Downloaded from jamia.bmj.com on March 30, 2013 - Published by group.bmj.com

Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project

Paul Avillach,[1,2] Preciosa M Coloma,[3] Rosa Gini,[4] Martijn Schuemie,[3] Fleur Mougin,[1] Jean-Charles Dufour,[2] Giampiero Mazzaglia,[5] Carlo Giaquinto,[6] Carla Fornari,[7] Ron Herings,[8] Mariam Molokhia,[9] Lars Pedersen,[10] Annie Fourrier-Réglat,[11] Marius Fieschi,[2] Miriam Sturkenboom,[3,12] Johan van der Lei,[3] Antoine Pariente,[1,2] Gianluca Trifirò,[3,13] on behalf of the EU-ADR consortium

► Additional data are published online only. To view these files please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2012-000933).

[1]LESIM, ISPED, University Bordeaux Segalen, Bordeaux, France
[2]LERTIM, EA 3283, Faculté de Médecine, Université Aix Marseille 2, Marseille, France
[3]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
[4]Regional Health Agency of ARS, Florence, Italy
[5]Health Search—Italian College of General Practitioners, Florence, Italy
[6]Pedianet—Societa' Servizi Telematici SRL, Padova, Italy
[7]Centre on Public Health, University of Milano-Bicocca, Milano, Italy
[8]PHARMO Coöperation UA, Utrecht, The Netherlands
[9]Department of Primary Care and Public Health Sciences, Kings College, London, UK
[10]Aarhus University Hospital, Århus Sygehus, Denmark
[11]INSERM U 657, University Bordeaux Segalen, Bordeaux, France
[12]Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands
[13]Department of Clinical and Experimental Medicine and Pharmacology, Section of Pharmacology, University of Messina, Messina, Italy

Correspondence to
Dr P Avillach, LESIM ISPED, University Bordeaux, 146 rue Leo Saignat 33076, Bordeaux, France; avillach@mac.com

PA and PMC contributed equally to this work.

Received 4 March 2012
Accepted 10 August 2012
Published Online First
6 September 2012

ABSTRACT

Objective Data from electronic healthcare records (EHR) can be used to monitor drug safety, but in order to compare and pool data from different EHR databases, the extraction of potential adverse events must be harmonized. In this paper, we describe the procedure used for harmonizing the extraction from eight European EHR databases of five events of interest deemed to be important in pharmacovigilance: acute myocardial infarction (AMI); acute renal failure (ARF); anaphylactic shock (AS); bullous eruption (BE); and rhabdomyolysis (RHABD).

Design The participating databases comprise general practitioners' medical records and claims for hospitalization and other healthcare services. Clinical information is collected using four different disease terminologies and free text in two different languages. The Unified Medical Language System was used to identify concepts and corresponding codes in each terminology. A common database model was used to share and pool data and verify the semantic basis of the event extraction queries. Feedback from the database holders was obtained at various stages to refine the extraction queries.

Measurements Standardized and age specific incidence rates (IRs) were calculated to facilitate benchmarking and harmonization of event data extraction across the databases. This was an iterative process.

Results The study population comprised overall 19 647 445 individuals with a follow-up of 59 929 690 person-years (PYs). Age adjusted IRs for the five events of interest across the databases were as follows: (1) AMI: 60–148/100 000 PYs; (2) ARF: 3–49/100 000 PYs; (3) AS: 2–12/100 000 PYs; (4) BE: 2–17/100 000 PYs; and (5) RHABD: 0.1–8/100 000 PYs.

Conclusions The iterative harmonization process enabled a more homogeneous identification of events across differently structured databases using different coding based algorithms. This workflow can facilitate transparent and reproducible event extractions and understanding of differences between databases.

INTRODUCTION

Spontaneous reporting of adverse drug reactions (ADRs) is currently the main source of data to monitor drug safety after licensing. It relies on healthcare professionals' ability and willingness to identify and report any suspected ADR to a centralized (nationwide or international) pharmacovigilance system.[1] However, the underreporting of ADRs remains an important limitation; it is estimated that only 1–10% of ADRs are reported through this channel[2 3] and it is likely to be subject to recording and ascertainment biases. It is increasingly being recognized that drug safety surveillance can benefit from the wide availability of healthcare databases to complement spontaneous reporting systems and overcome some of their shortcomings.[4–6]

In 2007, the USA Congress directed the Food and Drug Administration (FDA) to create a new post-marketing surveillance system, the Sentinel System.[7] The desired goal is to use, by 2012, electronic health data from 100 million subjects for the prospective and systematic safety monitoring of marketed medical products in real life settings.[8] The Observational Medical Outcomes Partnership (OMOP) initiative was created, also in the USA, to design a common framework with the aim of setting up a system for drug surveillance through data mining of electronic health records.[9] In Europe, several projects, such as the Pharmacoepidemiological Research on Outcomes of Therapeutics (PROTECT), have been recently funded to link healthcare databases throughout Europe under the umbrella organization of the Innovative Medicines Initiative (IMI).[10] Other European Union (EU) funded projects in which multiple healthcare databases are combined together to address specific safety issues include: non-steroidal anti-inflammatory drug related gastrointestinal and cardiovascular risks (SOS, http://www.sos-nsaids-project.org/), the arrhythmogenic risk of drugs (ARITMO—http://www.aritmo-project.org) and the safety of vaccines (VAESCO—http://www.vaesco.net).

In 2008, the EU funded project 'Exploring and understanding adverse drug reactions by integrative mining of clinical records and biomedical knowledge' (EU-ADR) was launched. The aim of this project was to design, develop and validate a computerized system to process data from eight electronic healthcare record (EHR) databases and several biomedical databases for drug safety signal detection.[11 12] Within this project, an event based

Downloaded from jamia.bmj.com on March 30, 2013 - Published by group.bmj.com

approach was adopted where a focused set of events of special interest in pharmacovigilance are evaluated for their association with all drugs captured in the EHR databases. Each of the eight databases in EU-ADR has unique characteristics depending on its primary objective and local function (ie, administrative claims or medical records) and contains medical information coded according to different languages and disease terminologies. For these reasons, queries for data extraction concerning potential adverse events have to be created based on local expertise. Due to structural, syntactic, and semantic heterogeneities of the databases participating in the EU-ADR project, it was not possible to construct a single query for data extraction that could be used as such in all databases. In this context of large scale drug safety monitoring using EHRs, the event data extraction from different databases requires a harmonization—that is, a process geared towards reaching a common definition and identification of events, which is both clinically sound and agreeable to all stakeholders. In this paper, we describe the harmonization process for the data extraction concerning five events deemed to be important in pharmacovigilance from eight different databases of the EU-ADR network.

