Journal of Biomedical Informatics 72 (2017) 77–84

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Text mining applied to electronic cardiovascular procedure reports to identify patients with trileaflet aortic stenosis and coronary artery disease

Aeron M. Small [a,1], Daniel H. Kiss [a,1], Yevgeny Zlatsin [b], David L. Birtwell [c], Heather Williams [c], Marie A. Guerraty [a], Yuchi Han [a], Saif Anwaruddin [a], John H. Holmes [b], Julio A. Chirinos [a], Robert L. Wilensky [a], Jay Giri [a], Daniel J. Rader [a,c,d,*]

[a] Department of Medicine and Cardiovascular Institute, University of Pennsylvania Perelman School of Medicine, PA, USA
[b] Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA
[c] Institute for Translational Medicine and Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[d] Department of Genetics, University of Pennsylvania Perelman School of Medicine, PA, USA

ARTICLE INFO

Article history:
Received 16 March 2017
Revised 21 May 2017
Accepted 12 June 2017
Available online 15 June 2017

Keywords:
Valvular heart disease
Coronary artery disease
Text mining
Administrative
Billing codes

ABSTRACT

Background: Interrogation of the electronic health record (EHR) using billing codes as a surrogate for diagnoses of interest has been widely used for clinical research. However, the accuracy of this methodology is variable, as it reflects billing codes rather than severity of disease, and depends on the disease and the accuracy of the coding practitioner. Systematic application of text mining to the EHR has had variable success for the detection of cardiovascular phenotypes. We hypothesize that the application of text mining algorithms to cardiovascular procedure reports may be a superior method to identify patients with cardiovascular conditions of interest.

Methods: We adapted the Oracle product Endeca, which utilizes text mining to identify terms of interest from a NoSQL-like database, for purposes of searching cardiovascular procedure reports and termed the tool "PennSeek". We imported 282,569 echocardiography reports representing 81,164 individuals and 27,205 cardiac catheterization reports representing 14,567 individuals from non-searchable databases into PennSeek. We then applied clinical criteria to these reports in PennSeek to identify patients with trileaflet aortic stenosis (TAS) and coronary artery disease (CAD). Accuracy of patient identification by text mining through PennSeek was compared with ICD-9 billing codes.

Results: Text mining identified 7115 patients with TAS and 9247 patients with CAD. ICD-9 codes identified 8272 patients with TAS and 6913 patients with CAD. 4346 patients with AS and 6024 patients with CAD were identified by both approaches. A randomly selected sample of 200–250 patients uniquely identified by text mining was compared with 200–250 patients uniquely identified by billing codes for both diseases. We demonstrate that text mining was superior, with a positive predictive value (PPV) of 0.95 compared to 0.53 by ICD-9 for TAS, and a PPV of 0.97 compared to 0.86 for CAD.

Conclusion: These results highlight the superiority of text mining algorithms applied to electronic cardiovascular procedure reports in the identification of phenotypes of interest for cardiovascular research.

1. Introduction

Electronic healthcare data are increasingly used for purposes of clinical and translational research [1,2]. Common sources for these types of data include administrative databases, which contain insurance or claims information for various procedures, diagnoses, or medications, and electronic health records (EHRs), which contain the clinical text and discrete measurements for inpatient and outpatient medical encounters [3]. The large number of patients in these databases often precludes a complete manual chart review: thus billing codes, commonly from the *International Classification of Diseases, 9th Revision, Clinical Modification* (ICD-9-CM), are used as a surrogate for phenotypes of interest [4]. However, the efficacy of this method is highly variable, depending

* Corresponding author at: Perelman School of Medicine, University of Pennsylvania, 11-125 Smilow Center for Translational Research, 3400 Civic Center Blvd, Philadelphia, PA 19104-5158, USA.
E-mail address: rader@mail.med.upenn.edu (D.J. Rader).
[1] These authors have contributed equally.

http://dx.doi.org/10.1016/j.jbi.2017.06.016
1532-0464/

greatly on the disease of interest and accuracy of the coding practitioner [5].

To improve the accuracy of phenotypes built from EHR and administrative databases, a variety of methods have been employed that combine available data using computational techniques and text mining applications. Text mining, a linguistic domain in computer science that ranges in function from free text searches to natural language processing (NLP), enables the automatic parsing of clinical free text to identify informative clinical concepts [6], and has been previously utilized to help answer a variety of clinical and cardiovascular research questions [7,8]. It is recognized that the addition of text mining and NLP have value over claims data alone in accurately identifying clinical phenotypes [9,10].

A frequent challenge encountered in clinical text mining tasks is managing the heterogenous nature of unstructured narrative text in the clinical record. The accurate detection of disease status from clinical text requires a nuanced understanding of patterns and key phrases in a subject's medical history, which can vary widely. A history of coronary artery disease (CAD), for example, may be hinted at by the presence of a cardiomyopathy, a history of an acute myocardial infarction, or a positive cardiac stress test. In 2014, the informatics for integrating biology and the bedside project (i2b2) challenged teams to determine risk factors related to CAD in diabetic patients using a combination of clinical free text and claims data. The CAD specific F1 scores, which were determined using expert adjudicated reports, did no better than 0.83 [11]. The authors point out that a likely contributing factor to this relatively low score was the wide variety of ways CAD indicators are described in the text.

In contrast to the highly heterogenous, free-form text present in patient notes and clinical correspondence, cardiovascular procedure report text, for example echocardiography or cardiac catheterization text, represents a highly stereotyped alternative which has been demonstrated to be easily parsed using text mining applications [12,13]. Additionally, cardiovascular procedure report text represents the gold standard in diagnosis for a number of cardiovascular conditions, including CAD. Despite its potential to improve the data quality in cardiovascular observational research, there has thus far been minimal adoption of text mining in this arena.

