Development and Validation of an Electronic Medical Record Algorithm to Identify Phenotypes of Rotator Cuff Tear

Chan Gao, MD PhD[1], **Run Fan, PhD**[2], **Gregory D. Ayers, MS**[2], **Ayush Giri, PhD**[4], **Kindred Harris, MD**[1,5], **Ravi Atreya, MD PhD**[3], **Pedro L. Teixeira, MD PhD**[3], **Nitin B. Jain, MD MSPH**[1,4,6]

[1.]Department of Physical Medicine and Rehabilitation, Vanderbilt University Medical Center

[2.]Department of Biostatistics, Vanderbilt University Medical Center

[3.]Dept of Biomedical Informatics, Vanderbilt University Medical Center

[4.]Division of Epidemiology, Vanderbilt University Medical Center

[5.]Faculty of Medicine, University of California (Los Angeles)

[6.]Department of Orthopaedics and Rehabilitation, Vanderbilt University Medical Center

Abstract

Background: A lack of studies with large sample sizes of patients with rotator cuff tears is a barrier to performing clinical and genomic research.

Objective: To develop and validate an electronic medical record (EMR)-based algorithm to identify individuals with and without rotator cuff tear.

Design: We used a de-identified version of the EMR of more than 2 million subjects. A screening algorithm was applied to classify subjects into likely rotator cuff tear and likely normal rotator cuff groups. From these subjects, 500 likely rotator cuff tear and 500 likely normal rotator cuff were randomly chosen for algorithm development. Chart review of all 1,000 subjects confirmed the true phenotype of rotator cuff tear or normal rotator cuff based on magnetic resonance imaging and operative report. An algorithm was then developed based on logistic regression and validation of the algorithm was performed.

Results: The variables significantly predicting rotator cuff tear included the number of times a Current Procedural Terminology code related to rotator cuff procedures was used (OR=3.3; 95% CI: 1.6–6.8 for ≥3 versus 0), the number of times a term related to rotator cuff lesions occurred in radiology reports (OR=2.2; 95% CI: 1.2–4.1 for ≥1 versus 0), and the number of times a term related to rotator cuff lesions occurred in physician notes (OR=4.5; 95% CI: 2.2–9.1 for 1 or 2 times versus 0). This phenotyping algorithm had a specificity of 0.89 (95% CI: 0.79–0.95) for rotator cuff tear, AUC of 0.842, and diagnostic likelihood ratios, DLR+ and DLR- of 5.94 (95%CI: 3.07–11.48) and 0.363 (95%CI: 0.291 – 0.453).

Corresponding Author: Nitin B. Jain, MD, MSPH, Department of Physical Medicine and Rehabilitation, Vanderbilt University Medical Center, 2201 Children's Way, Suite 1318, Nashville, TN 37212, United States of America, Phone : (615) 936-8508; Fax: (615) 322-7454, nitin.jain@vumc.org.

Conclusion: Our informatics algorithm enables identification of cohorts of individuals with and without rotator cuff tear from an EMR-based dataset with moderate accuracy.

Keywords

Algorithm; phenotyping; rotator cuff tear; electronic medical record; machine learning; logistic regression

LOE:

Level II

INTRODUCTION

Rotator cuff disease, a spectrum of disorders including tendinopathy, partial- and full-thickness tears, is estimated to affect 30% to 50% of the population older than 50 years [10]. Rotator cuff tears may lead to disabling shoulder pain, significant muscular atrophy, and humeral head migration[15]. The mechanisms and causes of symptomatic rotator cuff tears are multifactorial. There is likely a genetic predisposition to the etiology of rotator cuff tears [2,12,16]. A genome-wide association study (GWAS) that compares patients with rotator cuff tears to those without tears is needed to identify genetic predictors of rotator cuff tear. However, GWAS typically requires sample sizes of many thousands of individuals to make definitive conclusions. Recruitment of such a large cohort of patients is challenging and cost prohibitive. Hence, the use of de-identified electronic medical records (EMR) linked to biorepositories of patient genotype data is a more feasible alternative. An EMR-based informatics algorithm that identifies individuals with and without rotator cuff tear with good precision is needed for this purpose but is unavailable. EMR-based informatics algorithms have been developed to identify cases and controls for various other medical conditions [6].

The objective of our study was to develop an EMR-based algorithm to accurately identify phenotypes of patients with and without rotator cuff tears and validate it against a gold standard diagnosis of rotator cuff tear by chart review of information extracted from either magnetic resonance imaging (MRI) or operative reports. This algorithm could be applied to EMR settings and yield a large number of individuals with and without a tear.