METHODS

Data sources

The eight databases involved in the EU-ADR project contain information from the healthcare records of almost 20 million European citizens (table 1). Health Search/CSD Patient (HSD, Italy), Integrated Primary Care Information (IPCI, The Netherlands), Pedianet (Italy), and QRESEARCH (UK) are general practice (GP) databases where both clinical information and drug prescriptions are recorded. The Aarhus University Hospital Database (Aarhus, Denmark), PHARMO Network (The Netherlands), and the regional Italian databases of Lombardy and Tuscany are all comprehensive record linkage systems in which drug dispensing data of a well defined population are linked to a registry of hospital discharge diagnoses and various other registries. The databases are heterogeneous in both structure and content.[12 13] The respective scientific and ethics committees of each database approved the use of the data for this study.

For this analysis (which was done at the beginning of the project), databases contributed data from the period 1996–2007. Four disease terminologies are used to code the clinical events in the eight databases: the International Statistical Classification of Diseases and related health problems—9th and 10th revisions (ICD9-CM[14] and ICD10[15]); the International Classification of Primary Care (ICPC)[16]; and the READ CODE (RCD) classification.[17] Two GP databases also describe events in clinical notes using free text in either Dutch (IPCI) or Italian (HSD).

Linking disparate databases using a distributed network

A distributed database network approach was chosen in EU-ADR,[12] allowing database holders to maintain local control of their data, while reaching the goal of sharing data in a standardized manner. Without this control, database holders may be reluctant to participate in a large network, primarily because of concerns regarding privacy and data confidentiality.[18] The decentralized data storage avoids, or reduces, many of the security, proprietary, legal and privacy concerns of data owners at the institution and country levels. Moreover, this approach allows local database experts to keep the data within their protected environment and may easily and more effectively troubleshoot unexpected findings or data irregularities and inconsistencies.[19] Local experts are naturally in the best position to understand the context within which the data are recorded.

To deal with database heterogeneity, we defined a common database model (figure 1) that utilizes input files containing information on patient demographics and follow-up time, events, and drug exposures. The common input files required the combination of information from within each database, namely: (1) patient registration into the database system; (2) accounts of general practitioner (GP) visits, including diagnoses and referrals to specialists, in the case of primary care databases; (3) records of hospitalization and utilization of other healthcare services, in the case of administrative claims databases; (4) laboratory examinations; and (5) death and cause of death, when available. These common data files are created locally and subsequently managed by purpose built Java based software called Jerboa. The software queries patient level data in the different databases to create aggregated, de-identified statistics which are then sent in an encrypted format to a central repository for evaluation and further analyses.[12] The structure and content of the databases in EU-ADR is given in table 2.

A similar data model has likewise been developed by OMOP and the pilot FDA Sentinel System, Mini-Sentinel.[9 20] The main difference with the other data models is that in EU-ADR, the entries (tables) in the data model are homogeneous according to the setting within which the information was recorded (eg, during hospitalization, during a GP visit, at death, etc), while in OMOP the entries are conceived to be homogeneous in content, not in source (ie, one entry collects diagnoses and another collects procedures, while the encounter that gave rise to the information is recorded as an attribute).

Harmonization process for the event data extraction

The stepwise process adopted for the definition and harmonization of the queries for event data extraction is outlined in detail in figure 2. Within the EU-ADR project, 23 events of interest were identified as 'priority' events according to ranking criteria based on their relevance from a pharmacovigilance perspective.[16] For this paper, we describe the harmonization process for data extraction pertaining to the top five ranked events: (1) acute myocardial infarction (AMI); (2) acute renal failure (ARF); (3) anaphylactic shock (AS); (4) bullous eruptions (BE); and (5) rhabdomyolysis (RHABD). For each event, a clinical definition was first provided in a structured event definition form (EDF) using medical textbooks and published guidelines of diagnostic criteria from scientific societies concerning the events of interest. This definition was subsequently validated by medical specialists. The EDF for the event BE is shown as an example in appendix 1 (available online only).

Mapping of terminologies

To reconcile differences across terminologies, we built a shared semantic foundation for the definition of events by selecting disease concepts in the Unified Medical Language System (UMLS, V.2008AA),[21 22] based on the medical definitions reported in the EDF. As the four different terminologies encountered in the databases are part of the UMLS, the concepts could easily be projected into codes for the four disease coding terminologies used in the EU-ADR project. Table 3 shows the projection of UMLS concepts corresponding to the five events of interest into the various terminologies. This process of terminology mapping has been previously described for another event, upper gastrointestinal bleeding.[23]

The database holders were asked to construct initial queries for the identification of events of interest, using the recommended

Downloaded from jamia.bmj.com on March 30, 2013 - Published by group.bmj.com

Table 1 Characteristics of healthcare databases participating in the EU-ADR project

Characteristics	Pedianet* (Italy)	Health search/ CSD* (Italy)	Lombardy regional (Italy)	Tuscany regional (Italy)	IPCI (The Netherlands)	PHARMO (The Netherlands)	QRESEARCH (UK)	Aarhus (Denmark)
	Pedianet	HSD	Lombardy	ARS	IPCI	PHARMO	QRESEARCH	Aarhus
Current source population	129 742 children	771 907	9 924 758	3 585 560	479 585	1 280 805	1 515 116 (based on a 20% sample)	1 959 972
Type of database	General practice pediatric database	General practice database	Data warehouse record linkage system with (1) Registry of inhabitants (2) Regional drug dispensation records (3) Hospitalization claims	Data warehouse record linkage system with (1) Registry of inhabitants (2) Regional drug dispensation records (3) Hospitalization claims (4) Death registry	General practice database	Data warehouse record linkage system with (1) Registry of inhabitants (2) Regional drug dispensation records (3) Hospitalization claims (4) Laboratory values	General practice database	Data warehouse record linkage system with (1) Registry of inhabitants (2) Regional drug dispensation records (3) Hospitalization claims (4) Laboratory values (5) Death registry
Symptoms (yes/no)	Yes, as free text/ codes	Yes, as free text/ codes	No	No	Yes, as free text/ codes	Yes for some	Yes, as codes	No
Outpatient primary care diagnoses	Yes, as free text/ codes	Yes free text/ codes	No	No	Yes, as free text/ codes	No	Yes	No
Outpatient specialist care diagnoses	Yes, as free text/ codes	Yes	No	No	Yes	No	Yes	No
Hospital discharge diagnoses	Yes, as free text/ codes	Yes, as free text/ codes	Yes	Yes	Yes, as free text/ codes	Yes	Yes	Yes
Diagnosis coding scheme	ICD9-CM	ICD9-CM	ICD9-CM	ICD9-CM	ICPC	ICD9-CM	RCD v2 and v3	ICD10
Language of free text	Italian	Italian	No free text	No free text	Dutch	No free text	No free text	No free text
Diagnostic procedures	Yes	Yes	Yes	Yes	No	Yes for inhospital interventions	Yes	Yes, inhospital only
Laboratory tests	Yes	Yes	No	No	Yes	Yes (for a subset of patients)	Yes	Yes, inhospital only