Trileaflet aortic stenosis (TAS) and CAD are two common cardiovascular disorders, which are highly prevalent in the United States. Clinical studies on both entities have used diagnostic billing codes as surrogate diagnoses for these conditions. Both TAS and CAD are optimally diagnosed by cardiovascular procedures commonly recorded electronically: cardiac echocardiography for TAS and coronary angiography for CAD [14]. Using customized software which allows for keyword searches of clinical free text, we sought to ascertain whether text mining algorithms applied to cardiovascular procedure reports can accurately identify individuals with these disease processes, and additionally whether text mining alone is superior to ICD-9 based diagnostic code definitions previously used in the literature to identify individuals with these disorders.

2. Materials and methods

2.1. PennSeek

PennSeek is a custom implementation of Oracle's Endeca Information Discovery platform. Originally, this tool was developed for e-commerce applications, such as online shopping websites that require near-immediate response to users' search queries for retail products. We adapted the Endeca toolset to build an enhanced shopping-like application focused on clinical research, in the process branding it "PennSeek". By virtue of Endeca being web-based, it facilitates easy access through any modern browser or mobile device.

From the technical perspective, PennSeek allows for targeted keyword searches of unstructured or semi-structured medical documents currently residing in the Penn Medicine main EHR and ancillary systems. These keywords can range from simple constructs that any user can access without prior training, to complex programming that is manipulated by developers and data engineers. One of the challenges PennSeek solved was the difficulty of flexibly applying search algorithms against data sourced from multiple clinical systems at different levels of granularity. For example, patient registration systems may express the grain at the patient level, EHR systems may express the grain at the encounter level, and lab systems may express the grain at the order level. Without knowing in advance how a researcher intends to query these systems, an analytics engineer would have to either think of and design for every variation or limit the researcher to specific queries. PennSeek takes a different approach by allowing each clinical source system to retain its own grain instead of predefining a common grain across all. Each clinical source system's data is first stored and indexed in a NoSQL-like database. Overlays are then added to this dataset that allow segmentation by clinical category (e.g. encounters, medications, labs, etc.), by grain (e.g. patient, encounter, physician, etc.), or by linguistic components (e.g. word, sentence, etc.), adding structure to otherwise disorganized data. The database indexes nearly every attribute of a patient's discrete and unstructured clinical text, including individual words and phrases. This extensive indexing, combined with the ability to add overlays to the database in an in-memory architecture, allows PennSeek to query and retrieve data from across data sources orders of magnitude more quickly and flexibly than what is traditionally possible using relational databases and search functions built into clinical applications. This underlying architecture was key in supporting the prototyping, development, and testing of the text mining algorithms that were ultimately used to generate positive identifiers for TAS and CAD. An example of a keyphrase query for CAD in the PennSeek application is presented in Fig. 1.

2.2. Data sources/patient population

The study population for TAS included all 62,703 patients identified having at least one echocardiogram recorded in the Hospital of the University of Pennsylvania (HUP) electronic echocardiogram database (PROSOLV Cardiovascular, Indianapolis, IN) between January 2009 and October 2015. The study population for CAD included all 14,567 patients identified having at least one coronary angiogram report recorded in the HUP electronic catheterization database (McKesson Inc., San Francisco, CA) between January 2007 and October 2015. Echocardiography reports in PROSOLV and catheterization reports in McKesson were imported into PennSeek in order to facilitate text mining.

2.3. Classification of patients with trileaflet aortic stenosis

A diagnosis of TAS was determined by careful adjudication of the text summary accompanying echocardiography reports, which routinely includes a statement on the presence and degree of aortic valve stenosis. It was decided that the echocardiography text descriptions of the aortic valve, which are documented by the reading echocardiography attending, represent a more specific determination of disease status as compared to a disease definition determined by the aortic valve mean gradients or calculated valve areas accompanying the echo reports alone (there were many

