Multivariable prediction model for suspected giant cell arteritis: development and validation

ORIGINAL RESEARCH

Multivariable prediction model for suspected giant cell arteritis: development and validation

Edsel B Ing,[1] Gabriela Lahaie Luna,[2] Andrew Toren,[3] Royce Ing,[4] John J Chen,[5] Nitika Arora,[6] Nurhan Torun,[7] Otana A Jakpor,[8] J Alexander Fraser,[9] Felix J Tyndel,[10] Arun NE Sundaram,[10] Xinyang Liu,[11] Cindy TY Lam,[1] Vivek Patel,[12] Ezekiel Weis,[13] David Jordan,[14] Steven Gilberg,[14] Christian Pagnoux,[15] Martin ten Hove[2]

[1]Department of Ophthalmology and Vision Sciences, University of Toronto Medical School, Toronto, [2]Department of Ophthalmology, Queen's University, Kingston, ON, [3]Department of Ophthalmology, University of Laval, Quebec, QC, [4]Toronto Eyelid, Strabismus and Orbit Surgery Clinic, Toronto, ON, Canada; [5]Mayo Clinic, Department of Ophthalmology and Neurology, [6]Mayo Clinic, Department of Ophthalmology, Rochester, MN, [7]Department of Surgery, Division of Ophthalmology, Harvard Medical School, Boston, MA, [8]Harvard Medical School, Boston, MA, USA; [9]Department of Clinical Neurological Sciences and Ophthalmology, Western University, London, [10]Department of Medicine, University of Toronto Medical School, Toronto, ON, Canada; [11]Department of Medicine, Fudan University Shanghai Medical College, Shanghai, People's Republic of China; [12]Roski Eye Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA; [13]Departments of Ophthalmology, Universities of Alberta and Calgary, Edmonton and Calgary, AB, [14]Department of Ophthalmology, University of Ottawa, Ottawa, ON, [15]Vasculitis Clinic, Mount Sinai Hospital, Toronto, ON, Canada

Correspondence: Edsel B Ing
Department of Ophthalmology and Vision Sciences, University of Toronto Medical School, Michael Garron Hospital, 650 Sammon Avenue, K306, Toronto, ON M4C 5M5, Canada
Tel +1 416 465 7900
Fax +1 416 465 2035
Email edinglidstrab@gmail.com

Purpose: To develop and validate a diagnostic prediction model for patients with suspected giant cell arteritis (GCA).

Methods: A retrospective review of records of consecutive adult patients undergoing temporal artery biopsy (TABx) for suspected GCA was conducted at seven university centers. The pathologic diagnosis was considered the final diagnosis. The predictor variables were age, gender, new onset headache, clinical temporal artery abnormality, jaw claudication, ischemic vision loss (VL), diplopia, erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), and platelet level. Multiple imputation was performed for missing data. Logistic regression was used to compare our models with the non-histologic American College of Rheumatology (ACR) GCA classification criteria. Internal validation was performed with 10-fold cross validation and bootstrap techniques. External validation was performed by geographic site.

Results: There were 530 complete TABx records: 397 were negative and 133 positive for GCA. Age, jaw claudication, VL, platelets, and log CRP were statistically significant predictors of positive TABx, whereas ESR, gender, headache, and temporal artery abnormality were not. The parsimonious model had a cross-validated bootstrap area under the receiver operating characteristic curve (AUROC) of 0.810 (95% CI =0.766–0.854), geographic external validation AUROC's in the range of 0.75–0.85, calibration p_{H-L} of 0.812, sensitivity of 43.6%, and specificity of 95.2%, which outperformed the ACR criteria.

Conclusion: Our prediction rule with calculator and nomogram aids in the triage of patients with suspected GCA and may decrease the need for TABx in select low-score at-risk subjects. However, misclassification remains a concern.

Keywords: temporal artery biopsy, diagnosis, prediction rule, nomogram, giant cell arteritis, validation

Introduction

Giant cell arteritis (GCA) is the most common systemic vasculitis in the elderly, and may result in irreversible blindness, aortitis, myocardial infarction, stroke, or even death. De Smit et al suggest that the incidence of GCA will increase with our aging population with an estimated 3 million cases worldwide by the year 2050 as well as 500,000 patients with blindness at a cost of 76 billion dollars in the US alone.[1]

GCA can be a diagnostic conundrum, especially when it presents in an occult or atypical fashion. To date, there is no specific biomarker for GCA. Blood tests for inflammation have very poor specificity, and "seronegative" GCA can occur in up to 4% of the patients.[2] Temporal artery biopsy (TABx) remains the gold standard in the diagnosis of GCA, but is an invasive, time-consuming test with suboptimal sensitivity. Numerous articles[3–7] incorporate the 1990 American College of Rheumatology (ACR) classification criteria for GCA[8] to guide the decision for TABx. However, the ACR

criteria were not meant to be diagnostic criteria,[9] and without the TABx result, the ACR criteria only have a sensitivity of 29%. There are many clinical prediction rules in the diagnosis and management of patients with suspected GCA,[10–16] but few were developed using more than 500 TABx or 100 biopsy-positive GCA cases,[17] and few if any have external validation. Large collaborative studies can clarify the reliability and generalizability of prediction algorithms for patients with suspected GCA prior to TABx. We used a large multicenter dataset to develop and geographically validate a multivariable diagnostic prediction rule for GCA with an accompanying spreadsheet calculator and nomogram.

Methods

The consecutive records of subjects undergoing TABx for suspected GCA at secondary/tertiary care referral clinics were retrieved from four medical centers in ON, Canada; two from the US; and one from Switzerland (Table 1). This clinical audit was approved by the Michael Garron Hospital Research Ethics Board and Queen's Medical school, and was compliant with the Declaration of Helsinki and the TRIPOD guidelines.[18] Some of the data came from the de-identified records of prior research ethics board approved TABx projects (patient consent was not required by the ethics board),[19–22] and two centers conducted a chart review in July 2017 with patient consent. The chart review was not blinded.

This paper only considered cases of biopsy-proven GCA (BPGCA). As such the pathologic diagnosis was considered the final diagnosis. Healed arteritis was considered as positive for GCA. If the pathologic diagnosis was indeterminate, the record was considered negative for GCA.

Based on the literature review[15,17,23,24] and subject matter expertise, the candidate predictors for this study were age, gender, jaw claudication, new onset headache, temporal artery abnormality on physical examination (tenderness to palpation, decreased pulse, and scalp nodularity), diplopia, ischemia-related loss of visual acuity or field, or VL (a composite of ischemic optic neuropathy, retinal artery occlusion, or stroke), platelet level, C-reactive protein (CRP), and Westergren erythrocyte sedimentation rate (ESR) prior to glucocorticoid initiation.

Polymyalgia rheumatica (PMR) was not included as it can be a non-specific clinical manifestation, with overlapping age and acute phase response characteristics with GCA. The distinction of PMR from osteoarthritis flare can sometimes be difficult, and reports of joint X-rays were not uniformly available in this study. Except in patients on low-dose prednisone for PMR, bloodwork obtained after glucocorticoid initiation was excluded, but later patients were still considered for multiple imputation analysis. Abnormal ESR was defined as Westergren ESR >50 mm/hour. As there was variation in the CRP technique (highly sensitive versus rapid/regular) and upper limit of normal of CRP from different labs, each CRP was divided by the upper limit of normal to standardize the data.