*In Italy, children are cared for, until 14 years of age, by the family pediatrician and in the subsequent years by general practitioners.
EU-ADR, European Union-Adverse Drug Reaction; ICD9-CM and ICD10, International Classification of Diseases—9th revision Clinical Modification (CM) and 10th revision, respectively; ICPC, International Classification of Primary Care; IPCI, Integrated Primary Care Information; RCD, READ CODE Classification.

Downloaded from jamia.bmj.com on March 30, 2013 - Published by group.bmj.com

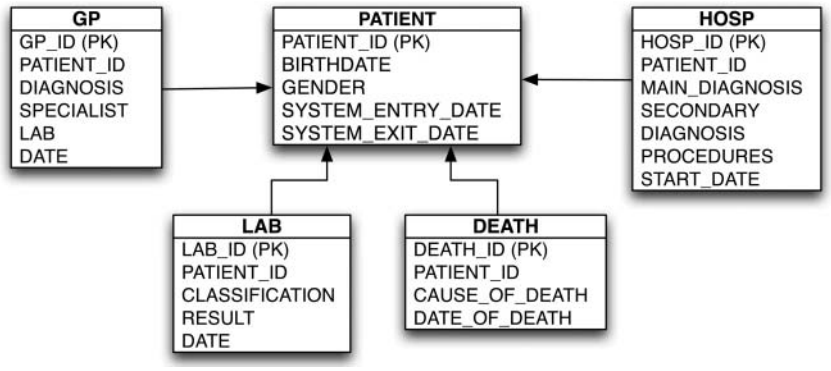

Figure 1 Common database model. HOSP, discharge summary from hospitalizations recorded by administrative/claims databases; DEATH, registry of death and causes of death; GP, information recorded by general practitioners during their clinical practice; LAB, information obtained from laboratory test results.

codes and terms from table 3. All queries from each database for each event were analyzed in content (ie, which codes or terms were used) and structure (ie, which record(s) from the database). These queries were subsequently compared across the different databases. Query analysis was aimed at assessing the consistency across different databases with respect to the use of similar information (ie, codes, free text, laboratory test results, query refinements) and to the search strategy within the same type of data (eg, primary and/or secondary hospital discharge diagnoses in claims databases). If major differences in the query analysis were identified, a consensus was reached among the respective database holders to adopt similar strategies for the event detection.

Evaluation of event data extraction

For each event and in each database, we calculated age specific and standardized incidence rates (IRs). Review of the literature was subsequently conducted to compare the event IRs obtained in the databases to what has been described in previous publications. Manual validation of the event extraction vis à vis medical charts was performed in the database Pedianet for all events. In those databases where it was possible to do so, validation by chart review and estimation of positive predictive values of the coding algorithms were conducted in a random sample of cases.[24 25]

RESULTS

The study population of the EU-ADR network for this analysis comprised overall 19 647 445 individuals with a follow-up of 59 929 690 person-years (PYs). Within this population, we identified overall the following number of events: (1) AMI=22 267; (2) ARF=2972; (3) AS=665; (4) BE=385; and (5) RHABD=1275.

Database holders ran several queries for each event, the recommended query consisting of a search in the primary hospital discharge diagnosis field of an administrative database record or the diagnosis field in a GP database record. Not all possible additional queries were relevant for each event (eg, no laboratory results were available to identify the event AS). The final agreement as to which query each database would adopt for each event data extraction was reached on the grounds of the following criterion: databases having a similar structure search in the same information sources but databases having broader sources of information can exploit them as much as possible, provided that the resulting IR is not inconsistent with what is described in the literature and with the IRs obtained from the other databases. Appendix 2 (available online only) shows the query analysis for AS (finalized after consensus discussions) as an example. The following are the age standardized IRs obtained using the harmonized queries for the five events of interest across the databases: (1) AMI: 60–148/100 000 PYs; (2) ARF: 3–49/100 000 PYs; (3) AS: 2–12/100 000

Table 2 Database content and attributes used for the event extraction process

Table	Fields	Aarhus	ARS	UNIMIB	HSD	IPCI	Pedianet	PHARMO	QRESEARCH
HOSP	Main diagnosis	ICD10	ICD9-CM	ICD9-CM				ICD9-CM	
	Secondary diagnosis	ICD10	ICD9-CM	ICD9-CM				ICD9-CM	
	Procedures	ICD10	ICD9-CM	ICD9-CM				ICD9-CM	
DEATH	Cause of death	ICD10	ICD9-CM						
GP	Diagnosis				ICD9-CM and free text (in Italian)	ICPC and free text (in Dutch)	ICD9-CM and free text (in Italian)		READ
	Specialist/hospital				Free text	Free text referrals	Free text/referrals		
	Lab				Free text/referrals				READ
	Death				Text/ICD9-CM	Text/ICPC	Text/ICD-9CM		READ
LAB	Classification	NPU				WCIA		WCIA	READ
	Result	NPU				numbers		numbers	numbers

HOSP: discharge summary from hospitalizations recorded by hospitals
- Main diagnosis: principal cause that led to the hospitalization episodes.
- Secondary diagnoses: five or more fields that contain diagnoses that refer either to pre-existing diseases or to complications that arose during the hospital stay.

DEATH: registry of death and causes of death
- Cause of death: principal cause of death.

GP: information recorded by general practitioners during their clinical practice
- Symptoms and/or physical examination and/or diagnosis.
- Specialist visit prescriptions and diagnoses.
- Laboratory results.

LAB: information obtained from laboratories
- Classification of laboratory analysis.