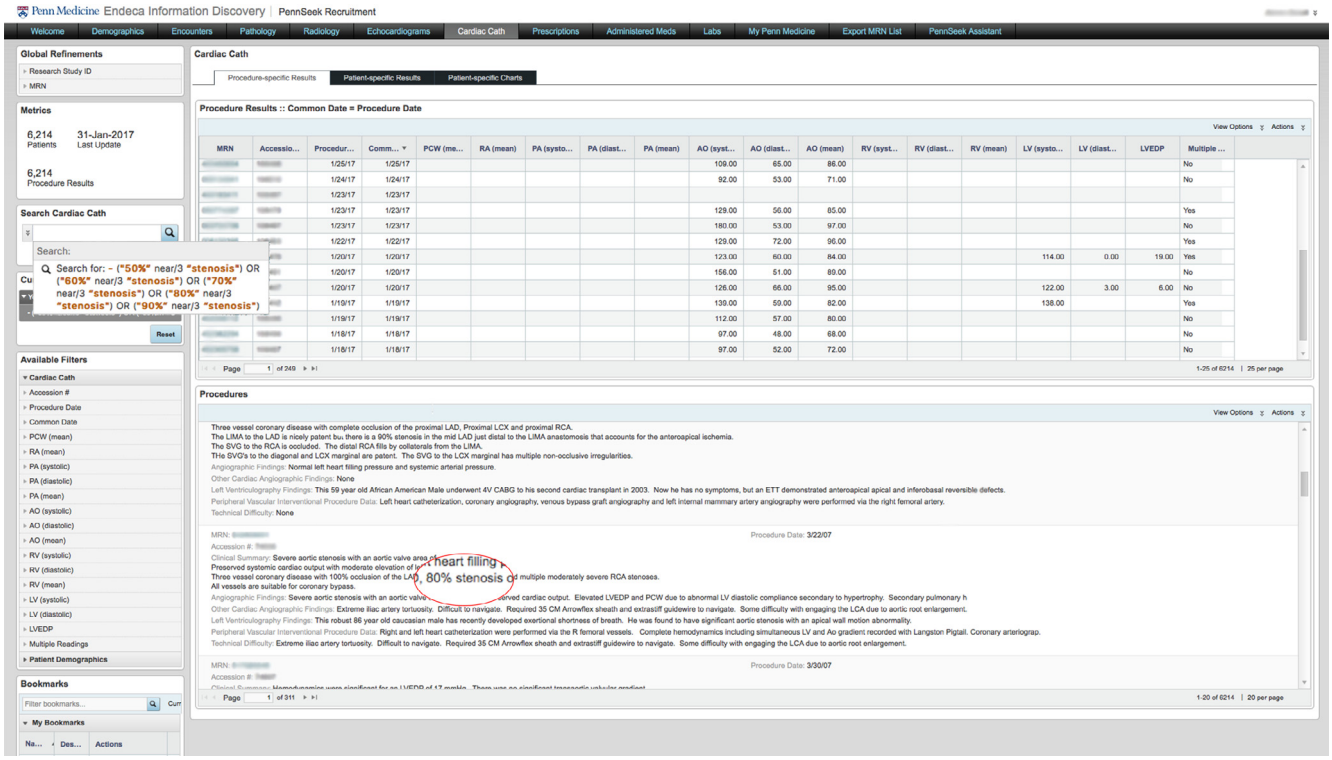

Fig. 1. Screenshot of a coronary artery disease key word query in PennSeek.

echocardiograms with a footnote that the exam was technically inadequate to determine the presence/severity of aortic valve disease). Individuals were considered disease positive if they had at least one echocardiogram with notation that there was mild (including low flow, low gradient) or more severe aortic stenosis in the echocardiogram text summary of the aortic valve. In the absence of a text description for the degree of aortic valve disease, we also considered individuals as disease positive if they met echocardiographic criteria for at least moderate aortic stenosis (aortic valve mean gradient $\geq$ 20 mmHg or maximum jet velocity $\geq$ 3.0 m/s) [14].

To identify individuals with TAS by text mining, we applied a discrete algorithm of keyword phrases with Boolean logic to the text of all echocardiography reports in PennSeek (inclusion and exclusion keyword phrases are reported in Table 3). To identify individuals with TAS by ICD-9 coding, we queried all patient charts having codes commonly referenced and validated in the literature for TAS without congenital valve abnormality [15–18] (specific codes and relevant diagnoses in Table 1).

Adjudication of disease status was completed by referencing the entirety of an individual's medical history, including detail from echocardiography reports performed outside of PROSOLV that were well notated in the medical record.

2.4. Classification of patients with coronary artery disease

We defined CAD as the presence of greater than 50% stenosis in any major epicardial vessel noted on a patient's coronary angiogram. Similar to our TAS definition, we used the text summary accompanying cardiac catheterization reports as the basis to determine the status of obstructive CAD. In consultation with board-trained cardiologists, we determined a set of key phrases that encapsulated obstructive CAD. These included the mention of greater than 50% stenosis in any vessel, "diffuse disease" of the coronary vasculature, or vessel "occlusion". We also considered a history of any surgical procedure to correct obstructive coronary disease (for example, the mention of a coronary stent or prior CABG) as an indicator of positive CAD status.

Table 1
Diagnostic codes for trileaflet aortic stenosis.

Code	Diagnosis	Number per Group (Percent of 8272)
Inclusion		
424.1	"aortic valve disorder"	8173 (98.8%)
395.0	"rheumatic aortic stenosis"	187 (2.3%)
395.2	"rheumatic aortic stenosis with insufficiency"	20 (0.2%)
396.0	"mitral valve stenosis and aortic valve stenosis"	27 (0.3%)
396.2	"mitral valve insufficiency and aortic valve stenosis"	37 (0.4%)
Exclusion		
Code	Diagnosis	
746.3	"congenital stenosis of aortic valve"	
746.4	"congenital insufficiency of aortic valve"	

To identify individuals with CAD by text mining, we first limited our patient population to only patients having a coronary angiogram by searching the text of all cardiac catheterization reports for the keywords: "left main", "LM", "RCA", "right coronary artery", "LCx", "left circumflex", LAD, or "left anterior descending". We then applied a discrete algorithm of keyword phrases felt to encapsulate the CAD phenotype to the text of all resulting catheterization reports in PennSeek (inclusion and exclusion keyword phrases are reported in Table 4). To identify individuals with CAD by ICD-9 coding we queried all patient charts for codes commonly referenced and validated in the literature for CAD [19–21] (specific codes and relevant diagnoses in Table 2). These included codes that were exact for coronary artery disease (414.01 – "coronary atherosclerosis of native coronary artery") in addition to codes

Table 2
Diagnostic codes for coronary artery disease.

Code	Diagnosis	Number per Group (Percent of 6910)
410	"acute myocardial infarction"	3 (0.04%)
411	"other acute and subacute forms of ischemic heart disease"	278 (4.0%)
412	"old myocardial infarction"	829 (12.0%)
413	"angina pectoris"	21 (0.3%)
414.00	"coronary atherosclerosis of unspecified type of vessel native or graft"	4700 (68.0%)
414.01	"coronary atherosclerosis of native coronary artery"	2305 (33.3%)
414.02	"coronary atherosclerosis of autologous vein bypass graft"	174 (2.5%)
414.03	"coronary atherosclerosis of nonautologous biological bypass graft"	1 (0.01%)
414.04	"coronary atherosclerosis of artery bypass graft"	107 (1.5%)
414.05	"coronary atherosclerosis of unspecified bypass graft"	35 (0.5%)
414.10	"aneurysm of heart wall"	45 (0.6%)
414.19	"other aneurysm of heart"	6 (0.08%)
414.2	"chronic total occlusion of coronary artery"	5 (0.07%)
414.3	"coronary atherosclerosis due to lipid rich plaque"	3 (0.04%)
414.4	"coronary atherosclerosis due to calcified coronary lesion"	23 (0.3%)
414.8	"other specified forms of chronic ischemic heart disease"	1063 (15.4%)
414.9	"chronic ischemic heart disease unspecified"	511 (7.4%)

implying a history of an ischemic insult to the heart (414.10 – "aneurysm of heart wall").

