BMJ open accessible medical research

# Identification of acute myocardial infarction from electronic healthcare records using different disease coding systems: a validation study in three European countries

Preciosa M Coloma,[1] Vera E Valkhoff,[1,2] Giampiero Mazzaglia,[3] Malene Schou Nielsson,[4] Lars Pedersen,[4] Mariam Molokhia,[5] Mees Mosseveld,[1] Paolo Morabito,[6] Martijn J Schuemie,[1] Johan van der Lei,[1] Miriam Sturkenboom,[1,7] Gianluca Trifirò,[1,6] on behalf of the EU-ADR Consortium

**To cite:** Coloma PM, Valkhoff VE, Mazzaglia G, *et al*. Identification of acute myocardial infarction from electronic healthcare records using different disease coding systems: a validation study in three European countries. *BMJ Open* 2013;**3**: e002862. doi:10.1136/bmjopen-2013-002862

▸ Prepublication history and additional material for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2013-002862).

Received 10 March 2013
Accepted 10 May 2013

For numbered affiliations see end of article.

**Correspondence to**
Dr Preciosa M Coloma;
p.coloma@erasmusmc.nl

## ABSTRACT

**Objective:** To evaluate positive predictive value (PPV) of different disease codes and free text in identifying acute myocardial infarction (AMI) from electronic healthcare records (EHRs).

**Design:** Validation study of cases of AMI identified from general practitioner records and hospital discharge diagnoses using free text and codes from the International Classification of Primary Care (ICPC), International Classification of Diseases 9th revision-clinical modification (ICD9-CM) and ICD-10th revision (ICD-10).

**Setting:** Population-based databases comprising routinely collected data from primary care in Italy and the Netherlands and from secondary care in Denmark from 1996 to 2009.

**Participants:** A total of 4 034 232 individuals with 22 428 883 person-years of follow-up contributed to the data, from which 42 774 potential AMI cases were identified. A random sample of 800 cases was subsequently obtained for validation.

**Main outcome measures:** PPVs were calculated overall and for each code/free text. 'Best-case scenario' and 'worst-case scenario' PPVs were calculated, the latter taking into account non-retrievable/non-assessable cases. We further assessed the effects of AMI misclassification on estimates of risk during drug exposure.

**Results:** Records of 748 cases (93.5% of sample) were retrieved. ICD-10 codes had a 'best-case scenario' PPV of 100% while ICD9-CM codes had a PPV of 96.6% (95% CI 93.2% to 99.9%). ICPC codes had a 'best-case scenario' PPV of 75% (95% CI 67.4% to 82.6%) and free text had PPV ranging from 20% to 60%. Corresponding PPVs in the 'worst-case scenario' all decreased. Use of codes with lower PPV generally resulted in small changes in AMI risk during drug exposure, but codes with higher PPV resulted in attenuation of risk for positive associations.

**Conclusions:** ICD9-CM and ICD-10 codes have good PPV in identifying AMI from EHRs; strategies are necessary to further optimise utility of ICPC codes and

## ARTICLE SUMMARY

### Article focus

- This article evaluates the positive predictive value (PPV) of different disease codes and free-text search in identifying cases of acute myocardial infarction (AMI) from population-based healthcare databases in three countries in Europe.

### Key messages

- The overall PPV of different disease coding systems for identifying AMI was good, ranging from a 'best-case scenario' PPV of 75% (International Classification for Primary Care (ICPC), Netherlands) to 95% (International Classification of Diseases 9th revision-Clinical Modification (ICD9-CM)) to 100% (ICD-10th revision (ICD-10), Denmark). These findings are consistent with PPV estimates for ICD9-CM and ICD-10 cited in the literature. Until now, there is no study describing the PPV of ICPC codes for identifying AMI.
- Use of free text alone had a lower PPV, ranging from 'best-case scenario' PPV of 20–60%. Strategies are necessary to optimise use of natural language processing in the identification of AMI in these electronic healthcare record (EHR) data.
- Misclassification of AMI cases resulting from the use of disease codes (or free text) with low PPV has corresponding implications in the estimation of incidence rates. Studies using EHR data to derive incidence rates of clinical events should thus correct for this potential misclassification.
- Use of more specific disease codes for identifying AMI during drug use may lead to a small but significant change in risk estimates and at the expense of decreased precision. Further studies are warranted to investigate the effect of different PPVs on outcome misclassification and should take into account the type of database as well as test more drug-event associations and control for other confounders.

## ARTICLE SUMMARY

### Strengths and limitations of this study

- Large healthcare databases covering a total population of over four million from three countries were investigated— a formidable challenge in itself because of the diversity in healthcare and disease coding practices. The implementation of a standardised validation questionnaire facilitated harmonised data collection and analysis across databases without compromising data protection. The opportunity to simultaneously evaluate different disease coding systems as well as free text also allowed the investigation of the effect of outcome misclassification.
- This study evaluated the accuracy of the codes using the PPV; however, there are other measures such as sensitivity and negative predictive value that could not be calculated.
- Despite the reasonable size of the random sample used in this validation study, it was not adequate to permit evaluation of some of the individual, less frequently occurring, codes.

free-text search. Use of specific AMI disease codes in estimation of risk during drug exposure may lead to small but significant changes and at the expense of decreased precision.

## INTRODUCTION

Cardiovascular diseases remain an important cause of morbidity and mortality worldwide and the conduct of disease surveillance has changed with the availability of secondary data sources as well as changes in disease coding terminologies. Information derived from multicountry databases containing electronic healthcare records (EHRs) is increasingly being used for drug safety surveillance, including drug-related adverse cardiovascular outcomes.[1–3] Cases of acute myocardial infarction (AMI) may be identified using electronic databases from different countries, which may differ not only in their healthcare systems, but also in their disease registration and coding procedures. Innovations in recent years have brought about discovery and subsequent clinical use of biomarkers that allow earlier recognition of disease as well as therapeutic interventions that reduce the extent of myocardial injury and mortality. Such developments have led to revisions in the definition of AMI and changes in diagnosis and prognosis.[4–6] Studies that estimate AMI incidence from EHR data must also then consider the implications of new diagnostic criteria on the disease coding practices of such databases.[7–10]

The accuracy of specific disease coding terminologies in identifying AMI from healthcare data has been evaluated in previous studies. These studies, mostly performed on data representing administrative/insurance claims, have derived positive predictive values (PPVs) of International Classification of Diseases-9th revision-Clinical Modifications (ICD9-CM) codes as well as diagnosis-related groups codes, used in billing.[11–14] A recent study evaluated the PPV of ICD-10th revision (ICD-10) diagnostic codes used to assess Charlson comorbidity index conditions, including myocardial infarction, in the population-based Danish National Registry of Patients.[15] Until now, there is no study that has evaluated the validity of International Classification of Primary Care (ICPC) codes and unstructured (free text) search, or the combination of diagnosis codes and free-text search, in the identification of AMI from electronic healthcare data. Furthermore, the opportunity to simultaneously evaluate different disease coding systems as well as free-text permits investigation of the effect of outcome misclassification.

