Accuracy of claim data in the identification and classification of adults with congenital heart diseases in electronic medical records

Sarah Cohen, Anne-Sophie Jannot, Laurence Iserin, Damien Bonnet, Anita Burgun, Jean-Baptiste Escudié

Accuracy of claim data in the identification and classification of adults with congenital heart diseases in electronic medical records

Performance des données de remboursement pour identifier et classer les adultes atteints de cardiopathie congénitale dans les dossiers médicaux électroniques

Abbreviated title: Identification of ACHD in electronic medical records

Sarah Cohen[a,*], Anne-Sophie Jannot[a,b], Laurence Iserin[c], Damien Bonnet[d], Anita Burgun[a,b], Jean-Baptiste Escudié[a,b]

[a] *INSERM-UMRS 1138, Team 22, Cordeliers Research Centre, Paris Descartes University, 75006 Paris, France*

[b] *Department of Medical Informatics and Public Health, Georges Pompidou European Hospital, AP-HP, 75015 Paris, France*

[c] *Adult Congenital Heart Disease Unit, Cardiology Department, M3C – Reference Centre for Complex Congenital Heart Diseases, Georges Pompidou European Hospital, AP-HP, 75015 Paris, France*

[d] *Department of Paediatric Cardiology, M3C – Reference Centre for Complex Congenital Heart Diseases, Hôpital Necker-Enfants Malades, AP-HP, 75015 Paris; Paris Descartes University Sorbonne Paris Cité, 75006 Paris, France*

* Corresponding author at: INSERM U1138, Equipe 22, Centre de Recherche des Cordeliers, 15 rue de l'Ecole de Médecine, 75006 Paris, France.

E-mail address: sarah.cohen.hegp@gmail.com (S. Cohen).

Summary

Background. – The content of electronic medical records (EMRs) encompasses both structured data, such as billing codes, and unstructured data, including free-text reports. Epidemiological and clinical research into adult congenital heart disease (ACHD) increasingly relies on administrative claim data using the International Classification of Diseases (9th revision) (ICD-9). In France, administrative databases use ICD-10, the reliability of which is largely unknown in this context.

Aims. – To assess the accuracy of ICD-10 codes retrieved from administrative claim data in the identification and classification of ACHD.

Methods. – We randomly included 6000 patients hospitalized at least once in 2000–2014 in a cardiology department with a dedicated specialized ACHD Unit. For each patient, the clinical diagnosis extracted from the EMR was compared with the assigned ICD-10 codes. Performance of ICD-10 codes in the identification and classification of ACHD was assessed by estimating sensitivity, specificity and positive predictive value.

Results. – Among the 6000 patients included, 780 (13%) patients with ACHD were manually identified from EMRs (107,092 documents). ICD-10 codes correctly categorized 629 as having ACHD (sensitivity 0.81, 95% confidence interval 0.78–0.83), with a specificity of 0.99 (95% confidence interval 0.99–1). The performance of ICD-10 codes in correctly categorizing the ACHD defect subtype depended on the defect, with sensitivity ranging from 0 (e.g. unspecified congenital malformation of tricuspid valve) to 1 (e.g. common arterial trunk), and specificity ranging from 0.99 to 1.

Conclusions. – Administrative data using ICD-10 codes is a precise tool for detecting ACHD, and may be used to establish a national cohort. Mining free-text reports in addition to coded administrative data may offset the lack of sensitivity and accuracy when describing the spectrum of congenital heart disease using ICD-10 codes.

Résumé

Contexte. – Les dossiers médicaux électroniques (DME) incluent des données structurées-codes de facturation- et des données non structurées sous forme texte libre. L'épidémiologie clinique concernant des adultes ayant une cardiopathie congénitale (ACC) repose de plus en plus sur des données administratives de remboursement utilisant la Classification internationale des Maladies,

neuvième révision (CIM-9). En France, ces bases de données utilisent la CIM-10 dont la fiabilité est inconnue dans ce contexte.

Objectifs. – Évaluer la performance des codes CIM-10 extraits des données de remboursement pour identifier et classer les patients ACC.

Méthodes. – 6000 patients adultes tirés au sort et hospitalisés au moins une fois en 2000–2014 ont été inclus. Le diagnostic clinique extrait du DME a été comparé aux codes CIM-10 assignés. La sensibilité, la spécificité et la valeur prédictive positive pour identifier et classer les patients ACC ont été évaluées.

Résultats. – 780/6000 (13 %) patients ACC ont été identifiés à partir des DME (107,092 documents). Les codes CIM-10 ont identifié 629 d'entre eux (sensibilité 0,81, IC95 % 0,78–0,83) avec une spécificité de 0,99 (IC95 % 0,99–1). La performance du codage pour catégoriser le type de cardiopathie dépendait de celui-ci avec une sensibilité allant de 0 à 1 et une spécificité de 0,99 et 1.

Conclusions. – Les données administratives utilisant la CIM-10 permettent d'identifier les ACC et pourraient servir à établir une cohorte nationale d'ACC. Les méthodes d'extraction automatique d'informations des comptes rendus cliniques pourraient compenser le manque de sensibilité et de précision de la CIM-10 à décrire le spectre des cardiopathies congénitales.

KEYWORDS

Adult congenital heart disease;

Diagnostic code;

Nomenclature;

Electronic medical records

MOTS CLÉS

Cardiopathie congénitale adulte ;

Code diagnostic ;

Nomenclature ;

Dossier médical électronique

Abbreviations: ACHD, adult with congenital heart disease; CDW, Clinical Data Warehouse; CHD, congenital heart disease; CM, Clinical Modification; EMR, electronic medical record; FN, false negative; FP, false positive; HEGP, Georges Pompidou European Hospital; ICD-9, International Classification of Diseases (9th revision); ICD-10, International Classification of Diseases (10th revision); NLP, natural language processing; NPV, negative predictive value; PFO, patent foramen ovale; PPV, positive predictive value; regex, regular expressions; TN, true negative; TP, true positive.

Background

Advances in medicine, paediatrics and surgery have resulted in the improved survival of patients with congenital heart disease (CHD) [1, 2], with > 85% of patients now reaching adulthood [3]. Accordingly, the population of patients with adult congenital heart disease (ACHD) is growing, and their number now exceeds the number of children with CHD [4]. Despite surgical correction, patients with ACHD retain a lifelong risk of late complications, arising from residual defects and clinical sequelae, resulting in high rate of resource utilization [5]. The growing healthcare needs of patients with ACHD required the creation of specialized ACHD care units. To improve and assess healthcare for this growing population, specific guidelines have been published [6, 7] and quality indices have been proposed [8]. Research in this domain benefits from the collection of a large quantity of accurate detailed longitudinal routine data. Over the past 10 years, health administrative databases, such as administrative claims databases, have been used to study differences in surgical outcomes [9], lifelong co-morbidities such as arrhythmia [10], stroke [11], infective endocarditis [12] and hospitalization rates [13], drawing important conclusions.

In most countries (e.g. North America), health administrative data have been coded until now using the International Classification of Diseases (9th revision) (ICD-9), with or without modifications. However, several studies comparing the accuracy of ICD-9 codes with medical records in the identification of patients with CHD found variations in agreement [14-16]. Since 2000 in France, administrative data used for the Diagnosis-Related Groups in hospitals have been based on the International Classification of Diseases (10th revision) (ICD-10), first published in 1993, the reliability of which is still largely unknown in this context. Moreover, many other countries have moved recently from ICD-9 to ICD-10.