MATERIALS AND METHODS

EMR database

An overview of our study design is shown in Figure 1. The synthetic derivative (SD) is the de-identified version of the EMR used at Vanderbilt University Medical Center (VUMC) [4]. Data of individuals from the SD as of 01/01/2017 since its inception in the mid-1990s was extracted following approval from the VUMC Institutional Review Board. Since we were interested in individuals with degenerative cuff tears, only those between the ages of 40 to 75 were included. The subjects were screened by simple algorithms and classified into two groups: likely rotator cuff tear and likely normal rotator cuff (Table S1). To be considered in the likely rotator cuff tear group, a subject had at least one of the following characteristics:

1. rotator cuff disorder-related International Classification of Disease 9th version (ICD-9), ICD 10th version (ICD-10), or Current Procedure Terminology (CPT) billing code (Table S2); 2. Use of the term "rotator cuff tear" or "rotator cuff repair" in either one of the radiology reports or one of the rehabilitation therapy notes. If patients did not satisfy the above criteria OR had explicit mention of "normal rotator cuff" in one of the radiology reports, physician note, or rehabilitation therapy note, they were categorized as a likely normal rotator cuff.

A total of 14,690 likely rotator cuff tear and 1,066,053 likely normal rotator cuff patients were extracted based on the above screening algorithm from SD. From this set of patients, 500 patients with likely rotator cuff tear and 500 patients with likely normal rotator cuff were randomly selected for algorithm development. The same number of likely rotator cuff tear and likely normal rotator cuff patients were randomly selected for algorithm validation. Patients in the algorithm development and validation sub-sets did not overlap.

Variables for algorithm development

Eight variables were selected for building the phenotyping model based on their high clinical relevance, availability across EMR systems, and non-overlapped utilization in documentation. These variables were derived from 3 categories: demographic features, billing codes and regular expression (RegEx). Demographic features were patient age (at the time of data extraction) and sex. The prevalence of rotator cuff tear increases with age [18,19]. Therefore, age was *a priori* included in our algorithm. Sex was included due to its biological importance and relationship with severity of symptoms of rotator cuff tear [8]. Billing codes included ICD-9 and CPT codes. ICD-10 codes were not included in algorithm development because the number of patients enrolled between 10/2015 (the time of transition to ICD-10 codes) and 12/2016 represented a small portion of the dataset. The number of times an ICD-9 code related to rotator cuff disorders appeared in a patient's medical record (ICD-9) and the number of times a CPT code related to rotator cuff repair was found in a patient's EMR (CPT) were calculated (Table S3). RegEx is a computer coding algorithm used to find terms such as specific phrases, text, or numbers within reports and notes in EMR (Table S3). In our study, RegEx was used to search physician notes, any of the radiology reports, shoulder MRI reports (a subset of radiology reports), and operative reports (a subset of physician notes). The MRI report was categorized independently even though it was included in radiology reports to assess its unique prediction value of rotator cuff tear. Similarly, operative report was assessed as an independent variable although it was included in physician notes. The specific terms that were searched in narrative documents are listed in the supplement (Table S3). The number of times that a given term appeared in the medical record was calculated for MRI report (MRI), physician notes (H&P), operative report (Op), and any of the radiology reports (Rad).

Chart Review

Individual chart review was used as the gold standard to ascertain whether a patient had a rotator cuff tear or not. The manual chart review was based on keyword searching of "supraspinatus", "infraspinatus", "subscapularis", "teres", "rotator", "MRI", and "operative" in the radiology reports (including MRI reports) and physician notes (including operative

reports). Phenotypes of rotator cuff tendons were determined by careful review of the paragraphs where the keywords were located. Patients with an MRI or operative report documenting rotator cuff tendon(s) tear or surgical history of rotator cuff repair were classified as true rotator cuff tear. Patients with documentation of MRI or operative findings of rotator cuff tear in physician notes, despite a lack of formal MRI or operative reports, were also classified as true rotator cuff tear. If the MRI or operative report documented an intact rotator cuff, such patients were classified as true normal rotator cuff. Imaging or operative documentation of an intact cuff is essential to classify patients as normal rotator cuff since imaging data shows that 40–65% of individuals aged 50 years without shoulder symptoms have a rotator cuff tear [13]. Hence, there is a high risk for misclassification if imaging confirmation is not sought to determine whether a rotator cuff is normal. Patients without documentation of shoulder MRI or operative findings in the EMR were excluded from our study since the status of the rotator cuff cannot be determined with high confidence in such individuals.