We conducted a validation study within the context of the EU-ADR Project (Exploring and Understanding Adverse Drug Reactions by Integrative Mining of Clinical Records and Biomedical Knowledge, http://www.euadr-project.org/). Funded by the European Commission under its Seventh Framework Programme, the EU-ADR Project has designed and developed a computerised integrative system that exploits EHR data from different countries (as well as biomedical data) to facilitate early detection of adverse drug reactions.[3] Databases contributing EHR data to the Project are part of the EU-ADR network and represent a huge resource for monitoring of drug safety in Europe. In this validation study, we evaluated and compared PPV of free-text search and disease codes from three different terminologies: ICPC; ICD9-CM and ICD-10 in identifying cases of AMI from population-based healthcare databases in Denmark, Italy and the Netherlands. We further assessed the effect of outcome misclassification on the estimation of risk of AMI during drug use.

## METHODS

### Data sources

The EU-ADR database network currently comprises anonymised demographic and clinical data of over 20 million individuals from eight population-based EHR databases in three European countries.[3] The data are pooled using a distributed network approach that allows data holders to maintain control over their protected data. Validation of AMI case identification was performed in three of these databases: (1) Integrated Primary Care Information (IPCI, the Netherlands); (2) Health Search/CSD Patient DB (HSD, Italy) and (3) Aarhus University Hospital Database (Aarhus, Denmark). IPCI and HSD are both general practice (GP) databases documenting patient consults, including referrals for hospitalisation or specialist care as well as prescriptions for medications. Aarhus is a comprehensive record-linkage database system in which drug dispensation data are linked to a registry of hospital discharge diagnoses and various other registries, including death registries. All these databases have been extensively used for epidemiological research.[16–18] A more detailed description of the characteristics of the databases has been previously published.[3 19] A table of database characteristics of the entire EU-ADR network is provided in online supplementary appendix 1. The three databases employ different disease coding terminologies: IPCI uses ICPC; HSD uses ICD9-CM and Aarhus uses ICD-10. Clinical narratives from general practitioners' notes in both HSD and IPCI are also recorded as unstructured text that can be used to identify medical events. Standardised data extraction was carried out using the Java-based software Jerboa, developed within the EU-ADR Project.[3]

### Cohort definition and follow-up time

To harmonise follow-up definitions across databases, we defined the eligibility period for each patient as starting on the date of registration in the database and ending on the date the patient transfers out of the system, with the last supply of data, occurrence of AMI (as described below) or on the patient's death, whichever is earlier. In order to be included in the study cohort, participants had to have at least 1 year of continuous and valid data.

### Identification of AMI

Potential cases of AMI were initially identified using harmonised and database-specific codes derived from hospital discharge diagnoses (in the case of Aarhus) or from general practitioner diagnoses (in the case of IPCI and HSD). These codes included the ICPC code K75 (IPCI), ICD9-CM codes 410/410.x/410.x0 (HSD), and the ICD-10 codes I21.x (Aarhus). IPCI and HSD also performed free-text search using specific key words. The ICD9-CM code 411.81 (corresponding to acute coronary occlusion) was specifically used in HSD, in combination with free text to refine the search. The free-text search strings employed in IPCI and HSD are given in online supplementary appendix 2. The process of mapping and harmonisation of event data extraction from different EHR databases in the EU-ADR project was based on medical concepts derived from the Unified Medical Language System.[20 21] We only considered the first occurrence (first diagnosis) of AMI in each patient.

### Case validation

Random sampling of cases for validation was carried out separately in each of the three databases using a specific module in the Jerboa software designed for this purpose. The module uses the standard random function in Java that generates random numbers. We required a sample size of 200 cases/database. Since the use of free-text search was known to be more extensive in IPCI, an additional 200 potential cases identified by free text were obtained in IPCI. A manual review of GP records and hospitalisation charts was performed by medically trained assessors using a standardised questionnaire, pilot-tested in the databases and reviewed by a panel of experts. Diagnostic criteria for AMI as prescribed in the current guidelines[4 7] were incorporated in the questionnaire, as well as information regarding cardiovascular risk factors and potential alternative diagnoses that could explain findings suggestive of AMI. For GP databases (IPCI and HSD), it was also determined whether the AMI diagnosis was made either directly by the general practitioner or by a medical specialist. The standardised questionnaire was then implemented as a computerised data entry algorithm using the custom-built software Chameleon. This software was installed locally in each database, allowing the data holders to keep patient-level data within their protected environment. The data entry algorithm is shown in figure 1 and the questionnaire in online supplementary appendix 3. On the basis of the information collected in the questionnaire, the potential AMI cases were classified as: (1) definite case; (2) non-case or (3) non-assessable case, if the available information was deemed to be insufficient for the case validation.

### Assessment of index date

We determined how the coded date of the event (which is detected automatically) was related to the actual date of diagnosis of AMI and to the date of onset of first symptoms, as derived from manual validation. In addition, for the administrative database Aarhus, we compared the coded date with the date of hospital admission related to the pertinent case.

### Statistical analyses

A. *PPV* and corresponding 95% CI were calculated overall in each database and specifically for each code or free-text search, using medical charts as the gold standard. PPV was calculated as the proportion of the number of confirmed AMI cases out of the total number of randomly selected potential cases. Non-assessable cases were not initially included in either the numerator or the denominator for the PPV calculation, under the assumption that these would not constitute significant bias. However, because the number of non-assessable and non-retrievable cases turned out to be unexpectedly high, we defined, a posteriori, two levels of PPV in order to account for the effects of both non-retrievable and non-assessable cases. We recalculated a 'worst-case scenario' PPV as the proportion of confirmed AMI cases out of the total number of randomly selected potential cases, this time including both non-retrievable and non-assessable cases. We retained as a 'best-case scenario' PPV the proportion of confirmed AMI cases out of the total number of cases that excluded both non-retrievable and non-assessable cases.
B. *Effect of outcome misclassification on AMI risk estimation during drug use.* To investigate the impact of outcome misclassification on estimation of risk of drug-related AMI, we evaluated the association between drug exposure and risk of AMI in the entire population covered by the three databases (ie, not only the randomly selected cases). Drug prescription and/or

