OPEN ACCESS

Billing code algorithms to identify cases of peripheral artery disease from administrative data

Jin Fan,[1,2] Adelaide M Arruda-Olson,[2] Cynthia L Leibson,[3] Carin Smith,[3] Guanghui Liu,[2] Kent R Bailey,[3] Iftikhar J Kullo[2]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-001827).

[1]Geriatric Cardiovascular Department, Chinese PLA General Hospital, Beijing, China
[2]Division of Cardiovascular Diseases and the Gonda Vascular Center, Mayo Clinic, Rochester, Minnesota, USA
[3]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA

Correspondence to
Dr Iftikhar J Kullo, Division of Cardiovascular Diseases, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA; kullo.iftikhar@mayo.edu

JF and AMA-O contributed equally.

Received 25 March 2013
Revised 4 October 2013
Accepted 9 October 2013
Published Online First 28 October 2013

ABSTRACT

Objective To construct and validate billing code algorithms for identifying patients with peripheral arterial disease (PAD).

Methods We extracted all encounters and line item details including PAD-related billing codes at Mayo Clinic Rochester, Minnesota, between July 1, 1997 and June 30, 2008; 22 712 patients evaluated in the vascular laboratory were divided into training and validation sets. Multiple logistic regression analysis was used to create an integer code score from the training dataset, and this was tested in the validation set. We applied a model-based code algorithm to patients evaluated in the vascular laboratory and compared this with a simpler algorithm (presence of at least one of the ICD-9 PAD codes 440.20–440.29). We also applied both algorithms to a community-based sample (n=4420), followed by a manual review.

Results The logistic regression model performed well in both training and validation datasets (c statistic=0.91). In patients evaluated in the vascular laboratory, the model-based code algorithm provided better negative predictive value. The simpler algorithm was reasonably accurate for identification of PAD status, with lesser sensitivity and greater specificity. In the community-based sample, the sensitivity (38.7% vs 68.0%) of the simpler algorithm was much lower, whereas the specificity (92.0% vs 87.6%) was higher than the model-based algorithm.

Conclusions A model-based billing code algorithm had reasonable accuracy in identifying PAD cases from the community, and in patients referred to the non-invasive vascular laboratory. The simpler algorithm had reasonable accuracy for identification of PAD in patients referred to the vascular laboratory but was significantly less sensitive in a community-based sample.

BACKGROUND

Lower extremity peripheral artery disease (PAD) is highly prevalent, affecting about eight million people aged ≥40 years in the US population.[1 2] Although common and associated with significant mortality and morbidity, PAD is a relatively understudied phenotype of atherosclerotic vascular disease. Ascertainment of cases of PAD based on the ankle-brachial index (ABI) may be biased owing to exclusion of patients diagnosed by other procedures, or those who could not tolerate lower extremity arterial evaluation. A reliable method for identifying patients with PAD from hospital-wide databases would be useful for genetic and epidemiologic studies of PAD.

Healthcare encounters are recorded for administrative and reimbursement purposes in the USA.[3] Administrative data are readily available, often for large populations, are inexpensive to acquire, and are computer readable. Given these advantages, such data are increasingly being used in medical research. Moreover, the growing use of the electronic medical records (EMRs) and the development and validation of informatics approaches to access diverse phenotypes from the EMR[4] may facilitate rapid phenotyping. The accuracy of billing codes for ascertaining disease conditions varies.[5–10] For example, the positive predictive value (PPV) of billing codes for acute myocardial infarction was reported to be 96.9%,[6] whereas the PPV for pelvic inflammatory disease was only 18.1%.[7]

Whether billing codes are useful in ascertaining cases of PAD is unknown. We investigated whether billing codes, which include International Classification of Diseases, 9th revision, clinical modification (ICD-9-CM), and Current Procedural Terminology, 4th revision (CPT-4) codes, could be used for reliable ascertainment of PAD cases from administrative databases. We developed and validated billing algorithms to ascertain PAD status at Mayo Clinic Rochester, Minnesota, using administrative data over an 11-year period.

METHODS

Study design and population

This historical cohort study was approved by the institutional review board of the Mayo Clinic. We used the Mayo Decision Support System database, which contains administrative billing information on every patient for each encounter (hospital inpatient, hospital outpatient, emergency department, office visit, and nursing home visit) and patient demographics (eg, age, sex, and geographic region). We excluded individuals who refused authorization for use of their medical records for research (n=1650).