To avoid overfitting, the minimum estimated sample size was found to be 500. With 10 candidate predictors, a minimum of 100 events (positive TABx) was required. Assuming a utility ratio of four negative TABx for each positive TABx, the minimum estimated sample size was found to be 500 subjects. Missing data at a rate of 10% was anticipated, suggesting that at least 550 records would require to be reviewed.

Statistical calculations were performed using Stata 14.2 (StataCorp LLC, College Station, TX, USA), and JMP Pro13 (JMP SAS Institute, Marlow, Buckinghamshire, UK)

Table 1 Characteristics of patients with negative versus positive temporal artery biopsy (n=530)

Factor	Negative biopsy	Positive biopsy	Univariate odds ratio	p-value	Range (low, high)
n	397	133			
Age in years, mean (SD)	72.9 (10.3)	76.8 (7.8)	1.044	<0.001	39, 96
Female	270 (68.0%)	88 (66.2%)	0.920	0.69	
New headache	289 (72.8%)	101 (75.9%)	1.179	0.48	
Temporal arterial abnormality	130 (32.7%)	64 (48.1%)	1.905	0.002	
Jaw claudication	73 (18.4%)	63 (47.4%)	3.994	<0.001	
Platelet level, mean (SD)	287 (102)	386 (139)	1.007	<0.001	53, 940
ESR, median (IQR)	35 (16, 57)	53 (34, 74)	1.014	<0.001	0.01, 240
CRP, median (IQR)	1.6 (0.48, 6.1)	7.8 (2.6, 16)	1.029	<0.001	0.025, 82.7
Vision loss	72 (18.1%)	47 (35.3%)	2.467	<0.001	
Diplopia	20 (5.0%)	8 (6.0%)	1.206	0.66	
Biopsy length in centimeter (n=482)	1.9 (0.7) n=376	1.9 (0.6) n=106	1.179	0.31	0.1, 4.6

Notes: CRP is divided by the upper limit of normal; ESR (Westergren) mm/1st hour.
Abbreviations: CRP, C-reactive protein; SD, standard deviation; IQR, interquartile range (25th–75th percentile).

and α=0.05 was used for statistical significance. Model misspecification was evaluated with Stata "linktest" and multicollinearity analyzed with Stata "collin" test.

Logistic regression (LR) does not require assumptions of normality, although multivariable normality provides a more stable solution. To optimize model fit, logarithmic transformation of any data that showed skewed distribution was examined. The best predictor subsets for the optimized full model, with and without log-transformed variables, were chosen based on clinical significance and statistical factors: *p*-values, confidence intervals, penalized-likelihood criteria to minimize Akaike information criterion (AIC), and minimize Bayesian information criterion (BIC), discrimination (area under the receiver operating characteristic curve [AUROC]), and calibration (Hosmer–Lemeshow goodness of fit, and Brier score with Spiegelhalter's *z*-statistic) (Table S1).

LR only analyzes complete cases and performs listwise deletion. As it cannot be assumed that data was missing completely at random ("Discussion" section) multiple imputation with 250 imputations was performed to discern possible bias, and to determine if there were any discrepancies in the confidence intervals of the predictor variables. Multiple imputation using chained equations (MICE) was performed on the full model without log transformations, as per convention.

As all covariates were clinically important, we retained the full model, but we developed a parsimonious model as per statistical convention (Tables 2 and 3). The statistically significant variables from the optimal full model were selected for the parsimonious model. A stepwise regression was performed in JMP Pro 13 software with 60% of the data for training, 20% for validation, and 20% for testing, using the forward direction and combined stopping rules to minimize AIC and BIC. Predictor(s) that were statistically significant

on MICE but not the complete case analysis were forced on to the parsimonious model for evaluation. An additional nested model excluding the two covariates with the highest *p*-value was made.

Internal validation of the final models was assessed by combined cross-fold validation and bootstrap techniques. After multivariable LR, 10-fold cross validation was performed, and the *c*-statistic corresponding to each fold was averaged. The cross-validated area under the receiver operator characteristics (ROC) curve was then bootstrapped to determine statistical inference. Three thousand computer-generated bootstrap samples, each including 530 patients from the study were refitted and the average odds ratio was obtained.

Geographic external validation was performed by holding out the data from each regional contributing center. Since large datasets are recommended for external validation,[25] if a regional dataset had fewer than 30 subjects, then it was placed in the combined group (Table 4). One-way analysis of variance (ANOVA) was performed to compare the patient characteristics in the different regions.

The actual performance of our models at the 5th and 95th percentile and Liu optimal cutoff points (Tables 5 and S2) were compared with the ACR model. JMP Pro 13 prediction profiler was used to compare our models using hypothetical examples.

An online spreadsheet calculator was made for both models, and a Kattan nomogram was made for the parsimonious model.

Length of biopsy was not a primary concern in our initial data collection. Recent literature suggests shorter specimen lengths are adequate for diagnosis ("Discussion" section) and bilateral TABx was routinely performed in patients with continued suspicion for GCA if the initial unilateral TABx

Table 2 Multivariable logistic regression, full model (n=530, pseudo R^2=0.256, AUROC =0.820, $p_{Hosmer–Lemeshow}$ =0.549, 530 jackknife replications, 3000 bootstrap replications, log likelihood −222.12)

Predictor	β	Standard error of β	Odds ratio	*p*-value	$β_{bootstrap}$	Standard error of $β_{bootstrap}$	$p_{bootstrap}$
Age in years	0.045	0.014	1.046	0.001	0.045	0.013	0.001
Sex (female)	−0.152	0.224	0.859	0.559	−0.152	0.269	0.572
New headache	0.231	0.365	1.259	0.426	0.231	0.302	0.446
Jaw claudication	1.296	0.964	3.656	<0.001	1.296	0.271	<0.001
TA abnormality	0.356	0.373	1.428	0.172	0.356	0.270	0.188
Vision loss	1.031	0.787	2.803	<0.001	1.031	0.286	<0.001
Diplopia	−0.229	0.399	0.795	0.648	−0.229	0.553	0.679
logESR	0.187	0.206	1.206	0.273	0.187	0.154	0.224
logCRP	0.290	0.131	1.337	0.003	0.290	0.091	0.001
Platelets	0.005	0.001	1.005	<0.001	0.005	0.001	<0.001
Constant	−9.967	0.0000751		0.000	−9.967	1.594	0.000

Notes: TA abnormality: temporal artery abnormality on clinical exam; logESR, natural logarithm of ESR; logCRP, natural logarithm CRP.
Abbreviations: CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; TA, temporal artery.