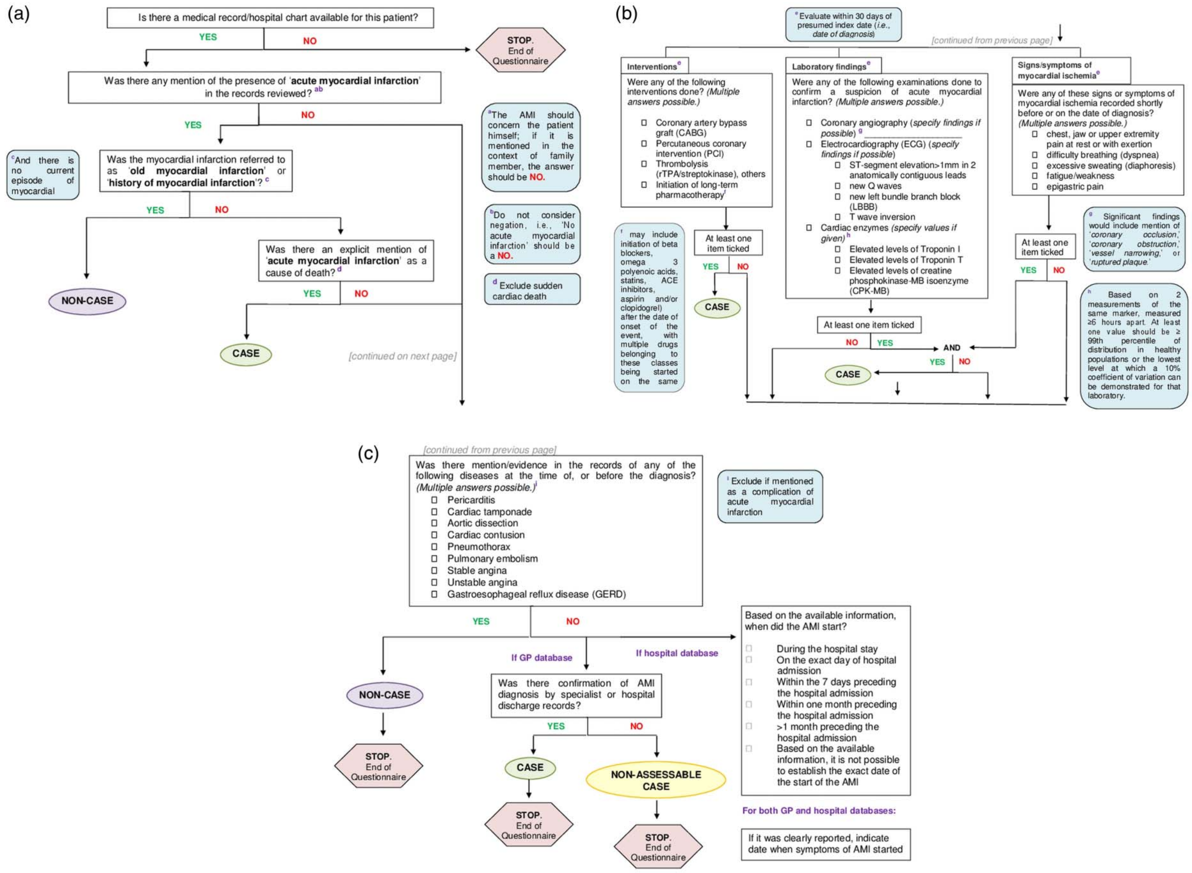

(a)
Is there a medical record/hospital chart available for this patient?
YES
NO
STOP. End of Questionnaire
Was there any mention of the presence of 'acute myocardial infarction' in the records reviewed? ab
YES
NO
Was the myocardial infarction referred to as 'old myocardial infarction' or 'history of myocardial infarction'? c
c And there is no current episode of myocardial
YES
NO
NON-CASE
Was there an explicit mention of 'acute myocardial infarction' as a cause of death? d
YES
NO
CASE
[continued on next page]
a The AMI should concern the patient himself; if it is mentioned in the context of family member, the answer should be NO.
b Do not consider negation, i.e., 'No acute myocardial infarction' should be a NO.
d Exclude sudden cardiac death
(b)
e Evaluate within 30 days of presumed index date (i.e., date of diagnosis)
[continued from previous page]
Interventions e
Were any of the following interventions done? (Multiple answers possible.)
□ Coronary artery bypass graft (CABG)
□ Percutaneous coronary intervention (PCI)
□ Thrombolysis (rTPA/streptokinase), others
□ Initiation of long-term pharmacotherapy f
f may include initiation of beta blockers, omega 3 polyenoic acids, statins, ACE inhibitors, aspirin and/or clopidogrel) after the date of onset of the event, with multiple drugs belonging to these classes being started on the same
At least one item ticked
YES
NO
CASE
Laboratory findings e
Were any of the following examinations done to confirm a suspicion of acute myocardial infarction? (Multiple answers possible.)
□ Coronary angiography (specify findings if possible) g
□ Electrocardiography (ECG) (specify findings if possible)
□ ST-segment elevation>1mm in 2 anatomically contiguous leads
□ new Q waves
□ new left bundle branch block (LBBB)
□ T wave inversion
□ Cardiac enzymes (specify values if given) h
□ Elevated levels of Troponin I
□ Elevated levels of Troponin T
□ Elevated levels of creatine phosphokinase-MB isoenzyme (CPK-MB)
At least one item ticked
NO
YES
AND
YES
NO
CASE
Signs/symptoms of myocardial ischemia e
Were any of these signs or symptoms of myocardial ischemia recorded shortly before or on the date of diagnosis? (Multiple answers possible.)
□ chest, jaw or upper extremity pain at rest or with exertion
□ difficulty breathing (dyspnea)
□ excessive sweating (diaphoresis)
□ fatigue/weakness
□ epigastric pain
At least one item ticked
YES
NO
g Significant findings would include mention of 'coronary occlusion,' 'coronary obstruction,' 'vessel narrowing,' or 'ruptured plaque.'
h Based on 2 measurements of the same marker, measured ≥6 hours apart. At least one value should be ≥ 99th percentile of distribution in healthy populations or the lowest level at which a 10% coefficient of variation can be demonstrated for that laboratory.
(c)
[continued from previous page]
Was there mention/evidence in the records of any of the following diseases at the time of, or before the diagnosis? (Multiple answers possible.) i
□ Pericarditis
□ Cardiac tamponade
□ Aortic dissection
□ Cardiac contusion
□ Pneumothorax
□ Pulmonary embolism
□ Stable angina
□ Unstable angina
□ Gastroesophageal reflux disease (GERD)
i Exclude if mentioned as a complication of acute myocardial infarction
YES
NO
NON-CASE
STOP. End of Questionnaire
If GP database
If hospital database
Was there confirmation of AMI diagnosis by specialist or hospital discharge records?
YES
NO
CASE
STOP. End of Questionnaire
NON-ASSESSABLE CASE
STOP. End of Questionnaire
Based on the available information, when did the AMI start?
□ During the hospital stay
□ On the exact day of hospital admission
□ Within the 7 days preceding the hospital admission
□ Within one month preceding the hospital admission
□ >1 month preceding the hospital admission
□ Based on the available information, it is not possible to establish the exact date of the start of the AMI
For both GP and hospital databases:
If it was clearly reported, indicate date when symptoms of AMI started