Electronic medical records (EMRs) are used as part of routine clinical care [17]; they have great potential to identify large cohorts, and serve as a rich data source for clinical and translational research. The information content encompasses both structured (i.e. coded) data and unstructured (i.e. narrative) data. Coded data are entered in a structured format, and usually include basic demographics and billing diagnostic codes, whereas narrative data are stored as free text in physician notes, and provide detailed information on a broad range of content, such as medical history or co-morbidities. Free-text reports contain information not captured by the structured data in EMRs [18]. Thereafter, we will use the term “administrative diagnosis” for a diagnosis derived from codified

data/billing codes while “clinical diagnosis” will be derived from narrative data. With the implementation of EMRs in most countries, strategies for automated large-scale free-text processing have emerged. Natural language processing (NLP) computational methods aim to extract information automatically from text. A straightforward option consists of using a rule-based NLP approach, which typically involves a list of relevant terms to extract. Thus, NLP could be used to extract CHD terms from EMRs, and hence detect patients with CHD in a corpus of medical records.

In the current study, our primary objective was to assess the accuracy of administrative data (i.e. ICD-10 billing codes) to detect ACHD, using the clinical diagnosis from narrative data as a reference, at a French adult hospital – the Georges Pompidou European Hospital (HEGP). HEGP is an 800-bed AP-HP hospital located in Paris that has benefited from a clinical information system since 2000, and has a dedicated specialized ACHD Unit within the General Cardiology Department. To establish this clinical diagnosis, an expert reviewed the EMRs manually with the assistance of the NLP tool. Our secondary objective was to assess the accuracy of administrative data to correctly classify ACHD into specific subtypes of CHD.

Methods

Overview of methodology

Fig. 1 shows an overview of our methodology. We randomly selected 6000 patients who had been hospitalized at least once in the General Cardiology Department within the HEGP, a tertiary academic medical centre. We then extracted all their documents between 2000 and 2014 from the Clinical Data Warehouse (CDW). We established a list of regular expressions (regex) to automatically identify potential mentions of CHD in patient records from the initial corpus of 6000 EMRs. EMRs with at least one match were reviewed manually by a clinician, to identify CHD lesions from the narrative reports. A web-based application, using information extraction methods, was used to assist clinicians in this task [19]. A clinical diagnosis was then attributed to each patient if at least one entity (i.e. at least one CHD lesion) was confirmed. In parallel, ICD-10 diagnosis codes for CHD were retrieved for the same set of patients. We then assessed the ability of the ICD-10 codes to detect ACHD by calculating sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV). We analysed the results globally, and for each subtype of CHD separately.

Data source

The HEGP is an 800-bed acute-care academic hospital in Paris; it was one of the first French teaching hospitals to implement a functional CDW [20]. The HEGP CDW relies on the i2b2 standard [21], and contains all information available from EMRs since the hospital opened in July 2000. As of 31 December 2014, it had integrated data from 808,869 patients [22]. The large majority of the data stored in the CDW are structured data, including administrative records (for in-patients and out-patients), diagnoses and procedures codes, structured observations, laboratory tests and drug prescriptions written in the hospital, but the CDW also integrates more than 3 million free-text reports in French, ranging from discharge summaries and letters to radiology and pathology reports.

Moreover, HEGP is the national reference centre for ACHD (http://hopital-georgespompidou.aphp.fr/offre-de-soins/centre-de-reference). Therefore, this CDW provides a reference cohort of patients with ACHD, who are treated as both in-patients and out-patients by ACHD specialists in a dedicated unit within the General Cardiology Department.

All free-text medical documents produced for each encounter with the medical team are stored in the CDW as free-text reports (letters for out-patient encounters, discharge summaries for in-patient encounters). Furthermore, for stays in hospital, ICD-10 codes corresponding to the principal and secondary diagnoses are provided for billing purposes, and stored in the CDW. Thus, diagnoses of CHD are potentially present in various forms: administrative diagnoses derived from coded data using ICD-10 for hospital stays only, and clinical diagnoses derived from narrative reports for in-patients and out-patients.

Data extraction from the CDW

We included a random sample of 5500 patients who had at least one hospitalization between 2000 and 2014 in the General Cardiology Department, and a random sample of 500 patients from the ACHD Unit. Motivation for selecting 500 patients from the ACHD Unit was to ensure a substantial prevalence of ACHD in our study population. For every patient included, we extracted all narrative reports produced between 2000 and 2014, including discharge summaries, out-patient reports, radiology reports, multidisciplinary expert meeting summaries, letters and all ICD-10 billing codes related to each hospital stay. The term corpus is used in the rest of the article to reference these 6000

EMRs. The corpus was automatically deidentified [20] to comply with the Institutional Review Board requirements.

The study protocol was approved by the Institutional Review Board of the hospital (IRB#00001072 Study #CDW_2014_0015), and the need for informed consent was waived.

Administrative diagnoses from codified data

Among the 14,400 existing ICD-10 codes, we identified 55 codes related to CHD (Q20–Q28.9). We used these ICD-10 codes to detect CHD from the billing data (Table A.1). Primary and secondary diagnoses were considered. Patients were classified as having ACHD according to administrative data if at least one ICD-10 CHD-related code was present in the EMR, or as not having ACHD if none was present.

Gold standard setting: Clinical diagnoses from narrative data

Data processing

Subset for linguistic resource

We established a list of medical terms used by cardiologists to denote the 55 CHD codes present in the ICD-10. More precisely, a trained ACHD cardiologist (S. C.) developed a list of CHD terms for each of these CHD subtypes, and translated them into regex (a pattern of characters matching specific strings in text). We used a subset of the ACHD Unit records to build this list of regex, following an iterative process. To account for variability in language usage, several regex could be used to target a given concept (e.g. in English, the patterns \btransposition\s+of\s+(the\s)?great\s+arteries and \bTGA\b to match for transposition of the great vessels). More examples are provided in Table 1 and Table A.1 (the latter in French). We decided to define regex that were broad enough to ensure high sensitivity scores, while reducing the burden of manual review. All the reports selected by the regex were then reviewed manually by an ACHD cardiologist.

Entity identification and corpus filtering

We applied the regex-based filtering method to the initial corpus of 6000 EMRs, to identify ACHD (Table 1, Table A.1). At the end of this process, we obtained two subsets: EMRs without any regex match were set aside, while EMRs with at least one match were kept for manual review.

Reviewing the CHD diagnoses

A trained ACHD cardiologist (S. C.) reviewed manually all EMRs in which at least one CHD regex was identified in its free-text reports, and then, if confirmed, extracted the clinical diagnosis (CHD subtype) to determine the reference diagnosis (gold standard). Therefore, a patient was considered as not having ACHD if no regex was recognised in their EMR or if none of the regex matches was confirmed by manual review. To facilitate the manual review, a browser-accessible application – FASTVISU [19] – was used, which highlights terms matching the regex in the text, and provides an interface to efficiently validate the presence/absence of each of the CHD subtypes (Fig. 2). A patient was considered as having ACHD regardless of the presence of surgical repair in their history, and native specific subtypes of CHD were still considered.

Gold standard validation

As shown in Fig. 1, two ACHD cardiologists (S. C. and L. I.) reviewed manually a subset of 2.5% randomly-selected EMRs (31 patients) containing at least one CHD regex, to evaluate the information extraction step performed by FASTVISU. Inter-reviewer agreements on the presence/absence of CHD lesion (primary objective) and on the presence/absence of specific CHD lesions (secondary objective) were estimated using Cohen's kappa.