Statistical Analysis

Among the 1,000 randomly selected patients for model development, 293 patients were assigned a definitive diagnosis of either true rotator cuff tear (n=193) or normal rotator cuff (n=100) according to the criteria described in the Chart Review section. The probability of rotator cuff tear was modelled using logistic regression. A redundancy analysis identifies variables that can be predicted from the remaining variables in the analysis. Redundancy analysis (elimination R^2=0.9) did not eliminate any of the original 8 candidate predictor variables. Predictor effects are presented as odds ratios with 95% confidence intervals. Internal model validation with bootstrap corrected standard diagnostic statistics (e.g. c-index) were conducted according to Harrell [14]. A full model with all the 8 variables and a more parsimonious clinical model with the 5 most significantly associated variables were estimated and internally validated in a similar manner. Nomograms for scoring individual patient probability of a rotator cuff tear were created.

In our validation test set, 263 patients were definitively diagnosed with either rotator cuff tear (n=193) or normal rotator cuff (n=70). The clinical prediction model developed above was applied to the validation test set to estimate the probability of a rotator cuff tear. Using individual chart review as the gold standard, calibration curves, the receiver operating characteristic curve (ROC), and area under the curve were evaluated for the predictive model using the validation test set. Sensitivity, specificity, and diagnostic likelihood ratio (DLR) were estimated at the maximum value of the Youden index along the ROC. The maximum value of the Youden index provides the optimal cut-off for the probability of a rotator cuff tear. DLR+ is defined as sensitivity/1-specificity and is equivalent to the probability of true positive/false positive. DLR- is defined as 1-sensitivity/specificity and is equivalent to the probability of false negative/true negative. DLRs quantify the post-test odds of disease and do not depend on the disease prevalence in the population. A perfect test has a (DLR+, DLR-) of (∞,0).

RESULTS

Chart review

There were 14,690 likely rotator cuff tear and 1,066,053 likely normal rotator cuff patients after the entire SD dataset was screened. A case-by-case chart review of the randomly selected 500 likely rotator cuff tear and 500 likely normal rotator cuff patients confirmed 193 with true rotator cuff tear, 100 with true normal rotator cuff, and 707 subjects where rotator cuff tear or normal rotator cuff status could not be definitively determined due to unavailability of imaging or operative information. Of the 193 true rotator cuff tear cases, 182 patients were derived from the likely rotator cuff tear sub-set and 11 from the likely normal rotator cuff sub-set. Of the 100 true normal rotator cuff individuals confirmed by chart review, 12 were derived from the likely normal rotator cuff sub-set and 88 from the likely rotator cuff tear sub-set. Of 707 excluded individuals, 230 were from the likely rotator cuff tear and 477 were from the likely normal rotator cuff. Demographics and other characteristics of rotator cuff tear and normal rotator cuff patients are presented in Table 1. A majority of true normal rotator cuff patients had tendinopathy (n=46) or their imaging/ operative report was normal/nonspecific (n=24) (Figure S1).

Informatics algorithm development

In the parsimonious clinical model, in addition to age and sex, CPT (OR=3.3; 95% CI=1.6–6.8 for 3 or more times versus none, and OR=2.8; 95% CI=1.1–7.5 for 1 or 2 times versus none), Rad (OR=2.2; 95%CI: 1.2–4.1 for 1 or more times versus none), and H&P (OR=4.5; 95%CI: 2.2–9.1 for 1 or 2 times versus none, and OR=23.1; 95%CI: 9.5–56.5 for 3 or more times versus none) were significant predictors of the rotator cuff tear phenotype. (Table S4). H&P had the highest contribution towards predicting rotator cuff tear followed by CPT (Figure 2). For ease of use in clinical settings, a nomogram was plotted with each of the significant variables in our clinical model to calculate the probability of rotator cuff tear (Figure S2). The clinical model had good internal validation characteristics and showed satisfactory overlap of predicted versus actual values for rotator cuff tear and a bootstrap corrected c-index (AUC) of 0.833 (Figure 3 and Table S5).

Informatics algorithm validation

We describe the value of the parsimonious clinical model, via validation in a different set of patients, using a panel of accuracy measures. AUC was 0.842 for the ROC curve. A candidate optimal cutoff probability for the diagnosis of rotator cuff tear at the maximum Youden index along the model ROC curve was at a sensitivity of 0.68 (95% CI: 0.61–0.74) and specificity of 0.89 (95% CI: 0.79–0.95) for rotator cuff tear (Figure 4). A perfect algorithm will predict disease perfectly with both sensitivity and specificity equal to 1. The diagnostic likelihood ratios, DLR+ and DLR-, were 5.94 (95% CI: 3.07–11.48) and 0.363 (95% CI: 0.291–0.453). Assuming the prevalence of a rotator cuff tear is 73% (validation data), the pre-test odds of disease is 0.73/0.27=2.7. With a DLR+=5.94, the post-test (algorithm) odds of a rotator cuff tear at the Youden cutoff of the modelling algorithm is 2.7×5.94=16.04.