**Figure 1** Data entry algorithm implemented based on a standardised questionnaire.

dispensation data were used to estimate the incidence rate of AMI during drug exposure. Drug prescriptions and dispensations are locally coded in each database (see online supplementary appendix 1), but these codes are linked to the Anatomical Therapeutic Chemical Classification (ATC, http://www.whocc.no/atc_ddd_index/) system, which is used as the common drug coding system in the EU-ADR network. Overlapping treatment episodes with the same drug (same ATC code) are combined into a single episode of drug use that starts when the first prescription begins and stops when the last prescription ends. When a patient uses more than one drug at a time, the corresponding person-time is labelled accordingly. Using individual data on the start date and end date of prescription or dispensation, those periods during which an individual is included in the study, but is not using any drug, are marked as unexposed. Events are then assigned to the episodes (ie, drug use/non-use) in which they occurred. The duration covered by each prescription or dispensation is estimated within each database, according to the legend duration (if dosing regimen is available), or is otherwise based on the defined daily dose. We estimated the incidence rate of AMI during the current use of six reference drugs: three drugs well known from the literature to be positively associated with AMI (positive controls: rofecoxib, rosiglitazone and levonorgestrel/oestrogen); and three other drugs, unlikely to be associated with AMI, based on the currently available literature (negative controls: ferrous sulfate, gemfibrozil, amoxicillin/clavulanic acid).[22] We employed the case definitions of AMI taking into account codes and free text with varying values of PPV: (1) 'AMI' included *all* eligible codes and free text to identify patients with AMI; (2) 'AMI50' included codes and free text having PPV ≥50% and (3) 'AMI75' included codes and free text having PPV of ≥75%. We calculated incidence rate ratios (IRRs) for AMI during drug exposure (with non-exposure to the specified drug as reference) for each of the six drugs, pooled across all the databases. A Mantel-Haenszel test was used to assess the differences between the incidence rates, corrected for age and sex.

## RESULTS

The three healthcare databases considered for this analysis comprised data from 4 034 232 individuals with 22 428 883 person-years of follow-up during the period 1996–2009. Within this population, a total of 42 774 potential cases of AMI were identified. From the random sample of 800 potential cases of AMI (200 cases/database plus an additional 200 cases for free text-identified cases in IPCI) selected for validation, the medical records/charts could be retrieved and reviewed for 748 (93.5%) of them. The hospital medical charts of 52 potential cases in Aarhus could not be accessed because no institutional agreement was in place to allow access to the medical charts. The demographic and clinical characteristics of the randomly selected 748 cases are shown in table 1. The mean age was 67 years across all the databases and the patients were predominantly men (62–70% overall). Chest pain at rest or with exertion was the most frequently reported symptom of AMI cases across all the databases (more than 50% of confirmed cases in both IPCI and Aarhus and 11% in HSD). Hypertension, cigarette smoking and dyslipidaemia were the most frequently recorded cardiovascular risk factors.

All 148 potential cases of AMI identified in Aarhus were confirmed by manual chart review. As regards IPCI, 93 (46.5%) potential ICPC-coded cases were confirmed, 31 (15.5%) were judged as non-cases and 76 (38%) cases were judged as non-assessable. From the 200 potential cases identified by free-text search, 26 (13%) cases were confirmed and 68 (34%) were considered non-assessable, while the remaining (106, 53%) were classified as non-cases. For HSD, 115 (57.5%) cases were confirmed and 79 (39.5%) were declared non-assessable. Table 2 shows the 'best-case scenario' PPV and 'worst-case scenario' PPV overall for the codes used to identify AMI in each database. In table 3, the percentage distribution and PPV for the specific diagnosis codes from each coding scheme and free-text search are given. All the ICD-10 codes used in Aarhus had the 100% 'best-case scenario' PPV. The PPVs in the 'worst-case scenario' all decreased, ranging from 66.7 (95% CI 1.3 to 100) to 75.0 (95% CI 58.7 to 91.3). Overall, the ICD9-CM codes had good PPV, with 410.9*, the most frequently reported code, having a 'best-case scenario' PPV of 96.9% (95% CI 93.5% to 100%) and a 'worst-case scenario' PPV of 59.9 (95% CI 50.1 to 69.6). The 'best-case scenario' PPV of free-text search alone in HSD was 60% (95% CI 17.1 to 100). In IPCI, the ICPC code K75 had a 'best-case scenario' PPV of 75% (95% CI 67.4% to 82.6%), while free-text search alone had a PPV of 19.7% (95% CI 12.9% to 26.5%). The 'worst-case scenario' PPVs were correspondingly lower. All validated cases of AMI in IPCI and HSD were supported with confirmation of the diagnosis by a medical specialist (ie, cardiologist).

The relationship between the coded date and the date of onset of symptoms across the three databases is shown in figure 2. There was not enough information for this assessment in 25 cases from Aarhus (16.9%). The lag time between coded date and date of symptom onset (as manually validated) ranged from 1 day before to more than 60 days before the automatically detected event date. The coded date for the majority of cases coincided with the onset of symptoms in all databases: Aarhus=72 cases (48.6%); HSD=110 cases (95.6%) and IPCI=67 cases (56.3%). For the administrative/claims database Aarhus, the characterisation of the coded index date with respect to hospitalisation is as follows: (1) date of hospital admission=100 cases (67.6%); (2) during hospital stay=9 cases (6.1%); (3) ≥7 days preceding hospitalisation 23 (15.5%) and (4) not possible to establish=16 (10.8%).

Figure 3 shows the IRRs, adjusted for age and sex, for six drugs across the different PPV categories. In general, although the number of AMI cases identified using all eligible codes ('AMI') was greater compared with case definitions based on code with ≥50% PPV or ≥75% PPV (ie, 'AMI50' and 'AMI75'), there was only a small change in the resulting IRRs. The clear exception is the positive control drug rosiglitazone, in which the IRR of 2.44 (95% CI 1.62 to 3.67) with the 'AMI definition' decreased to 1.62 (95% CI 0.77 to 3.41) with 'AMI75,' the risk then becoming insignificant; the IRR remained fairly stable at 1.64 (95% CI 0.78 to 3.45) with the 'AMI50' definition. The same trend was observed for rofecoxib and levonorgestrel/oestrogen: although the IRR changes corresponding to each definition were smaller compared with rosiglitazone, the risk disappeared with both the 'AMI75' and 'AMI50' definitions. For the negative controls (where the 95% CIs all included 1), the impact of using codes with different PPVs was less pronounced.