Adjudication of disease status was completed by referencing the entirety of an individual's medical history, including mention of cardiac catheterization reports outside of the HUP electronic catheterization database that were well notated in the medical record.

2.5. Manual chart review

For each disease entity, there were four possible outcome categories: patients uniquely identified as test positive by PennSeek alone, patients uniquely identified as test positive by ICD-9 alone, patients identified as test positive by both PennSeek and ICD-9, and patients not identified as test positive by either test. For validation, 250 patients were randomly selected from each of these four categories for CAD and 200 patients were randomly selected from each of these four categories for TAS. The echocardiography (for TAS) and cardiac catheterization reports (for CAD) of the selected patients were manually reviewed by a team of two researchers (AMS, DK) blinded to the patient's classification. For each patient review, AMS or DK longitudinally assessed for the disease of interest over all echocardiography or catheterization reports using the previously defined rule sets. Patients were then labeled as either disease positive, disease negative, or as having a questionable (for TAS) or mild (for CAD) disease status. For CAD, a status of mild CAD was assigned to reports having 30–50% stenosis in any major epicardial vessel. For TAS, a questionable disease status was attributed to one of: an echocardiographic report mentioning the presence or possibility of hemodynamically significant aortic stenosis but with a suboptimal exam precluding assignment of disease severity, or (for ICD-9 identified individuals) an echocardiographic report with a post-surgical aortic valve and no pre-surgical echocardiogram for comparison. An inter-annotator agreement study was performed by randomly selecting 25 charts reviewed by DK and 25 charts reviewed by AMS for each disorder, and comparing adjudication results by the opposite reviewer.

2.6. Statistical analysis

We calculated point estimates for the positive predictive value (PPV)/precision and negative predictive value (NPV) based on data collected from our manual review of patient charts. 95% credible intervals for the PPV (precision) and NPV were calculated using Bayesian inference with a uniform prior assumption and beta posterior. Sensitivity (recall) and specificity point estimates were calculated by applying Bayes' theorem to PPV/NPV and the proportion of test positive patients per test. Prevalence was estimated using conditional probability to sum the probabilities of disease status in the test positive and test negative groups. 95% credible intervals for sensitivity (recall) and specificity were approximated by extrapolating the covariance between PPV/precision and NPV for various correlation coefficients ranging from -1 to 1 and choosing the values providing the widest 95% credible interval range. All statistical calculations were performed using the R statistical software (R Foundation for Statistical Computing, Vienna, Austria).

3. Results

3.1. Timing of queries

Text mining queries in PennSeek were run in a stepwise fashion with inclusion and exclusion queries performed separately. Run times averaged 50.6 s (range: 11.5–42.3 s) per individual text mining term for TAS and 18.5 s (range: 48.3–52.7 s) per individual text mining term for CAD. The average run time to perform all keyword phrase queries and compile results was 202.2 s for TAS and 430.7 s for CAD. ICD-9 queries were performed in a single step for each disease, averaging 40.8 s for inclusion TAS codes (Table 1**A**), 6.1 s for exclusion TAS (Table 1**B**), and 6.3 s for CAD (Table 2).

3.2. Classification of patients with aortic stenosis

There were 62,703 individuals with an echocardiogram in PennSeek as of the October 2015 search date. Of those patients with an echocardiogram, 7115 individuals identified by PennSeek and 8272 individuals identified by ICD-9 codes met our qualification for TAS.

The number of patients per group and those overlapping are noted in a Venn diagram in Fig. 2. Patient allocation after manual review is shown in Table 5. All questionable diagnoses were considered as disease negative when identified by either test. Patients with a bicuspid valve were considered disease negative regardless of the degree of aortic stenosis, as our goal was to identify trileaflet disease. Following these rules, the total percentages of disease positive patients per group were: 98.5% (197 of 200 manually reviewed) for patients test positive by both PennSeek and ICD-9, 91.5% (183 of 200 manually reviewed) for patients uniquely test positive by PennSeek, 12.0% (24 of 200 manually reviewed) for patients uniquely test positive by ICD-9, and 0% (0 of 200 manually reviewed) for patients test negative by both PennSeek and ICD-9. An inter-annotator study of 25 randomly selected individuals for

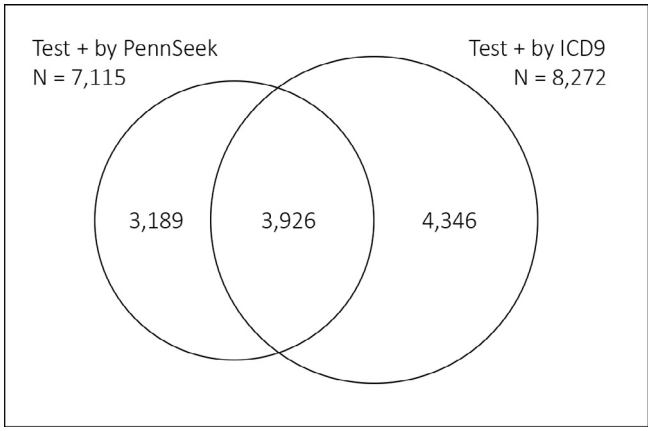

Fig. 2. Number of patients per group by test for trileaflet aortic stenosis.

Table 3
A: inclusion keyword phrases for trileaflet aortic stenosis. B: exclusion keyword phrases for trileaflet aortic stenosis.