A f*ull model* with all variables (age, sex, ICD-9, CPT, H&P, Rad, MRI, and Op) considered for model building achieved similar internal validation characteristics to the *clinical model* with a bootstrap corrected c-index (AUC) of 0.823 (Figures S3–S5 and Table S6).

DISCUSSION

We developed an informatics-based algorithm to identify individuals with the phenotypes of interest, rotator cuff tear and normal rotator cuff, from an EMR-based dataset. We present a final algorithm with 5 variables (age, sex, CPT, H&P, and Rad) that can be used to classify patients with and without tears in EMR with a high degree of confidence. Our clinical model showed good validation characteristics. Of note, our model showed stronger specificity (0.89, 95% CI: 0.79–0.95) than sensitivity (0.68, 95% CI: 0.61–0.74) which suggests a stronger capacity to "rule in" rotator cuff tear than "rule out" rotator cuff tear. This is further corroborated by the high value of DLR+ 5.94 (95% CI: 3.07–11.5). This algorithm can be used in future GWAS and clinical studies where large sample sizes of patients with rotator cuff tears are needed from the EMR. Similar informatics-based algorithms have been developed for various phenotypes including hypertension, systemic lupus erythematosus, and rheumatoid arthritis [1,5,17].

Predicting rotator cuff tear by data mining of clinical assessments has been reported [7]; however, data on informatics-based algorithms for rotator cuff tear and normal rotator cuff is lacking. Informatics-based algorithms extend beyond the simple use of ICD-9 or ICD-10 billing codes to ascertain cases and controls. ICD codes are often inaccurate and can lead to a high degree of misclassification of cases and controls [9]. ICD codes are commonly used by non-specialists leading to further misclassification. In contrast, CPT codes are surgical codes used by specialists and therefore may be more likely to be more accurate. CPT was found to be significantly associated with the prediction of rotator cuff tear in our analysis.

The narrative content of clinical notes was significantly associated with the diagnosis of rotator cuff tear in our study. Previous studies have shown the importance of narrative content of clinical notes in developing algorithms for other medical conditions [5,17]. Clinical notes that yielded significant narrative content in our final algorithm were physician notes and radiology reports. It was surprising that the narrative content of the MRI report and the operative report were not significant predictors of rotator cuff tear when assessed independently. This is likely because information from MRI reports was captured in the Rad variables and that for the Op variable was captured in physician notes. This implies that all radiology reports and physician notes should be assessed when searching for samples of patients with and without tears from EMR, and MRI and operative reports do not alone provide accurate results.

Our algorithm is based on records where an MRI or operative report was available. Hence, the possibility of spectrum bias exists where less severe cases of rotator cuff tears who did not require MRI/surgery would be excluded from our study. Although currently available data suggests that patient symptoms do not correlate with severity of structural findings on MRI (such as tear size or muscle atrophy) [3], we recognize this potential limitation of our study. However, it would be difficult to establish a gold standard for rotator cuff diagnosis

without documented MRI or operative findings. The algorithm has sub-optimal diagnostic accuracy for detecting patients with normal rotator cuff. To address this issue, a separate algorithm for "normal rotator cuff" should be created in future studies.

Our study has some other limitations. The etiology of rotator cuff tear can be traumatic or degenerative which implies different pathogenesis, clinical progression, and management [11]; however, the algorithm presented in this study does not account for traumatic versus non-traumatic etiology. When the study was initiated, ICD-10 codes were just introduced into clinical practice. Hence, very little data was available for the association of ICD-10 with the phenotype of rotator cuff and these codes were not included in our algorithm. Although we performed validation of our algorithm by using a different dataset, our model was not validated in a dataset outside our institution. Lastly, we were only able to confirm an intact cuff in patients classified as normal rotator cuff on one shoulder and the status of the contralateral shoulder was unknown.

We present a previously unavailable informatics algorithm for accurate identification of rotator cuff tear and normal rotator cuff from an EMR-based dataset with good validation characteristics. Age, sex, CPT, H&P, and Rad were significantly associated with the presence or absence of a rotator cuff tear. This algorithm can be used for rapid identification of patients with and without rotator cuff tears in EMR settings.