Non-invasive lower extremity arterial evaluation

Lower extremity arterial evaluation was performed using standardized protocols in the non-invasive vascular laboratory, which is certified by the Intersocietal Commission for the Accreditation of Vascular Laboratories. Systolic blood pressure was measured in each arm and in the dorsalis pedis and posterior tibial arteries bilaterally using a hand-held 8.3 MHz Doppler probe. The higher of the two arm pressures and the lower of the two ankle pressures were used to calculate the ABI in each leg.[11] The data were interpreted and reported by a vascular medicine specialist. Results of all non-invasive lower extremity arterial evaluations are entered into a database that becomes part of the Mayo EMR.

Open Access
Scan to access more free content

To cite: Fan J, Arruda-Olson AM, Leibson CL, *et al*. *J Am Med Inform Assoc* 2013;**20**: e349–e354.

Definition of PAD and normal ABI

PAD was defined as atherosclerotic occlusive arterial disease of the lower extremities, including arteries distal to the aortic bifurcation. The vascular laboratory diagnostic criteria for PAD were: (1) a resting/post-exercise ABI ≤0.9; or (2) the presence of poorly compressible arteries (ABI >1.4; or ankle blood pressure >255 mm Hg).[11] Patients with normal non-invasive vascular test(s) were classified as normal ABI.

PAD-related diagnostic and procedural billing codes

We identified all patients with any encounter at Mayo Clinic (Rochester, Minnesota, USA) between July 1, 1997 and June 30, 2008. We extracted all encounters and line item details including PAD-related billing codes (ICD-9-CM/CPT codes listed in online supplementary tables S1 and S2). Codes were assigned at discharge from each encounter. The encounter-level data contained a maximum of 16 diagnosis codes and a maximum of 16 procedural codes. We used the diagnosis codes assigned in any position of the claim (as primary or secondary diagnosis). We applied ICD-9-CM codes[12] created by the US National Center for Health Statistics. In addition, the line item detail data contained all CPT-4 codes[13] published and maintained by the American Medical Association, for procedures of line item detail provided during that encounter.

A list of PAD-related ICD-9-CM diagnosis/procedure and CPT-4 procedural codes was compiled through literature review[14–19] and manual searches of the ICD-9-CM/CPT code books.[12 13] All ICD-9-CM diagnosis codes relevant to lower extremity atherosclerotic disease (eg, 440.2×—atherosclerosis of native arteries of the extremities; 440.3×—atherosclerosis of bypass graft of the extremities; 440.8×—atherosclerosis of other specified arteries; 440.9×—generalized and unspecified atherosclerosis, including arteriosclerotic vascular disease not otherwise specified; 443.9×—peripheral vascular disease, unspecified), CPT-4 or ICD-9-CM procedural codes related to PAD (eg, non-invasive lower extremity arterial evaluation, arteriography, duplex ultrasonography, magnetic resonance angiography, and CT angiography of the lower extremity arteries), or surgical/revascularization procedures related to PAD (eg, lower extremity vascular bypass, percutaneous revascularization, amputation, etc) were included (see online supplementary table S1).

Exclusion codes

To exclude non-atherosclerotic causes of vascular disease (eg, vasculitis, Buerger's disease, etc) and procedures/surgeries (eg, amputation) not related to PAD, we also generated a list of exclusion codes (see online supplementary table S2). These codes were evaluated and finalized by the authors who had expertise in PAD, epidemiology, and informatics. A detailed description of all the billing and procedure codes used in this study is summarized in online supplementary table S3.

Statistical analysis

Patients evaluated in the non-invasive vascular laboratory were randomized in a 1:1 manner to create training and validation datasets. Using the training dataset, we performed multiple backward selection logistic regression analyses to estimate the joint association between each patient's PAD status and their billing codes using vascular laboratory criteria as the 'gold standard'. Dichotomous variables (ie, presence/absence of the specific code) were created for each billing code. For patients whose PAD status was negative, only billing codes until the last vascular laboratory examination were used as model inputs. For patients whose ultimate PAD status was positive, only billing codes after the first vascular laboratory examination that met PAD criteria were used as model inputs. This was done to ensure that the billing codes reflected the current PAD status of the patient.

In the final logistic regression model 13 billing codes independently predicted PAD status at a significance level of $p<0.001$. All had positive β coefficients. For each of these billing codes we assigned a weighted integer coefficient value. To generate the integer score, the β coefficient of each variable which independently predicted PAD status was rounded to the nearest multiple of 0.25.[20] To generate the integer we then assigned 1 point for each 0.25 unit increment in the β coefficient. The individual billing code score for each patient was calculated as the sum of these integers. The choice of 0.25, while somewhat arbitrary, was chosen to differentiate relative contributions of codes from each other.