GP, general practice; ICD9-CM and ICD10, International Classification of Diseases—9th revision Clinical Modification (CM) and 10th revision, respectively; ICPC, International Classification of Primary Care; IPCI, Integrated Primary Care Information; NPU, nomenclature, properties and units; WCIE, Werkgroep Coördinatie Informatisering en Automatisering reference model for GP information systems.

Downloaded from jamia.bmj.com on March 30, 2013 - Published by group.bmj.com

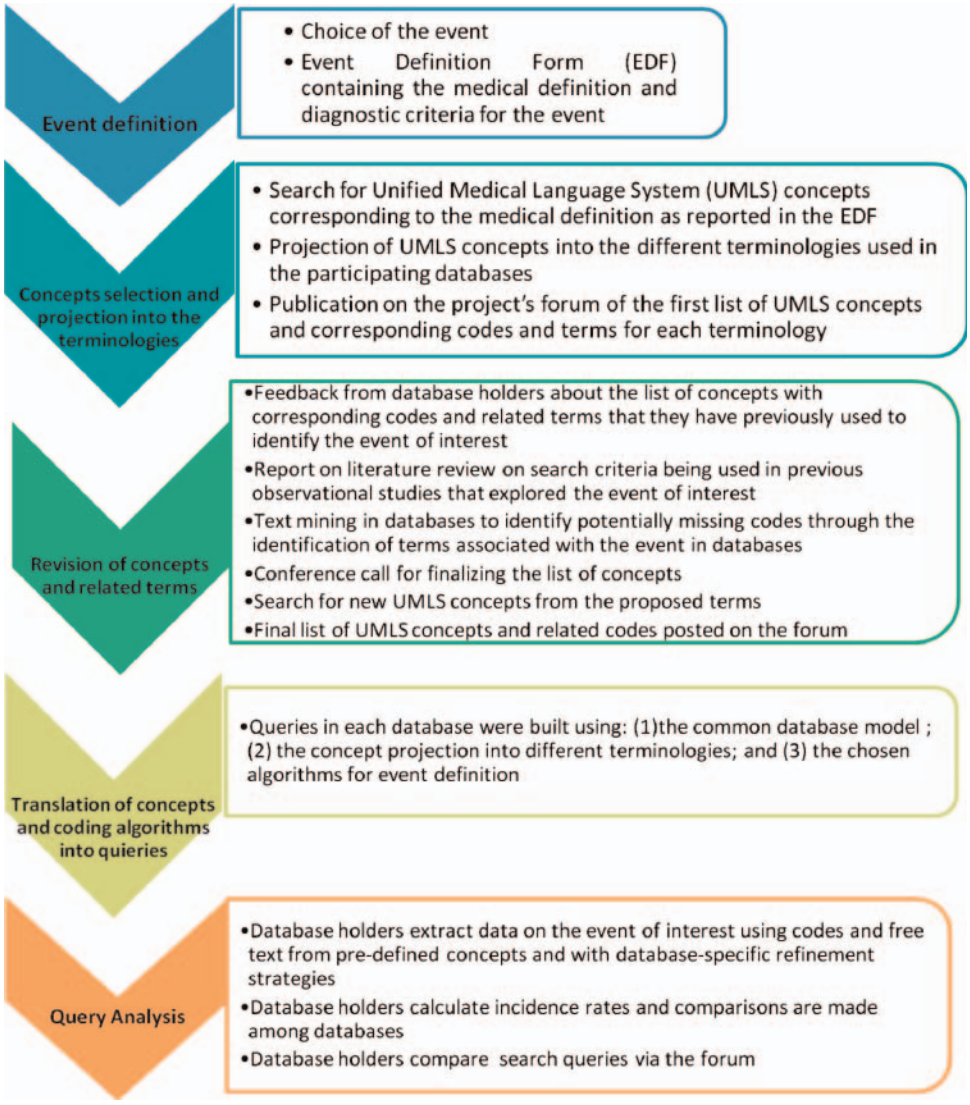

Figure 2 Workflow of the harmonization of event queries across the different databases. This figure is only reproduced in colour in the online version.

PYs; (4) BE: 2–17/100 000 PYs; and (5) RHABD: 0.1–8/100 000 PYs. The age specific IRs per database for all events are given in appendix 3 (available online only).

Benchmarking and harmonization of the event extraction processes changed the age standardized IRs across the databases to various extents. Analysis of the queries for the event data extraction revealed two main types of differences. The first is due to deviation from proposed concepts, resulting mainly from a database holder finding additional value in using other concepts that were found to be useful based on their own previous studies. Databases with free text information refined the queries according to ad hoc search algorithms while databases with information on laboratory test results used the numeric value associated with a diagnostic concept. As an example, 'anuria' was considered only if a value of serum creatinine was between 4 and 10 mg/dl and if the examination was done within 30 days of the date of diagnosis. Using the concept 'anuria' alone, without any laboratory results, would have decreased the specificity of the query. The second type of difference is due to modifications in the source of information considered in the query. In some databases, codes were searched for only in the field corresponding to 'primary hospital discharge diagnosis,' while in other databases, codes were searched for in both 'primary hospital discharge diagnosis' and 'secondary hospital discharge diagnosis'. The impact on the event rates of the revised queries incorporating additional information is summarized in table 4 for AMI, ARF, and AS. For AMI, an increase in the IRs resulted from the inclusion of records extracted from death registries in those databases for which this information was available: 25% and 15% increase in Aarhus and ARS, respectively. An increase in IR was similarly observed for ARF, when information from death registries was accounted for: 19% and 4% increase in Aarhus and ARS, respectively, although the introduction of a concept with refinement had a higher impact on the IR for PHARMO ('anuria,' with 52% increase). Information from death registries gave no additional contribution to the extraction of cases of AS in Aarhus or in ARS, while the search of specific codes within secondary discharge diagnoses had additional value for ARS (6% increase) and for Lombardy (27% increase).