Table 3 Multivariable logistic regression, parsimonious model (n=530, pseudo R^2=0.248, AUROC =0.816, $p_{\text{Hosmer–Lemeshow}}$ =0.812, 530 jackknife replications, 3000 bootstrap replications, log likelihood −224.51)

Predictor	β	Standard error of β	Odds ratio	p-value	$β_{\text{bootstrap}}$	Standard error of $β_{\text{bootstrap}}$	$p_{\text{bootstrap}}$
Age in years	0.044	0.013	1.044	0.001	0.044	0.012	<0.001
Jaw claudication	1.402	0.253	4.064	<0.001	1.402	0.256	<0.001
Vision loss	0.925	0.268	2.521	0.001	0.925	0.273	0.001
Platelets	0.005	0.001	1.005	<0.001	0.005	0.001	<0.001
Log CRP	0.358	0.080	1.431	<0.001	0.358	0.080	<0.001
Constant	−8.496	1.127	1.127	<0.001	−8.496	1.116	<0.001

Abbreviations: AUROC, area under the receiver operating characteristic curve; CRP, C-reactive protein.

was negative.[26,27] For completeness sake and to help guide the discussion, biopsy length was examined post hoc.

Results

Of the 688 TABx cases retrieved, 530 were complete records with 397 being negative and 133 being positive biopsies. The TABx dates from the various centers ranged from November 2005 to June 2017, and at least 56% of the TABx were done after 2010. Forty-eight percent of the patients were referred by ophthalmology, and the remainder was referred by rheumatology, internal medicine, or primary care centers.

The characteristics of the positive versus negative TABx are summarized in Table 1. Patients with positive TABx were older and had more jaw claudication, higher platelet level, higher ESR, higher CRP, and had more ischemic vision loss (VL) compared with the negative TABx group. The youngest patient with biopsy-proven GCA (BPGCA) was 54 years of age. GCA was more common in women, but on multivariable analysis, gender, new onset headache, temporal artery abnormality, ESR, diplopia, and biopsy length did not show a statistically significant difference between positive and negative biopsy groups.

Ten patients had BPGCA (10/133=7.5%) with normal platelet count (<400 per microliter), ESR <50 mm/hour,

and adjusted CRP ≤1. The subjects with "seronegative" BPGCA originated from five different regions, and each case was rechecked to ensure the absence of glucocorticoids prior to bloodwork. The seronegative BPGCA group had mean probability score of 0.108, median of 0.082, and less clinical temporal artery abnormality (p=0.012) than their seropositive counterparts, but other demographic features including age, gender, and biopsy length showed no statistically significant difference in the independent t-test.

Data on biopsy length was readily available for 482/530 (91%) patients that was used for LR. There was no statistically significant difference found with respect to the length of the specimen between the positive and negative biopsy groups on univariate LR (p=0.31). Bilateral biopsies were performed in 23% of the cases. One patient in the negative biopsy group had a TABx length of 0.1 cm, but this was a unique case.

Funduscopic findings were readily available for 32 out of 47 patients with BPGCA and VL. In this group, 23 (72%) patients had anterior ischemic optic neuropathy, 4 (12.5%) had central retinal artery occlusion, 4 (12.5%) had presumed posterior ischemic optic neuropathy, and 1 (3%) had a central retinal vein occlusion. We were able to retrieve the fundus findings in 26/72 patients with VL and

Table 4 Geographic external validation of full and parsimonious models by regional site

External validation set	n (events)	Training: validation split	Full training AUROC	Full validation AUROC	Parsimonious training AUROC	Parsimonious validation AUROC
Rochester, MN, USA	52 (14)	90:10	0.836	0.688	0.828	0.750
Toronto, ON, Canada	124 (21)	77:23	0.820	0.812	0.818	0.805
Ottawa, ON, Canada	119 (27)	77:23	0.822	0.794	0.816	0.804
Kingston, ON, Canada	172 (32)	68:32	0.836	0.824	0.832	0.845
Composite	63 (39)	88:12	0.816	0.782	0.812	0.793
London, ON, Canada	18 (13)	97:3	0.824	0.677	0.823	0.677
Boston, MA, USA	20 (10)	96:4	0.825	0.700	0.820	0.750
Zurich, CH, Switzerland	25 (16)	95:5	0.804	0.979	0.802	0.993

Notes: n, number of biopsies at each site; event, positive temporal artery biopsy; Composite, London, ON + Boston, MA + Zurich, CH.
Abbreviation: AUROC, area under the receiver operating characteristic curve.

Table 5 Model performance at 5th, 85th and 95th percentile

Performance at percentile	5th			95th		85th		
Model (probability)	Full (0.023)	Parsimonious (0.033)	ACR (0.150)	Full (0.782)	Parsimonious (0.790)	ACR (0.420)*	Full (0.514)	Parsimonious (0.501)
Sensitivity	1.000	0.993	0.901	0.173	0.158	0.256	0.429	0.437
Specificity	0.066	0.081	0.125	0.992	0.992	0.898	0.945	0.952
PPV	0.264	0.266	0.261	0.885	0.875	0.456	0.722	0.753
NPV	1.000	0.970	0.803	0.782	0.779	0.781	0.832	0.834
False negative rate	0	0.008	0.090	0.827	0.842	0.744	0.571	0.564
False positive rate	0.935	0.919	0.875	0.008	0.008	0.102	0.054	0.048
Misclassification rate	0.700	0.691	0.676	0.218	0.217	0.265	0.185	0.178

Notes: *The 95th percentile score for the ACR model is 0.443933, and corresponds with 0% sensitivity, 100%, specificity, unspecified PPV, 75% negative predictive value, no false positives, and 100% false negatives. (0.443933 is the maximum possible score and 14% of the data share this score). The next highest ACR probability score is 0.419872 which is the 85th percentile.
Abbreviations: ACR, American College of Rheumatology Classification non-histologic Criteria; PPV, positive predictive value.

a negative TABx, and all these patients had non-arteritic ischemic neuropathy.

The ESR and CRP levels had skewed distributions, but platelet values had a normal distribution. Although LR makes no assumptions of normality, model fitting with the log-transformed ESR and CRP yielded lower AIC and lower BIC than any combination of non-transformed/transformed ESR and CRP. Multivariable LR showed that age, jaw claudication, ischemic VL, platelets, and log-transformed CRP values were significantly predictive of positive TABx (Table 2) and these covariates were later used for the parsimonious model. There was no model specification error. There was no multicollinearity, with mean variance inflation factor (VIF) of 1.19 in the full model and maximum individual VIF of 1.45 (Supplementary material).

Twenty-three percent of the records had incomplete data, in which serology values were predominantly missing. Following were the major missing value patterns: 12% of the records had no serology values, 3% of the records had missing data regarding platelets and CRP, 3% had missing data regarding platelets alone, 2% had missing CRP values, and <1% had missing ESR values. MICE estimates of the non-transformed full model with 250 imputations showed little bias, if any, with the predictors that were statistically significant on complete case analysis, but the temporal artery abnormality predictor became statistically significant ($p_{original}$ =0.117, p_{MICE} =0.036) and was evaluated for the parsimonious model.

