the reasons of these differences.

The Surveillance Epidemiology and End Results (SEER) Programme of the National Cancer Institute was a population-based Registry established in 1972 and covering 11 separate geographic areas of the US.

The MONICA Registry for Acute Coronary events was established in 37 populations of 20 countries (16 European countries, Canada, USA, New Zealand, and China).

The MONICA Registry for Stroke events was established in 25 populations of 11 countries (10 European countries and China).

The Italian pilot Registry on Coronary events was implemented in 7 areas in the North, Centre and South of the country (1 Region, 4 Provinces, 1 metropolitan sub-area, and 1 multi-province area) which were representative of the overall population.

The Italian pilot Registry on Cerebrovascular events was implemented in 8 areas in the North, Centre and South of the country (2 Regions, 4 Provinces, 1 metropolitan sub-area, and 1 multi-province area), which were representative of the overall population.

The RIDI type-1 diabetes register was implemented in 9 Regions and 6 Provinces in the North, Center and South of the country.

As computer systems increase elaboration capacity, the quality of health administrative data improves (specifically mortality and HDR databases), and ethical and privacy issues are faced and solved for public health purposes by specific laws. In this framework, the implementation of a registry at national level can be taken into account. For example, in the Italian pilot registries of coronary and cerebrovascular events, the record-linkage to identify current events can be implemented by using mortality and HDR databases at national level, while Positive Predictive Values (PPVs) can be estimated by validating random samples of events in some well-defined areas of the country.

The availability of resident population should be checked at mid-year or in inter-census estimates, according to age groups (quinquennial or decennial) and sex. This is important to assess population health indicators such as incidence rate. The numbers of resident population are provided by municipal records.

A demographic characterisation shall be done for the population under surveillance and shall include a detailed description of this population; in particular, it should consider the following elements: demographic characteristics (age and gender distribution); socio-cultural characteristics (educational level, occupation, social group, unemployment rate, migration, immigrants with or without citizenship); characteristics of the healthcare system (specialised hospitals, GPs, rehabilitation clinics); macro (North, Centre, South) and micro areas (regions, or urban and rural areas). Disease frequency is often different in macro areas of the country; a description of differences in mortality and risk factors allows to select those areas to be included in the surveillance system. In a population-based surveillance, the phenomenon of immigration plays an important role, therefore immigrants coming from European and extra-European countries and resident in the study area should be enrolled. Geographical or administrative borders of the surveillance areas should be clearly defined.

**Duration** It should reflect the length of time during which the registry is expected to collect data in order to achieve its purpose and provide analyses of the data collected. For NCDs (i.e., Coronary heart disease, Stroke, Cancer) a long duration (about 10 years) is needed to estimate survival. In alternative, a biannual registry could be implemented at five-year or ten-year intervals to estimate attack rate and case fatality. This two-year duration is also sufficient to estimate PPVs to be applied every 5 years to estimate registry indicators. During the 5-year intervals, the estimated PPVs can be applied to the current events that occur every year, obtained through record-linkage of mortality and HDR, in order to estimate occurred events and provide estimates of attack rates and case fatality. In the following 2-year period, in addition to the record-linkage of mortality and HDR performed on an annual basis, complete information on a sample of



suspected events is collected from clinical records in order to validate the events and estimate the updated PPV to be applied in the following 5-year period. This procedure is then repeated on a regular basis.

**Setting** It refers to the specific setting through which the registry will collect data (e.g. administrative databases, clinical records, GPs records). In the simplified methodology, mortality records and Hospital Discharge Records databases related to a well-defined area are the starting point for data from which a Coronary events, or Cerebrovascular events, or Cancer registry can be implemented. In addition to these databases provided by the national/regional health system, accessibility is requested to clinical charts in the hospitals of the registry area, as well as administrative health databases and GP records.

### Data sources

*See the power point presentation 'Step 1 - Sources of Information' available on the web site [www.cuore.iss.it](http://www.cuore.iss.it); website address: <http://www.cuore.iss.it/eng/bridge/WP8/TrainingPopulation-basedRegistries.asp>*

The availability and accessibility of different data sources necessary to record-linkage, the identification of current events and the collection of all necessary information must be checked. In some regions, databases are available upon payment; in others, databases are free of charge and accessible upon presentation of a proposal to be submitted and approved by the ethical committee; usually mortality data are accessible at the National Institute of Statistics with a delay of few years, or at the National Public Health Institute. Other databases, such as HDR, are available at the Ministry of Health, which distributes them on request, in an anonymous form, or at regional level in a detailed form.

Existing hospital discharge and mortality data and other sources of health care data (e.g., emergency care, specialist visits, diagnostic service, drugs databases) must be analysed. Events occurring in the study area but involving non-residents or non-residents admitted to hospital in the study area do not qualify. Events involving residents of the study area but occurring out of the study area do qualify. Efforts must be made to find them or estimate their potential loss, and whether or not this loss could modify and interfere with the validity of the observed rate trends over the years.

Problems with these data must be identified: data coverage, ICD version and classification systems of health data, identification of events, procedures, unit of analysis (event or patient), PIN, availability of previously conducted epidemiological studies to assess data coherence. Data files are usually available in detailed form at regional level.

- *Data sources for population-based registries*

Data sources used in registries depend on the type of disease under surveillance and related-disease events; for example, to monitor NCDs in the general population, at least

the following sources of information should be available: mortality records with death certificates; HDRs with clinical records, autopsy registry, nursing home and clinic, emergency and ambulance services, GP clinical records, drug dispensing registry, exemptions.

### *Death certificates*

Death certificates provide complete data on fatal events and are collected in a systematic and continuous way in all countries. Mortality statistics are easily accessible but are usually published in a detailed and complete form after 2-4 years.

The format of death certificates varies across countries: it generally includes personal identification data, date and place of birth, date and place of death (i.e. nursing home, hospital, home), residence, and underlying cause of death (only few countries report underlying, immediate and contributing causes of death). Death is coded according to the International Classification of Disease and causes of death. Problems of temporal and geographical comparisons derive from the different versions of the ICD adopted over time (7<sup>th</sup>, 8<sup>th</sup>, 9<sup>th</sup>, 10<sup>th</sup> revision) and from different coding practices in each country. Furthermore, diagnostic criteria for coding death certificates are not defined at international level and the ICD versions are updated every 10 years by WHO.

The reliability of mortality data depends on the completeness and accuracy of the vital registration system, as well as on the completeness and accuracy of the registration and coding of causes of death. When the proportion of deaths coded as “unknown cause of death” is higher than 5%, cause-specific mortality data should be used with caution. The accuracy of the recorded causes of death depends on the autopsy rate. This rate varies largely among countries and over time. In some countries, the autopsy rate has declined in recent years, which may be a problem for the use of mortality statistics of some diseases. Data may be aggregated without detailed information on age and sex distribution and without separate diagnostic categories (e.g. all cardiovascular diseases, instead of AMI, stroke, heart failure, atrial fibrillation).

### *Hospital discharge and medical records*

HDRs give the number of hospitalisations in a definite period of time, which are important for purposes related to management of resources, costs and health services. Hospital discharge data are available in most EU countries; data may be aggregated without detailed information on age and sex distribution and without separate diagnostic categories (e.g. all cardiovascular diseases, instead of AMI, stroke, heart failure, atrial fibrillation).

HDRs include personal data, admission and discharge dates, type of hospitalisation (urgent, ordinary or transfer from other structures), main procedures (e.g., surgical intervention), discharge diagnoses, condition at discharge (at home, transfer to other structure, dead), and information used for reimbursement (e.g., Diagnosis-Related Group-DRG). Hospital discharge diagnoses are coded by ICD codes (currently ICD-9 or ICD-10). For some countries, only a limited number of diagnoses are coded. Use of HDRs and medical records can generate some difficulties: a) problems in assessing a specific

event may arise when an acute event is followed by a period of rehabilitation in a different hospital or clinic, or a transfer to other wards, as the event could be counted more than once; b) Clinical information and medical records reported in hospital documents must be seen and renewed for event validation; c) HDRs are not always validated on a routine basis, and validation studies are necessary to check their diagnostic quality; d) the validity of HDRs may vary according to the geographical region, the type of hospital or clinic and patient characteristics; e) hospital admission policies may vary over time and place, e.g. the registration of the most severe cases, or of the deaths occurred shortly after arrival to hospital may differ among hospitals. HDRs may also include patients not resident in the area under surveillance; f) the adoption of new diagnostic techniques may cause major changes in event rates estimated from HDRs; g) a further problem may derive from the use of DRG. In some countries, hospital reimbursement is based on the DRG tariff system, which is built on equal-resources criteria and aggregates events in major diagnostic categories; i) in order to assess the occurrence of events, HDRs from all hospital departments of the geographical area under surveillance should be used.

#### *Autopsy registry*

Not all countries routinely perform an autopsy on suspected or sudden deaths. Autopsy is performed on violent deaths or on deaths occurring in hospital when clinical diagnosis is undetermined. The former is performed by forensic medicine specialists, the latter by pathologists of the hospital where the death occurred. Data from autopsy registries refer therefore to a low percentage of deaths, but provide a more valid diagnosis to complement the information reported on the death certificates.

#### *Nursing homes and clinics*

Nursing home and clinics mainly provide data on cases among elderly patients who sometimes get care from these institutions without being admitted to a hospital. Therefore, information on events occurring in the nursing home may be critical, especially if the registry covers elderly patients up to 84 years of age.

In some countries, rehabilitation after an acute event is provided by the rehabilitation clinic, which may give information on patients who have received acute care outside the region.

#### *Emergency and ambulance services*

Data provided by emergency and ambulance services is useful to integrate information for registry implementation, since patients dying from sudden death or experiencing fatal events are not always able to reach the hospital. These services can provide data otherwise not obtainable, such as ECG during the acute phase of the event, blood pressure measurements, level of consciousness and muscular deficit at the time of event occurrence in pauci-symptomatic patients recurring to emergency services. The need for very urgent medical treatments often makes information partial, but the integration of this data with other data from other sources of information contributes to the implementation of the registry and event validation.

### *General practitioner medical records*

GPs provide information on those events that do not reach the hospital and for those patients who are hospitalised outside the area of their usual residence; GPs can provide clinical data and thus integrate information from other sources (HDR, death certificate, etc.). GPs datasets may also provide an adequate coverage for prevalence of NCDs. This network operates in a few countries (e.g., the Netherlands, UK, and Italy).

The GPs network may be affected by selection bias, as usually only volunteer GPs participate in studies. For this reason, data from GP network requires integration with other different sources of information and validation.

### *Drug dispensing registries*

In some countries, patients may receive comprehensive drug reimbursement under their national healthcare system, so drug prescriptions can serve to complete the information on disease.

### *Other sources*

Disease exemption, outpatient visits, laboratory tests, radiological and anatomic-pathological records. The more linked are the sources of information, the lower shall be the probability of missing cases.

## **Examples**

In the **Italian pilot registry of Coronary events**, the main sources of information are mortality databases, HDR databases for AMI/ACS event identification, and clinical records for AMI/ACS event validation. For fatal AMI/ACS events occurred out of hospital, questionnaires administered to families and GPs are used to collect information on the event validation.

In the **Italian pilot registry of Cerebrovascular events**, the main sources of information are mortality databases, HDR databases for stroke event identification and clinical records for stroke event validation. For fatal stroke events occurred out of hospital, questionnaires administered to families and GPs are used to collect information on the event validation.

In the **RIDI type-1 diabetes registry**, the main sources of information used to identify type 1 diabetes cases are 1) hospital discharge card or clinical record or card of the diabetes centre; 2) health card ticket for diabetes; 3) prescription of insulin syringes or other type of health care for homeopathic diabetes control; 4) reporting by the family doctor (GPs); 5) Patient Associations.

In any case, at least two independent sources must always be used. Two sources are said to be independent when the likelihood of recruiting a case with the primary source does not alter the likelihood of recruiting the same case with the secondary source: for example, source 1) clinic folder of diabetes centre and source 5) patient association.

- If entries in the association are made on a case by case basis by the diabetes centre, then, in this case, the two sources are not independent;
- If, on the other hand, memberships are made on the advice of the family physician and/or through population awareness campaigns activated by the Associations themselves, etc., then the two sources are independent.

## STEP 2 - Carry out record-linkage of administrative data

See the power point presentation ‘Step 2 - Record-linkage procedures’ available on the web site [www.cuore.iss.it](http://www.cuore.iss.it); website address: <http://www.cuore.iss.it/eng/bridge/WP8/TrainingPopulation-basedRegistries.asp>

In the Northern countries, where every citizen has a PIN included in national registries of hospital discharges and deaths, record-linkage for the identification of events is efficient and reliable. For countries that have not adopted the PIN, it may be much more difficult to perform this step. Files have to be organised with the same format and have to include the same personal variables needed to univocally identify subjects (family name, first name, date of birth, place of birth, residency).

**Core data set** A quality assurance plan (see section 5 ‘Quality control’) should be considered when developing a core dataset. In a registry, population, mortality and HDR databases are needed to identify current events in the resident population of well-defined geographical areas. In addition, detailed information from clinical records is needed for the validation of events.

Death Certificates and Hospital Discharge Record databases provide main information for record-linkage implementation.