## DISCUSSION

We examined PPV of primary hospital discharge diagnosis codes and general practitioner-recorded diagnoses for AMI in three European EHR databases. The overall 'best-case scenario' PPV for the coding scheme-based diagnoses was good, ranging from 75% (IPCI, ICPC coding) to 95% (HSD, ICD9-CM) to 100% (Aarhus, ICD-10). The use of free-text search was more extensive in IPCI compared with HSD, largely due to the lesser granularity of the ICPC coding system. The use of free text alone had a lower PPV, ranging from a 'best-case scenario' PPV of 20% in IPCI to 60% in HSD. Although 52 of the initially identified cases of AMI in Aarhus were missing and could not be validated, the inaccessibility of the corresponding medical charts was random and thus was deemed unlikely to introduce bias. However, to account for any potential bias introduced by these non-retrievable cases (as well as non-assessable cases), 'worst-case scenario' PPVs were calculated. The impact on the corresponding PPVs was high: for ICD-10 codes overall, PPV dropped to 74% while for ICD9-CM and ICPC codes PPV decreased to 60% and 46%, respectively. These findings reiterate the need for adequate case retrieval in outcome validation studies and, if necessary, to perform resampling and take into account the impact of missing cases in the analysis. More importantly, misclassification of AMI cases resulting from use of disease codes (or free text) with low PPV has analogous implications on the estimation of incidence rates. Studies using

**Table 1** Characteristics of patients in the random sample of potential AMI cases

| | IPCI | | HSD | | Aarhus | |
|---|---|---|---|---|---|---|
| | Total N: 400 (%) | Confirmed cases N: 93 (%) | Total N: 200 (%) | Confirmed cases N: 115 (%) | Total N: 148 (%) | Confirmed cases N: 148 (%) |
| Male sex (%) | 246 (61.5) | 87 (93.5) | 132 (66.0) | 80 (69.6) | 103 (69.6) | 82 (55.4) |
| Mean age (years) | 66 | 65 | 68 | 67 | 67 | 67 |
| Cardiovascular risk factors* | | | | | | |
| Family history of coronary heart disease | 86 (21.5) | 14 (15.1) | 11 (5.5) | 8 (7.0) | 19 (12.8) | 19 (12.8) |
| Dyslipidaemia | 73 (18.2) | 19 (20.4) | 98 (49.0) | 58 (50.4) | 19 (12.8) | 19 (12.8) |
| Diabetes mellitus | 62 (15.5) | 13 (14.0) | 66 (33.0) | 42 (36.5) | 17 (11.4) | 17 (11.4) |
| Hypertension | 126 (31.5) | 36 (38.7) | 113 (56.5) | 76 (66.0) | 44 (29.7) | 44 (29.7) |
| Obesity | 79 (19.8) | 23 (24.7) | 21 (10.5) | 16 (13.9) | 7 (4.7) | 7 (4.7) |
| Cigarette smoking | 111 (27.8) | 36 (38.7) | 39 (19.5) | 22 (19.1) | 50 (33.8) | 50 (33.8) |
| Clinical manifestations* | | | | | | |
| Chest, jaw or upper extremity pain at rest or with exertion | 102 (25.5) | 52 (55.9) | 23 (11.5) | 13 (11.3) | 118 (79.7) | 118 (79.7) |
| Difficulty breathing (dyspnoea) | 38 (9.5) | 15 (20.4) | 10 (5.0) | 6 (5.2) | 22 (14.8) | 22 (14.8) |
| Excessive sweating (diaphoresis) | 30 (7.5) | 19 (20.4) | 2 (1.0) | 0 (0) | 5 (3.4) | 5 (3.4) |
| Fatigue/weakness | 7 (1.8) | 0 (0) | 2 (1.0) | 1 (0.9) | 5 (3.4) | 5 (3.4) |
| Diagnostic workup performed* | | | | | | |
| Coronary angiography | 48 (12.0) | 39 (41.9) | 39 (19.5) | 38 (33.0) | 117 (79.1) | 117 (79.1) |
| ECG | | | | | | |
| ST-segment elevation | 27 (6.7) | 25 (30.1) | 7 (3.5) | 5 (4.4) | 59 (39.9) | 59 (39.9) |
| New Q waves | 7 (1.8) | 4 (4.3) | 1 (0.50) | 1 (0.9) | 9 (6.1) | 9 (6.1) |
| New left bundle branch block (LBBB) | 3 (0.8) | 0 (0) | 0 (0) | 0 (0) | 2 (1.4) | 2 (1.4) |
| T wave inversion | 17 (4.2) | 9 (9.7) | 8 (4.0) | 6 (5.2) | 19 (12.8) | 19 (12.8) |
| Other (ST segment depression, etc) | 19 (7.2) | 19 (20.4) | 7 (3.5) | 7 (6.1) | 9 (6.1) | 9 (6.1) |
| Cardiac enzymes | | | | | | |
| Elevated cardiac troponin I | 5 (1.2) | 41 (44.1) | 2 (1.0) | 2 (1.7) | 0 (0) | 0 (0) |
| Elevated cardiac troponin T | 6 (1.5) | 6 (9.7) | 0 (0) | 0 (0) | 123 (83.1) | 123 (83.1) |
| Elevated creatine phosphokinase (MB isoenzyme) | 23 (5.8) | 23 (24.7) | 5 (2.5) | 5 (4.4) | 86 (58.1) | 86 (58.1) |
| Other | 8 (2.0) | 5 (5.4) | 3 (1.5) | 3 (2.6) | 5 (3.4) | 5 (3.4) |
| Interventions performed* | | | | | | |
| Coronary artery bypass graft (CABG) | 15 (3.8) | 10 (10.8) | 17 (8.5) | 17 (14.8) | 7 (4.7) | 7 (4.7) |
| Percutaneous coronary intervention (PCI) | 78 (19.5) | 67 (72.0) | 41 (20.5) | 41 (35.6) | 93 (62.8) | 93 (62.8) |
| Thrombolysis (rTPA/streptokinase, others) | 6 (1.5) | 5 (5.4) | 4 (2.0) | 4 (3.5) | 9 (6.1) | 9 (6.1) |
| Initiation of long-term pharmacotherapy | 116 (29.0) | 93 (100) | 121 (60.5) | 90 (77.6) | 81 (56.8) | 81 (56.8) |
| Deaths with AMI identified as cause | 7 (1.8) | 7 (7.5) | 2 (1.0) | 2 (1.7) | 2 (1.4) | 2 (1.4) |
| Diagnosis confirmed by medical specialist† | 113 (28.2) | 93 (100) | 1 (0.5) | 1 (0.9) | NA | NA |

*Can add up to more than 100%.
†Only applicable for GP databases (HSD and IPCI).
AMI, acute myocardial infarction; HSD, Health Search/CSD Patient DB; IPCI, Integrated Primary Care Information.

EHR data to derive incidence rates of clinical events should thus correct for this potential misclassification.

