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

Edsel B Ing,1 Gabriela Lahaie Luna,2 Andrew Toren,1 Royce Ing,1 John J Chen,3 Nitika Arora,6 Nurhan Torun,7 Otana A Jakpor,8 J Alexander Fraser,9 Felix J Tyndel,10 Arun NE Sundaram,11 Xinyang Liu,11 Cindy TY Lam,1 Vivek Patel,12 Ezekiel Weis,13 David Jordan,14 Steven Gilberg,14 Christian Pagnoux,15 Martin ten Hove6

1Department of Ophthalmology and Vision Sciences, University of Toronto Medical School, Toronto; 2Department of Ophthalmology, Queen’s University, Kingston, ON; 3Department of Ophthalmology, University of Laval, Quebec, QC; 4Toronto Eyelid, Strabismus and Orbit Surgery Clinic, Toronto, ON, Canada; 5Mayo Clinic, Department of Ophthalmology and Neurology, Mayo Clinic, Department of Ophthalmology, Rochester, MN, 6Department of Surgery, Division of Ophthalmology, Harvard Medical School, Boston, MA; 7Mayo Clinic Medical School, Rochester, MN, Department of Clinical Neurological Sciences and Ophthalmology, Western University, London, 8Department of Medicine, University of Toronto Medical School, Toronto, ON, Canada; 9Department of Medicine, Fudan University Shanghai Medical College, Shanghai, People’s Republic of China; 10Roski Eye Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA; 11Departments of Ophthalmology, Universities of Alberta and Calgary, Edmonton and Calgary, AB; 12Department of Ophthalmology, University of Ottawa, Ottawa, ON; 13Vascilitis 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 claudiication, 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 $p_{cal}$ 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 articles3–7 incorporate the 1990 American College of Rheumatology (ACR) classification criteria for GCA5 to guide the decision for TABx. However, the ACR
criteria were not meant to be diagnostic criteria, 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, but few were developed using more than 500 TABx or 100 biopsy-positive GCA cases, 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. 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) 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 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)

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

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).
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 discrepan-
cies 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 sig-
nificant 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 comput-
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,29
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 parsi-
monious 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, $p_{\text{Hosmer–Lemeshow}}=0.549$, 530 jackknife
replications, 3000 bootstrap replications, log likelihood $−222.12$)**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$\beta$</th>
<th>Standard error of $\beta$</th>
<th>Odds ratio</th>
<th>$p$-value</th>
<th>$\beta_{\text{bootstr.}}$</th>
<th>Standard error of $\beta_{\text{bootstr.}}$</th>
<th>$p_{\text{bootstr.}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age in years</td>
<td>0.045</td>
<td>0.014</td>
<td>1.046</td>
<td>0.001</td>
<td>0.045</td>
<td>0.013</td>
<td>0.001</td>
</tr>
<tr>
<td>Sex (female)</td>
<td>−0.152</td>
<td>0.224</td>
<td>0.859</td>
<td>0.059</td>
<td>−0.152</td>
<td>0.269</td>
<td>0.572</td>
</tr>
<tr>
<td>New headache</td>
<td>0.231</td>
<td>0.365</td>
<td>1.259</td>
<td>0.042</td>
<td>0.231</td>
<td>0.302</td>
<td>0.446</td>
</tr>
<tr>
<td>Jaw claudication</td>
<td>1.296</td>
<td>0.964</td>
<td>3.656</td>
<td>$&lt;0.001$</td>
<td>1.296</td>
<td>0.271</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Ta abnormality</td>
<td>0.356</td>
<td>0.373</td>
<td>1.428</td>
<td>0.172</td>
<td>0.356</td>
<td>0.270</td>
<td>0.188</td>
</tr>
<tr>
<td>Vision loss</td>
<td>1.031</td>
<td>0.787</td>
<td>2.803</td>
<td>$&lt;0.001$</td>
<td>1.031</td>
<td>0.286</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Diplopia</td>
<td>−0.229</td>
<td>0.399</td>
<td>0.795</td>
<td>0.648</td>
<td>−0.229</td>
<td>0.553</td>
<td>0.679</td>
</tr>
<tr>
<td>logESR</td>
<td>0.187</td>
<td>0.206</td>
<td>1.206</td>
<td>0.273</td>
<td>0.187</td>
<td>0.154</td>
<td>0.224</td>
</tr>
<tr>
<td>logCRP</td>
<td>0.290</td>
<td>0.131</td>
<td>1.337</td>
<td>0.003</td>
<td>0.290</td>
<td>0.091</td>
<td>0.001</td>
</tr>
<tr>
<td>Platelets</td>
<td>0.005</td>
<td>0.001</td>
<td>1.005</td>
<td>$&lt;0.001$</td>
<td>0.005</td>
<td>0.001</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Constant</td>
<td>−9.967</td>
<td>0.0000751</td>
<td></td>
<td></td>
<td></td>
<td>−9.967</td>
<td>1.594</td>
</tr>
</tbody>
</table>

**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.
was negative.\textsuperscript{26,27} For completeness sake and to help guide the discussion, biopsy length was examined post hoc.