For example, the Core Data Set, from which to start for the identification of events by using the record-linkage, included:

**Table 4 - Death Certificates database**

Name of the field	Type of data	Dimension (characters)	Description and format availability at MoH (paper/electronic)
Family Name	Text	50	The record is in electronic format
First Name	Text	50	The record is in electronic format
Date of birth	Date	dd/mm/yyyy	The record is in electronic format
Place of Birth	Text	6	The data should be a code issued by the National Institute of Statistics. It is not recorder neither in paper nor in electronic format
Sex	Text	1	1=male; 2= female

Place of living	Text	6	The data should be a code issued by the National Institute of Statistics. It is not recorder neither in paper nor in electronic format
Date of death	Date/hour	dd/mm/yyyy	The record is in electronic format
Place of death	Text	6	The record is in electronic format as the full name of the place where the person died
Died at	Text	1	1=home; 2=health infrastructure (public or private hospital); 3=other The record can be obtained although it is not directly coded in the death certificate
Main cause of death	Text	4	ICD-9 or ICD-10 code of the main cause of death; the record is in electronic format
Secondary cause of death	Text	4	ICD-9 or ICD-10 code of the secondary cause of death; the record is in electronic format
Initial cause of death (*)	Text	4	Code of the initial cause of death
Intermediate cause of death (*)	Text	4	Code of the intermediate cause of death
Final cause of death (*)	Text	4	Code of the finale cause of death
Age	Numeric	Byte	The record is in electronic format
PIN (Personal Identifying Number)	Text	16	This individual code is registered in electronic format in the death certificate starting from August 1. 2012

(\*) The Italian death certificates specify the initial, intermediate, and final cause of death

**Table 5 - Hospital Discharge Record database**

Name of the field	Type of data	Dimension (characters)	Description and format availability at MoH (paper/electronic)
Hospital code	Text	6	The name of the hospital (not the hospital code) is in electronic format
Ward code / Med. Div. code	Text	3	Code of the in-patient ward; the record is in electronic format
Clinical record number	Text	8	Number of the clinical record; the record is in electronic format
Admission date	Date/hour	dd/mm/yyyy	The record is in electronic format

ID personal code	Text	16	The PIN is requested at admittance and recorded in paper format, but not in electronic format
Family Name	Text	50	The record is in electronic format
First Name	Text	50	The record is in electronic format
Sex	Text	1	1=male; 2= female
Place of birth	Text	6	The data is not recorded
Date of birth	Date/hour	dd/mm/yyyy	The data is not recorded in the Hospital Discharge Form. It may be obtained from the date of birth and the date of admission
Age	Numeric	Byte	The record is in electronic format
Place of living	Text	6	Patient's address record is in electronic format (not the code of the city/village he is living)
Kind of admission	Text	1	1=ordinary; 2=urgent;
Discharge date	Date/hour	dd/mm/yyyy	
Kind of discharge	Text	1	1=ordinary; 2=voluntary; 3=patient transferred to another hospital; 4=died
clind_1 (*)	Text	4	clinical diagnosis ICD-9 or ICD-10; main diagnosis
clind_2 (*)	Text	4	clinical diagnosis ICD-9 or ICD-10; 2° diagnosis
clind_3 (*)	Text	4	clinical diagnosis ICD-9 or ICD-10; 3° diagnosis
clind_4 (*)	Text	4	clinical diagnosis ICD-9 or ICD-10; 4° diagnosis

(\*) Usually Hospital Discharge Record includes also secondary, tertiary, and a fourth ICD-9 or ICD-10 diagnoses.

It is recommended to:

- explore the feasibility of record-linkage within Hospital Discharge Records-HDRs - *deterministic* or *probabilistic* approach based on personal variables or PIN use (within the same hospital, among hospitals of the area under surveillance, among hospitals at regional level). When hospital records are collected at national level, it is possible to include also those non-fatal events occurring out of the surveillance area. This activity is crucial to detect and fix all HDRs related to the same subject;
- explore the feasibility of record-linkage within mortality records - *deterministic* or *probabilistic* approach based on personal variables or using PIN within the area under surveillance or at regional level. When mortality records are collected at national

level, it is possible to include also those fatal events that occur out of the surveillance area. This activity is crucial to detect and fix possible duplication of death records related to the same subject;

- The record-linkage *deterministic* approach implies that all the variables that univocally identify a subject should exactly correspond, for the same subject, in each used source of information: e.g., in the comparison between mortality and HDR databases to identify hospitalised fatal cases, or to identify all the hospital discharges related to the same subject in the same HDR database;
- The record-linkage *probabilistic* approach implies that the identifying variables in different sources of information shall match, with the exception of a digit, or two digits, and so on. This means that the higher the number of digits we accept as exception, the greater will be the number of matching records, but the detection of current events will become more difficult and less reliable;
- a propaedeutic but unavoidable activity has to be conducted before record-linkage: it consists in accurately checking and cleaning both mortality and HDRs administrative databases from possible errors in identifying variables used for record-linkage (family name, first name, date of birth, place of birth, residency). There is no specific software allowing to compare some sets of specific variables in different records belonging to a same databases or different databases; or providing matching records and identifying similar records that differentiate for just 1, 2, or 3 digits; or specifying the kind and types of non-correspondence.

### Procedure to detect mistakes in subject identification variables and to eliminate duplications in databases

See the power point presentation ‘Step 2 - Detecting mistakes in subject identification variables and cleaning databases’ available on the web site [www.cuore.iss.it](http://www.cuore.iss.it); website address: <http://www.cuore.iss.it/eng/bridge/WP8/TrainingPopulation-basedRegistries.asp>

The procedure here suggested is time consuming but reliable; it is applicable when limited databases are used (e.g. databases referred to persons who are resident in a well-defined region, town, geographical area, or for a limited age range, instead of the whole country and overall population). The procedure should be repeated to compare the records included in the same database (e.g., singularly in mortality database to detect duplications; singularly in HDR database to identify all records pertaining to the same subject); then the procedure should be repeated to compare records between different databases (e.g., between mortality and HDR databases to identify deaths occurring in hospital, so that they will not be counted twice; or to identify the correct subject identifying information to be saved). The procedure can be described as follows:

a) in both mortality and HDR databases, all records should be ordered according to the variables used to identify the subject (family name, name, sex, date of birth,



residence). Family name and name will follow an alphabetic order; sex is dichotomic; date of birth and code of residence will follow a time and numeric order respectively;

b) all ordered records shall be reviewed to compare each record to the ones immediately preceding or following it; this will allow to easily identify 'micro' differences (incongruences) in the subject identification variables. Incongruences for the same person in different records can refer to a letter of the name or family name, a different gender code, a different day or month or year of birth, a different residential code;

c) when incongruences are limited to only one of the 5 variables used to identify the subject, it is more plausible that a mistake has occurred and the two compared records are referred to the same person; instead, when incongruences are more than one, it is probable that the compared two records are referred to different subjects. In this case all remaining variables of the two records should be compared to establish whether records are duplicated or whether they show two different episodes pertaining to the same subject or whether records are referred to different subjects. In the first case, one of the two records should be deleted; in the second case, both records should be maintained for the same subject; in the third case, both records should be maintained, but related to different subjects .

d) independent, external information should be used to establish the correct information about the identification variables of the subject (correct name, family name, date of birth, etc.) to be saved.

This procedure is necessary to avoid possible double counting of the same record or, on the contrary, to avoid a missing record-linkage between corresponding records due to possible errors in identifying variables. These kinds of errors can considerably bias results, since they influence the identification of the first event, the dates of the first and recurrent events, and consequently the number of events for the same subject and for the overall population included in the registry. Overall errors in mortality and HDR databases can range between 2% and 30% in different centres. This can considerably bias the results of the registry if databases are not accurately cleaned before performing the record-linkage. Quality and reliability of administrative databases can continuously improve by the use and checking of data.

The main limitations of record-linkage are the difficulty in obtaining administrative files for research purposes: mortality data files are usually available at the National Institute of Statistics, whereas hospital discharge data are available at the Ministry of Health. These kinds of data are anonymous and therefore do not allow record-linkage. Nominal files of both mortality and hospital discharge records are available at the regional level or at the health unit level. When combining databases, missing events are mainly explained by PIN or name errors and these errors lead to unsuccessful record-linkage. Record-linkage is important also to define and obtain minimal data sets (for mortality: PIN; family and first name; date and place of birth; gender; residence; date and place of death; underlying and secondary causes of death. For hospital discharge diagnosis, the

same variables should be considered, with the addition of admission date and hospital discharge diagnoses).

### **STEP 3 - Perform a pilot study and validate routine data**

*See the power point presentation 'Step 3 - Population-based registries of coronary and cerebrovascular events: the Italian pilot example' available on the web site [www.cuore.iss.it](http://www.cuore.iss.it); website address: <http://www.cuore.iss.it/eng/bridge/WP8/TrainingPopulation-basedRegistries.asp>*

Before starting a registry or a large scale use of linked administrative data, it is recommended to perform a pilot study in a small area on available hospital discharge and mortality data and on other available health care data.

Validation studies on available data include:

- Coverage estimates: comparison of different routine data sets (electronic or manual), number of patients treated in and out-of-area, hospital/mortality ratios, age and gender ratios, principal vs. secondary diagnoses and/or procedures;
- Validation of discharge diagnoses and every diagnosis reported in health care data sets, according to standard methodological/diagnostic criteria (including revision and abstraction of medical records) in a random sample or in all cases;
- Validation of mortality causes according to standard methodological/diagnostic criteria in a random sample or in all cases; analysis of the demography and representativeness of the area in comparison with the region or country;
- Selection of the age range of interest.

**Events** The definition of an event is of great importance. Standardised epidemiological definition of events should be adopted as far as possible, taking into account specific operative conventional agreements. Methods to ascertain events should be clearly established; diagnostic criteria, level of data details, and level of data validation should also be addressed. Event ascertainment methods must be considered when evaluating data registry, in particular as concerns the sensitivity indicator (the extent to which the methods identify all events of interest) and the external validity indicator (generalizability to similar populations).

### **Identification of events**

For many NCDs (e.g., Coronary Heart Disease, Stroke, Cancer), the event identification starts with the record-linkage of individual data from HDRs and mortality databases, according to the selected ICD codes reported in hospital discharge diagnoses and in the

causes of death; record-linkage is necessary to avoid double counting of deaths in hospital for which both HDR and death certificate are available. This implies that, in order to avoid erroneous or multiple identification of events, possible errors in individual identification variables must be removed before implementing the record-linkage.

**EUROCISS** recommends to identify **coronary major events** starting from record-linkage of mortality and HDRs databases:

- for mortality databases, by selecting the Underlying cause ICD-9 codes 410-414 (Ischaemic Heart Diseases) or 798-799 (Sudden death, Other ill-defined and unknown causes of morbidity and mortality) directly; or ICD-9 codes 250 (Diabetes), 401-404 (Hypertensive disease), 420-429 (Other forms of heart disease), 440-447 (Atherosclerosis, Aortic and other aneurysms, Other peripheral vascular disease, Arterial embolism and thrombosis, Polyarthritits nodosa and allied conditions, Other disorders of arteries and arterioles) if ICD-9 codes 410-414 are reported in at least one of the secondary causes of death (corresponding ICD-10 codes in the Underlying cause of death are I20-I25, R96-R99 directly selected; E10-E11, I11-I13, I30-I51, I70-I78 if ICD-10 codes I20-I25 are reported in at least one of the secondary causes of death) (Figure 5);
- for the HDR data base, selecting ICD-9 codes 410-414 or ICD-10 codes I20-I25 in any of the first 4 hospital discharge diagnoses. Day-hospital admissions and hospitalizations lasting less than 3 days are not considered (Figure 5).

If a patient experiences further acute symptoms suggestive of AMI/ACS within 28 days (where onset is day one and a new AMI/ACS occurring after 28 days is a new event) from the onset of a first episode, this second episode is not counted as a new AMI/ACS event. Equally, if a patient experiences further acute symptoms suggestive of AMI/ACS after 28 days (as stated above) from the onset of a first episode, this second episode is counted as a new event.

On these bases, events are identified as:

- Non-fatal AMI/ACS event, surviving at least 28 days from the onset of the AMI/ACS symptoms (Figure 5);
- Fatal AMI/ACS event, causing death within 28 days from AMI/ACS symptoms onset (Figure 5).

It should be noted that each event is registered separately.

**EUROCISS** recommends to identify **stroke major events** starting from record-linkage of mortality and HDRs databases:

- for mortality data base, selecting the Underlying cause ICD-9 codes 342 (Hemiplegia) or 430-434 (Subarachnoid haemorrhage, Intracerebral haemorrhage, Other and

unspecified intracranial haemorrhage, Occlusion and stenosis of precerebral arteries, Occlusion and stenosis of cerebral arteries) or 436-438 (Acute, but ill-defined, cerebrovascular disease; Other and ill-defined cerebrovascular disease; Late effects of cerebrovascular disease) directly; or ICD-9 codes 250 (Diabetes mellitus), 401-404 (Hypertensive disease), 427 (Cardiac dysrhythmias), 440 (Atherosclerosis) if ICD-9 codes 342 or 430-434 or 436-438 are reported in at least one of the secondary causes of death (corresponding ICD-10 codes in the Underline cause of death are G81 or I60-I69 directly; E10-E14, I10-I13, I46-I49, I70 if ICD-10 codes G81 or I60-I69 are reported in at least one of the secondary causes of death) (Figure 6);

- for the HDR data base, selecting ICD-9 codes 342, 430-434, 436-438 or ICD-10 codes G81, I60-I69 in any of the first 4 hospital discharge diagnoses. Day-hospital admissions and hospitalizations lasting less than 3 days are not considered (Figure 6).