Evaluation

Administrative diagnoses were compared with the clinical diagnoses extracted from narrative data, considered as the reference diagnosis (gold standard), with the following definitions: true positives (TPs) were ACHD (or CHD subtype) correctly identified by ICD-10 codes; false positives (FPs) were ACHD (or CHD subtype) incorrectly identified by ICD-10, with no confirmation by the narrative data; true negatives (TNs) were no ACHD (or CHD subtype) without ICD-10 CHD codes; false negatives (FNs) were true ACHD (or CHD subtype) according to narrative data, where the ICD-10 CHD code was missing.

We estimated sensitivity (TP/[TP + FN]), specificity (TN/[TN + FP]), NPV (TN/[TN + FN]), PPV (TP/[TP + FP]), accuracy (TP/Total) and their 95% confidence intervals. Each of these values was calculated: to assess the accuracy of administrative data in the detection of ACHD (primary objective); and for each specific CHD subtype (secondary objective). The HEGP CDW was queried using Structured Query Language (Oracle Server).

Results

Corpus

A total of 6000 patients (14.9%) who were hospitalized at least once in the HEGP General Cardiology Department between 2000 and 2014 were included. The subset of 500 records from the ACHD Unit led to an over-representation of this population, as they represented 8.3% of the corpus, whereas the patients from the ACHD Unit represent only 3.1% (1262/40,234) of the General Cardiology Department (Fig. 1). The free-text corpus corresponding to the 6000 patients comprised 107,092 documents, including discharge summaries, out-patient reports, multidisciplinary expert meeting summaries, letters and radiology reports. All records had at least one document. Patients had a median of 11 clinical documents [interquartile range 6; 23], with a maximum of 223.

Extraction of clinical and administrative diagnoses

Starting from the 6000 EMRs, 10,578 CHD regex matches were detected in 6272 documents from 1214 EMRs (Fig. 1). After a manual review of these EMRs, CHD was confirmed for 64% of the EMRs, corresponding to 780 patients (13% of the corpus).

After querying billing data from the corpus, 1122 CHD-related ICD-10 codes were detected in 677 patients out of 6000 (11.3% of the corpus) (Fig. 1). Of these potential patients with ACHD, 59.6% had only one ICD-10 code for CHD, 21.7% had two ICD-10 codes and 18.7% had at least three ICD-10 codes to describe their defect.

For gold standard retrieval validation, 31 patients selected randomly were reviewed to evaluate inter-reviewer agreement. We obtained 100% agreement between the two ACHD cardiologists for the presence/absence of CHD lesion (Cohen's kappa value of 1). The readers disagreed over 10 of the 169 items they viewed from the 31 patients, assigning a different specific CHD lesion. Therefore, the Cohen's kappa value for the secondary objective was 0.88 (95% confidence interval 0.81–0.95).

Accuracy of administrative data in the detection of ACHD

Among the 780 patients with confirmed ACHD (gold standard), 629 (81%) were correctly identified as having ACHD using the ICD-10 codes (sensitivity 0.81, 95% confidence interval 0.78–0.83). One-hundred and fifty-one patients (19%) did not have a CHD code, and thus could not be identified by administrative data. Of the 5220 patients identified as not having ACHD, 48 were wrongly categorized as having ACHD by ICD-10 codes. Table 2 summarizes the accuracy of administrative nomenclature (ICD-10) in the detection of ACHD, and its sensitivity, specificity, NPV and PPV.

Accuracy of administrative data in the identification of specific CHD

The most prevalent CHD lesions according to narrative data (clinical diagnoses) were "atrial septal defect" (Q211) (*n* = 223; 28.6% of the CHD population), "ventricular septal defect" (Q210) (*n* = 169; 21.7%), "bicuspid aortic valve and other congenital insufficiency of aortic valve" (Q231) (*n* = 147; 18.9%) and "tetralogy of Fallot" (Q213) (*n* = 110; 14.1%) (Table 3). Sensitivity, specificity, NPV and PPV of the administrative data (ICD-10 codes) are reported for each defect in Table 3. Sensitivity ranged from 0 to 1 and specificity from 0.99 to 1 (Fig. 3). "Common arterial trunk" (Q200), "congenital malformation of great vein, unspecified" (Q269), "tetralogy of Fallot" (Q213), "transposition of the great vessels" (Q203) and "Ebstein anomaly" (Q225) showed the highest sensitivity, whereas "congenital malformation of aortic and mitral valves, unspecified" (Q239) "other congenital malformations of tricuspid valve" (Q228), "congenital malformation of great arteries, unspecified" (Q259), "congenital malformation of tricuspid valve, unspecified" (Q229), "congenital heart block" (Q246) and "cor triatriatum" (Q242) had a sensitivity value of 0.

Discussion

In this study, we evaluated the performance (sensitivity, specificity, NPV and PPV) of ICD-10 codes in the identification of ACHD in a hospital population with a high proportion of patients with ACHD, and obtained a sensitivity of 0.81 and a specificity of 0.99. To our knowledge, this is the first study to investigate the performance of ICD-10 codes in this field. Frohnert et al. reported similar results (i.e. 86% sensitivity and 98% specificity) in the EMRs of infants using ICD-9-CM (Clinical Modification) codes from hospital discharge data and keywords in medical records at one hospital in Minneapolis,

MN, USA [15]. Although designed for billing, administrative databases coded with ICD codes have been used extensively in research related to CHD epidemiology and health services worldwide [1, 23, 24]. Nowadays, more and more countries have replaced ICD-9 with ICD-10 as the legacy coding system. In France, the French hospital discharge database (Programme de Médicalisation des Systèmes d'Information) contains a record for each acute in-patient stay. Nationwide record production has been mandatory since 1996, but it is limited to in-patient stays. The Programme de Médicalisation des Systèmes d'Information database has already been used in other clinical domains for disease monitoring or epidemiological purposes, particularly in cancer [25-28]. Thus, as no national comprehensive registry exists for ACHD, and given their coverage, administrative health databases may be an interesting tool for conducting ACHD studies on a national scale, and to help to build a French ACHD cohort. Moreover, record linkage in such databases gives access to longitudinal data, which may provide better information about patients lost to follow-up during translation from childhood to adulthood, for example.

Our analysis demonstrated that the performance of ICD-10 codes varied widely according to the CHD lesion, with a sensitivity ranging from 0 to 1. While ICD codes are ubiquitous in clinical practice, given their usage for billing purposes, several authors have shown that they have some limitations in describing accurately the spectrum of CHD. This trend was documented by Frohnert et al; in their study, only 41.2% of ICD-9-CM codes accurately reflected the cardiac defect diagnosed in infants [15]. Also using ICD-9, Cronk et al. found that state administrative databases exactly matched only half of the specific CHD diagnoses [14]. In a recent study of 2193 individuals with a CHD ICD-9 diagnosis code from a tertiary hospital, 1069 were confirmed with a CHD diagnosis by review, yielding overall accuracy of 48.7% [16]. However, when limited to those with moderate or complex lesions, accuracy reached 77% [16]. In our analysis, the highest sensitivity was reached for unambiguous ICD-10 codes, such as “Tetralogy of Fallot” (Q213), “transposition of the great vessels” (Q203) or “Ebstein anomaly” (Q225). On the other hand, unspecified ICD-10 codes, such as “congenital malformation of aortic and mitral valves, unspecified” (Q239) “other congenital malformations of tricuspid valve” (Q228), “congenital malformation of great arteries, unspecified” (Q259) and “congenital malformation of tricuspid valve, unspecified” (Q229), had the lowest sensitivity. This poor agreement between administrative codes and medical records in the classification of CHD may be explained by the lack of granularity of the ICD codes, even though ICD-10 has addressed some of the issues relating to ICD-9.