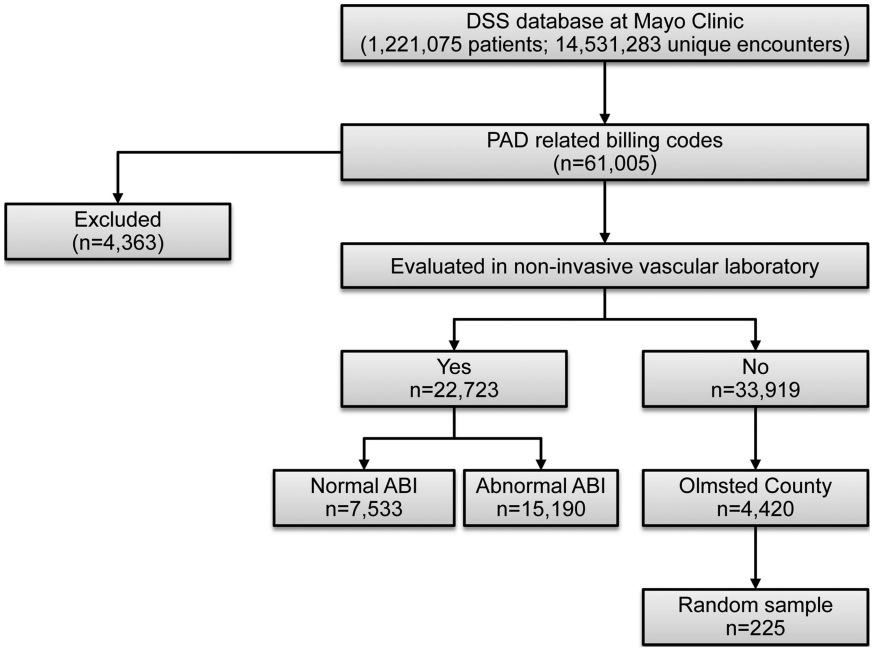

Figure 1 Flow diagram of patients in our study. ABI, ankle-brachial index; DSS, Decision Support System; PAD, peripheral arterial disease.

Table 1 Comparison of patients in the training and validation sets*

Variable	Training	Validation	p Value
Age (years)	68.0±12.5	68.1±12.6	0.53
Sex (male), n (%)	6681 (58.8)	6582 (58.0)	0.18
Race (white), n (%)	11 040 (97.2)	11 068 (97.5)	0.25

*n=11 356 for each dataset.

The overall discriminative ability of the final multiple logistic regression model was assessed with the c statistic (the area under the receiver operating characteristic (ROC) curve)[21 22] in training and validation datasets, respectively. Using ROC curve analysis, the optimal cut-off point of the billing code score was 8, as determined by a balance of sensitivity and specificity, yielding a sensitivity of 86% and a specificity of 83%. Sensitivity, specificity, PPV, and negative predictive value (NPV) were computed for the billing code score with a selection of possible cut-off points. Sensitivity—true positives/(true positives+ false negatives)—is the proportion of patients with true PAD (determined by the gold standard) classified correctly by the model, and specificity—true negatives/(true negatives+false positives)—is the proportion of patients without PAD (based on vascular laboratory criteria) classified correctly by the model. The proportion of positive calls who truly have PAD, is termed PPV=true positives/(true positives+false positives) and, correspondingly, the proportion of negative calls that are truly negative, is termed NPV=true negatives/(true negatives+false negatives). All the data were analyzed using the SAS 9.13 statistical package (SAS Institute, Cary, North Carolina, USA).

RESULTS

Study cohort characteristics

We identified all patients with any encounter at Mayo Clinic (Rochester, Minnesota, USA) between July 1, 1997 and June 30, 2008 (1 221 075 unique patients with 14 531 283 unique encounters) (figure 1). A total of 61 005 patients with potential PAD had at least one PAD-related diagnostic/procedural billing code. After exclusion of patients with codes indicating non-atherosclerotic vascular disease (n=4363), a total of 22 723 individuals were evaluated at the Mayo's non-invasive vascular laboratory and 33 919 were not. Among the patients evaluated in the vascular laboratory, 11 had inconclusive ABI results and were excluded from subsequent analysis. The remaining set of 22 712 patients was divided into training and validation sets of 11 356 patients each with 7617 (67.1%) and 7573 (66.7%) who met vascular laboratory criteria for PAD, respectively. The demographic characteristics of patients in these two subsets were similar (table 1).