Comparison of IRs of the events described in the literature revealed differences arising from broader event definitions being used by previous studies and use of other sources of information aside from coded hospital diagnoses, death registries, or

Downloaded from jamia.bmj.com on March 30, 2013 - Published by group.bmj.com

Table 3 Unified Medical Language System concepts projection into the four terminologies for five events of interest

Event	UMLS concept unique identifier	Preferred term	ICD9-CM	ICD10	RCD	ICPC
AMI	C0155626	Acute myocardial infarction	410.x	I21.x	G30z., XE0Uh	K75, K75002
	C0428953	ECG: myocardial infarction			323.., 323Z.	
	C0232320	ECG: antero-septal infarct			3233	
	C0428956	ECG: posterior/inferior infarct			3234	
	C0428955	ECG: subendocardial infarct			3235	
	C0232325	ECG: lateral infarction			3236	
	C0428953	ECG: myocardial infarction			323.., 323Z.	
	C0340324	Silent myocardial infarction			X200a	
	C0340283	Other acute and subacute ischemic heart disease NOS	411		G31.., G31yz	
	Only for refinement use*					
ARF	C0022660	Kidney failure, acute	584	N17		U99005
	C0022672	Kidney tubular necrosis, acute	584.5	N17.0	K040.	U05001
	C0003460	Anuria	788.5	R34	1AC0., R0851, R34	
	Only for refinement use*					
AS	C0002792	Anaphylactic shock		T78.2	SN50	A12004, A92005
	C0375697	Other AS	995.0			
	C0685898	AS due to adverse food reaction	995.6	T78.0	X70vm, X70w1	
	C0161840	AS due to serum	999.4	T80.5	SP34, X70vl	
	C0274304	AS, due to adverse effect of correct medicinal substance properly administered		T88.6	SN501	
BE	C0235818	Bullous eruption			XM05i	
	C0014742	Erythema multiforme	695.1	L51	M151z, XE1B0	S99007
	C0038325	Stevens–Johnson syndrome	695.1, 695.14, 695.15		M1517, X50CE	A12005, S99032
	C0014518	Toxic epidermal necrolysis	695.1, 695.14, 695.15	L51.2	M1518	
	C0085932	Skin diseases, bullous	694	L10-L14.9	M14.., M14z., My1.,	
RHABD	C0035410	Rhabdomyolysis	728.88	M62.8	X70AI	
	C1135344	Acute necrotizing myopathy	359.81			
	C1401301	Ischemia; muscle				
	C0027080	Myoglobinuria	791.3	R82.1	X709S, R113	

*These concepts and related codes were used only in some databases wherein it was possible to increase their specificity through a refinement of the query for the event data extraction, by including additional information (ie, laboratory test results, free text specifications).
AMI, acute myocardial infarction; ARF, acute renal failure; AS, anaphylactic shock; BE, bullous eruptions; ICPC, International Classification of Primary Care; NOS, not otherwise specified; RCD, READ CODE; ICD9-CM and ICD10, International Classification of Diseases—9th revision Clinical Modification (CM) and 10th revision, respectively; RHABD, rhabdomyolysis; UMLS, Unified Medical Language System.

free text. These differences are discussed in greater detail in the next section.

DISCUSSION

This study shows how event extractions may differ across databases and how different choices impact on the estimated incidence of a given event. We have described a workflow that has been successful for combining data across databases of various origins and constructs in the context of drug safety signal detection. The UMLS was used as the common terminological system to map events across different terminologies. The knowledge described in the various terminologies, which are included in the UMLS, was inadequate to define all of the clinical aspects of an event and so expert knowledge and experience from the database holders were necessary to build a more comprehensive definition of the event—that is, for some events (such as ARF and RHABD), disease codes alone were not sufficient to allow event identification, hence necessitating use of additional information from laboratory findings, or signs and symptoms from the free text narratives of GP records. Use of EHR databases requires an understanding of how the healthcare data are generated from the initial patient encounter all the way to completion of the database entry. Table 5 gives a summary of the main reasons for the differences in the extraction of events across the databases.

The small difference in the IR for AMI observed between the two Italian regional claims databases is probably explained by the fact that Lombardy could not contribute AMI related deaths occurring outside of the hospital while the lower IR observed in the Italian GP database is probably due to the non-routine recording of hospital deaths (including those attributed to AMI) by the GP. In line with this hypothesis, similar estimates were observed in another GP database, QRESEARCH from the UK (67.4/100 000 PYs). Despite using a search strategy similar to the one used by ARS, in the Danish claims database (Aarhus), a higher IR (126.5/100 000 PYs) was observed. This disparity may be due to inherent differences between the two underlying populations (Italian and Danish), as a consequence of the so-called south–north trend in cardiovascular diseases, largely attributable to the Mediterranean diet.[26–29] The same trend could probably explain the difference between Lombardy and PHARMO (93.4/100 000 PYs), which used the same query but have different populations. The highest IR for AMI was observed in the Dutch GP database IPCI. This is probably an overestimation since the IR observed in the Dutch administrative database PHARMO is much lower. This overestimation is most likely due to an extensive use of free text in the search strategy of IPCI.

The pattern of estimates for RHABD was noted to be quite different from the other events. This particular event is not captured by the administrative databases that are unable to link to a data source having laboratory results (ie, the two Italian regional databases, ARS and Lombardy). It is also important to note that RHABD is part of a spectrum of conditions (myopathy→RHABD→renal failure) and is more of a clinical manifestation than an actual diagnosis. Thus if patients are

Downloaded from jamia.bmj.com on March 30, 2013 - Published by group.bmj.com

Table 4 Comparison of age standardized incidence rates (per 100 000 person-years) for acute myocardial infarction, acute renal failure and anaphylactic shock, based on recommended queries and the effect of additional information

		Incidence rate based on recommended query		Incidence rate based on additional data (% increase)	
Event	**Database**	**HOSP-main**	**GP**	**Additional information from DEATH**	**Additional information from concept with refinement**
AMI	Aarhus	101.4		*126.5 (+25%)*	
	ARS	77.8		*90.2 (+15%)*	
	HSD		58.7		*59.1 (+0.5%)*
	IPCI		148.4		
	PHARMO	93.4			
	Lombardy	82.5			
ARF	Aarhus	6.3		7.1 (+19%)	*17.9 (+150%)*
	ARS	12.1		*12.6 (+4%)*	
	HSD		3.2		*3.3 (+3%)*
	IPCI		48.9		
	PHARMO	2.3			*3.6 (+52%)*
	Lombardy	*15.4*			
Event	**Database**	**HOSP-main**	**GP**	**Additional information from HOSP-sec**	**Additional information from DEATH**
AS	Aarhus	*5.7*		6.4 (+12%)	6.4 (+0%)
	ARS	*12.0*		12.7 (+6%)	12.8 (+0%)
	HSD		*5.2*		
	IPCI		7.9		
	PHARMO	*1.9*		2.4 (+26%)	
	Lombardy	*2.2*		2.8 (+27%)	

Italicized cells identify the final query.
HOSP-main: primary diagnosis that led to the hospitalization.
HOSP-sec: any one of secondary diagnoses; may refer to either pre-existing diseases or to complications that arose during the hospital stay.
DEATH: registry of death and causes of death.
GP: information recorded by general practitioners during patient visits.
AMI, acute myocardial infarction; ARF, acute renal failure; AS, anaphylactic shock; GP, general practice; IPCI, Integrated Primary Care Information; IRs, incident rates.

admitted to a hospital for RHABD due to whatever cause, by the time they are discharged, the etiology of the RHABD has already been found and this is what is recorded as the discharge diagnosis (eg, trauma, burns, sepsis, poisoning). Finally, if the RHABD progresses to ARF, then the case is likely to be registered as ARF.