CONCLUSION

Our informatics algorithm enables accurate and rapid identification of cohorts of individuals with and without rotator cuff tear from an EMR-based dataset. Our algorithm has good internal and external validity and hence can be applied to EMR in other settings and enable large genomic and clinical studies in this area.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

FUNDING and ACKNOWLEDGEMENT

References

1. Barnado A, Casey C, Carroll RJ, Wheless L, Denny JC, Crofford LJ. Developing Electronic Health Record Algorithms That Accurately Identify Patients With Systemic Lupus Erythematosus. Arthritis care & research. 2017;69(5):687–93. DOI: 10.1002/acr.22989 [PubMed: 27390187]

2. Dabija DI, Gao C, Edwards TL, Kuhn JE, Jain NB. Genetic and familial predisposition to rotator cuff disease: a systematic review. Journal of shoulder and elbow surgery. 2017;26(6):1103–12. DOI: 10.1016/j.jse.2016.11.038 [PubMed: 28162885]

3. Dunn WR, Kuhn JE, Sanders R, An Q, Baumgarten KM, Bishop JY, et al. Symptoms of pain do not correlate with rotator cuff tear severity: a cross-sectional study of 393 patients with a symptomatic

atraumatic full-thickness rotator cuff tear. Journal of bone and joint surgery (American volume). 2014; 96 (10): 793–800. DOI: 10.2106/JBJS.L.01304

4. Harris PA, Swafford JA, Edwards TL, Zhang M, Nigavekar SS, Yarbrough TR, et al. StarBRITE: the Vanderbilt University Biomedical Research Integration, Translation and Education portal. Journal of biomedical informatics. 2011;44(4):655–62. DOI: 10.1016/j.jbi.2011.01.014 [PubMed: 21310264]

5. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis care & research. 2010;62(8):1120–7. DOI: 10.1002/acr.20184 [PubMed: 20235204]

6. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ : British Medical Journal. 2015;350:h1885. doi: 10.1136/bmj.h1885 [PubMed: 25911572]

7. Lu HY, Huang CY, Su CT, Lin CC. Predicting rotator cuff tears using data mining and Bayesian likelihood ratios. PloS one. 2014;9(4): e94917 DOI: 10.1371/journal.pone.0094917 [PubMed: 24733553]

8. Maher A, Leigh W, Brick M, Young S, Millar J, Walker C, et al. Gender, ethnicity and smoking affect pain and function in patients with rotator cuff tears. ANZ journal of surgery. 2017;87(9):704–8. DOI: 10.1111/ans.13921 [PubMed: 28702950]

9. Moores KG, Sathe NA. A systematic review of validated methods for identifying systemic lupus erythematosus (SLE) using administrative or claims data. Vaccine. 2013;31 Suppl 10:K62–73. DOI: 10.1016/j.vaccine.2013.06.104 [PubMed: 24331075]

10. Motta Gda R, Amaral MV, Rezende E, Pitta R, Vieira TC, Duarte ME, et al. Evidence of genetic variations associated with rotator cuff disease. Journal of shoulder and elbow surgery. 2014;23(2):227–35. DOI: 10.1016/j.jse.2013.07.053 [PubMed: 24129055]

11. Pappou IP, Schmidt CC, Jarrett CD, Steen BM, Frankle MA. AAOS appropriate use criteria: optimizing the management of full-thickness rotator cuff tears. The Journal of the American Academy of Orthopaedic Surgeons. 2013;21(12):772–5. DOI: 10.5435/JAAOS-21-12-772 [PubMed: 24292934]

12. Roos TR, Roos AK, Avins AL, Ahmed MA, Kleimeyer JP, Fredericson M, et al. Genome-wide association study identifies a locus associated with rotator cuff injury. PloS one. 2017;12(12):e0189317 DOI: 10.1371/journal.pone.0189317 [PubMed: 29228018]

13. Sher JS, Uribe JW, Posada A, Murphy BJ, Zlatkin MB. Abnormal findings on magnetic resonance images of asymptomatic shoulders. The Journal of bone and joint surgery American volume. 1995;77(1):10–5. [PubMed: 7822341]

14. Steyerberg EW, Harrell FE Jr.: Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016, 69:245–7. [PubMed: 25981519]

15. Tashjian RZ. Epidemiology, natural history, and indications for treatment of rotator cuff tears. Clinics in sports medicine. 2012;31(4):589–604. DOI: 10.1016/j.csm.2012.07.001 [PubMed: 23040548]

16. Tashjian RZ, Granger EK, Farnham JM, Cannon-Albright LA, Teerlink CC: Genome-wide association study for rotator cuff tears identifies two significant single-nucleotide polymorphisms. Journal of shoulder and elbow surgery 2016, 25:174–9. [PubMed: 26350878]

17. Teixeira PL, Wei WQ, Cronin RM, Mo H, VanHouten JP, Carroll RJ, et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. Journal of the American Medical Informatics Association: JAMIA. 2017;24(1):162–71. DOI: 10.1093/jamia/ocw071 [PubMed: 27497800]