\textbf{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 \textless 50 mm/hour, and adjusted CRP \textless 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, and 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

\begin{table}
\centering
\caption{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 = $-224.51$)}
\begin{tabular}{lcccccc}
\hline
Predictor & $\beta$ & Standard error of $\beta$ & Odds ratio & p-value & $\beta_{\text{bootstrap}}$ & Standard error of $\beta_{\text{bootstrap}}$ & $p_{\text{bootstrap}}$ \\
\hline
Age in years & 0.044 & 0.013 & 1.044 & 0.001 & 0.044 & 0.012 & \textless 0.001 \\
Jaw claudication & 1.402 & 0.253 & 4.064 & \textless 0.001 & 1.402 & 0.256 & \textless 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 & \textless 0.001 & 0.005 & 0.001 & \textless 0.001 \\
Log CRP & 0.358 & 0.080 & 1.431 & \textless 0.001 & 0.358 & 0.080 & \textless 0.001 \\
Constant & -8.496 & 1.127 & 1.127 & \textless 0.001 & -8.496 & 1.116 & \textless 0.001 \\
\hline
\end{tabular}
\end{table}

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

\begin{table}
\centering
\caption{Geographic external validation of full and parsimonious models by regional site}
\begin{tabular}{lccccccc}
\hline
External validation set & n (events) & Training: validation split & Full training AUROC & Full validation AUROC & Parsimonious training AUROC & Parsimonious validation AUROC \\
\hline
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 \\
\hline
\end{tabular}
\end{table}

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.
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 ($p_{\text{original}} = 0.117$, $p_{\text{MICE}} = 0.036$) 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)

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

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

**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.

![ROC curves for full, parsimonious and ACR models.](https://www.dovepress.com/10.2147/OPTH.S158899)

**Figure 1 ROC curves for full, parsimonious and ACR models.**

**Notes:** Full model (n=530) $p_{\text{complete case}} = 0.549$. Parsimonious model (n=530) $p_{\text{complete case}} = 0.812$. ACR model = (n=525), $p_{\text{complete case}} = 0.0223$ (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.
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 \(\leq 0.07\) corresponded with a 95% chance of negative TABx and approximately 30% of the patients in our negative biopsy group had a positive TABx, suggesting that 7% of the TABx in this study could have been avoided. 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 \(\geq 0.89\) was not seen in patients with a negative biopsy.

**Discussion**

Several prediction algorithms for GCA diagnosis have been published,\(^{6,11,12,17,19,24,28}\) (Table S3) with the common goal of improving diagnostic accuracy and patient selection for TABx and for reducing patient morbidity and health care.
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