If a patient experiences further acute symptoms suggestive of stroke within 28 days (where the onset is day one and a new stroke occurring after 28 days is a new event) from the onset of a first episode, this second episode is not counted as a new stroke event. Equally, if a patient experiences further acute symptoms suggestive of stroke after 28 days (as stated above) from the onset of a first episode, this second episode is counted as a new event.

On these bases, events are identified as:

- Non-fatal stroke event, surviving for at least 28 days after the onset of the stroke symptoms (Figure 6);
- Fatal stroke event, causing death within 28 days from symptom onset (Figure 6).

It should be noted that each event is registered separately.

In the **RIDI type-1 diabetes registry**, diabetes cases are identified from different sources of information, as follows:

- diagnosis of insulin-dependent diabetes mellitus (IDDM) according to WHO classification
  - Fasting plasma glucose concentration  $\geq 126$  mg/dl (7.0 mmol/l), OR
  - 2hrPPG (2-Hours Post-Prandial Glucose)  $\geq 200$  mg/dl (11.1 mmol/l) after a 75-g glucose load 2-Hours Post-Prandial Glucose.

The date of first insulin administration should be considered as the date of diagnosis (1<sup>st</sup> insulin administration date = diagnosis date)

### Event validation

After identifying 'potential' events by using ICD codes and record-linkage, complete validation should be performed for a sample of them, in order to estimate the PPV for each included ICD code. Event validation is a process that allows to categorise events

according to standardized diagnostic criteria, based on all information on the event, collected from clinical charts (in case of hospitalization) or from family or GP interviews (in case of sudden death without hospitalization).

Here below, some examples of diagnostic criteria for Acute Myocardial Infarction, Stroke and type-1 Diabetes Mellitus are described.

### Diagnostic criteria for Acute Myocardial Infarction/Acute Coronary Syndromes (AMI/ACS)

#### 1. MONICA-MONItoring trends and determinants of Cardiovascular disease (1983-84)

[WHO Monica Project: MONICA manual. Part IV: Event Registration. <http://www.ktl.fi/publications/monica/manual/part4/iv-2.htm#s1-1>]

*Diagnostic Classification. There are the following categories:*

- (1) definite acute myocardial infarction*
- (2) possible acute myocardial infarction or coronary death*
- (3) ischaemic cardiac arrest with successful resuscitation not fulfilling criteria for definite or possible myocardial infarction*
- (4) no acute myocardial infarction or coronary death*
- (9) fatal cases with insufficient data, subsequently called "unclassifiable deaths" in collaborative MONICA publications.*

*The allocation of a diagnostic category must strictly follow the definitions provided. The criteria used for the diagnosis of "definite" and "possible" acute myocardial infarction are not necessarily those that would be used by a clinician, but rigid definitions are essential for event analysis.*

#### *(1) Definite acute myocardial infarction*

- a. Definite ECG or*
- b. Symptoms typical or atypical or inadequately described, together with probable ECG and abnormal enzymes, or*
- c. Symptoms typical and abnormal enzymes with ischaemic or non-codable ECG or ECG not available, or*
- d. Fatal cases, whether sudden or not, with naked-eye appearance of fresh myocardial infarction and/or recent coronary occlusion found at necropsy.*

#### *(2) Possible acute myocardial infarction or coronary death*

- a. Living patients: with typical symptoms whose ECG and enzyme results do not place them in category (1) and in whom there is not good evidence for another diagnosis for the attack, or*
- b. Fatal cases whether sudden or not (not in category (1)) where there is no good evidence for another cause of death, clinically or at autopsy.*
  - i. with symptoms typical or atypical or inadequately described, or*

- ii. *without typical or atypical or inadequately described symptoms but with evidence of chronic coronary occlusion or stenosis or old myocardial scarring at necropsy; or*
- iii. *with a good history of chronic ischaemic heart disease such as definite or possible myocardial infarction, or coronary insufficiency or angina pectoris in the absence of significant valvular disease or cardiomyopathy.*

*(3) Ischaemic cardiac arrest with successful resuscitation not fulfilling criteria for definite or possible myocardial infarction*

*Spontaneous cardiac arrest not provoked by medical intervention, electrocution, drowning or other gross physical insults, from presumed primary ventricular fibrillation secondary to ischaemic heart disease, in the absence of significant valvular disease or cardiomyopathy.*

*(4) No acute myocardial infarction*

*a. Living patients (not in category (1))*

- i. with combinations of symptoms and tests that do not qualify them for the definite category and who do not have typical symptoms that might place them in the possible category, or*
- ii. where illness episode has been explained by another diagnosis*

*b. Fatal cases, whether sudden or not, not in category (1) where another diagnosis has been made (clinically or at autopsy)*

*(9) Fatal cases with insufficient data*

*Cases with no autopsy, no history of typical or atypical or inadequately described symptoms, no previous history of chronic ischaemic heart disease and no other diagnosis. Living patients should not be allocated to this category. It is hoped that most centres will not need this category.*

#### **A. Electrocardiographic criteria**

**Code 1: Definite ECG**

*(I) The development of a diagnostic Q wave in serial records*

*-AND/OR-*

*(II) The evolution of a current injury which lasts more than one day.*

*The interpretation of a minimum of two or sometimes three ECG records is therefore necessary for the establishment of these categories.*

**I. Development of Q waves**

*Progression of Q codes from no Q to a diagnostic Q is sufficient, but change from no Q to an equivocal Q or from equivocal to diagnostic Q must be accompanied by deterioration in the ST segment or the T wave. A change in a Q code or in a 4, 5 or 9.2 code must occur within the same lead group but the Q can be in a different lead group to that in which the 4, 5 or 9.2 code is being followed. Note that Minnesota code 1.2.6 is equivalent to No Q code.*

1.1 No Q or QS code in the first ECG record followed by a record with a diagnostic Q or QS code (Minn. code 1.1.1 through 1.2.5 plus 1.2.7)

-OR-

1.2 An equivocal Q or QS code (Minn. code 1.2.8 or any 1.3 code) and no major ST segment depression (No Minn. code 4.1 or 4.2) in the first ECG record followed by a record with a diagnostic Q code PLUS a major ST segment depression (Minn. code 4.1 or 4.2)

-OR-

1.3 An equivocal Q finding and no ST segment elevation (No Minn. code 9.2) in the first ECG record followed by a record with a diagnostic Q code PLUS an ST segment elevation (Minn. code 9.2)

-OR-

1.4 An equivocal Q finding and no major T wave inversion (No Minn. code 5.1 or 5.2) in the first ECG record followed by a record with a diagnostic Q code PLUS a major T inversion (Minn. Code 5.1 or 5.2)

-OR-

1.5 No Q code and neither 4-1 nor 4-2 in the first ECG followed by a record with an equivocal Q code PLUS a 4.1 or 4.2

-OR-

1.6 No Q code and no 9.2 in the first ECG followed by a record with an equivocal Q code PLUS a 9.2

-OR-

1.7 No Q code and neither 5.1 nor 5.2 in the first ECG followed by a record with an equivocal Q code PLUS a 5.1 or 5.2

-OR-

II. Evolution of a current injury which lasts more than one day.

1.8 An ST segment Elevation (Minn. code 9.2) lasting more than one day (i.e. present on consecutive records of different dates)

AND

T wave progression on three or more records from 5.0 to 5.2 or from 5.3 to 5.1, with a more abnormal code present on consecutive records of different dates.

Note: The ST segment elevation does not have to be present in the same lead groups as the T progression, nor does it have to be exactly simultaneous. Q waves will often be present in the same graphs but they are not necessary to use this criterion for Definite ECG.

Code 2: Probable ECG

Evolution of repolarisation changes

2.1 No major ST segment depression in one ECG record (no 4.1 or 4.2) and another record with a major ST segment depression (Minn. code 4.1)

2.2 No ST segment elevation in one ECG record (no 9.2) and another record with an ST segment elevation (Minn. code 9.2)

2.3 No major T wave inversion in one ECG record (no 5.1 or 5.2) and another record with a major T wave inversion (Minn. code 5.1 or 5.2)

Note: Unlike the criteria in the previous classes, the evolution in this class can go in either direction, that is the codes can get better or worse.

Code 3: Ischaemic ECG (in one or more records)

Records not satisfying the above criteria which nonetheless show:

3.1 Minnesota codes 1.1.1 to 1.3.6 excluding 1.2.6 for Q and QS codes

-AND/OR-

3.2 Minnesota codes 4.1 through 4.3 for ST junction (J) and segment depression

-AND/OR-

3.3 Minnesota codes 5.1 through 5.3 for T wave items

-AND/OR-

3.4 Minnesota code 9.2 for ST segment elevation

Code 4: Other ECG

All other ECG findings, including normal ECG but note rules for not codifiable ECG below.

Code 5: Not codifiable ECG

The following Minnesota codes lead to suppression of all or most of these items, and a set of ECG records in which such findings are present in all records should be considered not codifiable (unless codifiable Q waves are present, for example in an ECG showing a 7.4)

6-1 Third degree A-V block, suppresses all 1,4,5 and 9.2

6-4-1 Persistent Wolff-Parkinson White Pattern, suppresses all other codes.

6-8 Artificial pacemaker, suppresses all other codes.

7-1-1 Complete left bundle branch block, suppresses 1.2.3,1.2.7,1.2.8, 1.3.2, 1.3.6 and all 4, 5 and 9.2 codes but the presence of a codifiable Q downgrades it to 7.4.

7-2-1 Complete right bundle branch block, suppresses 1.2.8, and all 4, 5 and 9.2 codes.

7-4 Intraventricular block suppresses all 4, 5, and 9.2 codes.

8-2-1 Ventricular fibrillation and asystole, suppress all other codes.

8-2-2 Idioventricular rhythm, suppresses all other codes.

8-4-1 Supraventricular tachycardia above 140/minute, suppresses all other codes.

Code 9: ECG absent

No ECG available or recorded. (Coded as 9, no data).

## **B. Symptoms**

1 Indicates typical; 2, atypical; 3, other; 4, none; 5, inadequately described; and 9, insufficient data.



*Code 1 (typical symptoms) when chest pain is present and characterized by (a) duration of more than 20 minutes and (b) no definite non-cardiac or cardiac non-atherosclerotic cause. If symptom duration is not stated, then code 5 (inadequately described) shall apply. The duration can be assumed to be 20 minutes if the history implies that the pain lasted while something else was going on, or until something else happened.*

*Code 2 (atypical) if symptoms were not typical but there was (a) one or more of atypical pain, acute left ventricular failure, shock, and syncope and (b) the absence of cardiac disease other than ischemic heart disease and (c) no definite non-cardiac or cardiac non-atherosclerotic cause.*

*Code 3 (other symptoms) when symptoms are well described but do not satisfy the criteria for typical or atypical. Symptoms due to a definite non-cardiac cause or to a definite non-atherosclerotic cardiac cause (e.g., pericarditis) should be coded 3.*

*Code 4 (no symptoms) in nonfatal cases if the patient reported no symptoms in the attack, and in fatal cases if the eyewitnesses of the fatal collapse state that the individual was completely normal and uncomplaining before the moment of death.*

*Code 5 (inadequately described symptoms) for cases otherwise satisfying criteria for typical pain but in which the duration of the pain is not described, so that it is not possible to classify the symptoms as typical.*

*Code 9 (insufficient data) if information on the presence or character of symptoms is inadequate.*

### **C. Serum Enzymes**

*1 Indicates abnormal; 2, equivocal; 3, nonspecific; 4, normal; 5, incomplete; and 9, insufficient data.*

*Each MCC should define, with the help of their local hospital laboratories, (a) the cardiac enzyme tests used and (b) the upper limit of normal for each test in each laboratory.*

*Code 1 (abnormal) if at least one reading is more than twice the upper limit of normal when measured within 72 hours or 3 calendar days of symptoms onset, admission to hospital, or any recurrence of symptoms.*

*Code 2 (equivocal) when serum enzyme levels are raised but to less than twice the upper limit of normal.*

*Code 3 (nonspecific) if serum enzyme levels are raised to more than twice the upper limit of normal but there are explanations other than myocardial infarction, such as liver disease, infections, defibrillation, or surgery.*

*Code 4 (normal) when the enzyme tests are done in time and are within the limits of normal.*

*Code 5 (incomplete) where tests are done >72 hours after the onset of acute symptoms or any recurrence.*

*Code 9 (insufficient data) when serum enzyme tests have not been done or results are unavailable.*

### **D. Necropsy Findings Summary**

1 Indicates definite; 2, equivocal; 4, negative; 8, alive at 28 days or no necropsy performed; and 9, insufficient data.

Code 1 (definite) if there visible to the naked eye was (a) myocardial infarction and/or (b) recent occlusion of a coronary artery (from ante-mortem thrombus or hemorrhage into a plaque or embolism).

Code 2 (equivocal) when the record does not show definite evidence or record any non-cardiac or cardiac, non-atherosclerotic disease causing death but there is (a) old myocardial infarction (scar) and/or (b) occlusion or severe stenosis (>50% reduction of lumen) by atheroma of one or more coronary arteries.