Variable selection for statistical modeling was based on the following clinical significance and statistical factors: p-values, the minimum AIC and BIC, discrimination, and calibration. The full model with log-transformed CRP and ESR had better discrimination and calibration than the

non-transformed models. There were no statistically significant interaction terms. The full model and the parsimonious model both had good discrimination (AUROC 0.82), and calibration (Figure 1; Table S1) with misclassification rate of 17.7%. However, the full model had a false negative rate of 54.1% and the parsimonious model had 56.4%. Bootstrap sensitivity analysis with 3,000 replications did not reveal any discrepancies. (Tables 2 and 3)

The gender and diplopia variables had the highest p-values, but when removed from the full model, the eight covariate nested model had poor calibration (Reduced model A, log transformed). Multiple imputation analysis suggested that the temporal artery abnormality variable is statistically significant, but its addition in the parsimonious model resulted in a poorly calibrated model (Reduced model B, log transformed). (Table S1)
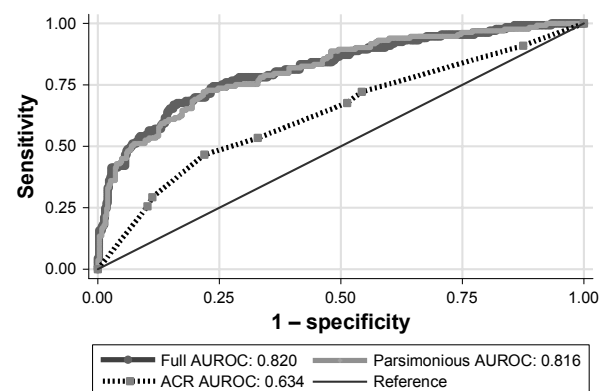
Figure 1 ROC curves for full, parsimonious and ACR models.
Notes: Full model (n=530) $p_{Hosmer-Lemeshow}$ =0.549. Parsimonious model (n=530) $p_{Hosmer-Lemeshow}$ =0.812. ACR model = (n=525). $p_{Hosmer-Lemeshow}$ =0.0223 (Five patients under the age of 50 years were excluded from logistic regression.).
Abbreviations: ROC, receiver operator characteristics; ACR, American College of Rheumatology Classification non-histologic Criteria.

A **External validation with 63:32 split training: validation**
Site: Kingston, ON (n=172 with 32 events)
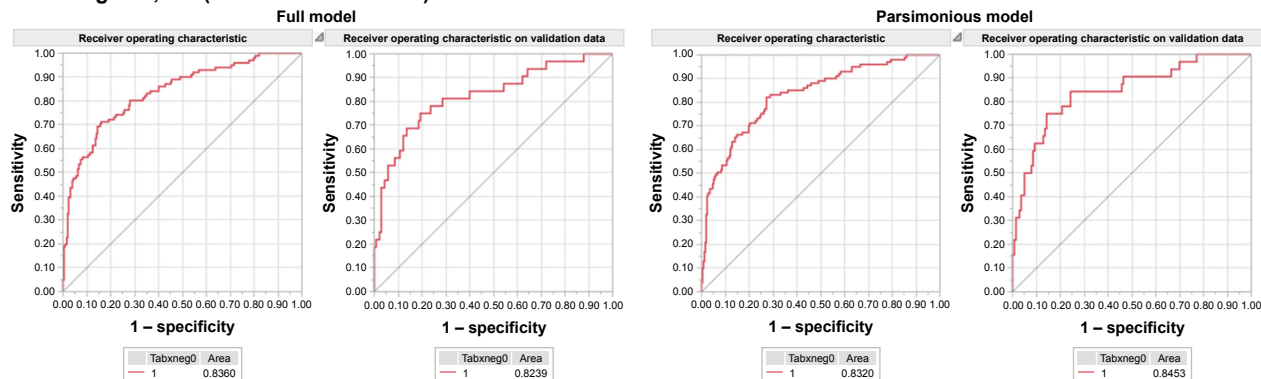
B **External validation with 90:10 split training: validation**
Site: Rochester, MN (n=52 with 14 events)
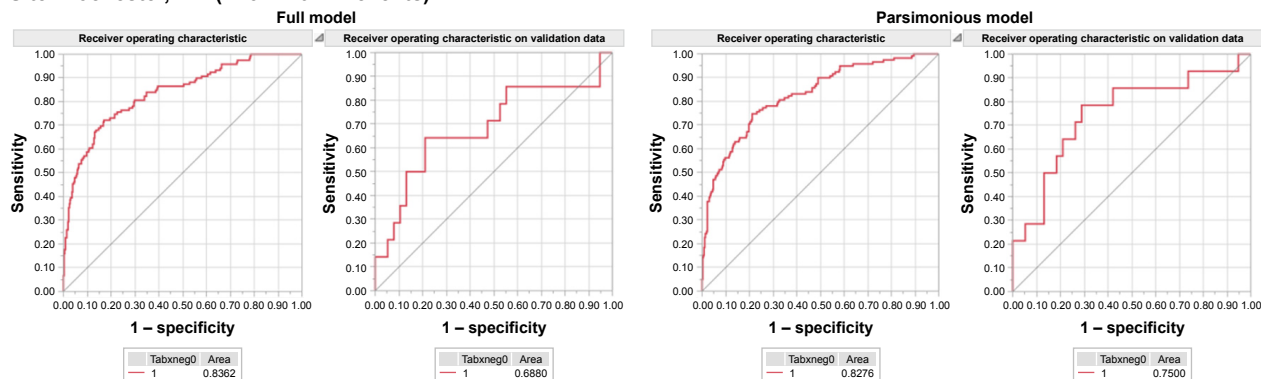

Figure 2 External geographic validation results of the highest (**A**) and lowest ranking datasets (**B**).

Internal validation with 10-fold cross validation and bootstrap technique showed the following *c*-statistics: 0.803 (95% CI =0.757–0.849) for the full model and 0.810 (95% CI =0.766–0.854) for the parsimonious model.

Five spatial external validations were performed with the largest datasets, and the *c*-statistics ranged from 0.688 to 0.824 for the full model and from 0.750 to 0.845 for the parsimonious model. (Table 4; Figure 2) ANOVA of the covariates for the regional datasets showed statistically significant difference (all at *p*<0.001) in clinical temporal arterial abnormality, platelets, ESR, CRP, ischemic VL, diplopia, and biopsy length between the different centers but not for age (*p*=0.534), gender (*p*=0.556), jaw claudication (*p*=0.239), or new headache (*p*=0.362). The post-hoc pairwise comparisons with Bonferroni correction are shown in supplementary material.

The full and parsimonious prediction models had similar performance, with almost overlapping ROC curves. (Figure 1) Compared to our study models, the ACR model has lower sensitivity, specificity, and greater misclassification error at almost all cutoff points except the 5th percentile.

(Tables 5 and S2). The output of the full, parsimonious, and the ACR models was compared using hypothetical examples (Table 6; Figure 3). The ACR model had a small range of probability outputs compared to the study models.

In the full model, no subject with probability score <0.027 had a positive TABx, suggesting that 7% of the TABx in this study could have been avoided. A probability score of ≤0.07 corresponded with a 95% chance of negative TABx and approximately 30% of the patients in our negative biopsy group had a probability score of ≤0.07. A probability score of 0.23 approximates the 25th percentile of the positive TABx group, and a score of 0.43 was the median value of the positive biopsy group, and was considered high risk for GCA. A probability score of ≥0.89 was not seen in patients with a negative biopsy.