Since the 1990s, members of the paediatric cardiology and cardiac surgery communities have expressed the need for improved nomenclature [29, 30]. In 2005, the International Paediatric and Congenital Cardiac Code was created, and is now recognized as the standard nomenclature within the field [31], More clinically-oriented coding systems, such as SNOMED CT, are used in several countries, and provide finer-grained codes for CHD. In the future, ICD-11 will be introduced, using the classic segmental sequential approach, linked to the International Paediatric and Congenital Cardiac Code [32], and will also integrate mapping to SNOMED CT [33].

However, even when the relevant code exists, there can be errors, as seen for coarctation of aorta (Q251; sensitivity 0.63), ventricular septal defect (Q210; sensitivity 0.58) or patent ductus arteriosus (Q250; sensitivity 0.50). In our study, reasons for misclassification included: misuse of congenital disease codes instead of acquired disease codes (e.g. acquired aortic valve disease coded as congenital aortic stenosis); non-exhaustiveness of description (e.g. Eisenmenger syndrome [Q218] resulting from a ventricular septal defect [Q210] coded as Eisenmenger syndrome only); misinterpretation of non-specific ICD-10 codes ("other/unspecified" terms); and inherent limitations with the lack of discriminatory detail in the ICD classification scheme. For example, ICD-10 fails to distinguish between atrial septal defect and patent foramen ovale (PFO), as they map to the same code (Q211). Similarly, in their study using ICD-9 codes, Khan et al. reported that the most common error was the erroneous classification of patients with PFO as having an atrial shunt [16] . Errors may occur at each step of the coding process: poorly described information may be given in the medical record; and physicians or administrative personnel in charge of coding may lack expertise in CHD terminology (especially in adult hospitals that are less familiar with CHD coding). In addition, as the primary purpose of these data is billing, coding may be financially driven. All of this may lead to considerable variations in the quality of administrative data in terms of diagnosis [14, 15].

To cope with the challenge of inaccuracy or incompleteness of ICD diagnosis codes, some investigators deployed algorithms, using both structured and unstructured data to find specific phenotypes in EMRs [34-36]. Outside the scope of CHD, Wei et al. evaluated the phenotyping performance of three major components of EMRs and their combination: billing codes (ICD-9); primary notes; and medications [37]. By working on a broad spectrum of phenotypes (atrial fibrillation, Alzheimer's disease, breast cancer, gout, human immunodeficiency virus infection, multiple sclerosis, Parkinson's disease, rheumatoid arthritis and diabetes mellitus), they demonstrated that combining the

three components results in superior phenotyping performance [37]. Much of the research into ACHD has relied largely upon administrative codes and, to our knowledge, none had the benefit of using additional EMR unstructured information. However, some initiatives are aimed at addressing these issues, such as merging multiple data sources. In Québec, Marelli et al. developed an algorithm to identify patients with CHD by using all available data for a given subject, including in-patient, out-patient, surgical procedural act and provider information [1]. Similarly, the addition of other information that is likely to be available in administrative data sets, such as age or encounter type, improved the accuracy of the data for CHD determination [16].

Clinical notes entered by physicians are valuable sources of patient information. In chronic disease such as CHD, repeated consultations are required, leading to an increasing amount of medical reports per patient. Therefore, NLP methods that automatically extract information from text are needed [38]. Shivade et al. reviewed 97 articles describing approaches to identifying patient phenotype cohorts using EMRs, of which 46 used an NLP-based approach [36]. Although used in general cardiology, particularly in the field of heart failure [39], NLP techniques have not been used for CHD documents. Unlike international classifications, such as ICD, which can benefit from shared algorithms, NLP techniques must be adapted to each language, and expert validation is required. Applications such as FASTVISU can be used to assist experts in validating the information that has been extracted automatically from text, and thus save them time in this task [19].

Clinical practice is nowadays largely supported by information technology. The widespread adoption of EMRs in hospitals enables large-scale secondary use of the data for research purposes. EMRs hold great potential in the context of ACHD, which is a relatively rare disease. Indeed, it provides an opportunity to aggregate longitudinal data and to increase sample size compared with transversal recruitment, which has a narrower window span. Selection of eligible patients for a clinical study is a tedious and costly task that would benefit from CDWs in the ACHD field [40-42]. Data extraction from EMRs is also a means of tracking quality indicators that are specific to the ACHD population [43]. Finally, in the USA, there have been several initiatives to create data integration models across other fields. PEDSnet is a clinical data research network in paediatrics, which harmonized data captured through the EMR systems of members using a common terminology, and uses open-source software to support data submission and aggregation [44]. The Cardiovascular Research Network consists of 15 geographically distributed healthcare delivery systems. Within this

network, data captured through the EMRs of each healthcare system are standardized across the virtual data warehouse at each site, with common data elements, naming conventions and definitions, to facilitate combining information in aggregate analyses. Then they are linked to multiple other electronic databases to conduct large-scale adult cardiovascular research more efficiently, including epidemiological studies, outcomes research, comparative-effectiveness studies and clinical trials [45]. These models may be useful examples to consider even more in the current era of harmonization of practices. The secondary use of EMRs opens new perspectives, and our ACHD community needs to fully leverage available and emerging data sources to support important investigations and conduct research most efficiently.

Study limitations

The generalizability of the study results regarding accuracy of ICD-10 codes obtained from discharge data is limited by the monocentric feature. However, ICD-10 is largely used worldwide as a legacy coding system for billing, thus the results may be extrapolated to other hospitals using the same coding system for the same purposes. This study included a high proportion of patients with ACHD (roughly 13%), much higher than in the general population or general hospitals. Our institution is the national reference centre for ACHD, and we assume that the staff – even non-CHD staff – are used to CHD terminology. Coding by providers less familiar with ACHD diagnoses may result in decreased coding accuracy and broader use of non-specific codes (such as Q223, Q228, Q229, Q239, Q248, Q254, Q259, Q264, Q268, Q269, Q278). Furthermore, in non-tertiary care centres, the proportion of less complex CHD subtypes, such as bicuspid aortic valve or PFO, is expected to be higher. In fact, atrial septal defect and PFO share the same ICD-10 code (Q211). But PFO is more common than all other CHDs combined, so it may overestimate the prevalence of the CHD. Bicuspid aortic valve (Q231) is another frequent lesion, and it may be inaccurately coded as an acquired disease, as it is less commonly a clinically relevant issue. Therefore, our results should be interpreted with caution, and additional validation work is needed.

Conclusions

Administrative data using ICD-10 codes is a useful tool for detecting ACHD, and may be used to establish a national cohort. However, the lack of accuracy in describing the spectrum of CHD may

affect the ability to precisely describe the CHD populations in terms of CHD subtypes. On the other hand, secondary use of EMRs and text-mining methods offer new opportunities to provide accurate and reliable information. While efforts are made to create a successful multicentre collaborative programme in ACHD, and to harmonize clinical data across centres, assessment of existing resources is essential. Thus, combining administrative data and free text may tend to enhance performance and open up new horizons of research and improvement, such as exploring lifelong co-morbidities, assessing treatment efficacy or implementing clinical decision support. Creative solutions are needed to link diverse systems, with the aim of leveraging the treasure trove of information from EMRs for clinical and epidemiological research.

Sources of funding

Dr. Sarah Cohen was supported by the Fondation pour la Recherche Médicale and AREMCAR (Hélène de Marsan Grant).