Routinely collected EHR data are increasingly being used in many areas of biomedical research and a recently identified promising area for EHRs is the proactive surveillance of potentially drug-induced outcomes. The validity of such surveillance activities depends, however, on the accuracy of the definitions of the outcomes being investigated, while at the same time preserving data confidentiality. The use of a standardised questionnaire implemented in an automated data entry validation algorithm facilitated harmonised data collection and analysis across different databases without compromising data protection. The procedure also enabled us to document recorded database information on cardiovascular risk factors such as a family history of coronary artery disease, hypertension, diabetes, dyslipidaemia, smoking and obesity. Such information may be useful in evaluating potential confounder effects when conducting epidemiological studies. Other information related to diagnostic procedures or interventions requiring hospitalisation (eg, coronary angiography) may not be consistently recorded in GP databases, unless provided with the discharge letters or referrals from specialists, hence the observed higher proportion of such information from reimbursement claims data (Aarhus). In the same way, the documentation of initiation of long-term pharmacotherapy for the management of AMI may not be as well documented in claims data as in GP data. It is important to note that information derived from GP databases are data recorded in the course of routine clinical care and provide a different perspective from those derived from databases documenting reimbursement claims for utilisation of healthcare services, which are more for auditing purposes.[23 24]

Since the context within which clinical events are recorded differs between GP databases and administrative/claims databases, there is often also an expected delay between onset of first symptoms (which are more likely to be documented in GP records) and diagnoses recorded upon hospital discharge (documented in reimbursement claims, and also in GP data if referral letters from a cardiologist are available). Our evaluation of the automatically detected index date shows that most of the time the coded event date coincided with the date of onset of first symptoms (and with the date of hospital admission for the administrative database), although there can be a wide range between these two dates.

For this validation study, we have chosen PPV as the relevant measure of accuracy for the codes used in identifying AMI from EHR. Such a metric enables the use of GP and claims records to determine the probability that an individual has an AMI, based on such data. PPV measurements are correlated with disease prevalence, however, and are strongly dependent on specificity. Specificity and other measures of validity, such as sensitivity (ie, how many cases of AMI are missed) and negative predictive value, cannot be calculated from our

**Table 2** Overall positive predictive value (PPV) for acute myocardial infarction (AMI) identification, according to database

| Source | Coding system | Number of cases sampled | Number of cases retrieved | Number of cases confirmed (%) | Number of cases considered non-assessable | PPV, best case scenario* (95% CI) | PPV, worst case scenario† (95% CI) |
|---|---|---|---|---|---|---|---|
| General practitioner (GP)/specialist diagnoses (IPCI, the Netherlands) | ICPC | 200 | 200 | 93 (46.5) | 76 | 75.0 (67.4 to 82.6) | 46.5 (37.7 to 55.3) |
|  | Free text | 200 | 200 | 26 (13.0) | 68 | 19.7 (12.9 to 26.5) | 13.0 (7.3 to 18.7) |
| GP/specialist diagnoses (HSD, Italy) | ICD9-CM | 187 | 187 | 112 (56.0) | 71 | 96.6 (93.2 to 99.9) | 59.9 (51.0 to 68.8) |
|  | free text | 13 | 13 | 3 (1.5) | 8 | 60 (17.1 to 100) | 23.1 (0 to 60.0) |
| Primary hospital discharge diagnoses (Aarhus, Denmark) | ICD-10 | 200 | 148 | 148 | 0 | 100 (100 to 100) | 74.0 (66.9 to 81.1) |

*Best-case scenario*: non-assessable and non-retrievable cases are not included in the numerator or denominator when the PPV is calculated.
†*Worst-case scenario*: both non-assessable and non-retrievable cases are included in the denominator when the PPV is calculated. For Aarhus, the number of cases that would have been retrieved per code was estimated based on the percentage of distribution of codes in the retrieved cases within the random sample.
ICPC, International Classification of Primary Care; IPCI, Integrated Primary Care Information.

**Table 3** Number and distribution of confirmed AMI cases by diagnostic code or free text

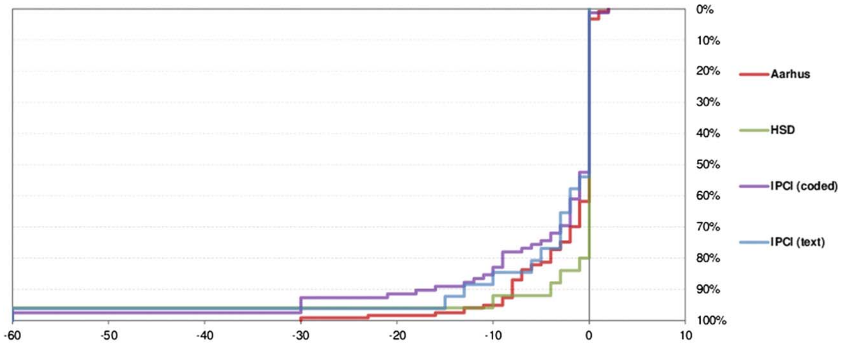
| Database/code | Code description | Number of records reviewed | Number of cases confirmed | Percentage of cases identified by such code or free text in random sample | PPV, best-case scenario† (*95% CI*) | PPV, worst case scenario‡ (*95% CI*) |
|---|---|---|---|---|---|---|
| *IPCI* | | | | | | |
| ICPC K75- | AMI | 200 | 93 | 100 | 75.0 (*67.4* to *82.6*) | 46.5 (*37.7* to *55.3*) |
| Free text | Specific key words* | 200 | 26 | 100 | 19.7 (12.9 to 26.5) | 13.0 (*7.3* to *18.7*) |
| *HSD* | | | | | | |
| ICD9-CM | | | | | | |
| 410 or 410.0 | AMI of anterolateral wall | 12 | 6 | 6.0 | 85.7 (*59.8* to *100*) | 50.0 (*13.0* to *87.0*) |
| 410.1 or 410.10 | AMI of other anterior wall | 4 | 4 | 2.0 | 100 | 100 |
| 410.20 | AMI of inferolateral wall | 1 | 1 | 0.5 | 100 | 100 |
| 410.3 | AMI of inferoposterior wall | 1 | 1 | 0.5 | 100 | 100 |
| 410.7 | Subendocardial infarction | 3 | 2 | 1.5 | 100 | 66.7 (*1.3* to *100*) |
| 410.9 or 410.90 | AMI, unspecified site | 157 | 94 | 78.5 | 96.9 (*93.5* to *100*) | 59.9 (*50.1* to *69.6*) |
| 410.9+Free text | AMI, unspecified site | 8 | 4 | 4.0 | 100 | 50.0 (*1.0* to *99.0*) |
| 411.81+Free text | Acute coronary occlusion without MI | 1 | 0 | 0.5 | 0 | 0 |
| Free text | Specific key words* | 13 | 3 | 6.5 | 60 (*17.1* to *100*) | 23.1 (*0* to *60.0*) |
| *Aarhus* | | | | | | |
| ICD-10 | | | | | | |
| I21.0 | Acute transmural MI of anterior wall | 20 | 20 | 13.5 | 100 | 74.1 (*54.9* to *93.3*) |
| I21.1 | Acute transmural MI of inferior wall | 17 | 17 | 11.5 | 100 | 73.9 (*53.0* to *94.8*) |
| I21.2 | Acute transmural MI of other sites | 2 | 2 | 1.4 | 100 | 66.7 (*1.3* to *100*) |
| I21.3 | Acute transmural MI of unspecified site | 26 | 26 | 17.6 | 100 | 74.3 (*57.5* to *91.1*) |
| I21.4 | Acute subendocardial MI | 56 | 56 | 37.8 | 100 | 73.7 (*62.2* to *85.2*) |
| I21.9 | AMI, unspecified | 27 | 27 | 18.2 | 100 | 75.0 (*58.7* to *91.3*) |

*See online supplementary appendix 2 for key words used.