Pedianet is an Italian nationwide database which contains medical information concerning children until 14 years of age, as recorded by a family pediatrician. Hence the IRs obtained from this database cannot be directly compared with those of the other databases, although the data can be explored to further study specific differences within the pediatric populations of each database.

Comparison with IRs in published literature

In the Spanish EPIC cohort study[30] involving over 33 000 individuals with ~300 000 PYs of follow-up, age standardized IRs of AMI were found to be in the range of 302–330/100 000 PYs in men and 60–114/100 000 PYs in women. In a large US community based population, the age and sex adjusted incidence of AMI was found to be 208/100 000 PYs.[31] Those figures are higher than those obtained in the EU-ADR project (59.1–148.4 per 100 000 PYs), most likely because these studies employed broader search queries and other sources of information aside from coded hospital diagnoses, death registries, and free text. In the Spanish study, ICD9CM diagnostic codes 410-414 and ICD10 I20-I25 codes (all codes for ischemic heart disease), as

Table 5 Main sources of differences in the extraction of events across databases in the European Union-Adverse Drug Reaction project

	Examples
(1) Differences in granularity of disease coding system used	ICPC coding generally less granular and less specific compared with the ICD system
(2) Availability of unstructured clinical information (ie, free text) to supplement and refine search query	Free text narratives, containing clinical signs and symptoms and other pertinent information, available in GP databases but not in administrative/claims databases
(3) Ability to link to information containing laboratory findings or procedures to supplement and refine search query	Administrative databases with record linkage provide more refined information for events that include laboratory findings as part of their definition (eg, acute renal failure)
(4) Availability of information on death due to the event of interest that occurs outside (and thus not usually recorded in) primary care practice or hospitalization	Out of hospital deaths due to acute myocardial infarction may not be routinely recorded by GPs but can be found in death registries
(5) Individual differences in recording practices among data contributors	GPs contribute data to HSD use the same data entry format and software, wherever they practice in Italy; GPs that contribute data to IPCI use various data entry formats and software, depending on the practice.
(6) Time span/severity of the clinical conditions that give rise to a record	Databases which collect data from GPs in a GP gatekeeper health system are likely to keep track of the majority of the clinical conditions that affect an individual during their time span in the study, while databases whose basic data source is hospital are expected to provide complete information on more severe events

GP, general practice; GPs, general practitioners; ICPC, International Classification of Primary Care; IPCI, Integrated Primary Care Information.

well as ICD9 procedure codes for stent placement and bypass operation (36.0 and 36.1, respectively), were used. Cases were also ascertained by means of self-reported questionnaires, population based specific AMI registries, and autopsy data. The US study included only individuals who were 30 years of age or older. There are conflicting views as to whether coronary heart disease, or associated cardiovascular risk factors, is more prevalent in North America or in Europe.[32–35] When defining the extraction strategy in EU-ADR, specificity of the query took precedence over sensitivity so as to avoid having too many false positive drug safety signals. For example, in identifying the event AMI, we did not use the concept 'myocardial infarction' but rather '*acute* myocardial infarction'. In other publications, the approach is to have a much broader definition of an event, including the whole spectrum of the acute coronary syndrome, which includes unstable angina.[36–38]

There are very few population based studies on the incidence of ARF; most are focused on special populations such as the elderly, patients in intensive care units, or those requiring renal replacement therapy.[39 40] It has been suggested that ARF is nearly as common as myocardial infarction.[41] Indeed, a population based study in Scotland obtained an IR of 181/100 000 PYs.[42] This study employed only laboratory values in the identification of cases (ARF defined as having baseline serum creatinine below the threshold (150 μmol/l in men or 130 μmol/l in women) that subsequently increased by a factor of 1.5 or more, or the glomerular filtration rate was reduced by at least 25%). The lower estimates we obtained in EU-ADR (3.3–48.9/100 000 PYs) are probably due to the fact that laboratory data were not uniformly available in all of the databases to supplement the coded diagnoses.

Downloaded from jamia.bmj.com on March 30, 2013 - Published by group.bmj.com

Most of the published studies on the incidence of AS are based on emergency room visits or hospitalizations, which makes direct comparisons difficult.[43–46] A Swiss study showed that the incidence of life threatening anaphylaxis ranged from 7.9 to 9.6/100 000 PYs, which is not far off from the estimates we obtained in EU-ADR (1.9–12.1/100 000 PYs).

Toxic epidermal necrolysis and Stevens–Johnson syndrome, which collectively comprise the event BE, are rare conditions occurring worldwide and most often in adults, women more likely than men. Its incidence is estimated at 2–3 cases/million population/year in Europe but is up to three times higher in the HIV infected population.[47] The IRs we obtained in EU-ADR (16–178 per million PYs) are clearly an overestimation of the true incidence, possibly due to inclusion of other forms of BE. Diagnosis of BE usually requires histologic confirmation,[48 49] information that is not uniformly available in the databases in EU-ADR.

As discussed previously, estimating the incidence of RHABD is difficult because it is part of a spectrum of conditions that start with myopathy progressing to muscle necrosis and renal failure, and most studies in the literature investigate the endpoint renal failure or the entire spectrum.[50 51] Thus use of diagnostic codes alone, without values for serum creatine phosphokinase (CPK), may underestimate the true incidence. At the same time, with case identification algorithms that employ CPK values, there is a need to distinguish elevations due to cardiac ischemia or infarction (usually detected by the CPK-MB isoenzyme) with those due to skeletal muscle injury (detected by the CPK-MM isoenzyme which, although predominantly found in skeletal muscle, is also the primary CPK isoenzyme present in heart muscle). We only employed CPK-total in the search algorithms.