Billing code score development and validation

Thirteen billing codes in the final logistic regression model independently predicted PAD status at p<0.001 (table 2). The billing codes with the largest ORs were: 440.24—atherosclerosis obliterans (ASO) extremities with gangrene (OR=45.4, 95% CI 16.5 to 125.0, p<0.0001); 440.21—ASO extremities with intermittent claudication (OR=29.9, 95% CI 24.8 to 36.0, p<0.0001); 440.23—ASO extremities with ulceration (OR=14.6, 95% CI 11.4 to 18.7, p<0.0001); 440.20—ASO extremities unspecified (OR=11.3, 95% CI 9.3 to 13.9, p<0.0001); 440.22—ASO extremities with rest pain (OR=9.4, 95% CI 5.9 to 14.8, p<0.0001); and 84.11—amputation of toe (OR=9.1, 95% CI 3.9 to 21.2, p<0.0001). The code 440.29—'Atherosclerosis of other native arteries of the extremities' did not reach statistical significance level (p<0.001), because only a small number of patients (74) in the entire dataset had this code (440.29). The model performed equally well in the training and validation sets with a c statistic of 0.91 for each. The ROC curve analysis showed that assignment of integer scores maintained their good performance in both datasets with c statistics of 0.91 (figure 2).

Estimates of sensitivity, specificity, PPV, and NPV using different cut-off points of billing code score are summarized in table 3. A billing code score ≥8 yielded the best sensitivity and specificity (86% and 83%, respectively) for ascertaining PAD status in the training set with nearly identical performance in the validation set (86% and 82%, respectively). The PPV of billing code score ≥8 was as high as 91% and the NPV was

Table 2 Thirteen billing codes selected by backward multivariate logistic regression analysis for ascertaining PAD status in the training dataset

Codes	Code description	Integer*	β±SE	OR	95% CI	p Value
Diagnosis codes (ICD-9-CM)						
440.24	ASO extremities with gangrene	15	3.82±0.52	45.4	16.5 to 125.0	<0.0001
440.21	ASO extremities with intermittent claudication	14	3.40±0.10	29.9	24.8 to 36.0	<0.0001
440.23	ASO extremities with ulceration	11	2.68±0.13	14.6	11.4 to 18.7	<0.0001
440.20	ASO extremities, unspecified	10	2.43±0.10	11.3	9.3 to 13.9	<0.0001
440.22	ASO extremities with rest pain	9	2.24±0.23	9.4	5.9 to 14.8	<0.0001
443.9	Peripheral vascular disease, unspecified	7	1.83±0.08	6.2	5.4 to 7.2	<0.0001
440.9	Generalized and unspecified ASO	6	1.62±0.10	5.1	4.2 to 6.1	<0.0001
Procedural codes (CPT-4 or ICD-9-CM)						
84.11	Amputation of toe	9	2.21±0.43	9.1	3.9 to 21.2	<0.0001
75716	Angiography, extremity, bilateral, radiological	8	1.91±0.32	6.7	3.6 to 12.5	<0.0001
73725	MRA lower extremity with or without contrast	7	1.78±0.44	6.0	2.5 to 14.2	<0.0001
75710	Angiography, extremity, unilateral, radiological	7	1.87±0.31	6.5	3.5 to 11.9	<0.0001
75635	CT angiogram—abdominal aorta and lower extremity runoff	6	1.58±0.29	4.8	2.8 to 8.5	<0.0001
93922	Non-invasive physiologic studies of lower extremity arteries	2	0.59±0.07	1.8	1.6 to 2.0	<0.0001

*1 point for each 0.25 unit increment in the β coefficient rounded to the nearest multiple of 0.25.
β, coefficient; ASO, atherosclerosis obliterans; ICD-9-CM, International Classification of Diseases, 9th revision, clinical modification; CPT-4, Current Procedural Terminology, 4th revision; MRA, magnetic resonance angiography; PAD, peripheral artery disease.

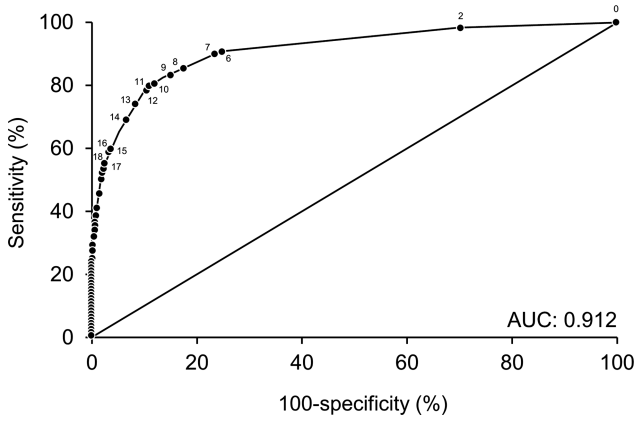

Figure 2 Receiver operating characteristic curve depicting the discriminatory performance of various integer scores. AUC, area under the curve.