†*Best-case scenario*: non-assessable and non-retrievable cases are not included in the numerator or denominator when the PPV is calculated.

‡ *Worst-case scenario*: both non-assessable and non-retrievable cases are included in the denominator when the PPV is calculated. For Aarhus, the number of cases that would have been retrieved per code was estimated based on the % distribution of codes in the retrieved cases within the random sample.

AMI, acute myocardial infarction; HSD, Health Search/CSD Patient DB; ICD9-CM, International Classification of Diseases 9th revision-clinical modification; ICPC, International Classification of Primary Care; IPCI, Integrated Primary Care Information.



**Note**: Y axis represents cumulative % of confirmed cases; X axis represents number of days

**Figure 2** Differences in automatically recorded date of acute myocardial infarction (AMI; time 0) and manually validated date of onset of AMI symptoms across the databases.

data, because the data extraction was based on searching for the codes/free text pertinent to the diagnosis of interest. Another limitation is that, in the estimation of drug-related IRR of AMI, we only adjusted for age and sex and did not consider other potential confounding factors.

The results we obtained in this study are consistent with the PPV estimates for ICD-10 and for ICD9-CM cited in the literature. The ICD-10 codes I21, I22 and I23 were found to have 98% PPV in a Danish study evaluating the accuracy of ICD-10-coded myocardial infarction as a component of the Charlson comorbidity index.[15] Previous studies evaluating earlier versions of ICD have also demonstrated the accurate coding practices in Danish administrative registries, including the Danish MONICA (Monitoring Trends and Determinants in Cardiovascular Disease) study where 93.5% of the patients in the Danish National Patient Register were found to have definite or possible AMI.[25] The PPV of the ICD9-CM code 410 to identify cases of AMI among records with a prior primary hospital discharge code in the Saskatchewan Hospital Automated Database was 97%.[12] In another study using Medicare claims, the PPV of several ICD9-CM codes (410.01, 410.11, 410.21, 410.31, 410.41, 410.51, 410.61, 410.71, 410.81 or 410.91) for identifying AMI in either primary or secondary hospital discharge diagnoses was 94.1%.[11] While ICPC codes are often used to estimate the incidence or prevalence of various clinical outcomes,[26–28] we are not aware of any published studies that have assessed the accuracy of ICPC codes in the identification of AMI in EHR data. A study in the Netherlands evaluated ICPC-coded diagnoses in GP records in the context of cardiovascular risk factor assessment after pre-eclampsia, but only the validity of an ICPC-coded pre-eclampsia diagnosis was determined.[29]

The available knowledge regarding the value of free-text mining in identifying outcomes from EHR data is an area of research that is gaining a lot of interest.[30 31] Our findings show that there is potential for the use of free-text search in identification of AMI from EHR databases, but that appropriate combination of key words and natural language processing techniques needs to be further evaluated and optimised.

Our investigation of the impact of outcome misclassification on estimation of AMI risk with drug use showed that the use of codes with lower PPV generally resulted in small changes in the estimated relative risks, but the use of codes with higher PPV may lead to attenuation or disappearance of risk for positive associations (non-differential misclassification biases the risk estimates towards the null). It is important to note that the change in the estimated risk of AMI during drug use when using more specific criteria is virtually due to the exclusion of AMI cases identified by free-text search: with AMI50, cases identified by free text in IPCI were excluded while with AMI75 all cases identified by free text in IPCI and HSD were excluded. The impact

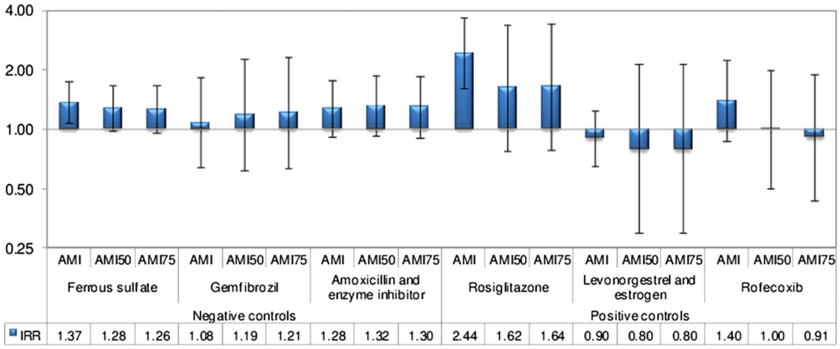


**Figure 3** Impact of codes and free text with different 'best-case scenario' positive predictive values on age-adjusted and sex-adjusted incidence rate ratio estimates for acute myocardial infarction during drug exposure (non-exposure to the same drug as reference).

analyses were performed on aggregated data, but should ideally be stratified according to database. This is because the impact on IRR is not only a function of using specific versus non-specific codes, but also a function of the database characteristics. Future studies should thus take into account the data source as well as test more drug-event associations, control for other confounders and increase sample size, especially since these estimates were based on relatively small numbers (as reflected in the wide CI). Although we considered only the 'best-case scenario' PPVs in the analyses for outcome misclassification, these findings suggest that similar implications would be expected with the 'worst-case scenario' PPVs.

## CONCLUSIONS

We have shown that a network of EHR databases from different countries with different disease coding systems can accurately identify patients with AMI and that adequate case retrieval remains an essential step in validation. The results obtained in this study are consistent with the PPV estimates for ICD9-CM and ICD-10 cited in the literature. Strategies are necessary to optimise the use of ICPC, in combination with free-text search, in the identification of AMI from EHR data. Use of more specific disease codes for identifying AMI during drug use may lead to a small but significant change in risk estimates and at the expense of decreased precision.

**Author affiliations**
[1]Department of Medical Informatics, Erasmus MC University Medical Center, Rotterdam, The Netherlands
[2]Department of Gastroenterology and Hepatology, Erasmus MC University Medical Center, Rotterdam, The Netherlands
[3]Department of Research, Health Search, Italian College of General Practitioners, Florence, Italy
[4]Department of Clinical Epidemiology, Aarhus University Hospital, Aarhus, Denmark
[5]Primary Care and Population Sciences, Kings College, London, UK
[6]Department of Clinical and Experimental Medicine and Pharmacology, University of Messina, Messina, Italy
[7]Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, The Netherlands

**Contributors** PMC, VEV, GT, MMol, MJS, MS and JvdL contributed to the conception and design of the study. MJS developed the Jerboa software for data extraction. MMos developed the Chameleon software for validation. PMC, VEV, MMol, PM and GT developed and tested the validation algorithms. GM, PM, MSN, LP and MS provided the data and performed local database analyses. PMC and VEV performed the aggregated analyses and PMC drafted the manuscript. All authors contributed to critically revising the manuscript for important intellectual content and approved the final version for submission.