Keyword phrase	Number per group (percentage of 7115)
(A)	
"mild aortic stenosis"	2514 (35.3%)
"mild-to-moderate aortic stenosis"	621 (8.7%)
"moderate aortic stenosis"	2016 (28.3%)
"moderate-to-severe aortic stenosis"	793 (11.1%)
"severe aortic stenosis"	3097 (43.5%)
"critical aortic stenosis"	1117 (15.7%)
"low flow, low gradient aortic stenosis"	116 (1.6%)
(B)	
"aortic valve is quadricuspid"	
"aortic valve is dysmorphic"	
"aortic valve prosthesis" OR "AV prosthesis"	
"is bicuspid"	

Table 4
Inclusion keyword phrases for coronary artery disease.

Keyword phrase	Number per group (percentage of 9247)
"CABG"	866 (9.4%)
"coronary artery bypass"	659 (7.1%)
"anastomosis"	215 (2.3%)
"LIMA"	1224 (13.2%)
"left internal mammary artery"	500 (5.4%)
"saphenous vein"	672 (7.2%)
"SVG"	1053 (11.4%)
"stent"	3568 (38.6%)
"restenosis" OR "re-stenosis"	767 (8.3%)
"diffusely diseased"	1620 (17.5%)
"occluded" OR "occlusion"	3073 (33.2%)
One of "(50, 55, 60, 65, 70, 76, 80, 85, 90, 95, 99, 100)%" within 3 words ("stenosis/stenosed/stenoses/lesion/lesions")	5366 (58.0%)

DK and 25 randomly selected individuals for AS resulted in 100% concordance.

For the 24 individuals identified as TAS-positive uniquely by ICD-9, these represented cases where the echocardiogram text lacked notation regarding the degree of aortic stenosis but echocardiographic criteria were consistent for moderate or more severe aortic stenosis. All 34 TAS questionable individuals uniquely categorized by ICD-9 represented cases having a post-surgical valve replacement echocardiogram without a pre-surgical comparison. For the 142 individuals who were incorrectly identified with TAS by ICD-9, 10 had a bicuspid valve without a diagnosis of TAS, 110 had a diagnosis of aortic insufficiency, and 22 were either completely free of aortic valvular pathology, or had aortic sclerosis without aortic stenosis.

In the category of individuals uniquely filtered by PennSeek, there were 11 patients who had a bicuspid valve morphology. These represented morphological variants or misspellings, which were not used as exclusion keywords in our PennSeek search. For example "the aortic valve. . .*appears bicuspid* with a raphe'" or "the aortic *valve bicuspid* with a raphe'". There were 5 patients with a questionable to probable status of TAS. These similarly represented morphological variants suggesting the possibility of true disease, for example "There is borderline mild aortic stenosis". For the single patient picked up by PennSeek as TAS positive who did not have disease, this individual had an aortic valve text summary including the following statement, "*No aortic stenosis is seen*. There is trace aortic regurgitation. Inadequate Doppler assessment to measure AVA. *There is probably mild aortic stenosis*." Review of several echocardiograms prior and after to this examination confirmed that the individual did not have disease.

3.3. Classification of patients with coronary artery disease

There were 14,567 individuals with a coronary angiogram in PennSeek between January 2007 and the search date, October 2015. Of all patients with a coronary angiogram, 9247 patients with CAD were identified by PennSeek, and 6910 patients were identified with CAD by ICD-9.

The number of patients per group is shown in Fig. 3, and the patient allocation after manual review is shown in Table 6. All mild CAD diagnoses were considered as CAD negative when identified by either test. Following this rule, the total percentages of disease positive patients per group were: 98% (245 of 250 manually reviewed) for patients test positive by both PennSeek and ICD-9, 94.8% (237 of 250 manually reviewed) for patients uniquely test positive by PennSeek, 5.2% (13 of 250 manually reviewed) for patients uniquely test positive by ICD-9, and 0.8% (2 of 250 manually reviewed) for patients test negative by both PennSeek and ICD-9. An inter-annotator study of 25 randomly selected individuals for DK and 25 randomly selected individuals for AS resulted in 100% concordance (see Table 7).

For the 79 individuals uniquely identified as having mild or true CAD by ICD-9 (i.e. negative by PennSeek), 66 patients had a mild diagnosis of CAD and only 13 had definite CAD. For the remaining 171 patients (of 250 manually reviewed) who were incorrectly identified with CAD by ICD-9, there were 52 individuals who had no angiographic evidence of coronary disease and 119 patients noted with only luminal irregularities on their coronary angiogram. For the 9 patients without CAD who were inappropriately picked up by our PennSeek search as having disease, these represented instances where a keyword phrase was used in a context not congruent with the left heart catheterization. For example, "the left internal iliac artery is *100% occluded*", or "bilateral patent aorto-iliac *stent* grafts". The 2 patients with definite CAD who were identified neither by PennSeek nor ICD-9, had in their cath reports morphological variants of CAD which were not included as keywords in our PennSeek search, for example "Selective left coronary angiography revealed a left main coronary artery with luminal irregularities, and *intermediate stenoses* in the proximal LAD and ostial diagonal lesion."

3.4. Comparison of tests

Table 5 shows the PPV/precision, NPV, sensitivity/recall, specificity, F-measure (F1 score), and corresponding credible intervals for both our text mining-based and ICD-9 based searches for the

Table 5
Results of manual review of echocardiograms for aortic stenosis.