Code 4 (negative) when there is recorded at necropsy (a) no definite evidence as described above and (b) evidence of non-cardiac or cardiac non-atherosclerotic disease causing death.

Code 8 (not relevant) if the patient is alive at 28 days or if necropsy was not done.

Code 9 (insufficient data) when the results of the necropsy were not obtained.

## 2. American Heart Association Criteria (2003)

[Luepker VR, Apple FS, Chistenson RH, Crow RS, Fortmann SP, Goff D, Goldberg RJ, Hand MM, Jaffe AS, Julian DG, Levy D, Manolio T, Mendis S, Mensah G, Pajank A, Prineas R, Reddy S, Roger V, Rosamond WO, Shahar E, Sharrett R, Sorlie P, Tunsall-Pedoe H. Case definitions for acute coronary heart disease in epidemiology and clinical research studies. *Circulation* 2003; 108: 2543-2549]

Case definitions for Acute Coronary Heart Disease in Epidemiology and Clinical Research Studies.

A statement from AHA Council of Epidemiology and Prevention; AHA Statistics Committee; World Heart Federation Council of Epidemiology and Prevention; European Society of Cardiology Working Group on Epidemiology and Prevention; Centers for Disease Control and Prevention; National Heart, Lung, and Blood Institute.

**Table 6 - Classification of AMI**

	<b>Biomarker Findings</b>							
	<b>Cardiac Symptoms or Signs Present</b>				<b>Cardiac Symptoms or Signs Absent</b>			
<b>ECG Findings</b>	<i>Diagnostic</i>	<i>Equivocal</i>	<i>Missing</i>	<i>Normal</i>	<i>Diagnostic</i>	<i>Equivocal</i>	<i>Missing</i>	<i>Normal</i>
<b>Evolvine diagnostic</b>	<i>Definite</i>	<i>Definite</i>	<i>Definite</i>	<i>Definite</i>	<i>Definite</i>	<i>Definite</i>	<i>Definite</i>	<i>Definite</i>
<b>Positive</b>	<i>Definite</i>	<i>Probable</i>	<i>Probable</i>	<i>No</i>	<i>Definite</i>	<i>Probable</i>	<i>Possible</i>	<i>No</i>
<b>Non specific</b>	<i>Definite</i>	<i>Possible</i>	<i>No</i>	<i>No</i>	<i>Definite *</i>	<i>Possible</i>	<i>No</i>	<i>No</i>
<b>Normal or other ECG findings</b>	<i>Definite</i>	<i>Possible</i>	<i>No</i>	<i>No</i>	<i>Definite *</i>	<i>No</i>	<i>No</i>	<i>No</i>

*Notes: Classification of case is at highest level allowed by combinations of 3 characteristics (cardiac signs and symptoms, ECG findings, biomarkers);*

*\* In absence of diagnostic troponin, downgrade to possible.*

### **Definitions of IHD**

*The definition of an Ischaemic heart Disease (IHD) event depends on symptoms, signs, biomarkers, and ECG and/or autopsy findings. These data may vary in quantity, quality, and timing. Definite, probable, and possible cases of fatal and nonfatal AMI, procedure-related events, and angina pectoris are defined on the basis of the extent and diagnostic quality of data. The recommendations emphasize biomarkers in a setting in which signs, symptoms, and/or ECG findings suggest acute ischemia.*

### **Definition of Terms**

#### **Cardiac Biomarkers**

*Cardiac biomarkers are blood measures of myocardial necrosis, specifically CK, CK-MB, CK-MBm, or troponin (cTn). The order of diagnostic value is cTn\_CK-MBm\_CK-MB\_CK.*

- A. Adequate set of biomarkers: At least 2 measurements of the same marker taken at least 6 hours apart*
- B. Diagnostic biomarkers: At least 1 positive biomarker in an adequate set (see A above) of biomarkers showing a rising or falling pattern in the setting of clinical cardiac ischemia and the absence of non-cardiac causes of biomarker elevation*
- C. Equivocal biomarkers: Only 1 available measurement that is positive, or a rising or falling pattern not in the setting of clinical cardiac ischemia or in the presence of non-ischemic causes of biomarker elevation*
- D. Missing biomarkers: Biomarkers not measured*
- E. Normal biomarkers: Measured biomarkers do not meet the criteria for a positive biomarker (see F below)*
- F. Positive biomarkers: At least 1 value exceeding the 99th percentile of the distribution in healthy populations or the lowest level at which a 10% coefficient of variation can be demonstrated for that laboratory.*

### **Cardiac Symptoms and Signs**

*Cardiac symptoms and signs are findings from patient interview and examination.*

- A. Cardiac symptoms: Presence of acute chest, epigastric, neck, jaw, or arm pain or discomfort or pressure without apparent non-cardiac source. More general, atypical symptoms, such as fatigue, nausea, vomiting, diaphoresis, faintness, and back pain, should not be used as a diagnostic criterion, although they are clinically useful in arriving at the correct diagnosis.*
- B. Cardiac signs: Acute congestive heart failure or cardiogenic shock in the absence of non-IHD causes.*

## **ECG Findings**

One or more ECG(s) may be collected in a possible cardiac event. These should be adjudicated or classified when possible.

The evolution of ECG findings may be demonstrated between the ECG(s) associated with the event or between a previously recorded ECG and the event ECG(s). In cases in which only a single event ECG is available, an evolving diagnostic ECG pattern can be recorded only if a previous study ECG is available (e.g., if there is no previous study ECG and only 1 event-related ECG, there can be no classification of “evolving diagnostic” ECG).

Precise measurement guidelines must be followed to measure wave onset and offset to determine wave duration and voltage. Most AMI-likely events occur in settings not controlled by epidemiology researchers, and therefore most ECG(s) will be hard copy, with varying levels of quality. The most extensively used measurement system for visual ECG findings is the Minnesota Code. These measurement guidelines should be coupled with validated biologically acceptable degrees of change in ECG wave forms to code an evolution of change. Another coding system that standardizes the measurement of ECG wave patterns is the Novacode (an extension of the Minnesota Code), which was designed for clinical trial ascertainment of AMI. More details on ECG coding are available on the Minnesota ECG Coding Center web site ([www.epi.umn.edu/ecg](http://www.epi.umn.edu/ecg)).

The categories are as follows:

- A. Evolving diagnostic ECG
- B. Positive ECG
- C. Nonspecific ECG
- D. ECG negative for ischemia

## **Postmortem Consistent With AMI**

Postmortem findings consistent with AMI are a cardiac pathology consistent with recent coronary occlusion or AMI  $\leq$  28 days old.

## **Case Classifications for IHD**

### **I. Nonfatal events**

#### **A. Definite AMI**

1. Evolving diagnostic ECG, or
2. Diagnostic biomarkers

#### **B. Probable AMI**

1. Positive ECG findings plus cardiac symptoms or signs plus missing biomarkers,  
or
2. Positive ECG findings plus equivocal biomarkers

#### **C. Possible AMI**

1. Equivocal biomarkers plus nonspecific ECG findings, or
2. Equivocal biomarkers plus cardiac symptoms or signs, or

3. *Missing biomarkers plus positive ECG*

D. *Unrecognized AMI*

1. *Appearance, in a non-acute setting, of a new diagnostic Q wave with or without ST-T- wave depression, or ST elevation*

E. *Medical procedure-related event*

1. *Cardiac events after (up to 28 days) a medical procedure (e.g., general surgery) with criteria for definite, probable, and possible AMI identical to those described above (I. A-C)*
2. *May be reported separately as procedure-related cardiac events or combined with overall event rates*
3. *If the medical procedure was performed for the treatment of acute ischemia (e.g., angioplasty, coronary bypass surgery), an event should be classified as described above (I.A-C) and not considered procedure-related*

F. *Unstable angina pectoris*

1. *New cardiac symptoms and positive ECG findings with normal biomarkers*
2. *Changing symptom pattern and positive ECG findings with normal biomarkers*

G. *Stable angina pectoris*

1. *Cardiac symptoms in a pattern that remains constant in presentation, frequency, character, and duration over time*

II. *Fatal events (hospitalised patients)*

A. *Definite fatal AMI*

1. *Death within 28 days of hospital admission in AMI cases defined in I.A*
2. *Postmortem findings consistent with AMI within 28 days*

B. *Probable fatal MI*

1. *Death within 28 days of hospital admission in cases defined in I.B*
2. *Death within 6 hours of hospital admission with cardiac symptoms and/or signs. Other confirmatory data (biomarkers, ECG) are absent or not diagnostic.*

C. *Possible fatal coronary event*

1. *Death within 28 days of hospital admission in cases defined in I.C, I.F, and I.G*
2. *Post-mortem findings show old infarct and/or  $\geq 50\%$  atherosclerotic narrowing of coronary arteries*

**3. Nomenclature for AMI/ACS proposed by the British Cardiac Society (2004)**

[Fox KAA, Birkhead J, Wilcox R, Knight C, Barth J. British Cardiac Society Working Group on the definition of myocardial infarction. *Heart* 2004; 90: 603-609]

*The clinical and cardiac marker manifestations are determined by the volume of myocardium affected and the severity of ischaemia. Despite the similarities in disease mechanism, the time course and severity of cardiac complications vary substantially across the spectrum of ACS. Similarly, treatment patterns differ.*

BCS proposes that the spectrum of ACS should be subdivided as follows:

- ACS with unstable angina
- ACS with myocyte necrosis
- ACS with clinical Acute Myocardial Infarction (AMI).

**Table 7 - Spectrum of Acute Coronary Syndrome (ACS)**

	<b>Markers</b>	<b>ECG</b>	<b>Pathology</b>
<b>ACS with unstable angina</b>	Troponin (TnT) and creatine-kinase (CKMB) undetectable	ST or T non-elevation or transient ST elevation or normal	Partial coronary occlusion (plaque disruption, intracoronary thrombus, microemboli)
<b>ACS with myocyte necrosis</b>	TnT elevation, < 1.0 ng/ml	ST or T elevation or transient ST elevation or normal	Partial coronary occlusion (plaque disruption, intracoronary thrombus, microemboli), more extended than that provoked by angina
<b>ACS with clinical myocardial infarction</b>	TnT elevation, > 1.0 ng/ml +/- CK-MB elevation	ST elevation or ST non-elevation or T inversion: may evolve Q waves	Complete coronary occlusion (plaque disruption, intracoronary thrombus, microemboli)

BCS recommends that the term “unstable angina” should be reserved for patients with a clinical syndrome, but with undetectable troponin or CK-MB markers.

Unstable angina requires supporting evidence of coronary disease (abnormal ECG or prior documented coronary disease).

The term “ACS with myocyte necrosis” should be reserved for patients with a typical clinical syndrome plus an increased troponin concentration below the diagnostic threshold (that is, troponin T < 1.0 ng/ml or AccuTnl < 0.5 ng/ml).

The term “clinical myocardial infarction” should be reserved for patients in the context of a typical clinical syndrome and a marker increase above the diagnostic threshold.

BCS proposes that the threshold to define clinical AMI shall be set at 1.0 ng/ml for troponin T or 0.5 ng/ml for AccuTnl (or equivalent threshold with other troponin I methods).

Therefore, BCS recommends that, in the context of a typical ACS, clinical myocardial infarction should be diagnosed when the maximum troponin T increase is > 1.0 ng/ml or AccuTnl > 0.5 ng/ml (and/or new Q waves develop on the ECG).



Individual laboratories that use other troponin I assays will need to estimate an equivalent troponin I concentration.

It is well recognised that the myocardium can be damaged after percutaneous coronary intervention (PCI) and cardiac markers may increase in up to a third of patients. It is

important to bear in mind, just as with spontaneous MI, that cardiac enzyme release after PCI should be integrated with clinical, angiographic, and ECG data to properly assess prognosis. Troponin concentrations should not be considered in isolation. BSC recommends systematic measurement of troponins after PCI (> 6 hours) as part of quality control standards.

Figure 2, below, describes the spectrum of acute coronary syndrome.