Discussion

Several prediction algorithms for GCA diagnosis have been published,[8,11,12,17,19,24,28] (Table S3) with the common goal of improving diagnostic accuracy and patient selection for TABx and for reducing patient morbidity and health care

Table 6 Hypothetical cases comparing the full, parsimonious, and American College of Rheumatology models

Clinical scenarios			Full model no vision loss	Full model with vision loss	Parsimonious model no vision loss	Parsimonious model with vision loss	ACR model
(Case I) **HA +** **Plat 475**	**90 yo** **TA Abn** **ESR 90**	F **JC +** **CRP 4×**	0.80	0.92	0.75	0.88	0.44
(Case II) **HA +** **Plat 399**	**80 yo** TA Abn **ESR 60**	F JC + **CRP 3×**	0.51	0.75	0.56	0.75	0.31
(Case III) **HA +** Plat 250	**80 yo** TA Abn **ESR 60**	F JC + **CRP 2×**	0.30	0.55	0.35	0.56	0.31
(Case IV) HA: No Plat 250	**80 yo** TA: No ESR 49	M **JC +** **CRP 2×**	0.28	0.52	0.35	0.56	0.16
(Case V) **HA +** Plat 250	**80 yo** **TA +** ESR 49	F No JC **CRP 2×**	0.14	0.31	0.11	0.22	0.26
(Case VI) **HA +** Plat 250	**80 yo** TA: No ESR 49	F No JC **CRP 2×**	0.10	0.24	0.11	0.22	0.16
(Case VII) **HA +** Plat 250	**65 yo** TA: No ESR 49	F No JC **CRP 2×**	0.06	0.14	0.06	0.13	0.16
(Case VIII) HA: No **Plat 390**	50 yo TA: No **ESR 55**	M No JC **CRP 2×**	0.05	0.14	0.06	0.14	0.31
(Case IX) HA: No Plat 250	50 yo TA: No **ESR 55**	M No JC **CRP 2×**	0.03	0.07	0.03	0.08	0.29
(Case X) HA: No Plat 250	50 yo TA: No ESR 49	F No JC **CRP 2×**	0.02	0.06	0.03	0.08	0.15
(Case XI) HA: No Plat 250	50 yo TA: No **ESR 55**	M No JC CRP 1	0.02	0.06	0.03	0.06	0.29

Notes: Parsimonious model: age, jaw claudication, platelets, logCRP, vision loss. The bold indicates the factor that changes as one moves upwards from the bottom of the chart. **Abbreviations:** +, present; No, absent; M, male; F, female; HA, new onset headache; TA, temporal artery; JC, jaw claudication; Plat, platelet level; ESR, erythrocyte sedimentation rate; ESR High, ESR ≥50 mm/hour; CRP, C-reactive protein.

expenditures. Compared to other prediction algorithms, the following are the strengths and distinguishing features of our study:

i) Its large size, validation, and generalizability. Our study had sufficient GCA events to support more than 10 candidate predictor variables with LR. The 0.80 (95% CI =0.76–0.85) *c*-statistic from combined internal bootstrap cross-validation and multiple imputations supports reproducibility of the prediction model. On geographic external validation, the *c*-statistic was found to range from 0.69 to 0.82 for the full model and even better for the parsimonious model. Generalizability is further enhanced by the collection of TABx results from seven different medical centers with an almost equal proportion of patients referred from ophthalmic and non-ophthalmic practices.

ii) Its design to independently predict the risk of GCA prior to TABx. Although TABx is usually a benign test, it is invasive and time-consuming. Ideally risk calculators should portend the risk of GCA prior to TABx to guide decision making. The ACR criteria[8] and other LR models[11,23] entreaty input of the TABx result or specimen length. The performance of our model was also directly compared against the 1990 ACR classification criteria.

iii) The employment of four statistically significant objective predictors (age, platelets, logCRP, and ischemic VL), the first three of which were maintained as continuous variables to preserve statistical power.[29] Prediction algorithms
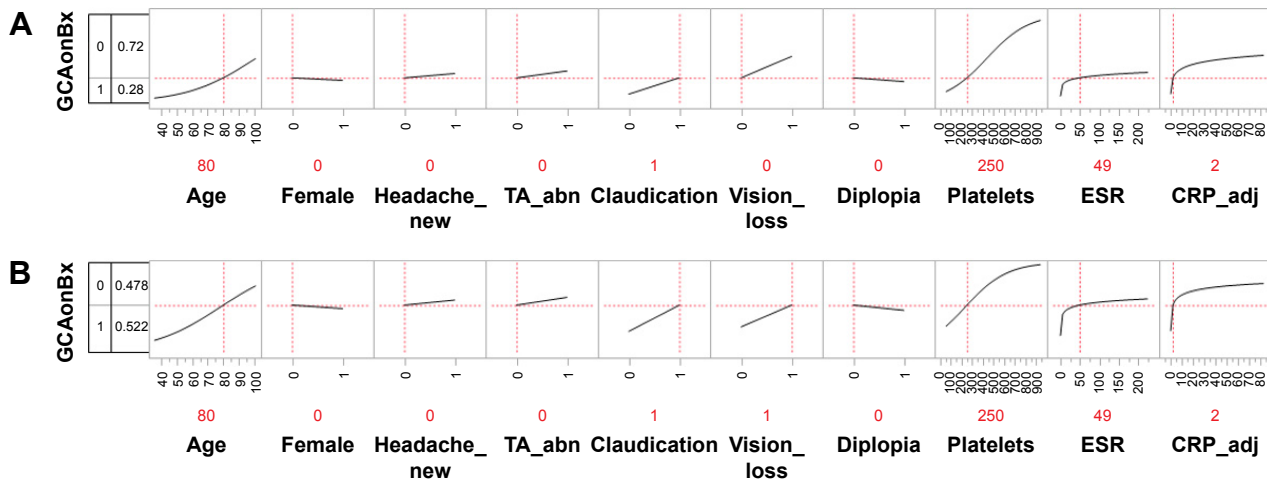
Figure 3 Prediction risk profile using the full model and Case 4 of Table 6.
Notes: Claudication, jaw claudication; CRP_adj, log (CRP divided by the upper limit of normal CRP). In this hypothetical case, an 80-year-old male has jaw claudication and CRP that is elevated twice normal, but no headache, temporal artery tenderness, or diplopia. The ESR is <50, and the platelet levels are normal. The risk of biopsy-proven GCA is 28% if there is no vision loss (**A**), but 52% in the setting of ischemic vision loss (**B**).
Abbreviations: CRP, C-reactive protein; GCA, giant cell arteritis; ESR, erythrocyte sedimentation rate; GCAonBx, biopsy-proven giant cell arteritis.

heavily based on patient symptoms[23] may be disadvantageous when the physician has cognitive or affective biases,[30] or when patient responses are ambiguous. Many guidelines or prediction rules do not incorporate CRP[15,17] and/or platelet count,[8,11] which are more accurate than ESR in the diagnosis of GCA.[31] Prediction rules that incorporate ESR, CRP, and platelet count are laudable[13] but can be improved by the addition of patient symptoms, such as jaw claudication.