**Funding** This research has been funded by the European Commission's Seventh Framework Programme (FP7/2007–2013) under grant no. 215847, The EU-ADR Project. The funding agency had no role in the design and conduct of the study, the collection and management of data, the analysis or interpretation of the data, and preparation, review or approval of the manuscript.

**Competing interests** None.

**Ethics approval** The respective scientific and ethics committees of each database approved the use of the data for this study.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** No additional data are available.

## REFERENCES

1. Go AS, Magid DJ, Wells B, *et al*. The Cardiovascular Research Network: a new paradigm for cardiovascular quality and outcomes research. *Circ Cardiovasc Qual Outcomes* 2008;1:138–47.
2. Stang PE, Ryan PB, Racoosin JA, *et al*. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med* 2010;153:600–6.
3. Coloma PM, Schuemie MJ, Trifiro G, *et al*. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Saf* 2011;20:1–11.
4. Thygesen K, Alpert JS, White HD, *et al*. Universal definition of myocardial infarction. *Circulation* 2007;116:2634–53.
5. Alpert JS, Thygesen K, White HD, *et al*. Implications of the universal definition of myocardial infarction. *Nat Clin Pract Cardiovasc Med* 2008;5:678–9.
6. Malasky BR, Alpert JS. Diagnosis of myocardial injury by biochemical markers: problems and promises. *Cardiol Rev* 2002;10:306–17.
7. Luepker RV, Apple FS, Christenson RH, *et al*. Case definitions for acute coronary heart disease in epidemiology and clinical research studies: a statement from the AHA Council on Epidemiology and Prevention; AHA Statistics Committee; World Heart Federation Council on Epidemiology and Prevention; the European Society of Cardiology Working Group on Epidemiology and Prevention; Centers for Disease Control and Prevention; and the National Heart, Lung, and Blood Institute. *Circulation* 2003;108:2543–9.
8. Graven T, Kruger O, Bronstad G. Epidemiological consequences of introducing new biochemical markers for detection of acute myocardial infarction. *Scand Cardiovasc J* 2001;35:233–7.
9. Kavsak PA, MacRae AR, Lustig V, *et al*. The impact of the ESC/ACC redefinition of myocardial infarction and new sensitive troponin assays on the frequency of acute myocardial infarction. *Am Heart J* 2006;152:118–25.
10. Sanfilippo FM, Hobbs MS, Knuiman MW, *et al*. Impact of new biomarkers of myocardial damage on trends in myocardial infarction hospital admission rates from population-based administrative data. *Am J Epidemiol* 2008;168:225–33.
11. Kiyota Y, Schneeweiss S, Glynn RJ, *et al*. Accuracy of Medicare claims-based diagnosis of acute myocardial infarction: estimating positive predictive value on the basis of review of hospital records. *Am Heart J* 2004;148:99–104.
12. Varas-Lorenzo C, Castellsague J, Stang MR, *et al*. Positive predictive value of ICD-9 codes 410 and 411 in the identification of cases of acute coronary syndromes in the Saskatchewan Hospital automated database. *Pharmacoepidemiol Drug Saf* 2008;17:842–52.
13. Pladevall M, Goff DC, Nichaman MZ, *et al*. An assessment of the validity of ICD Code 410 to identify hospital admissions for myocardial infarction: the Corpus Christi Heart Project. *Int J Epidemiol* 1996;25:948–52.
14. Petersen LA, Wright S, Normand SL, *et al*. Positive predictive value of the diagnosis of acute myocardial infarction in an administrative database. *J Gen Intern Med* 1999;14:555–8.
15. Thygesen SK, Christiansen CF, Christensen S, *et al*. The predictive value of ICD-10 diagnostic coding used to assess Charlson comorbidity index conditions in the population-based Danish National Registry of Patients. *BMC Med Res Methodol* 2011;11:83.
16. Risselada R, De Vries LM, Dippel DW, *et al*. Incidence, treatment, and case-fatality of non-traumatic subarachnoid haemorrhage in the Netherlands. *Clin Neurol Neurosurg* 2011;113:483–7.
17. Trifiro G, Morabito P, Cavagna L, *et al*. Epidemiology of gout and hyperuricaemia in Italy during the years 2005–2009: a nationwide population-based study. *Ann Rheum Dis* 2013;72:694–700.
18. Ehrenstein V, Antonsen S, Pedersen L. Existing data sources for clinical epidemiology: Aarhus University Prescription Database. *Clin Epidemiol* 2010;2:273–9.
19. Coloma PM, Trifiro G, Schuemie MJ, *et al*. Electronic healthcare databases for active drug safety surveillance: is there enough leverage? *Pharmacoepidemiol Drug Saf* 2012;21:611–21.
20. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. *Methods Inf Med* 1993;32:281–91.
21. Avillach P, Coloma PM, Gini R, *et al*. Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. *J Am Med Inform Assoc* 2013;20:184–92.

22. Coloma PM, Avillach P, Salvo F, *et al*. A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. *Drug Saf* 2013;36:13–23.
23. Suissa S, Garbe E. Primer: administrative health databases in observational studies of drug effects—advantages and disadvantages. *Nat Clin Pract Rheumatol* 2007;3:725–32.
24. Schneeweiss S. Understanding secondary databases: a commentary on "Sources of bias for health state characteristics in secondary databases". *J Clin Epidemiol* 2007;60:648–50.
25. Madsen M, Davidsen M, Rasmussen S, *et al*. The validity of the diagnosis of acute myocardial infarction in routine statistics: a comparison of mortality and hospital discharge data with the Danish MONICA registry. *J Clin Epidemiol* 2003;56:124–30.
26. Hak E, Rovers MM, Kuyvenhoven MM, *et al*. Incidence of GP-diagnosed respiratory tract infections according to age, gender and high-risk co-morbidity: the Second Dutch National Survey of General Practice. *Fam Pract* 2006;23:291–4.
27. Britt H, Meza RA, Del Mar C. Methodology of morbidity and treatment data collection in general practice in Australia: a comparison of two methods. *Fam Pract* 1996;13:462–7.
28. Esteban-Vasallo MD, Dominguez-Berjon MF, Astray-Mochales J, *et al*. Epidemiological usefulness of population-based electronic clinical records in primary care: estimation of the prevalence of chronic diseases. *Fam Pract* 2009;26:445–54.
29. Nijdam ME, Timmerman MR, Franx A, *et al*. Cardiovascular risk factor assessment after pre-eclampsia in primary care. *BMC Fam Pract* 2009;10:77.
30. Warrer PHE, Juhl-Jensen L, Aagaard L. Using text-mining techniques in electronic patient records to identify ADRs from medicine use. *Br J Clin Pharmacol* 2012;73:674–84.
31. Schuemie MJ, Sen E, t Jong GW, *et al*. Automating classification of free-text electronic health records for epidemiological studies. *Pharmacoepidemiol Drug Saf* 2012;21:651–8.