18. Tempelhof S, Rupp S, Seil R. Age-related prevalence of rotator cuff tears in asymptomatic shoulders. Journal of shoulder and elbow surgery. 1999;8(4):296–9. [PubMed: 10471998]

19. Yamamoto A, Takagishi K, Osawa T, Yanagawa T, Nakajima D, Shitara H, et al. Prevalence and risk factors of a rotator cuff tear in the general population. Journal of shoulder and elbow surgery. 2010;19(1):116–20. DOI: 10.1016/j.jse.2009.04.006 [PubMed: 19540777]

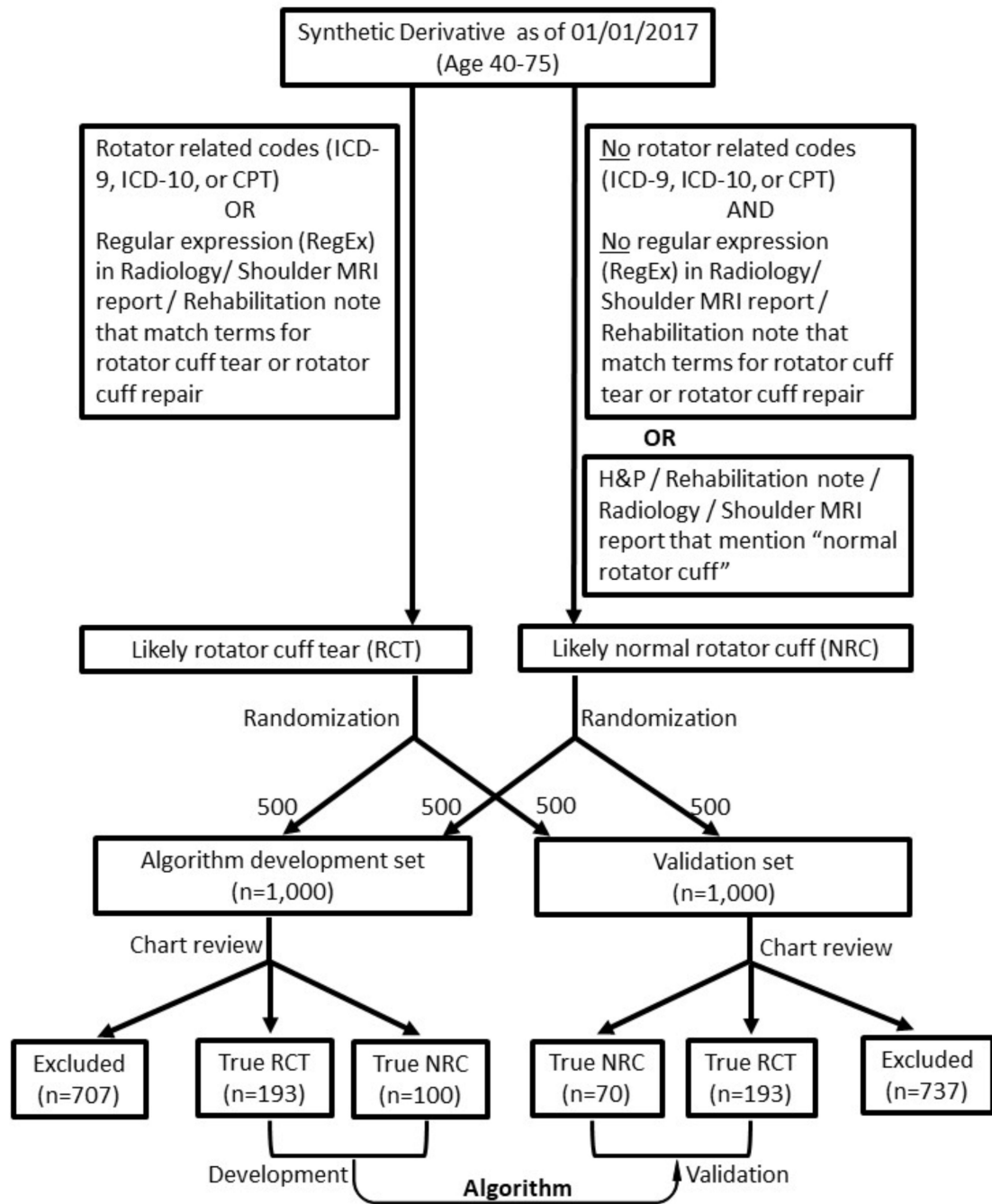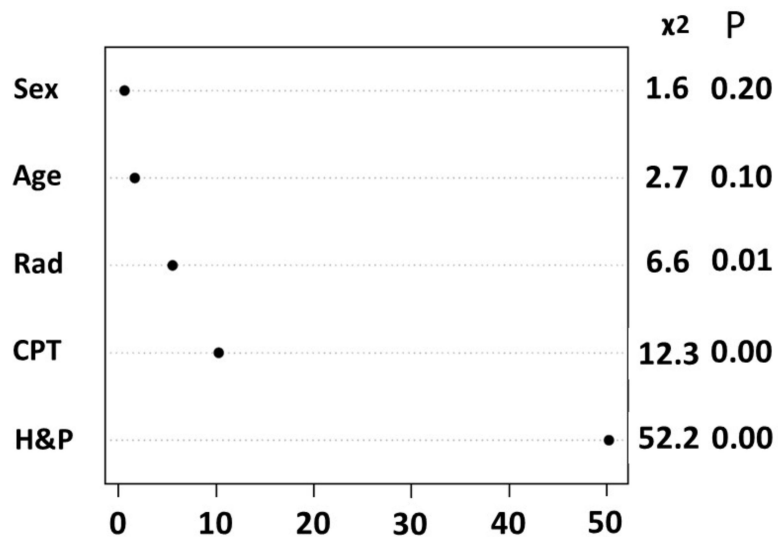**Figure 1.**
Study design