74% in training set. Predictive values of the billing code score in the validation set were similar to those in the training set.

Application of algorithms: patients evaluated in the vascular laboratory

We applied the model-based billing code algorithm to patients evaluated in the vascular laboratory and compared it with a simpler algorithm—that is, presence of one of the ICD-9 PAD billing codes 440.20– 440.29. Although the application of the model-based billing code algorithm provided a better NPV, the simpler algorithm was also reasonably accurate for identification of PAD in patients referred to the vascular laboratory, with lesser sensitivity and greater specificity (table 4). However, among patients who were negative by the simpler algorithm (10 236), there were 3511 (34.3%) who were truly positive for PAD (table 5). The model-based algorithm picked up 1314 of these missed cases, at a cost of 502 additional false positives, and provided highly significant incremental information ($p<0.0001$) within this group of patients. Since virtually all the patients who were positive by the simpler algorithm were also positive by the model-based algorithm, there was no additional information provided by the model-based algorithm in this subset.

Application of algorithms to a community-based sample

We used the resources of the Rochester Epidemiology Project (REP)[23] to identify a community-based sample of 4420 Olmsted County residents who were not referred to the Mayo vascular laboratory, but who were evaluated at the Mayo Clinic between (July 1, 1997 to July 31, 2008) (figure 1) and who had at least one of the PAD-related billing codes described in online supplementary table S1. We divided this sample into three groups: no PAD codes from the model-based algorithm (score=0) (n=2351); score between 1 and 7 (n=1458); score of eight or higher (n=611). Subsequently we randomly selected 75 subjects from each group (random sample n=225) for blinded manual abstraction which was conducted by a single experienced cardiologist (AMA-O). Of these, 13 had unknown PAD status after manual abstraction and were excluded from subsequent analysis. We applied the model-based algorithm to this community-based sample and demonstrated reasonable accuracy of the model-based billing code algorithm for identification of PAD within the sample (table 6).

When we applied the simpler algorithm to this same community-based sample, the sensitivity of the simpler algorithm was lower, and the specificity higher, than that of the model-based algorithm. Both these differences were statistically significant at $p<0.01$. Furthermore, as shown in table 5, the model-based algorithm was helpful in discriminating PAD status among the 172 subjects found to be negative by the simpler algorithm. Of these, 46 (27%) were truly positive based on manual abstraction. The model-based algorithm captured 22 of these 46 missed cases at a cost of only six false positives, and provided significant incremental information ($p<0.0001$) within this group of subjects. In this community-based sample, as in the vascular laboratory population, the model-based algorithm provided no additional information when the simpler algorithm yielded a positive result.

DISCUSSION

In this study we constructed and validated billing code algorithms to ascertain PAD status using administrative data. The model-based billing code algorithm had reasonable accuracy for identification of PAD cases and controls (without PAD) from the community, as well as in patients who underwent non-invasive vascular testing. The simpler algorithm based on the presence of one of the commonly used PAD billing codes had reasonable accuracy in identifying PAD in patients referred to the vascular laboratory, but was much less sensitive in detecting patients with PAD in a community-based sample.

The ABI is a commonly used test with relatively high sensitivity and specificity for diagnosing PAD.[24] Patients at Mayo Clinic with PAD or suspected of having PAD are typically referred for non-invasive evaluation in the vascular laboratory. Using the training dataset consisting of half of the patients referred to the non-invasive vascular laboratory, we derived a logistic regression model that was reasonably accurate in discriminating PAD status. To avoid overfitting and assure reproducibility of our results the model was validated in an independent validation dataset.

The final model comprised 13 billing codes, including eight ICD-9-CM diagnosis/procedure codes and five CPT-4 diagnostic procedural codes. The scores for diagnosis codes such as

Table 3 Accuracy for classification of PAD status using different cut-off points of billing code score in the training and validation sets*

	Sensitivity		Specificity		PPV		NPV	
Cut-off point	Training (%)	Validation (%)	Training (%)	Validation (%)	Training (%)	Validation (%)	Training (%)	Validation (%)
Billing score ≥3	91	91	76	75	88	88	80	80
Billing score ≥5	91	91	76	75	88	88	80	80
Billing score ≥8	86	86	83	82	91	91	74	74
Billing score ≥12	79	78	90	89	94	94	68	67

*n=11 356 for each dataset.
PAD, peripheral artery disease; PPV, positive predictive value; NPV, negative predictive value.