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

<table>
<thead>
<tr>
<th>Clinical scenarios</th>
<th>Full model no vision loss</th>
<th>Full model with vision loss</th>
<th>Parsimonious model no vision loss</th>
<th>Parsimonious model with vision loss</th>
<th>ACR model</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Case I)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA +</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TA Abn</td>
<td>90 yo</td>
<td>F</td>
<td>0.80</td>
<td>0.92</td>
<td>0.75</td>
</tr>
<tr>
<td>Plat 475</td>
<td>ESR 90</td>
<td>CRP 4×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Case II)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA +</td>
<td>80 yo</td>
<td>F</td>
<td>0.51</td>
<td>0.75</td>
<td>0.56</td>
</tr>
<tr>
<td>Plat 399</td>
<td>ESR 60</td>
<td>CRP 3×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Case III)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA +</td>
<td>80 yo</td>
<td>F</td>
<td>0.30</td>
<td>0.55</td>
<td>0.35</td>
</tr>
<tr>
<td>Plat 250</td>
<td>ESR 60</td>
<td>CRP 2×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Case IV)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA +</td>
<td>80 yo</td>
<td>M</td>
<td>0.28</td>
<td>0.52</td>
<td>0.35</td>
</tr>
<tr>
<td>Plat 250</td>
<td>ESR 49</td>
<td>CRP 2×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Case V)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA +</td>
<td>80 yo</td>
<td>F</td>
<td>0.14</td>
<td>0.31</td>
<td>0.11</td>
</tr>
<tr>
<td>Plat 250</td>
<td>ESR 49</td>
<td>CRP 2×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Case VI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA +</td>
<td>65 yo</td>
<td>F</td>
<td>0.10</td>
<td>0.24</td>
<td>0.11</td>
</tr>
<tr>
<td>Plat 250</td>
<td>ESR 49</td>
<td>CRP 2×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Case VII)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA +</td>
<td>50 yo</td>
<td>M</td>
<td>0.06</td>
<td>0.14</td>
<td>0.06</td>
</tr>
<tr>
<td>Plat 390</td>
<td>ESR 55</td>
<td>CRP 2×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Case IX)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA +</td>
<td>50 yo</td>
<td>M</td>
<td>0.05</td>
<td>0.14</td>
<td>0.06</td>
</tr>
<tr>
<td>Plat 250</td>
<td>ESR 55</td>
<td>CRP 2×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Case X)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA +</td>
<td>50 yo</td>
<td>F</td>
<td>0.03</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>Plat 250</td>
<td>ESR 55</td>
<td>CRP 2×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Case XI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA +</td>
<td>50 yo</td>
<td>M</td>
<td>0.02</td>
<td>0.06</td>
<td>0.03</td>
</tr>
<tr>
<td>Plat 250</td>
<td>ESR 55</td>
<td>CRP 2×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Case XII)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA +</td>
<td>50 yo</td>
<td>M</td>
<td>0.02</td>
<td>0.06</td>
<td>0.03</td>
</tr>
<tr>
<td>Plat 250</td>
<td>ESR 55</td>
<td>CRP 1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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.
heavily based on patient symptoms\textsuperscript{23} may be disadvantageous when the physician has cognitive or affective biases,\textsuperscript{30} or when patient responses are ambiguous. Many guidelines or prediction rules do not incorporate CRP\textsuperscript{15,17} and/or platelet count,\textsuperscript{8,11} which are more accurate than ESR in the diagnosis of GCA.\textsuperscript{31} Prediction rules that incorporate ESR, CRP, and platelet count are laudable\textsuperscript{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/1wlRFGeW2Vf-LlymY76KSTzlAflTrX5U_1770HbD1Y/edit?usp=sharing). Prior GCA studies have used univariate probability curves,\textsuperscript{31} theoretical decision analysis tables,\textsuperscript{15} scoring systems,\textsuperscript{13,20} or risk calculators,\textsuperscript{11} but many only provide odds ratios,\textsuperscript{12,16,17,24} or likelihood ratios\textsuperscript{14} that require extensive calculation to determine the output probability of GCA.

Figure 3 Prediction risk profile using the full model and Case 4 of Table 6.

Notes: Claudication, jaw claudication; CRP\textsubscript{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.

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, age, thrombocytosis and elevated CRP 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, in part due to epoch, lack of statistical power, and/or missing data. 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. 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, 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. 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.

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.

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/ increased 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. 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$_{c}$ (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 DCVAS57 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 al11 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 cm26 (n=1,520 TABx), 0.7 cm (n=966 TABx),2 or 1.5 cm46 (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 networks40 and support vector machine28 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>
<thead>
<tr>
<th>Table 7</th>
<th>Probability score cutoff points and risk of GCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed probability cutoff point</td>
<td>Full model observed percentile</td>
</tr>
<tr>
<td></td>
<td>Negative biopsy</td>
</tr>
<tr>
<td>2.7%</td>
<td>7th</td>
</tr>
<tr>
<td>7%</td>
<td>31st</td>
</tr>
<tr>
<td>14%</td>
<td>54th</td>
</tr>
<tr>
<td>23%</td>
<td>75th</td>
</tr>
<tr>
<td>43%</td>
<td>92nd</td>
</tr>
<tr>
<td>52%</td>
<td>95th</td>
</tr>
<tr>
<td>89%</td>
<td>99.9th</td>
</tr>
</tbody>
</table>

Notes: Results from the online calculator: https://docs.google.com/spreadsheets/d/1wRFGleV2WLVlLmY76KSTzA1TTrXSU17709Hd1Y/edit?usp=drf 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
We thank Drs Knecht and Bachmann for the use of their published data.

Disclosure
The authors report no conflicts of interest in this work.

References
20. Gabriel SE, O’Fallon WM, Achkar AA, Lie JT, Hunder GG. 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
We thank Drs Knecht and Bachmann for the use of their published data.

Disclosure
The authors report no conflicts of interest in this work.

References