Figure 2: Contribution of variables in predicting rotator cuff tear.
CPT=number of times a CPT code related to rotator cuff repair appeared in patient's medical record; H&P=number of times a rotator cuff lesion/rotator cuff repair-related term appeared in physician notes; Op= number of times a rotator cuff lesion/rotator cuff repair-related term appeared in patient's operative notes; Rad= number of times a rotator cuff lesion/rotator cuff repair-related term appeared in patient's imaging reports; Age in years; Sex (females versus males)
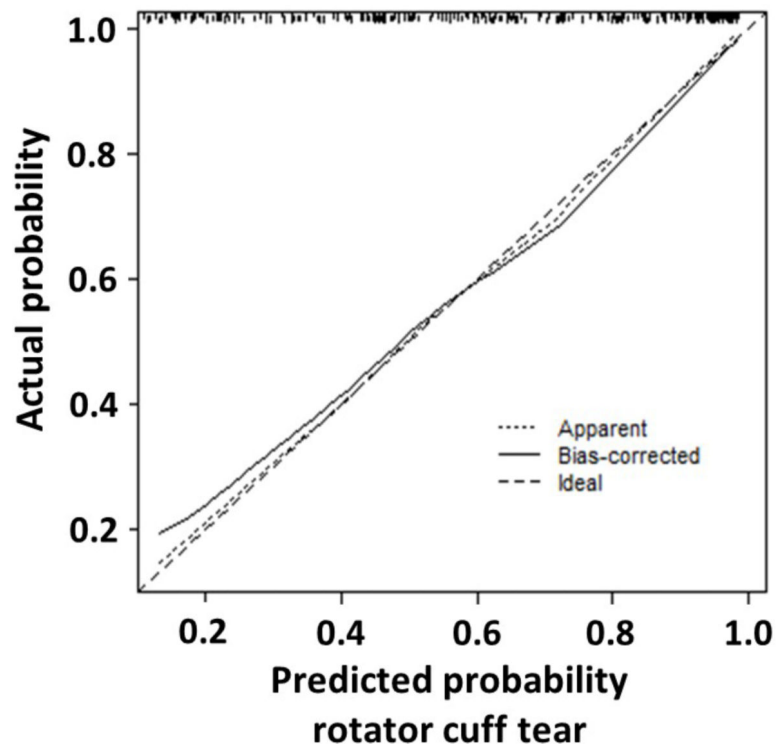
Figure 3: Validation of the clinical model.
Internal validation and calibration.

