Protocol for validating cardiovascular and cerebrovascular ICD-9-CM codes in healthcare administrative databases: the Umbria Data Value Project

Francesco Cozzolino, Iosief Abraha, Massimiliano Orso, Anna Mengoni, Maria Francesca Cerasa, Paolo Eusebi, Giuseppe Ambrosio, Alessandro Montedori

ABSTRACT

Introduction: Administrative healthcare databases can provide a comprehensive assessment of the burden of diseases in terms of major outcomes, such as mortality, hospital readmissions and use of healthcare resources, thus providing answers to a wide spectrum of research questions. However, a crucial issue is the reliability of information gathered. Aim of this protocol is to validate International Classification of Diseases, 9th Revision—Clinical Modification (ICD-9-CM) codes for major cardiovascular diseases, including acute myocardial infarction (AMI), heart failure (HF), atrial fibrillation (AF) and stroke.

Methods and analysis: Data from the centralised administrative database of the entire Umbria Region (910 000 residents, located in Central Italy) will be considered. Patients with a first hospital discharge for AMI, HF, AF or stroke, between 2012 and 2014, will be identified in the administrative database using the following groups of ICD-9-CM codes located in primary position: (1) 410.x for AMI; (2) 427.31 for AF; (3) 428 for HF; (4) 433.x1, 434 (excluding 434.x0), 436 for ischaemic stroke, 430 and 431 for haemorrhagic stroke (subarachnoid haemorrhage and intracerebral haemorrhage). A random sample of cases, and of non-cases, will be selected, and the corresponding medical charts retrieved and reviewed for validation by pairs of trained, independent reviewers. For each condition considered, case adjudication of disease will be based on symptoms, laboratory and diagnostic tests, as available in medical charts. Divergences will be resolved by consensus. Sensitivity and specificity with 95% CIs will be calculated.

Ethics and dissemination: Research protocol has been granted approval by the Regional Ethics Committee. Study results will be disseminated widely through peer-reviewed publications and presentations at national and international conferences.

INTRODUCTION

As computer technology continues to advance, use of digital administrative databases is increasingly growing in healthcare settings worldwide. These databases anonymously store patients' data about healthcare assistance they received, including birth, death or disease treatment. Usually, diagnosis of the disease is associated with a specific code from the International Classification of Diseases (ICD-9) 9th or 10th revision. The ICD is designed to map health conditions to corresponding generic categories along with specific variations. Merging of individual patient data from administrative databases with other sources (e.g., prescription and laboratory data) allows investigating a wide range of relevant and often unique public health questions monitoring population health status over time, and performing population-based pharmacoepidemiological research. These databases, therefore, have the potential to address important

Strengths and limitations of this study

The study will evaluate the validity of the International Classification of Diseases, 9th Revision—Clinical Modification (ICD-9-CM) codes for cerebrovascular and cardiovascular diseases (acute myocardial infarction, heart failure, atrial fibrillation and stroke), using the administrative database of Umbria Region.

The strength of this study is that it will review original source documentation available in medical charts to adjudicate case of cerebrovascular and cardiovascular diseases.

Once the administrative database will be validated for the considered diseases, it can be used for outcome research including pharmacoepidemiology, health service research and quality of care research.

Validation studies of administrative data are related to their context, and are not immediately generalisable to other settings.


© BMJ Open 2017. BMJ Publishing Group Ltd. For permission to reuse any of this content visit http://creativecommons.org/licenses/by/4.0/
issues in postmarketing surveillance\textsuperscript{3} epidemiology\textsuperscript{1} quality performance, and health services research.\textsuperscript{5} However, there is a concern that the power of administrative databases as a source of healthcare information cannot be fully exploited unless they are thoroughly validated.\textsuperscript{6–9} For instance, a systematic review\textsuperscript{10} of ICD-9 code validation in Italian administrative databases reported that only a few regional databases have been validated, and just for a limited number of ICD-9 codes of diseases.\textsuperscript{11–18}

While non-clinical information in healthcare databases, such as demographic and prescription data, are highly accurate,\textsuperscript{19} reliability of registered diagnoses and procedures is variable.\textsuperscript{20} Determining the accuracy of the latter two categories of clinical information is important to all potential users, and involves confirming the consistency of information within the databases with the corresponding clinical records of patients.\textsuperscript{19} Hence, it is imperative that health authorities systematically validate their databases for major diseases to productively use the information they contain.