Disclosure of interest

The authors declare that they have no conflicts of interest concerning this article.

References

[1] Marelli AJ, Mackie AS, Ionescu-Ittu R, Rahme E, Pilote L. Congenital heart disease in the general population: changing prevalence and age distribution. Circulation 2007;115:163-72.

[2] Khairy P, Ionescu-Ittu R, Mackie AS, Abrahamowicz M, Pilote L, Marelli AJ. Changing mortality in congenital heart disease. J Am Coll Cardiol 2010;56:1149-57.

[3] van der Linde D, Konings EE, Slager MA, et al. Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. J Am Coll Cardiol 2011;58:2241-7.

[4] Marelli AJ, Ionescu-Ittu R, Mackie AS, Guo L, Dendukuri N, Kaouache M. Lifetime prevalence of congenital heart disease in the general population from 2000 to 2010. Circulation 2014;130:749-56.

[5] Mackie AS, Pilote L, Ionescu-Ittu R, Rahme E, Marelli AJ. Health care resource utilization in adults with congenital heart disease. Am J Cardiol 2007;99:839-43.

[6] Baumgartner H, Bonhoeffer P, De Groot NM, et al. ESC Guidelines for the management of grown-up congenital heart disease (new version 2010). Eur Heart J 2010;31:2915-57.

[7] Warnes CA, Williams RG, Bashore TM, et al. ACC/AHA 2008 Guidelines for the Management of Adults with Congenital Heart Disease: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (writing committee to develop guidelines on the management of adults with congenital heart disease). Circulation 2008;118:e714-833.

[8] Gurvitz M, Marelli A, Mangione-Smith R, Jenkins K. Building quality indicators to improve care for adults with congenital heart disease. J Am Coll Cardiol 2013;62:2244-53.

[9] Kotowycz MA, Therrien J, Ionescu-Ittu R, et al. Long-term outcomes after surgical versus transcatheter closure of atrial septal defects in adults. JACC Cardiovasc Interv 2013;6:497-503.

[10] Bouchardy J, Therrien J, Pilote L, et al. Atrial arrhythmias in adults with congenital heart disease. Circulation 2009;120:1679-86.

[11] Lanz J, Brophy JM, Therrien J, Kaouache M, Guo L, Marelli AJ. Stroke in Adults With Congenital Heart Disease: Incidence, Cumulative Risk, and Predictors. Circulation 2015;132:2385-94.

[12] Cedars A, Benjamin L, Burns SV, Novak E, Amin A. Clinical predictors of length of stay in adults with congenital heart disease. Heart 2017.

[13] O'Leary JM, Siddiqi OK, de Ferranti S, Landzberg MJ, Opotowsky AR. The Changing Demographics of Congenital Heart Disease Hospitalizations in the United States, 1998 Through 2010. JAMA 2013;309:984-6.

[14] Cronk CE, Malloy ME, Pelech AN, et al. Completeness of state administrative databases for surveillance of congenital heart disease. Birth Defects Res A Clin Mol Teratol 2003;67:597-603.

[15] Frohnert BK, Lussky RC, Alms MA, Mendelsohn NJ, Symonik DM, Falken MC. Validity of hospital discharge data for identifying infants with cardiac defects. J Perinatol 2005;25:737-42.

[16] Khan A, Ramsey K, Ballard C, et al. Limited Accuracy of Administrative Data for the Identification and Classification of Adult Congenital Heart Disease. J Am Heart Assoc 2018;7.

[17] Wright A, Henkin S, Feblowitz J, McCoy AB, Bates DW, Sittig DF. Early results of the meaningful use program for electronic health records. N Engl J Med 2013;368:779-80.

[18] Escudie JB, Rance B, Malamut G, et al. A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease. BMC Med Inform Decis Mak 2017;17:140.

[19] Escudie JB, Jannot AS, Zapletal E, et al. Reviewing 741 patients records in two hours with FASTVISU. AMIA Annu Symp Proc 2015;2015:553-9.

[20] Zapletal E, Rodon N, Grabar N, Degoulet P. Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. Stud Health Technol Inform 2010;160:193-7.

[21] Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17:124-30.

[22] Jannot AS, Zapletal E, Avillach P, Mamzer MF, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. Int J Med Inform 2017;102:21-8.

[23] Riehle-Colarusso TJ, Bergersen L, Broberg CS, et al. Databases for Congenital Heart Defect Public Health Studies Across the Lifespan. J Am Heart Assoc 2016;5.

[24] Opotowsky AR, Siddiqi OK, Webb GD. Trends in hospitalizations for adults with congenital heart disease in the U.S. J Am Coll Cardiol 2009;54:460-7.

[25] Boudemaghe T, Belhadj I. Data Resource Profile: The French National Uniform Hospital Discharge Data Set Database (PMSI). Int J Epidemiol 2017.

[26] Remontet L, Mitton N, Couris CM, et al. Is it possible to estimate the incidence of breast cancer from medico-administrative databases? Eur J Epidemiol 2008;23:681-8.

[27] Maura G, Blotiere PO, Bouillon K, et al. Comparison of the short-term risk of bleeding and arterial thromboembolic events in nonvalvular atrial fibrillation patients newly treated with dabigatran or rivaroxaban versus vitamin K antagonists: a French nationwide propensity-matched cohort study. Circulation 2015;132:1252-60.

[28] Weill A, Paita M, Tuppin P, et al. Benfluorex and valvular heart disease: a cohort study of a million people with diabetes mellitus. Pharmacoepidemiol Drug Saf 2010;19:1256-62.

[29] Mavroudis C, Jacobs JP. Congenital Heart Surgery Nomenclature and Database Project: overview and minimum dataset. Ann Thorac Surg 2000;69:S2-17.

[30] Franklin RC, Jacobs JP, Krogmann ON, et al. Nomenclature for congenital and paediatric cardiac disease: historical perspectives and The International Pediatric and Congenital Cardiac Code. Cardiol Young 2008;18 Suppl 2:70-80.

[31] Pasquali SK, Jacobs JP, Farber GK, et al. Report of the National Heart, Lung, and Blood Institute Working Group: An Integrated Network for Congenital Heart Disease Research. Circulation 2016;133:1410-8.

[32] Houyel L, Khoshnood B, Anderson RH, et al. Population-based evaluation of a suggested anatomic and clinical classification of congenital heart defects based on the International Paediatric and Congenital Cardiac Code. Orphanet J Rare Dis 2011;6:64.

[33] Rodrigues JM, Robinson D, Della Mea V, et al. Semantic Alignment between ICD-11 and SNOMED CT. Stud Health Technol Inform 2015;216:790-4.

[34] Teixeira PL, Wei WQ, Cronin RM, et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. J Am Med Inform Assoc 2017;24:162-71.

[35] Anderson AE, Kerr WT, Thames A, Li T, Xiao J, Cohen MS. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States

population: A cross-sectional, unselected, retrospective study. J Biomed Inform 2016;60:162-8.

[36] Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21:221-30.

[37] Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016;23:e20-7.

[38] Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011;306:848-55.

[39] Garvin JH, DuVall SL, South BR, et al. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. J Am Med Inform Assoc 2012;19:859-66.

[40] Khairy P. Looking ahead: clinical trial design in adult congenital heart disease. Future Cardiol 2012;8:297-304.

[41] Gurvitz M, Burns KM, Brindis R, et al. Emerging Research Directions in Adult Congenital Heart Disease: A Report From an NHLBI/ACHA Working Group. J Am Coll Cardiol 2016;67:1956-64.