The hatch marks on top of the plot are the relative frequency of calibrated probabilities after dividing into 101 bins. The apparent line illustrates predicted versus observed values from the model. The bias-corrected line deviates more than the apparent line, showing the degree that the model is over-confident. The ideal line is the optimal scenario if the model had perfect prediction.
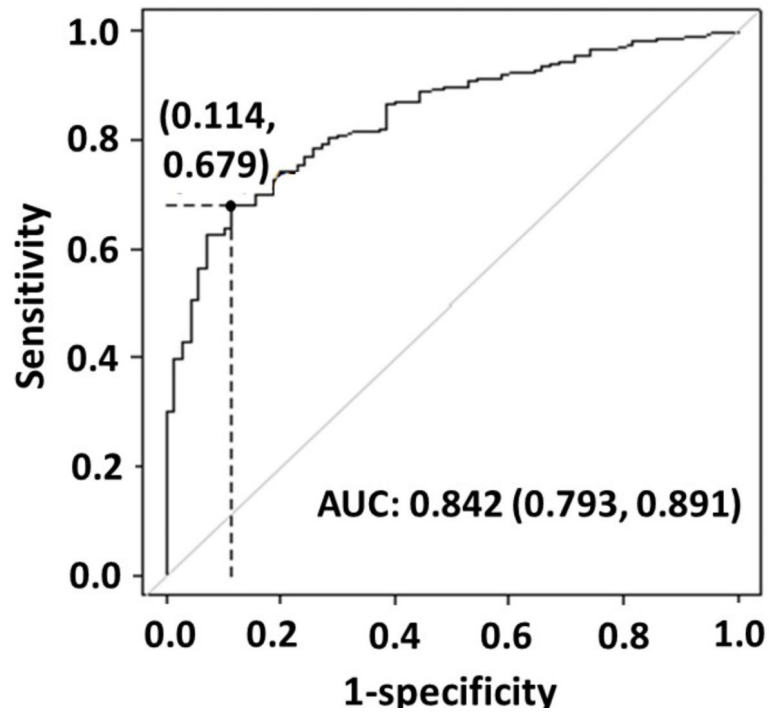
Figure 4: Optimal cut-off probability for diagnosis of rotator cuff tear.
A candidate optimal cutoff probability for the diagnosis of rotator cuff tear at the maximum Youden index along the model ROC curve was at a sensitivity of 0.68 and specificity of 0.89.

Table 1.

Demographics and other characteristics of the patients with and without rotator cuff tears.

PARAMETERS		Training Set			Validation Set		
		Normal Rotator Cuff (N=100)	Rotator Cuff Tears (N=193)	Combined (N=293)	Normal Rotator Cuff (N=70)	Rotator Cuff Tears (N=193)	Combined (N=263)
Age (mean ± standard deviation)		57.9 ± 8.1	60.4 ± 8.8	59.5 ± 8.6	56.3 ± 8.1	61.9 ± 8.7	60.4 ± 8.9
Sex							
	Female	52% (52)	40% (78)	44% (130)	50% (35)	45% (87)	46% (122)
	Male	48% (48)	60% (115)	56% (163)	50% (35)	55% (106)	54% (141)
ICD-9*							
	0	12% (12)	5% (10)	8% (22)	1% (1)	5% (9)	4% (10)
	1 or2	40% (40)	28% (55)	32% (95)	56% (39)	23% (44)	32% (83)
	3 or4	27% (27)	15% (29)	19% (56)	21% (15)	20% (38)	20% (53)
	5	21% (21)	51% (99)	41% (120)	21% (15)	53% (102)	44% (117)
CPT*							
	0	72% (72)	44% (85)	54% (157)	71% (50)	46% (89)	53% (139)
	1 or2	9% (9)	13% (25)	12% (34)	16% (11)	10% (19)	11% (30)
	3	19% (19)	43% (83)	35% (102)	13% (9)	44% (85)	36% (94)
MRI*							
	0	84% (84)	70% (136)	75% (220)	26% (18)	28% (54)	27% (72)
	1	16% (16)	30% (57)	25% (73)	73% (52)	72% (139)	73% (191)
HP*							
	0	70% (70)	23% (45)	39% (115)	57% (40)	16% (31)	27% (71)
	1 or2	23% (23)	24% (47)	24% (70)	37% (26)	29% (56)	31% (82)
	3	7 (7)	52% (101)	37% (108)	6% (4)	55% (106)	42% (110)
OR*							
	0	96% (96)	66% (128)	76% (224)	96% (67)	65% (126)	73% (193)
	1	4% (4)	34% (65)	24% (69)	4% (3)	35% (67)	27% (70)
RR*							
	0	68% (68)	47% (90)	54% (158)	59% (41)	50% (96)	52% (137)
	1	32% (32)	53% (103)	46% (135)	41% (29)	50% (97)	48% (126)

ICD-9=number of times a rotator cuff lesion related ICD-9 code appeared in the patient's medical record; CPT=number of times a CPT code related to rotator cuff repair appeared in patient's medical record; MRI=number of times a rotator cuff lesion/rotator cuff repair related term appeared in patient's MRI reports; H&P=number of times a rotator cuff lesion/rotator cuff repair related term appeared in physician notes; Op= number of times a rotator cuff lesion/rotator cuff repair related term appeared in patient's operative notes; Rad= number of times a rotator cuff lesion/rotator cuff repair related term appeared in patient's imaging reports