In Western countries, cardiovascular and cerebrovascular diseases are the leading cause of mortality and morbidity, and represent a major social and economic problem. In this context, most of the burden is due to occurrence of acute myocardial infarction (AMI),\textsuperscript{22} heart failure (HF),\textsuperscript{23} atrial fibrillation (AF)\textsuperscript{24} and stroke.\textsuperscript{17} Administrative healthcare databases can provide a comprehensive understanding of the burden of these diseases in terms of important outcomes such as mortality, hospital readmissions and use of other healthcare resources. In addition, such databases can aid in monitoring adherence to essential drug therapies, including the use of evidence-based therapies.

The Regional Health Authority of Umbria has started a research regarding case definitions of diseases\textsuperscript{26–28} as well as validating ICD-9 codes for diseases.\textsuperscript{29} The objective of the present protocol is to evaluate the accuracy of the ICD-9-CM codes related to AMI, HF, AF, ischaemic and haemorrhagic stroke (intracerebral and subarachnoidal haemorrhage) in the administrative database of the Regional Health Authority of Umbria.

**Methods**

**Setting and data source**

**Administrative database**

The Regional Administrative Database of Umbria gathers information regarding all hospital admission medical records on all 910 000 residents, including personal demographics, hospital admission and discharge dates, vital status, hospital department, primary and secondary diagnoses and surgical or diagnostic procedures. In addition, the database records all drug prescriptions listed in the National Drug Formulary, and it allows identification of the prescriber. In Italy, all hospital medical doctors, including those who fill discharge diagnoses, are salaried.

The Regional Administrative Database has been used for pharmacoepidemiology and drug-related outcome researches.\textsuperscript{2} Each resident has a unique national identification code, with which it is possible to link the various types of information, corresponding to each person, within the database. In Italy, healthcare assistance is covered almost entirely by the Italian National Health System (NHS), therefore most residents’ significant healthcare information can be found within the healthcare databases.

**Source population**

Source population will be represented by permanent residents aged 18 or above of Umbria Region. Any resident that has been discharged alive (with exclusion of voluntary discharge and interhospital transfer of patients) from a hospital with a diagnosis of AMI, AF, HF and stroke (ischaemic stroke, intracerebral haemorrhage, subarachnoidal haemorrhage) will be considered. Residents that have been hospitalised outside the regional territory of competence will be excluded from analysis.

**Case selection and sampling method**

Patients with the first occurrence of below diagnosis located in primary positioning of the administrative database between 2012 and 2014 will be identified using the following groups of ICD-9-CM codes: (1) 410 for AMI; (2) 427.31 for AF and 427.32 for atrial flutter; (3) 428 for HF; (4) 433.x1, 434 (excluding 434.x0), 436 for ischaemic stroke, 430 and 431 for haemorrhagic stroke (subarachnoidal haemorrhage and intracerebral haemorrhage). It is worth mentioning that according to the Italian legislation, the principal diagnosis coding needs to be based on the condition identified at the end of hospitalisation, which constitutes the main cause of the need of treatment and/or diagnostic tests, and is mainly responsible for the use of resources.\textsuperscript{33} Table 1 displays the description of the ICD-9-CM codes for each disease of interest.

**Validation criteria**

Criteria used for validation of the considered ICD-9-CM codes will be those derived from international guidelines if available, or systematic reviews published on these topics.