Group:	Test positive only by PennSeek N = 3189	Test positive only by ICD-9 N = 4346	Test positive both PennSeek and ICD-9 N = 3926	Test negative both PennSeek and ICD-9 N = 69,703
Number adjudicated:	200	200	200	200
Definite trileaflet aortic stenosis by manual review	183	24	197	0
Questionable to probable trileaflet aortic stenosis by manual review	5	34	2	3
Presence of a bicuspid valve	11	10	1	1
No trileaflet aortic stenosis by manual review	1	132	0	196

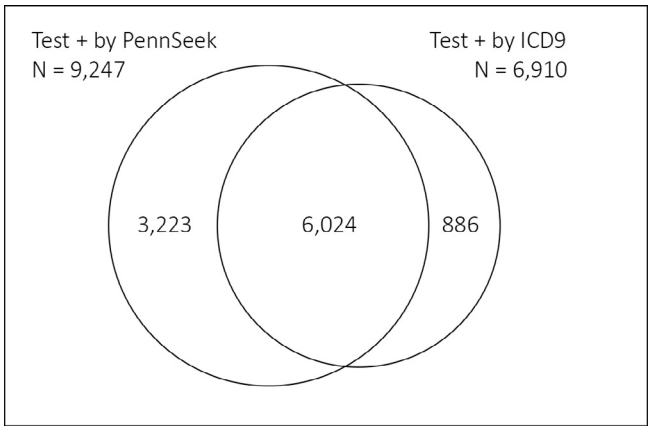

Fig. 3. Number of patients per group by test for coronary artery disease.

2 diseases. For TAS, PennSeek achieved a PPV of 0.95 compared with a PPV of 0.53 for ICD-9, and the F1 score for PennSeek was 0.97 compared with 0.67 for ICD-9 codes. For CAD, PennSeek achieved a PPV of 0.97 compared with 0.86 for ICD-9, and the F1 score for PennSeek was 0.98 compared with 0.75 for ICD-9 codes.

4. Discussion

With the advent of large biobanks and new technologies enabling access to biospecimens linked to "big data" for cardiovascular research, there is increased interest in abstracting clinical information from large retrospective patient cohorts for research purposes. Billing codes, an appealing option due to their ubiquitous nature and ease of use, are frequently utilized to identify patients with phenotypes of interest [22–24]. While this method has limitations, it remains in wide use due to the lack of a suitable alternative. The migration to ICD-10 promised to improve the accuracy of billing codes, however initial studies have demonstrated comparable accuracy to ICD-9 codes [25,26]. As biobanks linked to EHR data are increasingly being used for large-scale genomics and biomarker studies, efforts to better validate phenotypes of interest are vitally important.

We demonstrate that a novel database management platform, PennSeek, employing a version of text mining to cardiovascular procedure reports, has a higher PPV/precision and allows for greater discrimination than queries of the EHR utilizing only administrative codes for two common cardiovascular conditions, TAS and CAD. PennSeek searches more accurately identified individuals having these diseases than ICD-9 searches. Moreover, we demonstrated that for these two cardiovascular phenotypes, ICD-9 searching had a relatively low PPV/precision and may inappropriately identify patients as candidates for study who do not have the phenotype of interest.

Our data represents the first example of text mining approaches to identify CAD using cardiac catheterization text. Previous NLP and text mining approaches to identify CAD in retrospective EHR cohorts have been limited to the use of clinical narrative text. For example, the 2014 i2b2 challenge required users to identify key history components related to CAD using only text from the clinical encounter [11,27]. A group led by Katherine Liao et al. in 2015 similarly applied NLP to clinical notes from three different and diverse patient populations to determine CAD status, achieving a PPV of 90% [28]. From a biological perspective, we believe that determin-

Table 6
Results of manual review of coronary angiograms for coronary artery disease.

Group:	Test positive only by PennSeek N = 3223	Test positive only by ICD-9 N = 886	Test positive both PennSeek and ICD-9 N = 6024	Test negative both PennSeek and ICD-9 N = 4434
Number adjudicated:	250	250	250	250
Definite CAD by manual review	237	13	245	2
Mild CAD by manual review	4	66	2	0
No CAD by manual review	9	171	3	248

Table 7
Summary statistics per diagnostic test.

Disease	Test	PPV (precision)	NPV	Sensitivity	Specificity (recall)	F 1 Score
Aortic stenosis	PennSeek	0.95 (0.93–0.97)	0.99 (0.97–0.99)	0.92 (0.82–0.99)	0.99 (0.992–0.99)	0.97
	ICD-9	0.53 (0.48–0.58)	0.96 (0.93–0.98)	0.61 (0.43–0.79)	0.95 (0.94–0.95)	0.68
Coronary artery disease	PennSeek	0.97 (0.95–0.98)	0.94 (0.91–0.96)	0.99 (0.98–0.99)	0.94 (0.91–0.97)	0.98
	ICD-9	0.86 (0.84–0.88)	0.60 (0.55–0.64)	0.66 (0.63–0.69)	0.54 (0.51–0.56)	0.75

ing case status using gold standard measurement should provide the most accurate phenotype for use in retrospective cohort designs. Furthermore, CAD descriptive language in cardiac catheterization reports are generally more highly stereotyped than relevant history in the clinical narrative, explaining our improved F1 score of 0.98 relative to other groups' attempts.

In contrast to CAD, there are a variety of successful efforts in the application of text mining and NLP to echocardiography reports for aortic stenosis. It was demonstrated in 2005 that the highly stereotyped nature of echocardiography text presents an advantage for simple language extraction tasks, with an effort to characterize 'aortic stenosis' status from a collection of 703 echocardiogram reports at the Massachusetts General Hospital achieving a high precision (99%) and recall (78%) [13]. More recently, a group led by Chinmoy Nath at Northwestern University developed an NLP application titled 'EchoInfer', which can automatically extract cardiovascular data from echocardiography reports. EchoInfer achieved a PPV of 91.67 and an F1 score of 89.80 for description of the degree of aortic stenosis. Our method similarly takes advantage of the stereotyped nature of echocardiography reports to determine case status for a biologically important phenotype, trileaflet aortic stenosis, achieving comparable PPV and F1 scores to the published data.