Coding changes in international disease classification have posed new challenges for the comparability of indicators for various diseases. For example, in the worldwide MONICA Project which studied fatal and non-fatal coronary events through population based registers, different versions of the ICD were used in different countries.[33] This is in addition to the innovations made in the past decade with respect to diagnostic technologies that have enabled diagnosis at earlier stages (such as novel biomarkers).[52–54] Healthcare systems and physician practices that vary between countries may also play a role in how clinical outcomes are captured in databases.[55] All of these factors need to be considered when analyzing trends in disease frequency, severity, prognosis and subsequent variations in medical practice. It is hoped that the harmonization process we have described will enable the identification of outcomes across differently structured databases using compatible definitions and facilitate a better comparison of IRs among various data sources and countries.

Limitations and future directions

While the aim of our study was not to estimate the most accurate IRs for each event, the benchmarking and harmonization process enabled estimation and evaluation of rates of events with a common definition across databases having a similar structure. This work is based on currently available data that do not capture sources of bias and residual differences, including the effects of immigration and ethnic variation. Case validation using manual review of hospitalization records and GP records remains an important part of the process; work is currently ongoing in EU-ADR to determine how the accuracy of database queries influences the estimation of risks, in the context of drug safety surveillance.

CONCLUSION

No single data source is likely to be sufficient to meet all of the expected needs for drug safety surveillance; hence, it is valuable to assess the feasibility and utility of analyzing multiple data sources concurrently. We have provided an external shared semantic basis in content and structure for the creation of queries adapted to the heterogeneous EHR databases within the EU-ADR network. The iterative harmonization process enabled a more homogeneous identification of events across differently structured databases using different coding schemes. This workflow can facilitate transparent and reproducible extraction of events using EHR databases as well as a better understanding of differences between databases.

Acknowledgments The authors would like to thank Julia Hippisley-Cox for sharing the data from the QREASEARCH Database and her comments on the paper.

Contributors PA, PMC, RG, MS, GT: contributed to the conception, design, and drafting of the article. FM, J-CD, GM, CG, CF, RH, MM, LP, AF-R, MF, MS, JvdL, AP: revised the paper critically. Final approval of the version to be published was obtained from all authors.

Funding This work was supported by the European Commission Seventh Framework Programme (FP7/2007–2013) under grant No 215847—The EU-ADR Project.

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 The data remain with the database owner. Each database owner can be contacted to have access to their data.

REFERENCES

1. **Lindquist M.** VigiBase, the WHO global ICSR database system: basic facts. *Drug Inf J* 2008;**42**:409–419.
2. **De Bruin ML,** van Puijenbroek EP, Egberts AC, *et al*. Non-sedating antihistamine drugs and cardiac arrhythmias—biased risk estimates from spontaneous reporting systems? *Br J Clin Pharmacol* 2002;**53**:370–4.
3. **Begaud B,** Martin K, Haramburu F, *et al*. Rates of spontaneous reporting of adverse drug reactions in France. *JAMA* 2002;**288**:1588.
4. **Nadkarni PM.** Drug safety surveillance using de-identified EMR and claims data: issues and challenges. *J Am Med Inform Assoc* 2010;**17**:671–4.
5. **Reisinger SJ,** Ryan PB, O'Hara DJ, *et al*. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *J Am Med Inform Assoc* 2010;**17**:652–62.
6. **Sturkenboom MCJM.** Other databases in Europe for the analytic evaluation of drug effects. In: Mann R, Andrews E, eds. *Pharmacovigilance*. 2nd edn. Chichester, UK: John Wiley & Sons, Ltd, 2007.
7. **FDA Sentinel Initiative.** http://www.fda.gov/Safety/FDAsSentinelInitiative (accessed 20 Apr 2012)
8. **Platt R,** Wilson M, Chan KA, *et al*. The new Sentinel Network—improving the evidence of medical-product safety. *N Engl J Med* 2009;**361**:645–7.
9. **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.
10. **Weeber MVR,** Klein H, De Jong-Van Den Berg LT, *et al*. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *J Am Med Inform Assoc* 2003;**10**:252–9.
11. **Trifiro G,** Pariente A, Coloma PM, *et al*. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiol Drug Saf* 2009;**18**:1176–84.
12. **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.
13. **Avillach P,** Mougin F, Joubert M, *et al*. A semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European eu-ADR project. *Stud Health Technol Inform* 2009;**150**:190–4.