**Acute myocardial infarction**

For validation of AMI, we will consider the Guidelines of the European Society of Cardiology,\textsuperscript{35} using the relative set of guidelines available at the time of each patient’s discharge. To adjudicate the event as an AMI, in addition to troponin release (with at least one peak level greater than the upper limit of reference range), at least one of the following criteria of the reference standard needs to be met:

A. Cardiac symptoms at rest (eg, such as angina, chest pain, chest discomfort, chest heaviness, or substernal
chest pain), consistent with coronary artery disease/AMI.

B. Evidence for a new AMI on the presenting ECG, as demonstrated by any one of these:
1. ST elevation (≥1 mm) in two contiguous leads;
2. new Q waves;
3. new left bundle branch block.

C. Presence of myocardial infarction/injury (exclude old myocardial infarction) on any ECG during index hospital stay.

D. Imaging evidence of new loss of viable myocardium, or new regional wall motion abnormality.

E. Identification of an intracoronary thrombus by angiography.

Atrial fibrillation
To validate ICD-9 code 427.31 (index test) related to AF, we will require at least one ECG tracing documenting the presence of AF when present in the medical record. However, in the medical literature, alternative

---

Table 1 Description of ICD-9-CM codes

<table>
<thead>
<tr>
<th>ICD-9-CM diagnosis code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute myocardial infarction</td>
<td>410.0 Anterolateral wall</td>
</tr>
<tr>
<td>Atrial fibrillation and flutter</td>
<td>427.31 Atrial fibrillation</td>
</tr>
<tr>
<td>Heart failure</td>
<td>428.0 Congestive heart failure, unspecified</td>
</tr>
<tr>
<td>Haemorrhagic stroke</td>
<td>430 Subarachnoid haemorrhage</td>
</tr>
<tr>
<td>Ischaemic stroke</td>
<td>433 Occlusion and stenosis of precerebral arteries with cerebral infarction</td>
</tr>
<tr>
<td>Stroke</td>
<td>433.01 Occlusion and stenosis of basilar artery</td>
</tr>
<tr>
<td>Stroke</td>
<td>433.11 Occlusion and stenosis of carotid artery</td>
</tr>
<tr>
<td>Stroke</td>
<td>433.21 Occlusion and stenosis of vertebral artery</td>
</tr>
<tr>
<td>Stroke</td>
<td>433.31 Occlusion and stenosis of multiple and bilateral</td>
</tr>
<tr>
<td>Stroke</td>
<td>433.81 Other specified occlusion and stenosis of precerebral artery</td>
</tr>
<tr>
<td>Stroke</td>
<td>433.91 Unspecified occlusion and stenosis of precerebral artery</td>
</tr>
<tr>
<td>Stroke</td>
<td>434 Occlusion of cerebral arteries with cerebral infarction</td>
</tr>
<tr>
<td>Stroke</td>
<td>434.01 Cerebral thrombosis with cerebral infarction</td>
</tr>
<tr>
<td>Stroke</td>
<td>434.11 Cerebral embolism with cerebral infarction</td>
</tr>
<tr>
<td>Stroke</td>
<td>434.91 Cerebral artery occlusion, unspecified with cerebral infarction</td>
</tr>
</tbody>
</table>
criteria have been used to identify cases of AF: (1) any mention of current AF, without a reference ECG; (2) two ECG measurements documenting the presence of AF in the same hospital admission; (3) the presence of at least two ECG documenting the presence of AF in more than one hospital admissions for AF in the same period. Accordingly, we will also record and analyse charts according to these different criteria for subsequent comparison, in order to understand which algorithm gives the best diagnostic yield.

Heart failure
For the validation of the ICD-9-CM 428.x relating to HF, we will consider the European Society of Cardiology Heart Failure Guidelines published at the time of patient’s discharge. Those guidelines consider the algorithm for the diagnosis of HF in the non-acute setting and in the acute setting. Since the ICD-9 codes do not distinguish between the acute and non-acute setting, we will combine the clinical presentation of both settings.