[42] Khairy P, Hosn JA, Broberg C, et al. Multicenter research in adult congenital heart disease. Int J Cardiol 2008;129:155-9.

[43] Broberg C, Sklenar J, Burchill L, Daniels C, Marelli A, Gurvitz M. Feasibility of Using Electronic Medical Record Data for Tracking Quality Indicators in Adults with Congenital Heart Disease. Congenit Heart Dis 2015;10:E268-77.

[44] Forrest CB, Margolis PA, Bailey LC, et al. PEDSnet: a National Pediatric Learning Health System. J Am Med Inform Assoc 2014;21:602-6.

[45] 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.

Figure legends

Figure 1. Overview of the methodology. ACHD: adults with congenital heart disease; CHD: congenital heart disease; EMR: electronic medical record; HEGP: Georges Pompidou European Hospital; ICD-10: International Classification of Diseases (10th revision); PPV: positive predictive value.

Figure 2. Screenshot of manual review using FASTVISU. FASTVISU presents all documents from a patient in chronological order in the left screen panel. Keywords and expressions detected by the regular expressions (regex) modules are highlighted. These highlighted concepts are also summarized on the right screen panel of the application, and provided with a clickable link to the corresponding occurrence in the text. A “yes” or “no” vote for the presence or the absence of a congenital heart disease lesion is also available on the right screen panel.

Figure 3. Sensitivity and specificity value for each International Classification of Diseases (10th revision) code. CHD: congenital heart disease.

Table 1 Examples (translated into English) of congenital heart disease lesions with their associated patterns of regular expressions for entity recognition and their corresponding International Classification of Diseases (10th revision) codes.[a]

CHD lesion	Regular expressions	ICD-10 code
Transposition of the great vessels	\btransposition\s+of\s+(the\s)?great\s+arteries	Q203
	\bTGA\b	
Ventricular septal defect	\bventricular\s+septal\s+defect\b	Q210
	\bvsd\w*\b	
	\bvsd\s*([IViv123]+[ab]?)\b	
Tetralogy of Fallot	\bFallot\b	Q213
Atrial septal defect	\batrial\s+septal\s+defect\b	Q211
	\basd\w*\b	
	\bostium\s+secondum\s+[defect]?\b	
	\bforamen\s+ovale\b	
	\bpfo\b	
	\bsinus\s+ve[i]?nosus\b	
Atrioventricular septal defect	\bcommon\s+atrioventricular\s+canal\b	Q212
	\batrioventricular\s+septal\s+defect\	
	\batrioventricular\s+canal\s+defect\b	
	\bavsd\b	

	\bavcd\b	
	\bav\s+defect\b	
	\bendocardial\s+cushion\s+defect\b	
Ostium primum atrial septal defect	\bostium\s+primum(\s+atrial\s+septal\s+defect)?\b	Q212
	\bostium\s+primum\s+a(trial)?(\s)?s(eptal)?(\s)+d(efect)?\b	
Coarctation of aorta	\bcoarctation\s+of\s+(the\s)?aorta\b	Q251
	\baortic\s+coarctation\b	
	\bcoarctation\b	

CHD: congenital heart disease; ICD-10: International Classification of Diseases (10th revision).

[a] The comprehensive list of CHD lesions, regular expressions and ICD-10 codes is available in Table A.1 (in French).

Table 2. Sensitivity, specificity and positive and negative predictive values for International Classification of Diseases (10th revision) codes in the identification of adult congenital heart disease.

From administrative data (ICD-10 codes)	From narrative data (clinical diagnosis confirmed by expert)			
	ACHD	No ACHD	Total	
ACHD	629 (TP)	48 (FP)	677	PPV = 0.93 (0.91–0.95)[a]
No ACHD	151 (FN)	5172 (TN)	5323	NPV = 0.97 (0.96–0.98)[a]
Total	780	5220	6000	
	Sens = 0.81 (0.78–0.83)[a]	Spe = 0.99 (0.99–1)[a]		Acc = 0.97 (0.96–0.98)[a]

Acc: accuracy; ACHD: adult with congenital heart disease; FN: false negative; FP: false positive; ICD-10: International Classification of Diseases (10th revision) ; NPV: negative predictive value; PPV: positive predictive value; Sens: sensitivity; Spe: specificity; TN: true negative; TP: true positive.

[a] Numbers in brackets are 95% confidence intervals.

Table 3 Sensitivity, specificity and positive and negative predictive values for International Classification of Diseases (10th revision) codes in the identification of specific congenital heart disease lesions[a].

ICD-10 code	Lesion	Clinical diagnoses[b]	Sensitivity	Specificity	PPV	NPV
	All CHDs	780	0.81	0.99	0.93	0.97
Q211	Atrial septal defect	223 (28.6)	0.54	0.99	0.79	0.98
Q210	Ventricular septal defect	169 (21.7)	0.58	1.00	0.84	0.99
Q231	Congenital insufficiency of aortic valve; bicuspid aortic valve	147 (18.9)	0.36	1.00	0.82	0.98
Q213	Tetralogy of Fallot	110 (14.1)	0.92	1.00	0.80	1.00
Q221	Congenital pulmonary valve stenosis	90 (11.5)	0.49	1.00	0.70	0.99
Q203	Transposition of the great vessels; discordant ventriculoarterial connection	83 (10.6)	0.87	1.00	0.73	1.00
Q254	Other congenital malformations of aorta: congenital aneurysm or dilatation of aorta; aneurysm of sinus of Valsalva	68 (8.7)	0.13	1.00	0.45	0.99
Q204	Single ventricle	58 (7.4)	0.66	1.00	0.84	1.00
Q251	Coarctation of aorta	52 (6.7)	0.63	1.00	0.92	1.00
Q261	Persistent left superior vena cava	44 (5.6)	0.05	1.00	1.00	0.99
Q241	Laevocardia	43 (5.5)	0.09	1.00	1.00	0.99
Q220	Pulmonary valve atresia	37 (4.7)	0.78	1.00	0.81	1.00

Q243	Pulmonary infundibular stenosis	37 (4.7)	0.05	1.00	1.00	0.99
Q212	Atrioventricular septal defect	34 (4.4)	0.79	1.00	0.75	1.00
Q240	Dextrocardia	28 (3.6)	0.14	1.00	1.00	1.00
Q230	Congenital aortic atresia; congenital aortic stenosis	27 (3.5)	0.22	1.00	0.60	1.00
Q268	Other congenital malformations of great veins: absence of (inferior or superior) vena cava; azygos continuation of inferior vena cava; persistent left posterior cardinal vein; Scimitar syndrome	27 (3.5)	0.19	1.00	0.71	1.00
Q239	Congenital malformation of aortic and mitral valves, unspecified	26 (3.3)	0.00	1.00	0.00	1.00
Q250	Patent ductus arteriosus	26 (3.3)	0.50	1.00	0.68	1.00
Q264	Anomalous pulmonary venous connection, unspecified	26 (3.3)	0.27	1.00	0.58	1.00
Q238	Other congenital malformations of aortic and mitral valves	25 (3.2)	0.08	1.00	0.50	1.00
Q245	Malformation of coronary vessels; congenital coronary (artery) aneurysm	25 (3.2)	0.52	1.00	0.76	1.00
Q205	Discordant atrioventricular connection; corrected transposition	24 (3.1)	0.79	1.00	0.76	1.00
Q218	Other congenital malformations of cardiac septa; Eisenmenger defect	23 (2.9)	0.43	1.00	0.50	1.00
Q263	Partial anomalous pulmonary venous connection	23 (2.9)	0.43	1.00	0.91	1.00
Q224	Congenital tricuspid stenosis; tricuspid atresia	22 (2.8)	0.59	1.00	0.81	1.00
Q244	Congenital subaortic stenosis	22 (2.8)	0.27	1.00	0.43	1.00
Q201	Double outlet right ventricle	21 (2.7)	0.24	1.00	0.56	1.00