Table 4 Comparison of model-based billing code algorithm versus simpler algorithm: patients evaluated in the vascular laboratory

Gold standard=ABI		
	Model-based algorithm	Simpler algorithm (codes 440.20–440.29)
Sensitivity, % (95% CI)	85.5 (85.0 to 86.1)	76.9 (76.2 to 77.6)
Specificity, % (95% CI)	82.6 (81.7 to 83.5)	89.3 (88.6 to 90.0)
PPV, % (95% CI)	90.8 (90.4 to 91.3)	93.5 (93.1 to 94.0)
NPV, % (95% CI)	73.9 (72.9 to 74.8)	65.7 (64.8 to 66.6)

CI calculated by binomial exact distribution.
ABI, ankle-brachial index; NPV, negative predictive value; PPV, positive predictive value.

Table 5 Comparison of model-based algorithm versus simpler algorithm in A) patients who underwent testing in the vascular laboratory; and B) patients from the community

A. Simpler algorithm negative (n=10 236)		
Model-based algorithm negative n=8420 (82.3%)	**Model-based algorithm positive n=1816 (17.7%)**	
6223 (73.9%)	502 (27.6%)	**Vascular laboratory negative n=6725 (65.7%)**
2197 (26.1%)	1314 (72.4%)	**Vascular laboratory positive n=3511 (34.3%)**
Simpler algorithm positive (n=12 487)		
Model-based algorithm negative n=3 (0.02%)	**Model-based algorithm positive n=12 484 (99.98%)**	
0 (0.0%)	808 (6.5%)	**Vascular laboratory negative n=808 (6.5%)**
3 (100.0%)	11 676 (93.5%)	**Vascular laboratory positive n=11 679 (93.5%)**
B. Simpler algorithm negative (n=172)		
Model-based algorithm negative n=144 (83.7%)	**Model-based algorithm positive n=28 (16.3%)**	
120 (83.3%)	6 (21.4%)	**Manual abstraction negative n=126 (73.3%)**
24 (16.7%)	22 (78.6%)	**Manual abstraction positive n=46 (26.7%)**
Simpler algorithm positive (n=40)		
Model-based algorithm negative n=0 (0.0%)	**Model-based algorithm positive n=40 (100.0%)**	
0 (0.0%)	11 (27.5%)	**Manual abstraction negative n=11 (27.5%)**
0 (0.0%)	29 (72.5%)	**Manual abstraction positive n=29 (72.5%)**

Percentages provided in each cell represent column percentages.
PAD, peripheral arterial disease.

Table 6 Comparison of model-based algorithm versus simpler algorithm: community-based sample not evaluated in the vascular laboratory

Gold standard=Manual chart abstraction		
	Model-based algorithm	Simpler algorithm (code 440.20 to 440.29)
Sensitivity, % (95% CI)	68.0 (56.2 to 78.3)	38.7 (27.6 to 50.6)
Specificity, % (95% CI)	87.6 (80.9 to 92.6)	92.0 (86.1 to 95.9)
PPV, % (95% CI)	75.0 (63.0 to 84.7)	72.5 (56.1 to 85.4)
NPV, % (95% CI)	83.3 (76.2 to 89.0)	73.3 (66.0 to 79.7)

CI calculated by binomial exact distribution.
NPV, negative predictive value; PPV, positive predictive value.

'arteriosclerosis obliterans with gangrene of extremities', 'intermittent claudication', 'ulceration', or 'rest pain' were higher than for procedural codes. Although most revascularization procedures of lower extremity were done for PAD, no such code remained in the final model, probably because the frequency of these procedures in the vascular laboratory population was low.

To evaluate the generalizability of our findings we applied both the model-based billing code algorithm and the simpler algorithm consisting of PAD diagnostic codes to a community-based sample of Olmsted County residents who were not referred to the Mayo vascular laboratory. We used the resources of the REP[23] to identify this sample. The REP enables population-based studies of disease risk factors, incidence, and outcomes.[23 25 26] Results from studies in Olmsted County have been consistent with national data.[26] Age, sex, and ethnic characteristics of Olmsted County residents were similar to those of the state of Minnesota and the upper Midwest. However, Olmsted County residents were less ethnically diverse than the entire US population, more highly educated, and wealthier.[26] These insights may guide the generalization of our findings. However, we should emphasize that no single community in the USA is completely representative of the entire USA. Therefore no specific geographic unit can claim any better level of representativeness.[26]