iv) Provision of an output probability nomogram (Figure 4) and online calculator for the risk of GCA (https://docs.google.com/spreadsheets/d/1wlRFGleW2Vf-LlylmY76KSTzIAf1TrX5U_1770HhD1Y/edit?usp=sharing). Prior GCA studies have used univariate probability curves,[31] theoretical decision analysis tables,[15] scoring systems,[13,20] or risk calculators,[11] but many only provide odds ratios,[12,16,17,24] or likelihood ratios[14] that require extensive calculation to determine the output probability of GCA.
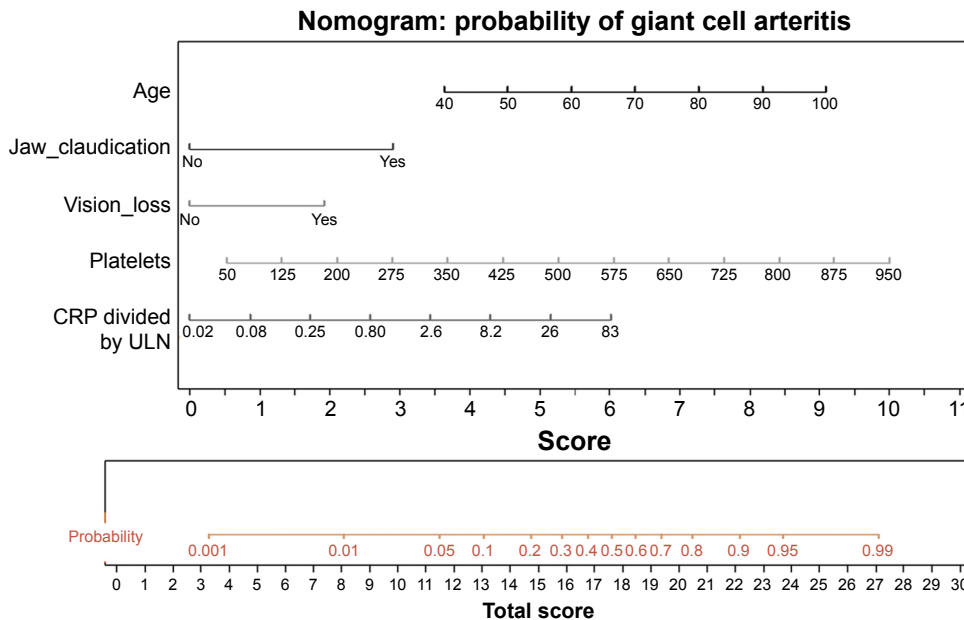

Figure 4 Nomogram of parsimonious model.
Notes: The length and location of each nomogram scale indicates the relative importance of the predictor variable. A vertical line is drawn down from the value of each covariate to determine the score. The sum of the scores is used to determine the probability for a positive temporal artery biopsy.
Abbreviations: CRP, C-reactive protein; ULN, upper limit of normal.

The length and location of our nomogram scales visually communicate the statistical importance of each covariate and the probability for GCA is enumerated from simple addition, rather than odds ratios or likelihood ratios.

Our work agrees with previous studies that have shown jaw claudication,[12,16,17,23] age,[23] and thrombocytosis and elevated CRP[31,32] to be statistically significant predictors for GCA. The odds ratio of 1.005× for platelet level seems outwardly small, but platelets were a continuous variable with a wide range. For a 50 unit increase in platelets, the odds ratio for positive TABx was found to be 1.29×, and for a 100 unit increase in platelets, the odds ratio was found to be 1.66×.

We also found that log CRP and ischemic VL were useful predictors for GCA. Few prediction rules incorporate CRP,[31,32] in part due to epoch, lack of statistical power, and/or missing data.[23] In our study, 20% of the patients had missing CRP data as it was sometimes not requisitioned prior to glucocorticoid initiation, and some practitioners only requisition the ESR and not the CRP values in patients with suspected GCA or vice versa. In some institutions, the result of CRP test takes longer to return than the ESR test, and may not be available or recorded prior to referral for consideration of biopsy. Some private labs did not offer CRP testing. The health care facility where the patient was initially assessed may differ from the location where TABx was performed, making it more difficult to find the results retrospectively. As CRP and other predictors may not have been missing completely at random, multiple imputation was performed, which did not suggest bias of note in the missing data.

VL is one of the most feared complications of GCA, and absent from most rheumatology-based prediction schemas. In our study, half of the patients were referred by ophthalmologists; disc edema and retinal artery occlusion proved to be compelling predictors for GCA.

In contrast to other reports,[12,24,33] diplopia and new onset headache were not statistically significant predictors in this study. This may be because VL was a more common eye finding, and patients with monocular VL have little or no binocular diplopia. Six subjects had diplopia and ischemic VL, but only one had BPGCA. Since half of our patients originated from ophthalmologists, the complaint of diplopia should have been well scrutinized, and this may also account for bias compared to some rheumatology studies.

Headache is a common complaint in the elderly with up to 51% of the individuals at 65 years of age or older have this symptom.[34] Although ANOVA did not support geographic heterogeneity in the frequency of cephalgia, a standardized definition for the new onset headache of GCA may render headache a more discriminating predictor. The International Classification of Headache Disorders' criteria specifies headache in close temporal relation to other signs and symptoms of GCA, worsening of headache in parallel to worsening GCA, and improvement of headache after 3 days of high dose glucocorticoids.[35]

Statistical significance should be but one consideration in predictive modeling. Although parsimonious models save time and facilitate ease of use with nomograms, the spreadsheet calculator was generated for the full model; each of our study covariates is referenced in the literature as clinically significant, and as such, the full model may better control for confounding and bias. Although gender was not statistically significant, it is an expected control variable in most medical studies. The temporal arterial abnormality predictor variable became statistically significant on multiple imputation estimates. Predictors associated with a particular hypothesis can be retained, even if they are not statistically significant. It was hypothesized that if VL was an important predictor of GCA, there would be fewer tendencies for binocular diplopia. Our sample was large enough such that the covariates with $p > 0.05$ had a negligible effect on the statistical degrees of freedom. Another important reason for covariate retention is because variables with high statistical significance are not necessarily highly predictive, due to different properties of their underlying distribution. Sets of variables with predictive power above a certain threshold may differ from variable modules identified by statistical significance-based criterion such as the chi-square test.[36]

Although our study appears to be the largest TABx prediction rule study to date, and the only one with external validation, the limited size of our external validation (EV) sets is a potential weakness. ANOVA showed that six of the covariates were statistically significant regional case-mix, which likely accounts for the heterogeneous discrimination scores. The Rochester group had the lowest EV c-statistic, and the lowest proportion of temporal artery tenderness/decreased pulse, average platelet values, and training validation ratio (10%). The Mayo series is more likely to be a referral cohort, with possible atypical presentations of GCA.[17] The three smallest individual datasets, which comprised the "combined" EV set, had a higher proportion of positive TABx and may reflect referral bias or selection bias. The fair to good EV c-statistics $AUROC_{EV}$ (0.688–0.824 for the full model and 0.750–0.845 for the parsimonious model) in the setting of diverse regional case-mix suggests that

our model is transportable. As our data came from seven different centers, the AUROC confidence intervals for the bootstrapped 10-fold internal validation (0.757–0.849) for the full model and 0.766–0.854 for the parsimonious model may be more representative than those from the geographic validation. Further collaborative, international studies such as the DCVAS[37] may achieve the minimum size validation sets of 100 events and 100 non-events suggested for EV of LR prediction rules.[25]

Our study had some limitations, which includes its retrospective nature with missing data, the constraint to BPGCA, and misclassification rate. Retrospective studies performed at different institutions may not have uniform definitions of jaw claudication, clinical temporal arterial abnormality, and recent onset headache, which can be inherently subjective assessments. With 10 predictor variables, missing data was not unexpected in a retrospective study. Multiple imputation analysis of the missing data showed minimal bias.