Q225	Ebstein anomaly	20 (2.6)	0.85	1.00	0.94	1.00
Q232	Congenital mitral stenosis; congenital mitral atresia	20 (2.6)	0.30	1.00	0.60	1.00
Q233	Congenital mitral insufficiency	20 (2.6)	0.15	1.00	0.25	1.00
Q223	Other congenital malformations of pulmonary valve	14 (1.8)	0.21	1.00	0.18	1.00
Q228	Other congenital malformations of tricuspid valve: dysplastic tricuspid valve, straddling, overriding	14 (1.8)	0.00	1.00	0.00	1.00
Q229	Congenital malformation of tricuspid valve, unspecified	14 (1.8)	0.00	1.00	0.00	1.00
Q256	Stenosis of pulmonary artery; supravalvular pulmonary stenosis	14 (1.8)	0.07	1.00	0.08	1.00
Q259	Congenital malformation of great arteries, unspecified	14 (1.8)	0.00	1.00	0.00	1.00
Q234	Hypoplastic left heart syndrome; mitral valve atresia	12 (1.5)	0.33	1.00	1.00	1.00
Q206	Isomerism	11 (1.4)	0.18	1.00	0.67	1.00
Q226	Hypoplastic right heart syndrome	11 (1.4)	0.09	1.00	0.17	1.00
Q200	Common arterial trunk	8 (1.0)	1.00	1.00	1.00	1.00
Q257	Other congenital malformations of pulmonary artery: aberrant pulmonary artery, agenesis or aneurysm or hypoplasia of pulmonary artery; pulmonary arteriovenous aneurysm	8 (1.0)	0.13	1.00	0.07	1.00
Q255	Atresia of pulmonary artery	6 (0.8)	0.17	1.00	0.33	1.00
Q262	Total anomalous pulmonary venous connection	4 (0.5)	0.25	1.00	1.00	1.00

Q202	Double outlet left ventricle	3 (0.4)	0.67	1.00	0.50	1.00
Q208	Other congenital malformations of cardiac chambers and connections	3 (0.4)	0.33	1.00	0.17	1.00
Q253	Stenosis of aorta; supravalvular aortic stenosis	3 (0.4)	0.33	1.00	0.17	1.00
Q242	Cor triatriatum	1 (0.1)	0.00	1.00	-	1.00
Q246	Congenital heart block	1 (0.1)	0.00	1.00	0.00	1.00
Q269	Congenital malformation of great vein, unspecified	1 (0.1)	1.00	1.00	1.00	1.00
Q214	Aortopulmonary septal defect	0	-	-	-	-
Q222	Congenital pulmonary valve insufficiency	0	-	-	-	-
Q248	Other specified congenital malformations of heart; diverticulum of left ventricle; Uhl's disease	0	-	-	-	-
Q252	Atresia of aorta	0	-	-	-	-
Q260	Congenital stenosis of vena cava	0	-	-	-	-
Q278	Other specified congenital malformations of peripheral vascular system	0	-	-	-	-

ACHD: adult with congenital heart disease; CHD, congenital heart disease; ICD-10, International Classification of Diseases (10th revision); NPV: negative predictive value; PPV: positive predictive value.

[a] Lesions are ranked in decreasing order of number of cases.

[b] Data are expressed as number (%).