Additional work is needed to assess the performance of algorithms for identifying PAD cases and controls at other institutions and to assess the incremental value of additional data sources to enhance the PPV and NPV of EMR-based methods to ascertain PAD. For example, we have previously shown that natural language processing can be used to mine the text in radiology reports to identify cases of PAD.[27]

Strengths and limitations

Our study has several strengths. These include the large Mayo Clinic administrative dataset, and the availability of results of non-invasive lower extremity arterial testing performed in the accredited vascular laboratory at Mayo. The initial dataset included all the encounters and line item details of patients with potential PAD over an 11-year period in one medical center. All PAD-related ICD-9-CM diagnosis/procedure codes and CPT-4 procedural codes with updates of new treatment procedures in the time of interest were included, minimizing the possibility of underdiagnosis. As is true for most common diseases, PAD is not always diagnosed using uniform criteria in the present coding systems. This might lead to labeling of patients with different pathological processes as having the same disease. We identified non-atherosclerotic causes of vascular disease and lower extremity treatment procedures unrelated to PAD using relevant billing codes.

Limitations of this study need to be noted. First, diagnosis/procedural codes are assigned for billing and reimbursement. While codes are accurate and complete for these purposes, they

may be limited for clinical and research purposes. As for other studies that identify diseases using billing codes and retrospective data over a long time period, the accuracy might have been influenced by bias in coding, modification of disease definition and coding system, the accuracy of recorded data, and potential distortion of information for non-clinical reasons.[28] Second, the data for this study were retrieved from databases of a single medical center and the algorithm is likely to perform differently in other medical settings as a result of regional and institutional variations in coding accuracy and thoroughness. The logistic regression model may have to be calibrated at each center.

CONCLUSION

In conclusion, we describe an approach to searching administrative datasets for PAD status using billing codes. The model-based billing code algorithm had reasonable accuracy for identification of PAD status of those in the community, and in patients who were assessed in the non-invasive vascular laboratory. The simpler algorithm based on the presence of any one of the commonly used ICD-9 codes for PAD had reasonable accuracy for identification of PAD in patients evaluated in the vascular laboratory but was significantly less sensitive in a community-based sample. Our approach provides a strategy for rapidly ascertaining PAD status for genetic association and epidemiological studies of PAD.

Contributors IJK, CLL and KRB defined the research theme. IJK designed the methods, obtained funding, interpreted the results, and drafted the manuscript. JF participated in the study design, interpreted the results and drafted the manuscript. AMA-O contributed to the analysis, methods, and interpretation of the results and drafted the revised manuscript. GL and CS performed the statistical analysis. CLL, CS and KRB co-designed the methods, discussed analyses, interpretation, and presentation. All authors have contributed to, seen and approved, the manuscript.

Funding This work was supported by grant HG06379 from the National Human Genome Research Institute. JF was supported by Chinese PLA General Hospital Young Scientist Training Program. This study was made possible using the resources of the Rochester Epidemiology Project, which is supported by the National Institute on Aging of the National Institutes of Health under Award Number R01AG034676. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests None.