**Figure 2 - Spectrum of Acute Coronary Syndrome (ACS)**

ACS with unstable angina	ACS with myocyte necrosis	ACS with clinical myocardial infarction
<p>Marker: Tn (Troponin) and CKMB (creatinine kinase) undetectable</p>	<p>Marker: Troponin elevated TnT &lt; 1.0 ng/ml</p>	<p>Marker: Tn (Troponin) and CKMB (creatinine kinase) undetectable</p>
<p>ECG: ST↑ or T↓ or transient ST↑ or normal</p>		<p>ECG: ST↑ or ST↓ or T inversion: may evolve Q waves</p>
<p>Risk of death (from hospitalisation to 6 months): 5-8%</p>	<p>Risk of death (from hospitalisation to 6 months): 8-12%</p>	<p>Risk of death (from hospitalisation to 6 months): 12-15%</p>
<p>Pathology (plaque disruption, intracoronary thrombus, micro-emboli): partial coronary occlusion</p>		<p>Pathology (plaque disruption, intracoronary thrombus, micro-emboli): complete coronary occlusion</p>
<p>Left Ventricular function: no measurable dysfunction</p>		<p>Left Ventricular function: systolic dysfunction, LV dilatation</p>

#### 4. European Society of Cardiology/American College of Cardiology Criteria (2000)

[The Joint European Society of Cardiology/American College of Cardiology Committee. Myocardial infarction redefined. A consensus document of The Joint European Society of Cardiology/American College of Cardiology Committee for the Redefinition of Myocardial Infarction. Eur Heart J 2000; 21: 1502-1513]

Criteria for definition of acute, evolving or recent myocardial infarction

*Any one of the following criteria satisfies the diagnosis for an acute, evolving or recent myocardial infarction:*

- (1) Typical rise and gradual fall (troponin) or more rapid rise and fall (CK-MB) of biochemical markers of myocardial necrosis with at least one of the following:
  - (a) ischemic symptoms;*
  - (b) development of pathologic Q waves on the ECG;*
  - (c) ECG changes indicative of ischemia (ST segment elevation or depression); or*
  - (d) coronary artery intervention (e.g., coronary angioplasty).**
- (2) Pathologic findings of an acute MI.*

*Criteria for established myocardial infarction*

*Any one of the following criteria satisfies the diagnosis for established myocardial infarction:*

- (1) Development of new pathologic Q waves on serial ECGs. The patient may or may not remember previous symptoms. Biochemical markers of myocardial necrosis may have normalized, depending on the length of time that has passed since the infarct developed.*
- (2) Pathologic findings of a healed or healing myocardial infarction.*

### **Diagnostic criteria for Stroke**

#### **MONICA - MONItoring trends and determinants of Cardiovascular disease**

[World Health Organization: WHO Monica Project: MONICA manual. Part IV: Event Registration. <http://www.ktl.fi/publications/monica/manual/part4/iv-2.htm#s1-1>]

*Stroke is defined as rapidly developed clinical signs of focal (or global) disturbance of cerebral function lasting more than 24 hours (except in cases of sudden death or if the development of symptoms is interrupted by a surgical intervention), with no apparent cause other than a vascular origin: it includes patients presenting clinical signs and symptoms suggestive of subarachnoid haemorrhage, intracerebral haemorrhage or cerebral ischaemic necrosis. Global clinical signs are accepted only in cases of subarachnoid haemorrhage or in patients with deep coma. Brain lesions detected by CT-scan but not accompanied by acute focal signs are not accepted as stroke, nor are extradural and subdural haemorrhages. This definition does not include TIA or stroke events in cases of blood disease (e.g. leukemia, polycythaemia vera), brain cancer or brain metastases.*

*Secondary stroke caused by trauma should also be excluded.*

*The diagnostic classification follows:*

#### **(1) Definite focal signs**

- unilateral or bilateral motor impairment (including dyscoordination)*



- *unilateral or bilateral sensory impairment*
- *aphasia/dysphasia (non-fluent speech)*
- *hemianopia (half-sided impairment of visual fields)*
- *diplopia*
- *forced gaze (conjugate deviation)*
- *dysphagia of acute onset*
- *apraxia of acute onset*
- *ataxia of acute onset*
- *perception deficit of acute onset.*

## **(2) Not acceptable as sole evidence of focal dysfunction**

*Although strokes can present themselves in the following way, these signs are not specific and therefore cannot be accepted as definite evidence for stroke.*

- *dizziness, vertigo*
- *localized headache*
- *blurred vision of both eyes*
- *dysarthria (slurred speech)*
- *impaired cognitive function (including confusion)*
- *impaired consciousness*
- *seizures*

*On the basis of the background information, each event may be classified into:*

*Definite stroke*

*Not stroke*

*Insufficient data*

*Insufficient data should be mainly used for fatal cases, especially for cases of sudden death without necropsy.*

*Cerebrovascular lesions discovered at autopsy are considered for diagnostic category.*

*All patients having insufficient supporting evidence of stroke, but for whom the diagnosis of stroke cannot be entirely excluded, should be classified as insufficient data, e.g. cases with no necropsy, no documented history of focal neurologic deficits and no other diagnosis. Living patients can be classified into this category if:*

- *it is impossible to say whether the symptoms were from stroke or from some other disease, e.g. epilepsy, or*
- *patients with symptoms and clinical findings otherwise typical for a stroke but the duration remaining uncertain.*

## **Subtype definition**

*Cases identified as 'definite stroke' were classified into stroke subtypes.*

*The MONICA subtype definition of stroke has to be confirmed by CT scan, examination or autopsy.*

***Subarachnoid Haemorrhage (ICD-8 or ICD-9 430; ICD-10 I60)***

*Symptoms:*

*Abrupt onset of severe headache or unconsciousness or both. Signs of meningeal irritation (stiff neck, Kernig and Brudzinski signs). Focal neurological deficits are usually not present.*

*Findings:*

*At least one of the following must be present additional to typical symptoms.*

- 1. Necropsy - recent subarachnoid haemorrhage and an aneurysm or arteriovenous malformation*
- 2. CT-scan - blood in the Fissura Sylvii or between the frontal lobes or in the basal cistern or in cerebral ventricles*
- 3. CSF (liquor) bloody (>2,000 rbc per cm<sup>3</sup>) and an aneurysm or an arteriovenous malformation found on angiography*
- 4. CSF (liquor) bloody (>2,000 rbc per cm<sup>3</sup>) and xanthochromic and the possibility of intracerebral haemorrhage excluded by necropsy or CT-examination*

***Intracerebral haemorrhage (ICD-8 or ICD-9 431; ICD-10 I61)***

*Symptoms:*

*Usually sudden onset during activities. Often rapidly developing coma, but small haemorrhage presents no consciousness disturbance.*

*Findings:*

*CSF often, but not always bloody or xanthochromic. Often severe hypertension is present.*

*Haemorrhage must be confirmed by necropsy or by CT-examination.*

***Brain infarction due to occlusion of precerebral arteries (ICD-8 432; ICD-9 433; ICD-10 I63.2)***

*Symptoms:*

*May vary.*

*Findings:*

*The occlusion must be confirmed by angiography or ultrasound or necropsy.*

***Brain infarction due to cerebral thrombosis (ICD-8 433; ICD-9 434; ICD-10 I63.0, I63.3, I63.6)***

*Symptoms:*

No severe headache, if at all. Onset acute, sometimes during sleep. Often gradual progression of focal neurologic deficits. Usually, no, or only slight, disturbance of consciousness. TIA can often be detected in history. Often other symptoms of atherosclerosis (IHD, peripheral arterial disease) or underlying diseases (hypertension, diabetes) can be detected.

**Findings:**

Brain infarction in the necropsy or in the CT-examination and no evidence for an embolic origin.

OR

CT-scan of satisfactory quality shows no recent brain lesion although clinical criteria of stroke are fulfilled.

**Embolic brain infarction (ICD-8 434; ICD-9 434; ICD-10 I63.1, I63.4)**

**Symptoms:**

Abrupt onset, usually completion of the neurologic deficits within a few minutes. Disturbance of consciousness absent or only slightly present at the onset.

**Findings:**

As in brain infarction due to cerebral thrombosis, but in addition a source of the embolus must be detectable. The most common origins are:

- arrhythmia (atrial flutter and fibrillation)
- valvular heart disease (mitral)
- recent Acute Myocardial Infarction (AMI) (within previous 3 months)

**Remarks**

If it is impossible to assign one of these sub-categories to a definite stroke event, the subcategory ‘Acute, but ill-defined cerebrovascular disease’ should be recorded (ICD-9 436; ICD-10 I64). If the clinical criteria for a stroke are fulfilled, but a CT-scan (of satisfactory technical quality) fails to reveal a brain lesion of recent origin, the patient has in all probability suffered an ischaemic stroke. In this case, the type of stroke should be coded as ICD-9 434 (Infarction), ICD-10 I63.9 (Cerebral infarction, unspecified).

**Table 8 - MONICA subtype**

<b>Type of stroke</b>	<b>ICD Code</b>
Subarachnoid haemorrhage	ICD-8 or ICD-9 430; ICD-10 I60
Intracerebral haemorrhage	ICD-8 or ICD-9 431; ICD-10 I61
Brain infarction due to occlusion of precerebral arteries	ICD-8 432; ICD-9 433;

	ICD-10 I63.2
<i>Brain infarction due to cerebral thrombosis</i>	ICD-8 433; ICD-9 434; ICD-10 163.0, 163.3, 163.6
<i>Embolic brain infarction</i>	ICD-8 434; ICD-9 434; ICD-10 I63.1, I63.4

### Criteria for Type-1 Diabetes Mellitus cases identification

[World Health Organization: definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: Report of a WHO/IDF Consultation. Geneva, World Health Org., 2006;

Carle F, Gesuita R, Bruno G, Coppa GV, Falorni A, Lorini R, Martinucci ME, Pozzilli P, Prisco F, Songini M, Tenconi MT, Cherubini V; RIDI Study Group. Diabetes incidence in 0- to 14-year age-group in Italy: a 10-year prospective study. *Diabetes Care* 2004; 27: 2790-2796]

In the RIDI Type 1 Diabetes Registry, diabetes cases are defined and identified as follows:

- diagnosis of insulin-dependent diabetes mellitus (IDDM) according to WHO classification
  - Fasting plasma glucose concentration  $\geq$  126 mg/dl (7.0 mmol/l), OR
  - 2hrPPG (2-Hours Post-Prandial Glucose)  $\geq$  200 mg/dl (11.1 mmol/l) after a 75-g glucose load
- The date of first insulin administration should be considered as the date of diagnosis (1st insulin administration date = diagnosis date)
- The date of diagnosis should follow the date of activation of the register, which represents the starting point of the detection
- Age patient should be  $<$  30 years at the 1<sup>st</sup> insulin administration
- Patient has resided in the geographical area covered by the register since at least six months before the diagnosis (Non-resident cases must be registered and notified to the Central Coordinator who will report them to the relevant registers).

All registries report newly diagnosed insulin-treated children using a special form that includes patients' personal identification number, date of birth, sex, date of diagnosis (defined as the date when the first insulin injection was given), and municipality of residence.

Cases diagnosed as type-2 diabetes or other specific types were excluded.

Each registry used at least two independent data sources for case ascertainment, including hospital discharges, prescription registries, personal national health system cards needed by each patient to obtain syringes and strips free of charge, summer camp

*rosters for diabetic children, membership lists of patient associations, and records of diabetes centers.*

*The completeness of ascertainment of each registry has been estimated by using the capture-recapture method.*

### **Validation of sources for the Italian RIDI - Type-1 Diabetes Mellitus registry**

*Validation of sources were automatically carried out by the dedicated software RIDI-PROG through the 'capture-recapture method', also taking into account the number of sources used. The programme provided the percentage values of the completeness of the individual sources and thus the overall ability to assess the local register.*

### **STEP 4 - Set up of a population-based registry**

Based on Step 2 and 3, it is possible to set up a population-based register following A (record-linkage between routine administrative data-based registries) or B (disease-specific data collection).

A. Register based on record-linkage between routine administrative data:

- when the linkage procedure between hospital discharge and mortality records is feasible, it is important to define the demographic characteristics of the population included in the registry (age, sex, residence, etc.), to identify the event and its duration, as well as how to handle transfer between hospitals with difference in the diagnoses between the admitting hospital and the hospital where the patient is transferred, how to define new events, recurrent events, fatal and non-fatal events, etc.;
- validation of diagnostic information is recommended in a sufficient sample size of the identified events randomly selected during a period of the year or during same days each month (in order to trace seasonal variation), or consecutively from the beginning of the year, with the estimation of sensitivity, specificity and PPV of defined events;
  - in the **Italian Pilot Study of Registry of Coronary Events**, 500 coronary events were selected (consecutively from the beginning of the year) for each year of registration in each of the 7 local registers; about 7,000 AMI/ACS events were overall validated in two years of registration in order to estimate PPVs. In some areas the amount of 500 events, to be yearly validated, was reached including events registered in few months (e.g., Brianza register in the North of Italy); while in Caltanissetta (South of Italy), it was necessary to enlarge the area of registration to other adjacent areas to reach the requested amount of 500 events per year. Out of about 7,000 events, 3,020 (43%) were validated as coronary events according to MONICA diagnostic criteria. On this basis, PPVs were estimated and applied to the overall population to estimate coronary events and, consequently, attack rates and case fatality;

- in the **Italian pilot registry of Cerebrovascular events**, 500 stroke events were selected (consecutively from the beginning of the year) for each year of registration in each of the 8 local registers; about 8,000 stroke events were overall validated in two years of registration in order to estimate PPVs. In some areas the amount of 500 events, to be yearly validated, was reached including events registered in few months (e.g., Veneto region register in the North of Italy); in Caltanissetta (South of Italy), it was necessary to enlarge the area of registration to other adjacent areas to reach the requested amount of 500 events per year as well as in coronary registry. Out of about 8,000 events, 3,099 (39%) were validated as stroke events according to MONICA diagnostic criteria. On this basis, PPVs were estimated and applied to the overall population to estimate stroke events and, consequently, attack rates and case fatality;
- when the validation of a sample of events is not possible and area-specific PPVs cannot be estimated, available PPVs, drawn from other registries or studies, can be applied to the identified current events in order to estimate the number of occurred events in the population;
- population data by age and sex of the area under surveillance are needed to estimate incidence (attack rates, when it is not possible to distinguish between first or recurrent event), case fatality, and mortality rates;
- periodical validation of samples of events should be performed in order to update PPVs estimates. For example, for coronary and cerebrovascular diseases, for which survival has been continuously increasing in recent years, less severe non-fatal events are expected to increase and, consequently, changes in PPVs of ICD codes are also expected.