This study targets BPGCA. With the exception of Grossman,[24] most studies do not incorporate biopsy-negative GCA (BNGCA). Patients with BNGCA may have more headaches and polymyalgia rheumatica but less visual complications and jaw claudication than BPGCA and may require a different set of decision rules.[24,38] TABx is the gold standard for BPGCA, but "there are no independent validating criteria to determine whether giant cell arteritis is present when a temporal artery biopsy is negative".[39] The schema of Ellis and Ralston,[40] was utilized by Vilaseca et al[41] for BNGCA, but has not been widely applied. Unless imaging studies show evidence of vessel abnormality, the diagnosis of BNGCA relies on clinical judgment, exhaustive anamnesis,[23] and amelioration with systemic glucocorticoids in the absence of neoplasm. BNGCA may result

from inadequate specimen length and skip areas, but routine bilateral biopsies are not strongly advocated and specimen lengths of 1.5 cm appear to be adequate.[42–44] A review of 240 TABx found that specimen length was not associated with the diagnostic yield of TABx.[45] Others report fixed TABx length of 0.5 cm[26] (n=1,520 TABx), 0.7 cm (n=966 TABx),[27] or 1.5 cm[46] (n=538 TABx) as the possible optimum length threshold TABx length to predict GCA and avoid false negative TABx. There was no statistically significant difference in the lengths of TABx in the positive or negative biopsy groups in our study, 90% of which had a fixed length >1 cm in both groups.

Although our prediction model outperformed the non-histologic ACR classification criteria, at a probability cutoff point of 0.5, there remained an 18.1% misclassification rate with a sensitivity and specificity of 45.9% and 94.2%, respectively. To improve future models, large prospective studies or "big datasets" with standardized predictor definitions, additional clinical criteria (eg, neck pain, weight loss, fever), and objective predictors such as ocular pulse amplitude,[21] OCT ultrasound, MRI of the arteries, HLA-DRB1*04,[47] and genetic markers should be considered. Alternative prediction schemas such as neural networks[10] and support vector machine[28] can be compared with LR models.

In patients with suspected GCA whose blood results have not been clouded by high dose glucocorticoids, a possible clinical interpretation of the probability values from our cohort of 530 patients is summarized in Table 7. Since no subject with probability score <0.024 had BPGCA, TABx can probably be avoided in these patients. With GCA probability scores <0.07, the clinician and patient may contemplate deferral of TABx and glucocorticoids with close observant management. Patients with probability scores between 0.7 and 0.23 are at low to moderate risk of

Table 7 Probability score cutoff points and risk of GCA

Observed probability cutoff point	Full model observed percentile		Parsimonious model observed percentile		Risk of GCA (biopsy-proven)
	Negative biopsy	**Positive biopsy**	**Negative biopsy**	**Positive biopsy**	
2.7%	7th	I	4th	I	Very low
7%	31st	5th	26th	5th	Low
14%	54th	16th	53rd	12th	Intermediate
23%	75th	25th	73rd	26th	Moderate
43%	92nd	50th	91st	49th	High
52%	95th	55th	95th	59th	Very high
89%	99.9th	96th	99.9th	97th	Exceedingly high

Notes: Results from the online calculator: https://docs.google.com/spreadsheets/d/1wlRFGIeW2Vf-LlyImY76KSTzIAf1TrX5U_1770HhD1Y/edit#gid=0 should be interpreted with the cutpoint values in this table.
Abbreviation: GCA, giant cell arteritis.

GCA and should be considered for TABx and glucocorticoid treatment. Probability scores in the range of 0.24–0.43 are at moderate to high risk of GCA, and scores ≥0.43 are at high risk of GCA. Although some may argue that TABx could be avoided with a ≥0.89 probability score, the authors endorse pathologic confirmation, given the side effects of prolonged glucocorticoid treatment and the occasional alternative diagnoses obtained from TABx.

Conclusion

We developed and validated a LR prediction model for BPGCA. Jaw claudication, platelet levels, log CRP, ischemic VL, and age were statistically significant predictors for positive TABx. Prediction models are not infallible and cannot substitute for clinical acumen or pathologic confirmation. However, they organize decision making and help systematize the decision to perform TABx.

Acknowledgment

Disclosure

The authors report no conflicts of interest in this work.