Ethics approval Ethics approval was provided by the institutional review board of the Mayo Clinic.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Allison MA, Ho E, Denenberg JO, *et al*. Ethnic-specific prevalence of peripheral arterial disease in the United States. *Am J Prev Med* 2007;32:328–33.
2 Hirsch AT, Criqui MH, Treat-Jacobson D, *et al*. Peripheral arterial disease detection, awareness, and treatment in primary care. *JAMA* 2001;286:1317–24.
3 Klabunde CN, Warren JL, Legler JM. Assessing comorbidity using claims data: an overview. *Med Care* 2002;40(8 Suppl):IV-26–35.
4 Kullo IJ, Fan J, Pathak J, *et al*. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *J Am Med Inform Assoc* 2010;17:568–74.
5 Hsia DC, Krushat WM, Fagan AB, *et al*. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. *N Engl J Med* 1988;318:352–5.
6 Petersen LA, Wright S, Normand SLT, *et al*. Positive predictive value of the diagnosis of acute myocardial infarction in an administrative database. *J Gen Intern Med* 1999;14:555–8.
7 Ratelle S, Yokoe D, Blejan C, *et al*. Predictive value of clinical diagnostic codes for the CDC case definition of pelvic inflammatory disease (PID): implications for surveillance. *Sexually Transm Dis* 2003;30:866.
8 Anaya DA, Becker NS, Richardson P, *et al*. Use of administrative data to identify colorectal liver metastasis. *J Surg Res* 2012;176:141–6.
9 Ducharme R, Benchimol EI, Deeks SL, *et al*. Validation of diagnostic codes for intussusception and quantification of childhood intussusception incidence in Ontario, Canada: a population-based study. *J Pediatr* 2013;163:1073–9 (e3).
10 Lix LM, Yogendran MS, Leslie WD, *et al*. Using multiple data features improved the validity of osteoporosis case ascertainment from administrative databases. *J Clin Epidemiol* 2008;61:1250–60.
11 Arain FA, Ye Z, Bailey KR, *et al*. Survival in patients with poorly compressible leg arteries. *J Am Coll Cardiol* 2012;59:400–7.
12 International classification of diseases, ninth revision, clinical modification: U.S. Dept. of Health and Human Services, Centers for Disease Control and Prevention, Centers for Medicare and Medicaid Services. 2008.
13 Association. AM. CPT. *Physicians' current procedural terminology*. Chicago: American Medical Association, 1997–2008.
14 Pippenger M, Holloway RG, Vickrey BG. Neurologists'use of ICD-9CM codes for dementia. *AAN Enterprises* 2001;1206–9.
15 Logar CM, Pappas LM, Ramkumar N, *et al*. Surgical revascularization versus amputation for peripheral vascular disease in dialysis patients: a cohort study. *BMC Nephrol* 2005;6:3.
16 Eslami MH, Zayaruzny M, Fitzgerald GA. The adverse effects of race, insurance status, and low income on the rate of amputation in patients presenting with lower extremity ischemia. *J Vasc Surg* 2007;45:55–9.
17 Gasse C, Jacobsen J, Larsen AC, *et al*. Secondary medical prevention among Danish patients hospitalised with either peripheral arterial disease or myocardial infarction. *Eur J Vasc Endovasc Surg* 2008;35:51–8.
18 Kezerashvili A, Delaney J, Schaefer MJ, *et al*. The impact of co-morbid peripheral vascular disease on mortality in a racially balanced cohort with congestive heart failure. *J Card Fail* 2008;14(6S):83.
19 Filippi A, Paolini I, Innocenti F, *et al*. Blood pressure control and drug therapy in patients with diagnosed hypertension: a survey in Italian general practice. *J Hum Hypertens* 2009;23:758–63.
20 Singh M, Rihal CS, Selzer F, *et al*. Validation of Mayo Clinic risk adjustment model for in-hospital complications after percutaneous coronary interventions, using the National Heart, Lung, and Blood Institute dynamic registry. *J Am Coll Cardiol* 2003;42:1722–8.
21 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.
22 Beck JR, Shultz EK. The use of relative operating characteristic (ROC) curves in test performance evaluation. *Arch Pathol Lab Med* 1986;110:13–20.
23 St Sauver JL, Grossardt BR, Yawn BP, *et al*. Data resource profile: the Rochester Epidemiology Project (REP) medical records-linkage system. *Int J Epidemiol* 2012;41:1614–24.
24 Hirsch AT, Haskal ZJ, Hertzer NR, *et al*. ACC/AHA 2005 Practice Guidelines for the management of patients with peripheral arterial disease (lower extremity, renal, mesenteric, and abdominal aortic): a collaborative report from the American Association for Vascular Surgery/Society for Vascular Surgery, Society for Cardiovascular Angiography and Interventions, Society for Vascular Medicine and Biology, Society of Interventional Radiology, and the ACC/AHA Task Force on Practice Guidelines (Writing Committee to Develop Guidelines for the Management of Patients With Peripheral Arterial Disease): endorsed by the American Association of Cardiovascular and Pulmonary Rehabilitation; National Heart, Lung, and Blood Institute; Society for Vascular Nursing; TransAtlantic Inter-Society Consensus; and Vascular Disease Foundation. *Circulation* 2006;113:e463–654.
25 Rocca WA, Yawn BP, St Sauver JL, *et al*. 3rd. History of the Rochester Epidemiology Project: half a century of medical records linkage in a US population. *Mayo Clin Proc* 2012;87:1202–13.
26 St Sauver JL, Grossardt BR, Leibson CL, *et al*. Generalizability of epidemiological findings and public health decisions: an illustration from the Rochester Epidemiology Project. *Mayo Clin Proc* 2012;87:151–60.
27 Savova GK, Fan J, Ye Z, *et al*. Discovering peripheral arterial disease cases from radiology notes using natural language processing. *AMIA Annu Symp Proc* 2010;2010:722–6.
28 Tirschwell DLMD, Longstreth WTJMD. Validating administrative data in stroke research. *Stroke* 2002;33:2465–70.