Downloaded from jamia.bmj.com on March 30, 2013 - Published by group.bmj.com

14. **ICD-9-CM**. International Classification of Diseases, 9th edition, Clinical Modification: update. Official authorized addendum, effective October 1, 1986. *J Am Med Rec Assoc* 1986;**57**(Suppl) 1–32.
15. **Pavillon G,** Maguin P. The 10th revision of the International Classification of Diseases. *Rev Epidemiol Sante Publique* 1993;**41**:253–5.
16. **Lamberts H,** Wood M, eds. *ICPC: international classification of primary care*. Oxford: Oxford University Press, 1987.
17. **O'Neil M,** Payne C, Read J. Read Codes Version 3: a user led terminology. *Methods Inf Med* 1995;**34**:187–92.
18. **Moore KM,** Duddy A, Braun MM, *et al*. Potential population-based electronic data sources for rapid pandemic influenza vaccine adverse event detection: a survey of health plans. *Pharmacoepidemiol Drug Saf* 2008;**17**:1137–41.
19. **Administration UFaD.** Adverse Event Reporting System (AERS). http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm (accessed 11 Nov 2011).
20. **Platt R,** Carnahan RM, Brown JS, *et al*. The US Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf* 2012; **21**(Suppl 1):1–8.
21. **Humphreys BL.** The 1994 unified medical language system knowledge sources. *Health Libr Rev* 1994;**11**:200–3.
22. **Lindberg DA,** Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med* 1993;**32**:281–91.
23. **Avillach P,** Joubert M, Thiessard F, *et al*. Design and evaluation of a semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European EU-ADR project. *Stud Health Technol Inform* 2010;**160**:1085–9.
24. **Valkhoff VE,** Coloma PM, Lapi F, *et al*. Positive predictive value for upper gastrointestinal bleeding in four healthcare databases using different coding systems in the EU-ADR project. Presented at the Digestive Disease Week, San Diego, California, May 19–22, 2012.
25. **Coloma PM,** Valkhoff VE, Mazzaglia G, *et al*. Accuracy of coding-based algorithms in identification of acute myocardial infarction in multi-country electronic healthcare records databases. In: Coloma PM**.** *Mining electronic healthcare record databases to augment drug safety surveillance (PhD thesis)*. Rotterdam, The Netherlands: Erasmus Universiteit Rotterdam, 2012.
26. **Barchielli A,** Balzi D, Pasqua A, *et al*. Incidence of acute myocardial infarction in Tuscany, 1997–2002: data from the Acute Myocardial Infarction Registry of Tuscany (Tosc-AMI). *Epidemiol Prev* 2006;**30**:161–8.
27. **Yusuf S,** Reddy S, Ounpuu S, *et al*. Global burden of cardiovascular diseases: part I: general considerations, the epidemiologic transition, risk factors, and impact of urbanization. *Circulation* 2001;**104**:2746–53.
28. **Menotti A.** Food patterns and health problems: health in southern Europe. *Ann Nutr Metab* 1991;**35**(Suppl 1):69–77.
29. **Menotti A,** Lanti M, Puddu PE, *et al*. Coronary heart disease incidence in northern and southern European populations: a reanalysis of the seven countries study for a European coronary risk chart. *Heart* 2000;**84**:238–44.
30. **Larranaga N,** Moreno C, Basterretxea M, *et al*. Incidence of acute myocardial infarction in the Spanish epic cohort. *An Sist Sanit Navar* 2009;**32**:51–9.
31. **Yeh RW,** Sidney S, Chandra M, *et al*. Population trends in the incidence and outcomes of acute myocardial infarction. *N Engl J Med* 2010;**362**:2155–65.
32. **Tunstall-Pedoe H,** Kuulasmaa K, Amouyel P, *et al*. Myocardial infarction and coronary deaths in the World Health Organization MONICA Project. Registration procedures, event rates, and case-fatality rates in 38 populations from 21 countries in four continents. *Circulation* 1994;**90**:583–612.
33. **Stewart AW,** Kuulasmaa K, Beaglehole R, for the WHO MONICA Project. Ecological analysis of the association between mortality and major risk factors of cardiovascular disease. The World Health Organization MONICA Project. *Int J Epidemiol* 1994;**23**:505–16.
34. **Higgins M.** Patients, families and populations at high risk for coronary heart disease. *Eur Heart J* 2001;**22**:1682–90.
35. **Wolf-Maier K,** Cooper RS, Banegas JR, *et al*. Hypertension prevalence and blood pressure levels in 6 European countries, Canada, and the United States. *JAMA* 2003;**289**:2363–9.
36. **Perugini E,** Maggioni AP, Boccanelli A, *et al*. [Epidemiology of acute coronary syndromes in Italy]. *G Ital Cardiol (Rome)* 2010;**11**:718–29.
37. **Ruff CT,** Braunwald E. The evolving epidemiology of acute coronary syndromes. *Nat Rev Cardiol* 2011;**8**:140–7.
38. **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.
39. **Uchino S.** The epidemiology of acute renal failure in the world. *Curr Opin Crit Care* 2006;**12**:538–43.
40. **Hoste EA,** Schurgers M. Epidemiology of acute kidney injury: how big is the problem? *Crit Care Med* 2008;**36**(4 Suppl):S146–51.
41. **Srisawat N,** Kellum JA. Acute kidney injury: definition, epidemiology, and outcome. *Curr Opin Crit Care* 2011;**17**:548–55.
42. **Ali T,** Khan I, Simpson W, *et al*. Incidence and outcomes in acute kidney injury: a comprehensive population-based study. *J Am Soc Nephrol* 2007;**18**:1292–8.
43. **Moro Moro M,** Tejedor Alonso MA, Esteban Hernandez J, *et al*. Incidence of anaphylaxis and subtypes of anaphylaxis in a general hospital emergency department. *J Investig Allergol Clin Immunol* 2011;**21**:142–9.
44. **Harduar-Morano L,** Simon MR, Watkins S, *et al*. A population-based epidemiologic study of emergency department visits for anaphylaxis in Florida. *J Allergy Clin Immunol* 2011;**128**:594–600 e591.
45. **Koplin JJ,** Martin PE, Allen KJ. An update on epidemiology of anaphylaxis in children and adults. *Curr Opin Allergy Clin Immunol* 2011;**11**:492–6.
46. **Tang ML,** Osborne N, Allen K. Epidemiology of anaphylaxis. *Curr Opin Allergy Clin Immunol* 2009;**9**:351–6.
47. **Fritsch P.** European Dermatology Forum: skin diseases in Europe. Skin diseases with a high public health impact: toxic epidermal necrolysis and Stevens–Johnson syndrome. *Eur J Dermatol* 2008;**18**:216–17.
48. **Kaufman DW.** Epidemiologic approaches to the study of toxic epidermal necrolysis. *J Invest Dermatol* 1994;**102**:31S–3S.
49. **French LE.** Toxic epidermal necrolysis and Stevens Johnson syndrome: our current understanding. *Allergol Int* 2006;**55**:9–16.
50. **Bollaert PE,** Frisoni A. Epidemiology, mechanisms and clinical features of rhabdomyolysis. *Minerva Anestesiol* 1999;**65**:245–9.
51. **Bosch X,** Poch E, Grau JM. Rhabdomyolysis and acute kidney injury. *N Engl J Med* 2009;**361**:62–72.
52. **Madsen M,** Gudnason V, Pajak A, *et al*. Population-based register of acute myocardial infarction: manual of operations. *Eur J Cardiovasc Prev Rehabil* 2007; **14**(Suppl 3):S3–22.
53. **Di Pasquale G,** Lombardi A, Casella G. The redefinition of acute myocardial infarction. *Ital Heart J* 2004;**5**(Suppl 6):9S–18S.
54. **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.
55. **von dem Knesebeck O,** Bonte M, Siegrist J, *et al*. Country differences in the diagnosis and management of coronary heart disease—a comparison between the US, the UK and Germany. *BMC Health Serv Res* 2008;**8**:198.

Downloaded from jamia.bmj.com on March 30, 2013 - Published by group.bmj.com

Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project

Paul Avillach, Preciosa M Coloma, Rosa Gini, et al.

J Am Med Inform Assoc 2013 20: 184-192 originally published online September 6, 2012
doi: 10.1136/amiajnl-2012-000933

Updated information and services can be found at:
http://jamia.bmj.com/content/20/1/184.full.html

These include:

Data Supplement *"Web Only Data"*
http://jamia.bmj.com/content/suppl/2012/09/06/amiajnl-2012-000933.DC1.html

References This article cites 48 articles, 9 of which can be accessed free at:
http://jamia.bmj.com/content/20/1/184.full.html#ref-list-1

Email alerting service Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/