An important novel finding in our study was that for both TAS and CAD there exist a large number of individuals identified by ICD-9 as having disease who do not have evidence for active disease by gold standard diagnosis. There are several explanations for the dissimilar accuracy in these data when identifying cardiovascular disease by PennSeek versus diagnostic codes, which highlight concerns regarding the utilization of administrative databases for research purposes. First, billing codes are not generated for purposes of clinical research and often do not meet the rigorous standards that one would apply to a scientific study. Administrative data is subject to errors of omission, commission, and definition. For example, a 65 year-old individual with luminal irregularities on her cardiac catheterization may warrant the assignment of a diagnostic code for CAD to justify the use of a statin. However, for research purposes, luminal irregularities generally do not meet the definition of CAD used in a scientific study. Additionally, billing codes may be erroneously assigned in an effort to indicate what pathology is being evaluated. For example, a physician may assign a patient the code of 395.0 corresponding to the diagnosis 'rheumatic aortic stenosis' in an effort to indicate the patient has aortic valve disease when in fact the patient's disease is not secondary to a rheumatic process. Finally, billing codes may be non-specific. The most commonly reported ICD-9 code for aortic stenosis is V 424.1, which corresponds to the diagnosis 'aortic valve disorder'. While aortic stenosis is the most common reason a patient would warrant this diagnosis, less common pathologies involving the aortic valve, most prominently aortic insufficiency in our data set, may be included in this analysis. While the migration to ICD-10 may improve the accuracy of billing codes, initial studies have demonstrated comparable accuracy to ICD-9 codes.

The strengths of this study are that we included two different and common cardiovascular conditions that are often the subject of large-scale database research [16,29]. Additionally, whereas prior studies have only discussed the merits or flaws of utilizing billing codes to identify patients for study, we developed and utilized a novel database management system incorporating text mining with access to the text from cardiovascular procedure reports to improve upon cardiovascular research. To our knowledge, this is the first description of such a text mining platform that utilizes cardiovascular procedure reports. Finally, it is important to note that our use of this software was not a "niche" application. A similar technique could readily be adopted at other sites: the Oracle Endeca based software is widely available, and could be applied at any institution that uses an EHR.

The limitations of our study include that this was a single center study. It is possible that other institutions are more accurate in their use of billing codes, although there is a wealth of evidence in the literature that other institutions have experienced comparable issues [5]. Additionally, we did not evaluate other cardiovascular conditions, which may have a higher fidelity in terms of assigned billing codes. Finally, we recognize that certain conditions, such as heart failure, may be more difficult to identify via this technique, as it requires synthesis of complex clinical data.

In conclusion, we demonstrate that the application of customized text mining to cardiovascular procedure reports, a novel use of text mining in the clinical domain, was superior to administrative codes for two common cardiovascular conditions. This approach has the potential to greatly improve the accuracy of cohort generation using the EHR for a wide range of diseases for purposes of clinical research.

Conflicts of interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgements

Financial support for this work was received from a Doris Duke Clinical Research Mentorship Award (A. Small).

References

[1] L.V. Rasmussen, W.K. Thompson, J.A. Pacheco, A.N. Kho, D.S. Carrell, J. Pathak, P. L. Peissig, G. Tromp, J.C. Denny, J.B. Starren, Design patterns for the development of electronic health record-driven phenotype extraction algorithms, J. Biomed. Inform. 51 (2014) 280–286.
[2] S. Yu, K.P. Liao, S.Y. Shaw, V.S. Gainer, S.E. Churchill, P. Szolovits, S.N. Murphy, I. S. Kohane, T. Cai, Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources, J. Am. Med. Inform. Assoc. 22 (2015) 993–1000.
[3] N. Peek, J.H. Holmes, J. Sun, Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics, Yearb Med. Inform. 9 (2014) 42–47.
[4] J.C. Denny, M.D. Ritchie, M.A. Basford, J.M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D.R. Masys, D.M. Roden, D.C. Crawford, PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations, Bioinformatics 26 (2010) 1205–1210.
[5] E. Birman-Deych, Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors, Med. Care 43 (2005) 480–485.
[6] P.M. Nadkarni, L. Ohno-Machado, W.W. Chapman, Natural language processing: an introduction, J. Am. Med. Inform. Assoc. 18 (2011) 544–551.
[7] G. Karystianis, A. Dehghan, A. Kovacevic, J.A. Keane, G. Nenadic, Using local lexicalized rules to identify heart disease risk factors in clinical notes, J. Biomed. Inform. 58S (2015) S183–S188.
[8] P. Warrer, E.H. Hansen, L. Juhl-Jensen, L. Aagaard, Using text-mining techniques in electronic patient records to identify ADRs from medicine use, Br. J. Clin. Pharmacol. 73 (2012) 674–684.
[9] Li L. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening – a case study. AMIA symposium, 2008, 405.
[10] Pakhomov P. Serguei, Electronic medical records for clinical research-application to the identification of heart failure, Am. J. Managed Care 13 (2007).
[11] A. Stubbs, C. Kotfila, H. Xu, O. Uzuner, Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task Track 2, J. Biomed. Inform. (2015).
[12] J.H. Garvin, S.L. DuVall, B.R. South, B.E. Bray, D. Bolton, J. Heavirland, S. Pickard, P. Heidenreich, S. Shen, C. Weir, M. Samore, M.K. Goldstein, 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. 19 (2012) 859–866.
[13] Chung J. Concept-value pair extraction from semi-structured clinical narrative- a case study using echocardiogram reports. AMIA Symposium, 2005.

[14] R.A. Nishimura, C.M. Otto, R.O. Bonow, B.A. Carabello, J.P. Erwin 3rd, R.A. Guyton, P.T. O'Gara, C.E. Ruiz, N.J. Skubas, P. Sorajja, T.M. Sundt 3rd, J.D. Thomas, G. American College of Cardiology/American Heart Association Task Force on Practice, AHA/ACC guideline for the management of patients with valvular heart disease: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines, J Am Coll Cardiol. 2014 (63) (2014) e57–e185.