#### B. Register based on disease-specific data collection:

- set up a pilot population-based register with proven standardised protocols and evaluate the pilot study results (coverage, completeness of information and diagnostic validity);
- based on the results of the pilot study, a full scale register should be set up, if feasible, and it should be decided whether to use hot or cold pursuit; access and availability of clinical records, needed for the event validation, in the hospitals of the geographical area of the registry should be checked;
- then, a full-scale register (target population, data collection methods and validation procedures), if feasible, shall be designed.

#### To set up a full scale register:

- one or more populations representative for the region or the country shall be selected;

- for each selected population, a population-based register with approved standardised protocols shall be set up;
- a detailed protocol shall be written for data collection, and shall include identification of events and validation procedures as reported in Step 2 and Step 3;
- information coverage, representativeness and completeness shall be evaluated;
- If relevant, the results from the register shall be used to validate administrative data.

### Description of the software for the implementation of the population-based registry of Coronary events (used in the EuroMed project for the coronary registry of Zagabria population)

*See the power point presentation 'Step 4 - Software for the implementation of the population-based registries of Coronary and Cerebrovascular events' available on the web site [www.cuore.iss.it](http://www.cuore.iss.it); website address: <http://www.cuore.iss.it/eng/bridge/WP8/TrainingPopulation-basedRegistries.asp>*

As previously mentioned, the availability of a software to support the implementation of a population-based registry is fundamental; this software should be specifically planned and elaborated for the disease or the condition under surveillance. Today many softwares are available, improving day by day the different processes with increasing technological advances; this means that all the softwares become obsolete in a short time, then to explain the underlying methodology is fundamental.

A summary description of the steps implemented in the software used for the Italian pilot study of the population-based registries of coronary events and of cerebrovascular events, how the software works, and outcomes that can be elaborated, is reported here below as an example.

The same software was also implemented in the framework of the EuroMed Programme (launched by the European Commission in 2008 and funded by the Italian Ministry of Health) for the implementation of the Coronary event registry in the population of the town of Zagreb in Croatia.

The methods derive from EUROCISS as previously described. The software is downloadable from the web site of the Progetto CUORE ([www.cuore.iss.it](http://www.cuore.iss.it)). The user can install the software following a stand-alone scheme, where both the software and database are located on the same computer, or a client-server scheme, where the database is installed on one server and the software is installed on one or more computers connected to this server. Some operative information to perform both administrative activities and user activities of the software are summarized and described here.

The main administration features that the software administrator can perform include user management, events loading, events generation and management; the main user

activities include the validation procedure, the estimation of Positive Predictive Values (PPVs) from a sample of validated events, or the inclusion and use of an already estimated set of PPVs, the estimation of the number of events, and the elaboration of attack rates and case fatality.

#### *Administrative activities:*

User management functions consent to manage user authentication by username and password, add a new user, modify user, and delete the user.

Events loading functions permit to load the mortality and hospital discharge diagnoses records that are necessary to run the event generation. Before performing this activity, it is very important to check that the format of all data complies with the defined format provided by the software.

After loading mortality and hospital discharge diagnoses records, it is possible to add a new event or change or delete a selected event. An internal check, based on comparison of birth and death or discharge dates, consents to fix ages of loaded subjects for both mortality and hospital discharge diagnoses records.

Events generation and management: after selecting residence codes, calendar years, age ranges, ICD code version, duration of hospitalization (establishing the number of days under which the record shall be considered as 'day-hospital', and then not included), and the number and modality (consecutive or random) of sample selection for the events to be validated, the software will implement the record-linkage and will generate the separate record lists of first Coronary (CE) Events (current events) and the record lists of the extracted sample of CE to be validated.

The sample size of events to be validated can be freely managed in the software and should be chosen in order to have statistically significant PPVs estimates:

- in the **Italian pilot registry of Coronary events**, 500 coronary events were selected (consecutively from the beginning of the year) for each year of registration in each of the 7 local registers; about 7,000 AMI/ACS events were overall validated in two years of registration in order to estimate PPVs.

#### *User activities:*

Event validation procedure: for each event to be validated, included in the extracted CE sample, it is possible to fill in and save specific forms, including all the medical information drawn from the clinical records, necessary to the validation of the event.

EC validation can be implemented following MONICA diagnostic criteria (electrocardiogram codified by Minnesota code, symptoms, enzymes, history of IHD, necropsy) and 'new' ESC/ACC diagnostic criteria (mainly based on troponin values).

On the basis of input information, it is possible to assign the MONICA diagnostic category (Definite, Possible, No IMA, Insufficient data) to each validated event. For Fatal CE, the



MONICA aggregate is given by summing Definite, Possible and Insufficient data categories; for Non-Fatal CE, the MONICA aggregate is given by summing Definite and Possible categories.

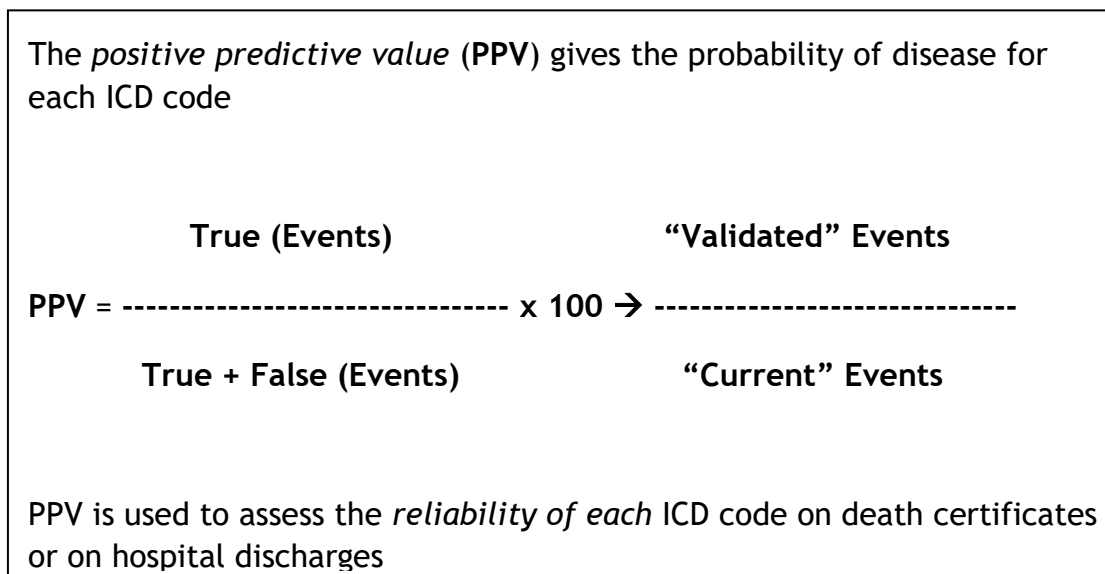
The list of all CE events included in the samples extracted for validation are displayed and divided in two different families: ‘already validated events’ and ‘events still to be validated’.

The user can filter the content of these lists by opening the search procedure, before introducing one of the search key (medical records, family name, name or hospital code), or selecting the options ‘Show CE only’.

Once all CE selected events are validated, it is possible to estimate the PPV for each ICD code as underlying cause of death (fatal event) or first discharge diagnosis (non-fatal event) (Figure 3) and to display and export them as an excel file.

*See the power point presentation ‘Step 4 - Estimation of Positive Predictive Values (PPV)’ available on the web site [www.cuore.iss.it](http://www.cuore.iss.it); website address: <http://www.cuore.iss.it/eng/bridge/WP8/TrainingPopulation-basedRegistries.asp>*

**Figure 3 - Positive predictive value for an identified ICD code in the Italian pilot registry of Coronary events**



The software consents to apply PPV to the overall Coronary current events through the corresponding ICD code in order to estimate the number of fatal and non-fatal CE, and to display and export them as an excel file (Figure 4).

*See the power point presentation ‘Step 4 - Estimation of coronary and cerebrovascular events on the basis of specific PPVs’ available on the web site [www.cuore.iss.it](http://www.cuore.iss.it); website address: <http://www.cuore.iss.it/eng/bridge/WP8/TrainingPopulation-basedRegistries.asp>*

**Figure 4 - Number of estimated events for an identified ICD code (fatal and non-fatal events separately)**

$$N_{EE} = N_{CE} * \sum (PPV_i * Pr_i)$$

where:

$N_{EE}$  = Number of estimated events

$N_{CE}$  = Number of current events by record linkage

$PPV_i$  = Positive predictive value for the 'i'-identified ICD code (proportion of events with an identified ICD code validated as positive over the number of total events with the same ICD code)

$Pr_i$  = Prevalence of the 'i'-identified ICD code

Once the estimated number of CE events is found and an excel file describing population by age group and sex (according to the stratification chosen during the 'Event generation' procedure) is uploaded, the software allows to estimate case fatality indicators, as the ratio between fatal and total events, and attack rates, as the ratio between fatal or non-fatal events and population, and to display and export them as an excel file.

Both indicators can be estimated by age group and sex and can be successively age-standardized according to a well-defined standard population (e.g. European Standard Population).

In alternative, if 'external' PPVs are already available, it is possible to upload a file including PPV by ICD code for fatal and non-fatal events and - by-passing the procedure to estimate PPV from the validation of the extracted sample of events - use them directly to estimate case fatality and attack rate indicators.

### Description of the software for the population-based registry of Cerebrovascular events

*See the power point presentation 'Step 4 - Software for the implementation of the population-based registries of Coronary and Cerebrovascular events' available on the web site [www.cuore.iss.it](http://www.cuore.iss.it); website address: <http://www.cuore.iss.it/eng/bridge/WP8/TrainingPopulation-basedRegistries.asp>*

A specific software was built as a tool to implement the Italian pilot register of Cerebrovascular events according to a simplified methodology adopted on the basis of

the EUROCISS recommendations as previously described. The software is downloadable from the web site of the Progetto CUORE ([www.cuore.iss.it](http://www.cuore.iss.it)). The user can install the software following a stand-alone scheme, where both the software and the database are located on the same computer, or a client-server scheme, where the database is installed on one server and the software is installed on one or more computers connected to this server. Some operative information to perform both administrative activities and user activities of the software are summarized and described here. The main administration features that the software administrator can perform include user management, events loading, events generation and management; the main user activities include the validation procedure, the estimation of PPVs from a sample of validated events, or the inclusion and use of an already estimated set of PPVs, the estimation of the number of events, and the elaboration of attack rates and case fatality.

#### *Administrative activities:*

User management functions consent to manage user authentication by username and password, add a new user, modify user, and delete the user.

Events loading functions permit to load the mortality and hospital discharge diagnoses records that are necessary to run the event generation. Before performing this activity, it is very important to check that the format of all data complies with the defined format provided by the software.

After loading mortality and hospital discharge diagnoses records, it is possible to add a new event or change or delete a selected event. An internal check, based on comparison of birth and death or discharge dates, consents to fix ages of loaded subjects for both mortality and hospital discharge diagnoses records.

Events generation and management: after selecting residence codes, calendar years, age ranges, ICD code version, duration of hospitalization (establishing the number of days under which the record shall be considered as 'day-hospital'), the number and modality (consecutive or random) of sample selection for the events to be validated, the software will implement the record-linkage and will generate the separate record lists of first Cerebrovascular (ACV) Events (current events) and the record lists of the extracted sample of ACV to be validated.

The sample size of events to be validated can be freely managed in the software and should be chosen in order to have statistically significant PPVs estimates:

- in the **Italian pilot registry of Cerebrovascular**, 500 stroke events were selected (consecutively from the beginning of the year) for each year of registration in each of the 8 local registers; about 8,000 stroke events were overall validated in two years of registration in order to estimate PPVs;

#### *User activities:*

Event validation procedure: for each event to be validated, included in the extracted ACV sample, it is possible to fill in and save specific forms, including all the medical information drawn from the clinical record, necessary to the validation of the event.

ACV validation can be implemented following MONICA diagnostic criteria (based on clinical signs of focal (or global) disturbance of cerebral function lasting more than 24 hours).

On the basis of input information, it is possible to assign the MONICA diagnostic category (Definite, Definite ACV with CE, No ACV, Insufficient data) to each validated event. For Fatal ACV, the MONICA aggregate is given by summing Definite, Definite ACV with CE, and Insufficient data categories; for Non-Fatal ACV, the MONICA aggregate is given by summing Definite and Definite ACV with CE categories. A new category was created as 'No MONICA, Yes Technology' when a No ACV MONICA event resulted as Definite thanks to new technologies (CAT-Computed Axial Tomography and NMR-Nuclear Magnetic Resonance).

The list of all ACV events included in the sample extracted for validation are displayed and divided in two different families: 'already validated events' and 'events still to be validated'.

The user can filter the content of these lists by opening the search procedure, before introducing one of the search key (medical records, family name, name, or hospital code), or selecting the options 'Show ACV only'.

Once all the ACV selected events are validated, it is possible to estimate the PPV for each ICD code as underlying cause of death (fatal event) or first discharge diagnosis (non-fatal event) (Figure 3) and to display and export them as an excel file.

*See the power point presentation 'Step 4 - Estimation of Positive Predictive Values (PPV)' available on the web site [www.cuore.iss.it](http://www.cuore.iss.it); website address: <http://www.cuore.iss.it/eng/bridge/WP8/TrainingPopulation-basedRegistries.asp>*

The software consents to apply PPVs to the overall Cerebrovascular current events according to their corresponding ICD code in order to estimate the number of fatal and non-fatal ACV, to display, and to export them as an excel file (Figure 4).