Diagnosis of HF will be adjudicated when, in addition to the presence of symptoms (such as dyspnoea, orthopnoea), or presence of signs at physical examination (rales, bilateral ankle oedema, increased jugular venous pressure, apical beat), at least one of the following conditions are found in the medical chart:
1. any abnormality in resting ECG (ie, sinoatrial disease, atrioventricular block, or abnormal intraventricular conduction);
2. plasma natriuretic peptides (NPs) concentration (elevated levels of NPs are considered brain natriuretic peptide (BNP) ≥35 pg/mL, and/or N-terminal pro-brain natriuretic peptide (NT-proBNP) ≥125 pg/mL);
3. echocardiography abnormalities (ventricular and atrial volumes and function) attributable to heart failure.

Stroke
For the validation of stroke, we will consider the ICD-9 codes related to stroke valid when both of the following conditions will be present:
1. any abnormality in resting ECG (ie, AMI classified as positive by the administrative database and medical record review) and all cases deemed positive by medical chart review. Sensitivity expresses the proportion of ‘true positives’ (ie, AMI identified as negative by the administrative database and medical record review), and with all cases deemed negative by medical chart review. For both sensitivity and specificity, 95% CIs will be calculated. We account for a 10% of missing charts for sensitivity and specificity. Positive and negative predicting values will also be reported along with their 95% CI.
Reporting
Complete, transparent and accurate reporting is essential in research, because it allows readers to assess internal validity as well as to evaluate the generalisability and applicability of results. To ensure quality and thoroughness of reporting, any dissemination or publication of the results from the present study will follow recommended guidelines based on the criteria published by the Standards for Reporting of Diagnostic accuracy (STARD) initiative for the accurate reporting of investigations of diagnostic studies. The project will be completed within 2 years.

DISCUSSION
Administrative databases constitute a valid instrument to get insights into epidemiology and health management, and to get ‘real-world’ knowledge, of major disease under situations in which randomised trials are not able to provide the required evidence, for practical or economic reasons, or because of selective inclusion criteria. Epidemiological studies are frequently based on administrative claims databases to identify cases of specific diseases, such as AMI, and often contain ICD-9 codes. These codes have the advantage of being widely available and require lower effort and cost than consulting medical charts. However, their effective exploitation hinges on the prerequisite that coding does reliably reflect the clinical condition of a given patient.

In this protocol, we present the approach to be used to analyse the validity of ICD-9-CM codes for important cardiovascular diseases, such as AMI, HF, AF, and stroke using the Regional Health database of Umbria.

We seek to verify whether an acceptable level (at least 80% for sensitivity and 90% for specificity) of operating characteristics is in fact achieved for each identified ICD-9-CM codes. In case of inadequate level of validation, algorithms will be considered to explore potential elements to enhance the level of validation. Elements of an algorithm may include the type of the data source, number of years of administrative data, other diagnostic codes, medication use or laboratory data. The algorithms selected for evaluation will be based on the literature review and consultations with clinicians, health services researchers. If the desired accuracy is not reached despite the use of algorithms, results of our research will be used to identify potential areas for improvement for future validation.

DISSEMINATION
Study results will be disseminated widely through peer-reviewed publications and presentations at national and international conferences.

Twitter Follow Paolo Eusebi @paoloeusebi

Contributors IA, FC, MO, AMo and GA conceived the study. IA, FC, MO, GA, AMe, MFC, PE and AMo were responsible for designing the protocol. IA, FC, MO, GA and AMo drafted the protocol manuscript. IA, FC and MO developed the search strategy. IA, FC, MO, GA, AMe, MFC, PE and AMo critically revised the successive versions of the manuscript and approved the final version.

Funding This review protocol was funded by the Regional Health Authority of Umbria. [Progetto Data-Value: valorizzazione del dato sanitario regionale per la Ricerca dei Servizi Sanitari (Health Services Research)].

Disclaimer The study funder was not involved in study design or writing of the protocol.

Competing interests None declared.

Ethics approval Ethics approval has been obtained from the Regional Ethics Committee of Umbria (CEAS).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Study results will be disseminated widely through peer-reviewed publications and presentations at national and international conferences.

Open Access This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

REFERENCES