[15] A.O. Badheka, V. Singh, N.J. Patel, S. Arora, N. Patel, B. Thakkar, S. Jhamnani, S. Pant, A. Chothani, C. Macon, S.S. Panaich, J. Patel, S. Manvar, C. Savani, P. Bhatt, V. Panchal, N. Patel, A. Patel, D. Patel, S. Lahewala, A. Deshmukh, T. Mohamad, A.A. Mangi, M. Cleman, J.K. Forrest, Trends of Hospitalizations in the United States from 2000 to 2012 of patients >60 years with aortic valve disease, Am. J. Cardiol. 116 (2015) 132–141.

[16] G. Thanassoulis, C.Y. Campbell, D.S. Owens, J.G. Smith, A.V. Smith, G.M. Peloso, K.F. Kerr, S. Pechlivanis, M.J. Budoff, T.B. Harris, R. Malhotra, K.D. O'Brien, P.R. Kamstrup, B.G. Nordestgaard, A. Tybjaerg-Hansen, M.A. Allison, T. Aspelund, M.H. Criqui, S.R. Heckbert, S.J. Hwang, Y. Liu, M. Sjogren, J. van der Pals, H. Kalsch, T.W. Muhleisen, M.M. Nothen, L.A. Cupples, M. Caslake, E. Di Angelantonio, J. Danesh, J.I. Rotter, S. Sigurdsson, Q. Wong, R. Erbel, S. Kathiresan, O. Melander, V. Gudnason, C.J. O'Donnell, Post WS and group CECW genetic associations with valvular calcification and aortic stenosis, New Engl. J. Med. 368 (2013) 503–512.

[17] M.A. Clark, S.V. Arnold, F.G. Duhay, A.K. Thompson, M.J. Keyes, L.G. Svensson, R. O. Bonow, B.T. Stockwell, D.J. Cohen, Five-year clinical and economic outcomes among patients with medically managed severe aortic stenosis: results from a Medicare claims analysis, Circ. Cardiovasc. Qual. Outcomes. 5 (2012) 697–704.

[18] J.P. Vavalle, H.R. Phillips, S.A. Holleran, A. Wang, C.M. O'Connor, P.K. Smith, G.C. Hughes, J.K. Harrison, M.R. Patel, Analysis of geographic variations in the diagnosis and treatment of patients with aortic stenosis in North Carolina, Am. J. Cardiol. 113 (2014) 1874–1878.

[19] E. Zakynthinos, N. Pappa, Inflammatory biomarkers in coronary artery disease, J. Cardiol. 53 (2009) 317–333.

[20] L. Yang, M. Yu, S. Gao, Prediction of coronary artery disease risk based on multiple longitudinal biomarkers, Stat. Med. (2015).

[21] M.H. Wu, H.C. Chen, S.J. Yeh, M.T. Lin, S.C. Huang, S.K. Huang, Prevalence and the long-term coronary risks of patients with Kawasaki disease in a general population <40 years: a national database study, Circ. Cardiovasc. Qual. Outcomes. 5 (2012) 566–570.

[22] L.B. Goldstein, Accuracy of ICD-9-CM coding for the identification of patients with acute ischemic stroke, Stroke (1998).

[23] G. Chen, N. Khan, R. Walker, H. Quan, Validating ICD coding algorithms for diabetes mellitus from administrative data, Diabetes Res Clin Pract. 89 (2010) 189–195.

[24] L. Tamariz, T. Harkins, V. Nair, A systematic review of validated methods for identifying ventricular arrhythmias using administrative and claims data, Pharmacoepidemiol. Drug Saf. 21 (Suppl 1) (2012) 148–153.

[25] L.K. Jorgensen, L.S. Dalgaard, L.J. Ostergaard, N.S. Andersen, M. Norgaard, T.H. Mogensen, Validity of the coding for herpes simplex encephalitis in the Danish National Patient Registry, Clin. Epidemiol. 8 (2016) 133–140.

[26] R.J. Jolley, H. Quan, N. Jette, K.J. Sawka, L. Diep, J. Goliath, D.J. Roberts, B.G. Yipp, C.J. Doig, Validation and optimisation of an ICD-10-coded case definition for sepsis using administrative health data, BMJ Open 5 (2015) e009487.

[27] V. Kumar, A. Stubbs, S. Shaw, O. Uzuner, Creation of a new longitudinal corpus of clinical narratives, J. Biomed. Inform. 58 (Suppl) (2015) S6–S10.

[28] K.P. Liao, A.N. Ananthakrishnan, V. Kumar, Z. Xia, A. Cagan, V.S. Gainer, S. Goryachev, P. Chen, G.K. Savova, D. Agniel, S. Churchill, J. Lee, S.N. Murphy, R. M. Plenge, P. Szolovits, I. Kohane, S.Y. Shaw, E.W. Karlson, T. Cai, Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts, PLoS One. 10 (2015) e0136651.

[29] S.S. Panaich, A.O. Badheka, S. Arora, N.J. Patel, B. Thakkar, N. Patel, V. Singh, A. Chothani, A. Deshmukh, K. Agnihotri, S. Jhamnani, S. Lahewala, S. Manvar, V. Panchal, A. Patel, N. Patel, P. Bhatt, C. Savani, J. Patel, G.T. Savani, S. Solanki, S. Patel, A. Kaki, T. Mohamad, M. Elder, A. Kondur, M. Cleman, J.K. Forrest, T. Schreiber, C. Grines, Variability in utilization of drug eluting stents in United States: Insights from nationwide inpatient sample, Catheter. Cardiovasc. Interv. (2015).