*See the power point presentation 'Step 4 - Estimation of coronary and cerebrovascular events on the basis of specific PPVs' available on the web site [www.cuore.iss.it](http://www.cuore.iss.it); website address: <http://www.cuore.iss.it/eng/bridge/WP8/TrainingPopulation-basedRegistries.asp>*

Once the estimated number of ACV events is found and an excel file describing population by age group and sex (according to the stratification chosen during the 'Event generation' procedure) is uploaded, the software allows to estimate case fatality indicators, as the ratio between fatal and total events, and attack rates, as the ratio between fatal or non-fatal events and population, and to display and export them as an excel file.

Both indicators can be estimated by age group and sex and can be successively age-standardized according to a well-defined standard population (e.g. European Standard Population).

Alternatively, if 'external' PPVs are already available, it is possible to upload a file including PPV by ICD code for fatal and non-fatal events and - by-passing the procedure to estimate PPV from the validation of the extracted sample of events - use them directly to estimate case fatality and attack rate indicators.

### Description of the software 'RIDI-PROG' for the Italian RIDI - Type-1 Diabetes Mellitus registry

- The dedicated software RIDI-PROG is usually delivered to local registers upon a specific request to the Coordination Centre, sited at the University of Ancona (Italy), when a Local Register Operative Protocol is presented;
- the RIDI-PROG is distributed free of charge to local registers;
- RIDI-PROG includes an online help service and consent to implement 3 essential activities:
  - a) to perform data entry;
  - b) to allow transmission of data;
  - c) to elaborate basic analyses of recorded data.
- the data entry function includes two levels:
  - 1) demographic data;
  - 2) case identification data.
- data entry function will allow to enter, verify, and fix loaded data.
- the transmission function allows to transmit data to the Coordination Center.
- the analysis function allows to elaborate data from local registers for:
  - a) assessing the capacity of cases ascertainment from the different recording sources and the completeness of the register;
  - b) calculating specific incidence rates, by sex and age group, standardized rates with relative confidence intervals and seasonality.

### STEP 5: Analyses and dissemination of results

*See the power point presentation 'Step 5 - Elaboration of main indicators for population-based registries of coronary and cerebrovascular events' available on the*

Registry outcomes are mainly summarized in indicators that measure the occurrence of the disease in a population; for a chronic disease, such as CVDs and Cancer, the major indicators of a population-based registry are incidence rate, case fatality rate, survival rate, and attack rate:

**Incidence rate** - it is calculated as the number of first events occurred in the resident population of a well-defined geographic area during a specific period of time (e.g., 1 year). This indicator can be assessed only if information on first events is available; to be sure of selecting only 'first' events in the calculation of the incidence, and not include 'previous events' occurred before the establishment of the registry, a long period (about 5/10 years) of retrospective observation is needed (e.g. in Northern European countries, an event is defined as first event if in the previous 7 years there is no hospital discharge with the event as primary or secondary diagnosis in HDRs).

The Incidence rate is the ratio between the number of first events occurred in 1 year in the population included in the registry and in the population itself (the population at the beginning or at the end of the period, or the mean population in the period, can be used alternatively as denominator);

$$I = \frac{\text{Number of 1}^{\text{st}} \text{ events in the period } t}{\text{Population at risk}} \times 10^n$$

The Incidence rate can be graphically represented by using histograms by age, gender, or calendar period.

**Attack rate** - it is the number of events (first and recurrent events) occurred in the resident population under surveillance in a well-defined period of time (e.g., 1 year); this indicator is used when the population under surveillance is large and it is not possible to distinguish between first and recurrent events; as survival improves, each individual can experience more than one event, so it would be important to identify not only the first but also all recurrent events. It is the ratio between the number of first and recurrent events occurred in 1 year in the population included in the registry and the population itself (the population at the beginning or at the end of the period, or the mean population in the period, can be used alternatively as denominator).

$$AR = \frac{\text{Number of 1}^{\text{st}} \text{ and recurrent events in the period } t}{\text{Population at risk}} \times 10^n$$

Attack rates can be graphically represented using histograms by age, gender, or calendar period.

Since incident events (new cases) can be registered only once for each person, they also identify the persons experiencing first ever disease event, while attack rate includes both incident event and recurrent events occurring in the same person.

For example, in type-1 diabetes, basically incidence rate is the indicator measuring occurrence of the disease, since it is not possible to have recurrent events; similarly in cancer disease, where it is very rare to have recurrent cancer cases in the same site (new cancer cases in different sites are considered different incident events). Differently, in coronary and cerebrovascular diseases, both indicators, incidence and attack rates, are of primary importance for describing occurrence of disease, since multiple events are possible for both diseases. In particular, incidence rate is fundamental for aetiological studies and research, while attack rate is of first concern for health care performance studies, where health complications and consequences are evaluated.

**Case fatality** - it is the proportion of fatal events in relation to the overall events (fatal and non-fatal events) occurred in the resident population of a well-defined geographical area by a specific period of time (e.g., 1 hour, 28 days); all the in- and out-of-hospital fatal and non-fatal events should be considered in the denominator.

$$CF (t\text{-period}) = \frac{\text{Number of FATAL events in the period } t}{\text{Number of FATAL + NON FATAL events in the period } t} \times 100$$

Case fatality can be graphically represented using turtle graphics by age, gender, calendar period.

**Survival rate** - it is the proportion of subjects which are included in the registry and still did not experience a fatal or non-fatal event at different time periods (e.g., 28 days, 6 months, 1 year, 5 years).

They can be graphically represented using Kaplan-Meyer curves by age, gender, or calendar period.

All these indicators can be stratified by sex, age group, calendar year of registration, geographical area, and other variables, such as, for instance, educational level, according to the availability of information, the duration and the geographical extent of the registry.

## Planning a fruitful dissemination of results

See the power point presentation 'Step 6 - Graphic description of main indicators for population-based registries of coronary and cerebrovascular events' available on the web site [www.cuore.iss.it](http://www.cuore.iss.it); website address: <http://www.cuore.iss.it/eng/bridge/WP8/TrainingPopulation-basedRegistries.asp>

- Definition of a strategy to analyse data and disseminating results to decision-makers, stakeholders and the population at large;
- Yearly web publishing of incidence/attack rate, case fatality and survival rate indicators, according to gender and age-standardised rates with the European population as reference (35 to 74 and 35 to 84);
- Utilisation of data for research. This is very important to ensure a high quality of the registry over time. And high quality registers can be the basis for good research.

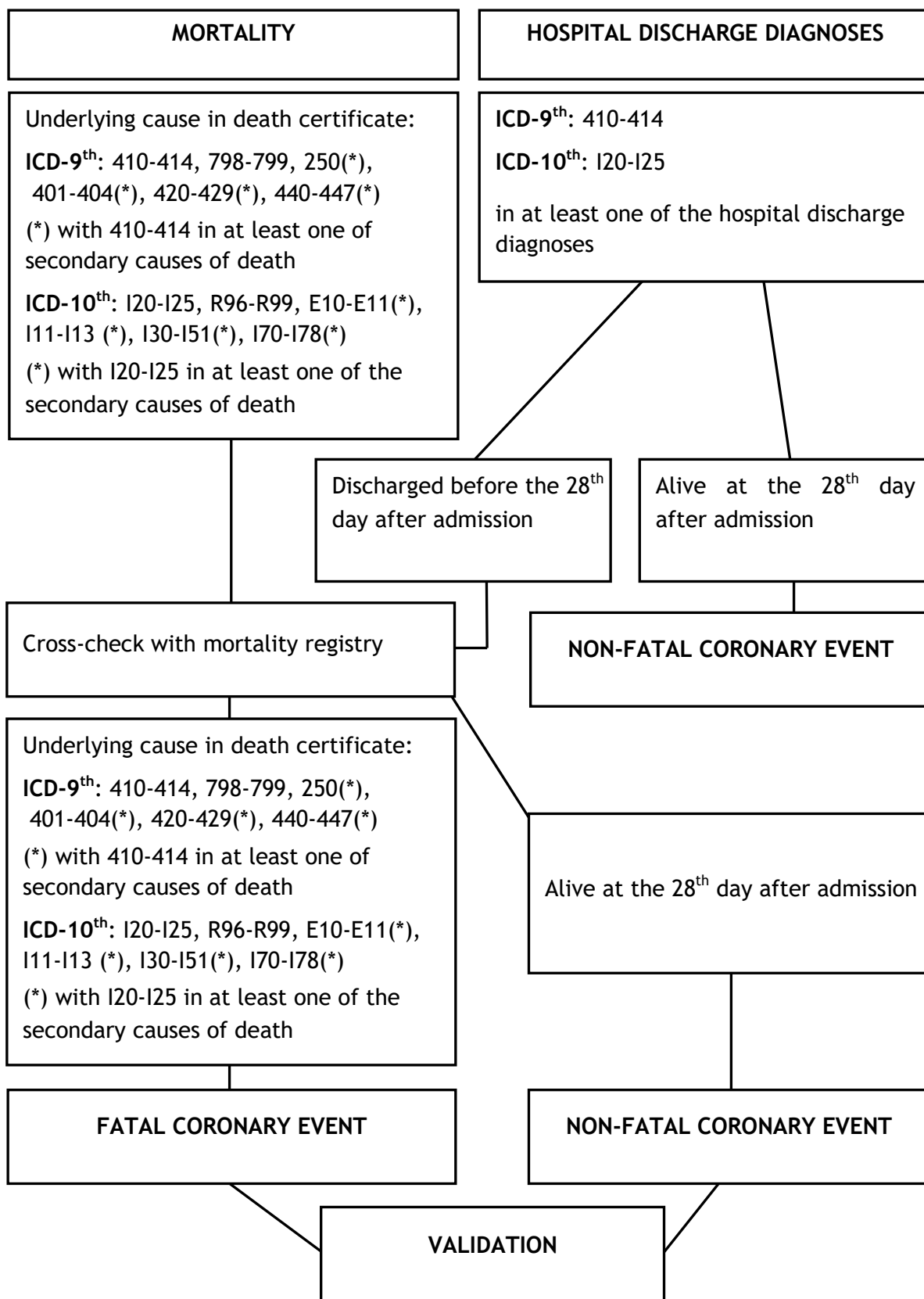
## 4. STEPWISE PROCEDURE: THE EXAMPLE OF CORONARY AND CEREBROVASCULAR REGISTRIES

At national level, the Italian pilot registries of Coronary and Cerebrovascular events were implemented to cover fatal and non-fatal coronary and cerebrovascular events in the general population aged 35-74 years. They were launched in Italy in 2000, following the MONICA and EUROCISS experiences, with the aim of periodically estimating attack rates and case fatality rates of coronary and cerebrovascular events in several geographical areas representative of the country, in order to monitor time trends of most impacting CVD in adult population. Current events were assessed through record-linkage between two main sources of information: death certificates and hospital discharge diagnosis records; events were identified through the International Classification of Diseases (ICD) codes and duration. Figure 5 shows the flow-chart describing the registry process: the ICD codes used in each source of information to select fatal and non-fatal coronary events, and the methods to identify conventional duration  $\leq 28$  days for each event; a random sample of current events were validated applying the MONICA diagnostic criteria; sample events were classified as having definite, possible, and insufficient data based on presence and duration of symptoms, ECG read by Minnesota code, cardiac enzymes, history of IHD, autopsy. Validated fatal coronary events corresponded to the aggregate of definite, possible, and insufficient data; validated non-fatal coronary events corresponded to the aggregate of definite and possible data. Validated events consented to assess the positive predictive value (PPV) for each ICD code of the main cause of death in fatal events, and for each ICD code of the first hospital discharge diagnosis in non-fatal events (Figure 3). To calculate the number of estimated events, the number of current events was multiplied by the PPV of each specific mortality or discharge ICD code derived from the validation of the random sample of current events (Figure 4).