References

1. De Smit E, Palmer AJ, Hewitt AW. Projected worldwide disease burden from giant cell arteritis by 2050. *J Rheumatol*. 2015;42(1):119–125.
2. Kermani TA, Schmidt J, Crowson CS, et al. Utility of erythrocyte sedimentation rate and C-reactive protein for the diagnosis of giant cell arteritis. *Semin Arthritis Rheum*. 2012;41(6):866–871.
3. Davies CG, May DJ. The role of temporal artery biopsies in giant cell arteritis. *Ann R Coll Surg Engl*. 2011;93(1):4–5.
4. Pieri A, Milligan R, Hegde V, Hennessy C. Temporal artery biopsy: are we doing it right? *Int J Health Care Qual Assur*. 2013;26(6):559–563.
5. Quinn EM, Kearney DE, Kelly J, Keohane C, Redmond HP. Temporal artery biopsy is not required in all cases of suspected giant cell arteritis. *Ann Vasc Surg*. 2012;26(5):649–654.
6. Hussain O, McKay A, Fairburn K, Doyle P, Orr R. Diagnosis of giant cell arteritis: when should we biopsy the temporal artery? *Br J Oral Maxillofac Surg*. 2016;54(3):327–330.
7. Cristaudo AT, Mizumoto R, Hendahewa R. The impact of temporal artery biopsy on surgical practice. *Ann Med Surg (Lond)*. 2016;11:47–51.
8. Hunder GG, Bloch DA, Michel BA, et al. The American college of rheumatology 1990 criteria for the classification of giant cell arteritis. *Arthritis Rheum*. 1990;33(8):1122–1128.
9. Hunder GG. The use and misue of classification and diagnostic criteria for complex diseases. *Ann Intern Med*. 1998;129(5):417–418.
10. Astion ML, Wener MH, Thomas RG, Hunder GG, Bloch DA. Application of neural networks to the classification of giant cell arteritis. *Arthritis Rheum*. 1994;37(5):760–770.
11. González-López JJ, González-Moraleja J, Rebolleda G, Muñoz-Negrete FJ. A calculator for temporal artery biopsy result prediction in giant cell arteritis suspects. *Eur J Intern Med*. 2014;25(8):e98–e100.
12. Rodriguez-Valverde V, Sarabia JM, González-Gay MA, et al. Risk factors and predictive models of giant cell arteritis in polymyalgia rheumatica. *Am J Med*. 1997;102(4):331–336.
13. Weis E, Toren A, Jordan D, Patel V, Gilberg S. Development of a predictive model for temporal artery biopsies. *Can J Ophthalmol*. Epub 2017 June 28.
14. Belliveau MJ, Ten Hove MW. Giant cell arteritis. *CMAJ*. 2011; 183(5):581.
15. Niederkohr RD, Levin LA. Management of the patient with suspected temporal arteritis. a decision-analytic approach. *Ophthalmology*. 2005; 112(5):744–756.
16. Rieck KL, Kermani TA, Thomsen KM, Harmsen WS, Karban MJ, Warrington KJ. Evaluation for clinical predictors of positive temporal artery biopsy in giant cell arteritis. *J Oral Maxillofac Surg*. 2011;69(1): 36–40.
17. Gabriel SE, O'Fallon WM, Achkar AA, Lie JT, Hunder GG. The use of clinical characteristics to predict the results of temporal artery biopsy among patients with suspected giant cell arteritis. *J Rheumatol*. 1995;22(1):93–96.
18. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;162:55–63.
19. Toren A, Weis E, Patel V, Monteith B, Gilberg S, Jordan D. Clinical predictors of positive temporal artery biopsy. *Can J Ophthalmol*. 2016; 51(6):476–481.
20. Knecht PB, Bachmann LM, Thiel MA, Landau K, Kaufmann C. Ocular pulse amplitude as a diagnostic adjunct in giant cell arteritis. *Eye (London)*. 2015;29(7):860–865.
21. Ing E, Pagnoux C, Lutchman C, et al. Dynamic contour tonometry to measure ocular pulse amplitude in patients with suspected giant cell arteritis. Paper presented at: North American Neuro-ophthalmology Society 43rd Annual Meeting; April 4; 2017; Washington, DC.
22. Chen JJ, Leavitt JA, Fang C, Crowson CS, Matteson EL, Warrington KJ. Evaluating the incidence of arteritic ischemic optic neuropathy and other causes of vision loss from giant cell arteritis. *Ophthalmology*. 2016; 123(9):1999–2003.
23. González-López JJ, González-Moraleja J, Burdaspal-Moratilla A, Rebolleda G, Núñez-Gómez-Álvarez MT, Muñoz-Negrete FJ. Factors associated to temporal artery biopsy result in suspects of giant cell arteritis: a retrospective, multicenter, case-control study. *Acta Ophthalmol*. 2013;91(8):763–768.
24. Grossman C, Barshack I, Koren-Morag N, Ben-Zvi I, Bornstein G. Baseline clinical predictors of an ultimate giant cell arteritis diagnosis in patients referred to temporal artery biopsy. *Clin Rheumatol*. 2016; 35(7):1817–1822.
25. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol*. 2005;58(5):475–483.
26. Mahr A, Saba M, Kambouchner M, et al. Temporal artery biopsy for diagnosing giant cell arteritis: the longer, the better? *Ann Rheum Dis*. 2006;65(6):826–828.
27. Ypsilantis E, Courtney ED, Chopra N, et al. Importance of specimen length during temporal artery biopsy. *Br J Surg*. 2011;98(11):1556–1560.
28. Lee M, De Smit E, Wong Ten Yuen A, Sarossy M. The use of statistical modeling to predict temporal artery biopsy outcome from presenting symptoms and laboratory results. *Acta Ophthalmol*. 2014;92(S253).
29. Cumberland PM, Czanner G, Bunce C, Dore CJ, Freemantle N, Garcia-Finana M; Ophthalmic Statistics Group. Ophthalmic statistics note: the perils of dichotomising continuous variables. *Br J Ophthalmol*. 2014;98(6):841–843.
30. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. *BMC Med Inform Decis Mak*. 2016;16(1):138.
31. Walvick MD, Walvick MP. Giant cell arteritis: laboratory predictors of a positive temporal artery biopsy. *Ophthalmology*. 2011;118(6): 1201–1204.
32. Costello F, Zimmerman MB, Podhajsky PA, Hayreh SS. Role of thrombocytosis in diagnosis of giant cell arteritis and differentiation of arteritic from non-arteritic anterior ischemic optic neuropathy. *Eur J Ophthalmol*. 2004;14(3):245–257.

33. Smetana GW, Shmerling RH. Does this patient have temporal arteritis? *JAMA*. 2002;287(1):92–101.

34. Prencipe M, Casini AR, Ferretti C, et al. Prevalence of headache in an elderly population: attack frequency, disability, and use of medication. *J Neurol Neurosurg Psychiatry*. 2001;70(3):377–381.

35. International Headache Society, Headache Classification Committee. 6.4.1 Headache attributed to giant cell arteritis (GCA). *The International Classification of Headache Disorders 3rd edition (Beta version)*. 2016. Available from: https://www.ichd-3.org/6-headache-attributed-to-cranial-or-cervical-vascular-disorder/6-4-headache-attributed-to-arteritis/6-4-1-headache-attributed-to-giant-cell-arteritis-gca/. Accessed August 16, 2017.

36. Lo A, Chernoff H, Zheng T, Lo S. Why significant variables aren't automatically good predictors. *Proc Natl Acad Sci U S A*. 2015;112(45): 13892–13897.

37. Robson J, Watts R, Grayson P. AB0757 EULAR/ACR diagnostic and classification criteria of systemic vasculitis (DCVAS) study update. *Ann Rheum Dis*. 2013;71 (Suppl 3):681.

38. Gonzalez-Gay MA, Garcia-Porrua C, Llorca J, Gonzalez-Louzao C, Rodriguez-Ledo P. Biopsy-negative giant cell arteritis: clinical spectrum and predictive factors for positive temporal artery biopsy. *Semin Arthritis Rheum*. 2001;30(4):249–256.

39. Nesher G. The diagnosis and classification of giant cell arteritis. *J Autoimmun*. 2014;48–49:73–75.

40. Ellis ME, Ralston S. The ESR in the diagnosis and management of the polymyalgia rheumatica/giant cell arteritis syndrome. *Ann Rheum Dis*. 1983;42(2):168–170.

41. Vilaseca J, González A, Cid MC, Lopez-Vivancos J, Ortega A. Clinical usefulness of temporal artery biopsy. *Ann Rheum Dis*. 1987;46(4): 282–285.

42. Mukhtyar C, Guillevin L, Cid MC, et al; European Vasculitis Study Group. EULAR recommendations for the management of large vessel vasculitis. *Ann Rheum Dis*. 2009;68(3):318–323.

43. Bienvenu B, Ly KH, Lambert M, et al; Groupe d'Étude Français des Artérites des gros Vaisseaux, under the Aegis of the Filière des Maladies Auto-Immunes et Auto-Inflammatoires Rares. Management of giant cell arteritis: recommendations of the French study group for large vessel vasculitis (GEFA). *Rev Med Interne*. 2016;37(3):154–165.

44. Dasgupta B, Borg FA, Hassan N, et al; BSR and BHPR Standards, Guidelines and Audit Working Group. BSR and BHPR guidelines for the management of giant cell arteritis. *Rheumatology (Oxford)*. 2010; 49(8):1594–1597.

45. Grossman C, Ben-Zvi I, Barshack I, Bornsstein G. Association between specimen length and diagnostic yield of temporal artery biopsy. *Scand J Rheumatol*. 2017;46(3):222–225.

46. Oh LJ, Wong E, Gill AJ, McCluskey P, Smith JE. Value of temporal artery biopsy length in diagnosing giant cell arteritis. *ANZ J Surg*. Epub 2016 Nov 1.

47. Mackie SL, Taylor JC, Haroon-Rashid L, et al; UK GCA Consortium; UKRAG Consortium. Association of HLA-DRB1 amino acid residues with giant cell arteritis: genetic association study, meta-analysis and geo-epidemiological investigation. *Arthritis Res Ther*. 2015;17(1):195.