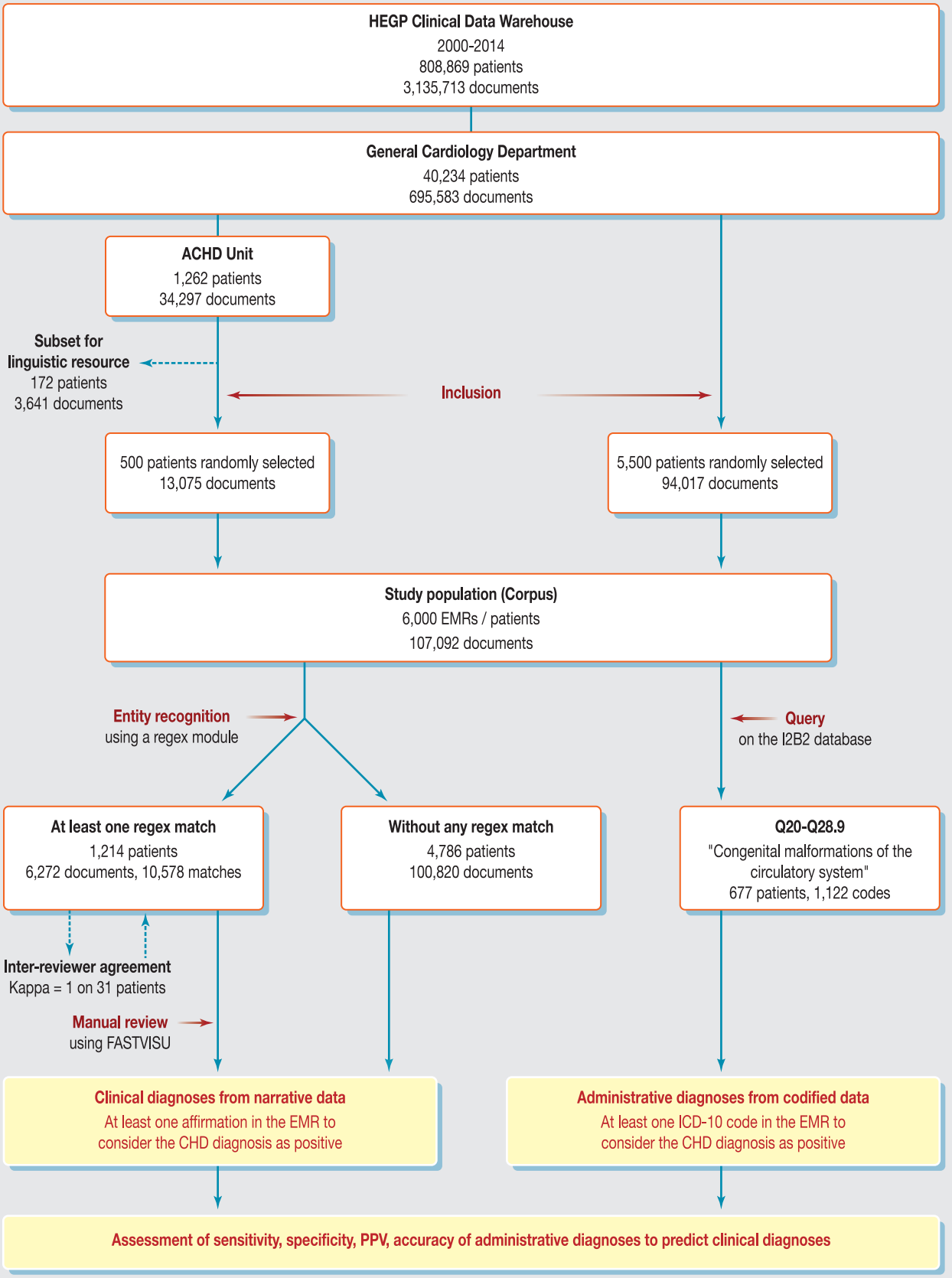
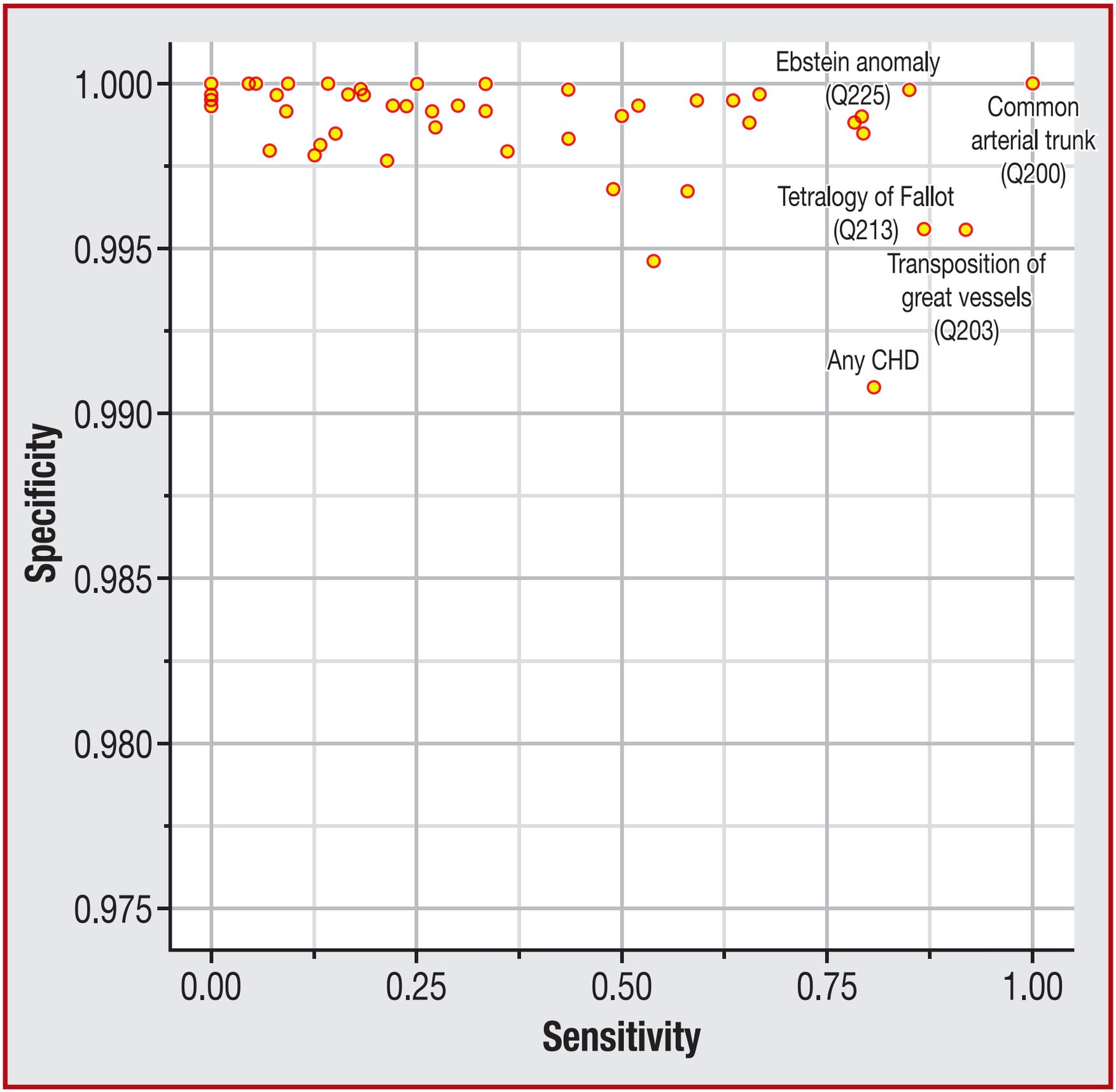

HEGP Clinical Data Warehouse
2000-2014
808,869 patients
3,135,713 documents
General Cardiology Department
40,234 patients
695,583 documents
ACHD Unit
1,262 patients
34,297 documents
Subset for linguistic resource
172 patients
3,641 documents
Inclusion
500 patients randomly selected
13,075 documents
5,500 patients randomly selected
94,017 documents
Study population (Corpus)
6,000 EMRs / patients
107,092 documents
Entity recognition
using a regex module
Query
on the I2B2 database
At least one regex match
1,214 patients
6,272 documents, 10,578 matches
Without any regex match
4,786 patients
100,820 documents
Q20-Q28.9
"Congenital malformations of the circulatory system"
677 patients, 1,122 codes
Inter-reviewer agreement
Kappa = 1 on 31 patients
Manual review
using FASTVISU
Clinical diagnoses from narrative data
At least one affirmation in the EMR to consider the CHD diagnosis as positive
Administrative diagnoses from codified data
At least one ICD-10 code in the EMR to consider the CHD diagnosis as positive
Assessment of sensitivity, specificity, PPV, accuracy of administrative diagnoses to predict clinical diagnoses

Figure 2.


Select a serie
SC_CC_CC_tranche003 - 2.cardiopathi
Select a patient
11) 2004886071 (14)
Select a regex catalog
2.cardiopathies congénitales v(
R
Select a poll
SC_CC
Select allowed vote values
OUI/NON
Select a voter
DG1 (2522)
Show categories without match
<<
R
>>
- VOTE -
COMPTE RENDU D'USIC
Relevant aggregation level
NOM : NOM
PRENOM : PRENOM
Née le : DATENAIS
Hospitalisation du 16/06/2004 au 18/06/2004 en USIC puis jusqu'au 23/06/2004 en Cardio1
Fait le 16 juin 2004 Interne : Dr XXXXXX, Dr XXXXXX
REF : NS CCA/PH : Dr XXXXXX, Dr XXXXXX
Destinataires :
Dr XXXXXX. Cardiologue. Cardiologie pédiatrique. Necker
MOTIF D'HOSPITALISATION
Patiente âgée de 62 ans hospitalisée pour malaise et dyspnée
ANTECEDENTS
Filtering by text processing and highlighting regex
Cardiovasculaires
CAV avec HTAP fixée (syndrome d'Eisenmenger)
Médicaux
HTA paroxystique
Maladie d'alzheimer
Episodes d'hémoptysie en septembre 2001 et avril 2003
Oxygène à domicile
Cyanose chronique
Chirurgicaux
Kystectomie ovarienne à 18 ans
FACTEURS DE RISQUES CARDIO-VASCULAIRES
Hypertension Artérielle , sous bithérapie, mal équilibrée (Sectral et amlor)
Hérédité Coronarienne (mère hypertendue, IDM à 55 ans)
Pas de Tabagisme
Pas de Surcharge pondérale (Poids 50Kg Taille 164cm BMI=18)
MODE DE VIE
Q218 - syndrome d.eisenmenger
Oui
Non
Validation of the diagnosis by a voting system (physician chart review)
Q221 - stenose congenitale de la valve pulmonaire
Oui
Non
Q222 - Insuffisance congénitale de la valve pulmonaire
Oui
Non


Ebstein anomaly
(Q225)
Common
arterial trunk
(Q200)
Tetralogy of Fallot
(Q213)
Transposition of
great vessels
(Q203)
Any CHD
1.000
0.995
0.990
0.985
0.980
0.975
Specificity
0.00
0.25
0.50
0.75
1.00
Sensitivity