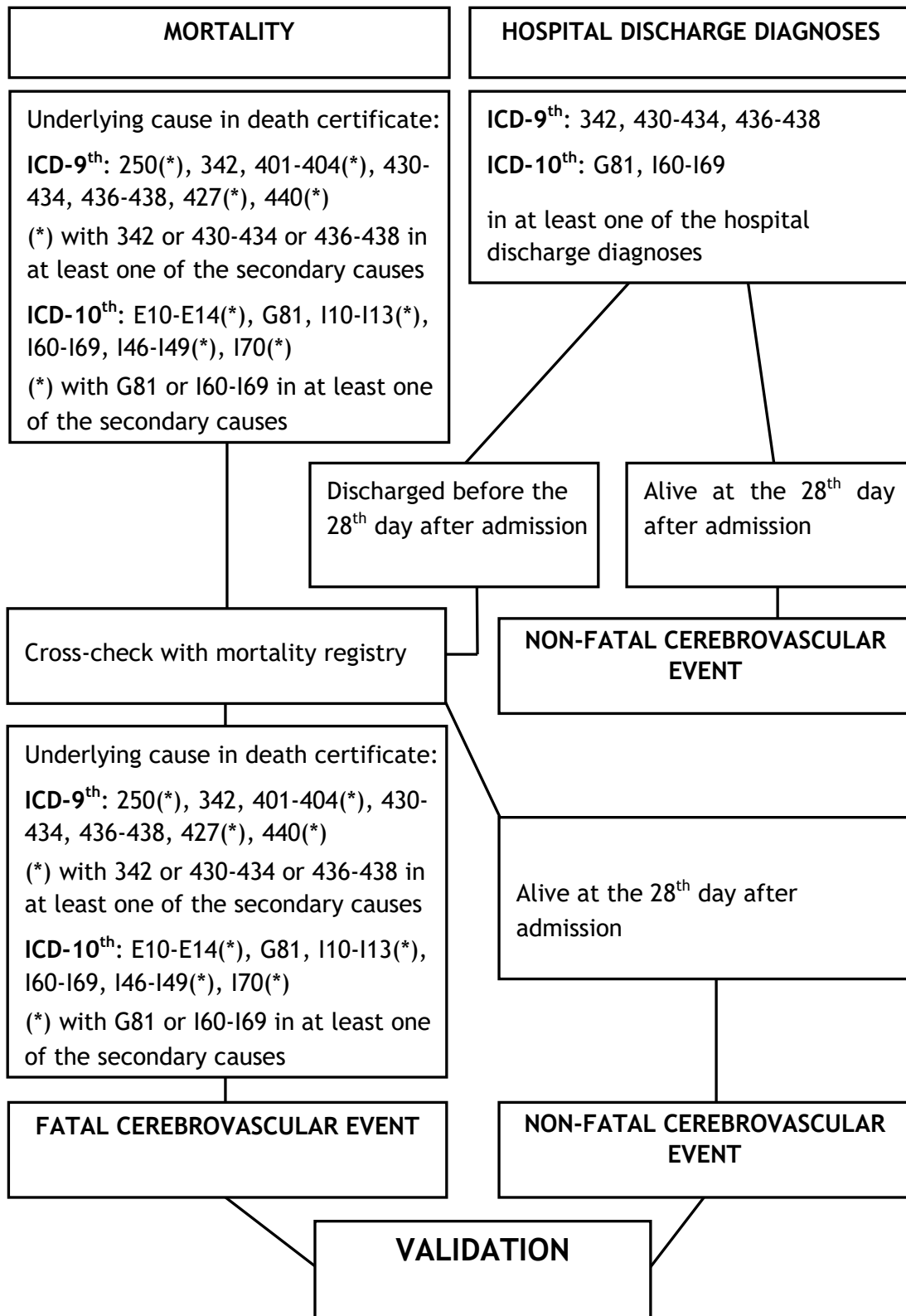


Attack rates, which included first and recurrent events, in the age range 35-74 years, were then calculated dividing the number of estimated events by the resident population and by 10-years age range, separately by gender, and standardized by direct method using the European Standard Population; case fatality rates at the 28<sup>th</sup> day were determined by the ratio between estimated fatal events and total events, separately by gender, and standardized by direct method using the European Standard Population. A similar methodological path was applied to identify fatal and non-fatal stroke events and to estimate related attack rates and case fatality in the population (Figures 6, 3-4).

**Figure 5** - Flow-chart summarizing the methodological path to select fatal and non-fatal *coronary* events starting from mortality and hospital discharge diagnoses-HDRs databases in the Italian pilot Registry of Coronary events.



**Figure 6** - Flow-chart summarizing the methodological path to select fatal and non-fatal *cerebrovascular events* starting from mortality and hospital discharge diagnoses-HDRs databases in the Italian pilot Registry of Cerebrovascular events



## 5. QUALITY CONTROL

See the power point presentation 'Quality control' available on the web site [www.cuore.iss.it](http://www.cuore.iss.it/eng/bridge/WP8/TrainingPopulation-basedRegistries.asp); website address: <http://www.cuore.iss.it/eng/bridge/WP8/TrainingPopulation-basedRegistries.asp>

Quality assessment is extremely important for a correct use of registry data.

The term "quality" can be applied to registries to describe the confidence that the registry design, management, and data analysis can offer against bias and errors in inference, which means erroneous conclusions drawn from the registry.

In determining the utility of a registry for research or decision-making, it is critical to understand the quality of the procedures used to obtain the data and the quality of the data stored in the database, which is extremely important also for a valid monitoring and comparison between regions and countries. Data management - i.e. the integrated system for registry data collection, cleaning, storage, monitoring, reviewing and reporting - determines data utility to achieve the goals of the registry. All quality control steps are influenced by registry hypothesis and objectives.

Registry protocols should consider how to ensure quality to a sufficient level for the intended purposes and should also consider how to develop appropriate quality assurance plans. Registry managers should assess and report on those quality assurance activities; data collection and validation procedures have better quality if performed in accordance with international guidelines or standards.

The quality of registry data is evaluated according to data completeness, validity and timeliness.

Quality evaluations of any registry must be done with respect to the registry's purposes and should consider the disease, the type and the purpose of the registry, its feasibility and affordability.

### *Data quality assurance*

Quality assurance aims at affirming that data was collected in accordance with established procedures and that data meets the requested quality standards to accomplish the registry intended purposes and the intended use of data. Quality assurance activities generally fall under three main categories: 1. Data quality assurance, 2. quality assurance of registry procedures, and 3. quality assurance of computerized systems. Since many registries are large, the level of quality assurance that can be obtained may be limited by budgetary constraints.

1. Requirements for **quality assurance should be defined during the registry planning**. Because certain requirements may have significant cost implications, a quality assurance plan is recommended. It should be based on **identifying the most important sources of error or potential bias in procedures** that may affect the quality of the registry in the context of its intended purpose.

Clarity of description and standardization of definitions are essential to data quality assurance. Data collection procedures for each registry should be clearly defined and described in a detailed manual. Event inclusion and exclusion criteria and the validation procedures should be reported. A data dictionary should also be done, with a detailed description of each variable used by the registry, including the source of the information, coding information, if used, and normal ranges, if relevant. Data definitions include ranges and acceptable values for each individual data element. For example, logic checks for data capture validity may be created for data elements that should be mutually exclusive. Data managers should develop formal data review manuals to be used by reviewers and data entry personnel. Registries to evaluate events should use predetermined and defined uniform and systematic methods of data collection, all data-related procedures—including the permitted sources of data, data elements and their definitions and data quality requirements.

2. Generally, registry manuals of operations includes instructions to search data that will go into the registry (e.g. specific diagnoses or laboratory results). Data codification could include either standardized codes from a codebook (e.g. the ICD-10 code) corresponding to a text diagnosis in a chart, or codes that may be unique to the registry (e.g. a scale from 1 to 4 for definite, possible, insufficient data, non-event). All abstraction and coding instructions must be carefully documented and incorporated into a data dictionary for the registry. Because of the “noise” in unstructured, hard-copy documents (e.g. spurious marks or illegible writing) and the lack of precision in natural language, clinical data processed from the same documents by different abstracters may differ. This is a potential source of error in a registry. Manuals should also include data validation rules referring to the logical checks on data entered into the database against predefined rules for either value ranges or logical consistency, with respect to other data fields for the same event.

**Data validation rules** should also include information on how to handle missing data; invalid entries (e.g. multiple selections in a single-choice field, alphabetic data in a numeric field); erroneous entries (e.g. patients of the wrong gender answering gender-based questions); and inconsistent data (e.g. an answer to one question contradicting the answer to another one). Guidelines should consider procedures to remedy these data problems. For example, a data error on an interview form could require to query the interviewer or the patient, or to refer to other data sources that may be able to solve the problem. Documentation of any data review activity and remediation efforts, including dates, times, and results of the query, should be maintained.

3. **Data cleaning** refers to the correction of data problems, including missing values, incorrect or out-of-range values, responses that are logically inconsistent with other responses in the database, and duplicate patient records. How and to what level data will be cleaned should be addressed upfront in a data manual identifying data elements to be cleaned, describing data validation rules or logical checks for out-of-

range values, explaining how to handle missing values and values that are logically inconsistent, and discussing how to identify and manage duplicate patient records. Ideally, automated data checks are pre-programmed into the database for presentation at the time of data entry. These data checks are particularly useful for cleaning data at the site level, when the patient or medical record is readily accessible. Even relatively simple edit checks, such as range values for laboratories, can have a significant effect on improving data quality. The automated mechanisms of numeric checks and alerts can improve validity and reliability of data collected. Data managers shall carefully review the data, using both data extracts analysed by algorithms and hand review, to identify discrepancies and generate “queries” to send to the sites for solution. It is very difficult to foresee all potential data discrepancies at the time of development of the data management manual and edit checks. Therefore, even with the use of automated data validation parameters, some manual cleaning is often still performed.

The creation of explicit **data definitions for each variable to be collected is essential to select data elements**. This is important to ensure the internal validity of the proposed study, so that all participants in the data collection acquire the requested information in the same reproducible way. When deciding on data definitions, it is important to determine which data elements are required and which elements may be optional. This is particularly true in cases where the registry may collect a few additional “nice to know” data elements. The determination will differ depending on whether the registry uses existing medical record documentation to obtain a particular data element or whether the clinician is asked directly. However, if clinicians are asked to provide this information prospectively, they can immediately do it. Moreover, accounting for missing or unknown data should be taken into consideration. In some cases, a data element may be unknown or not documented for a particular patient, and a follow-up with the patient to answer the question may not be possible. Therefore, the form should contain an option for “not documented” or “unknown” data, so that the person filling in the case report form could provide a response to each question rather than leaving some unanswered. Depending on the analysis plans for the registry, the distinction between undocumented data and missing data may be important.

As regards reduction of **false-positive cases**, strategies can foresee that some evidence in the patient record of medical procedures (e.g. cholecystectomy for gallstone disease or podiatry examination for type 1 diabetes) or interventions (e.g. insulin or glucose-lowering medications for type 1 diabetes) could provide greater confidence in the validity of the event definition. Such an approach often results in a reduced number of cases included and a reduced precision, but provides improved validity. **Registration of causes of death may be incorrect** and may need to be validated, and the collection of information on deaths occurring outside the area of residence has to be ensured. It is to be expected that some events occur outside hospital. **Duplicate registration** of the same case should also be avoided by paying careful attention to record-linkage during the registration process. When the event is defined (codes and duration), it may be easy

to identify duplicate coding and take out information for quality control purposes. Duplicate codes may include events transferred from one ward to another. In some cases the duration of the admission is very short (<2 days) either because the patient is transferred elsewhere or because of diagnosis misclassification. These cases may also be picked up for validation.

As regards accuracy improvement of data exposure, one frequent shortcoming of epidemiologic research is to compare the occurrence of disease in an index group with the occurrence of disease in all other groups who do not satisfy the index group definition. Such methods are easily applicable in administrative databases, due to the abundance of participants who do not meet the index group definition. This “all others” reference group is therefore usually a poorly defined mixture of individuals. For example, if a pharmaceutical registry is used to compare the incidence of a disease in statin users with the incidence of disease in those who do not use statins, the nonusers’ reference group will contain individuals with indications for statin use but who have not been prescribed statins, as well as individuals without indications for statin use. Nonusers also differ from users in the frequency of contact with medical providers, and this raises the potential for differential accuracy ascertainment of health outcomes. It is therefore preferable to first ensure that the nonusers’ reference group contains individuals who have indications for treatment, and who, if possible, receive alternative therapies for the same indication. If the different categories of statins - such as hydrophilic and hydrophobic statins - are assigned to patients according to the patients’ biological characteristics, then a comparison between users of hydrophilic statins and users of hydrophobic statins is often more valid. On the basis of these definitions, only individuals with indications for statins, and treated with statins, are included in the analysis, and therefore the possibility of confounding is reduced in comparison to a follow up carried out on other indications.

**It is always important to clearly define the registry objective, the patient population, as well as potential confounders and modifiers. Researchers must also understand the conditions under which data were originally collected.** The selection of data elements requires balancing factors such as their importance for registry integrity and for the analysis of primary outcomes, their reliability, their contribution to the overall burden for respondents, and the incremental costs associated with their collection. Selection begins with the identification of relevant domains. Specific data elements are then selected, with a focus on established clinical data standards, common data definitions, and the use of patient identifiers. It is important to determine which elements are absolutely necessary, and which ones are desirable but not essential. **Overall, the choice of data elements should be guided by parsimony, validity, and a focus on achieving the registry purpose.** Information on behavioural and lifestyle factors (e.g. tobacco use, alcohol drinking, exercise habits, and diet) is infrequently captured or is poorly measured in many databases. Some databases can provide proxy measurements of these behavioural factors. For example, poor lung function or diagnosis of chronic obstructive pulmonary disease is a proxy marker for tobacco smoking history; alcohol-related diseases, such as cirrhosis, or prescriptions for disulfiram can be used as

proxy markers for alcohol abuse, and medically diagnosed obesity may be a proxy marker for poor diet and lack of exercise. However, none of these proxies provides a reliable measure of the actual concept.

To prevent the low validity of a registry, and, in particular, the loss of generalizability, it is important to consider to what extent an area is representative of the whole country (representativeness): it could be representative according to mortality rates, the distribution of risk factors (socioeconomic status and health behaviour), and the distribution of health services (specialized hospital and GPs). In some countries, it might be better to start implementing a register with high-risk areas. The population to be monitored should be selected to produce estimates of disease rates that are sufficiently robust from a statistical point of view, so that trends can be established and data comparability ensured. In general, it is necessary to select more than one area representative of socioeconomic or ethnic differences, so to have a comprehensive picture for the whole country, and it is recommended to establish a coordinating body between the areas to ensure comparability.