Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development and pilot temporal validation of the PREST score for predicting non-response to intravenous glucocorticoids in active moderate-to-severe thyroid eye disease: a retrospective cohort study

Seunggeun Kim
Catholic University of Korea
Jungyul Park
ophjyp@naver.com

The Catholic University of Korea Seoul St. Mary's Hospital
Suk-Woo Yang
The Catholic University of Korea Seoul St. Mary's Hospital

Research Article

Keywords: Thyroid eye disease, Graves' orbitopathy, Methylprednisolone, Glucocorticoids, Treatment failure, Thyrotropin receptor antibody, Thyroid-stimulating immunoglobulin, Clinical prediction model

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11290848/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Development and pilot temporal validation of the PREST score for predicting non-response to intravenous glucocorticoids in active moderate-to-severe thyroid eye disease: a retrospective cohort study

Seunggeun Kim[1], Jungyul Park[2*], Suk-Woo Yang[2*]

[1] College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul 06591, Republic of Korea

[2] Department of Ophthalmology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul 06591, Republic of Korea

***Corresponding authors:**

Jungyul Park, Department of Ophthalmology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul 06591, Republic of Korea. Email: ophjyp@naver.com

Suk-Woo Yang, Department of Ophthalmology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul 06591, Republic of Korea. Email: yswoph@hanmail.net

Email addresses: SK, ksg6289@catholic.ac.kr; JP, ophjyp@naver.com; SWY, yswoph@hanmail.net

Abstract

Background: Intravenous glucocorticoids are the first-line treatment for active moderate-to-severe thyroid eye disease, but 20% to 60% of patients respond inadequately, and lower response rates have been reported in Asian cohorts. We aimed to develop and pilot-validate a simple score based on routine baseline measures to predict non-response.

Methods: In this retrospective single-center cohort study, adults with active (clinical activity score ≥3) moderate-to-severe thyroid eye disease who completed 12 weeks of intravenous methylprednisolone (cumulative dose 4.5 g) were divided into a development cohort (2016–2020; n = 89) and a temporal validation cohort (2021–2024; n = 24). Non-response was defined as early treatment escalation for predefined clinical reasons or improvement in fewer than two of five response components. Predictors were selected with logistic regression and converted to integer points. Discrimination (area under the receiver operating characteristic curve, AUC), calibration (Brier score), and decision curve analysis were assessed, and the score was compared with five machine-learning classifiers.

Results: Non-response occurred in 19 of 89 patients (21.3%) in the development cohort and 6 of 24 (25.0%) in the validation cohort. The PREST score assigned 3 points for thyrotropin receptor antibody ≥4.59 IU/L, 2 points for clinical activity score ≥4, 2 points for proptosis ≥19 mm, and 1 point for thyroid-stimulating immunoglobulin ≥406%. The AUC was 0.809 (95% confidence interval 0.696–0.909) in development and 0.870 (0.721–0.974) in validation. At a cutoff of 6 points, sensitivity and specificity were 73.7% and 74.3% in development and 66.7% and 77.8% in validation. The Brier score was 0.159, and decision curve analysis showed net benefit at threshold probabilities of about 10% to 20%. In development, the response rate fell from 96.3%

for scores of 0–2 to 41.2% for scores ≥7 (P for trend < 0.001). Machine-learning models discriminated less well than PREST in validation.

Conclusions: A four-item score based on routinely available serologic and ophthalmic measures stratified the risk of non-response to intravenous glucocorticoids in this single-center pilot study. Validation in larger multicenter cohorts is needed before clinical use.

Keywords: Thyroid eye disease; Graves' orbitopathy; Methylprednisolone; Glucocorticoids; Treatment failure; Thyrotropin receptor antibody; Thyroid-stimulating immunoglobulin; Clinical prediction model

Background

Thyroid eye disease (TED), also called Graves' orbitopathy, is the most common extrathyroidal manifestation of Graves' disease and can cause pain, diplopia, and sight-threatening complications [1-3]. Its annual incidence in the United States has been estimated at about 9 to 16 per 100,000 women and 1 to 3 per 100,000 men [4, 5]. The disease usually passes through an active inflammatory phase before it enters an inactive fibrotic phase. The active phase is often described as lasting about 12 to 18 months, although its duration varies between patients, and impaired quality of life can persist after inflammation has subsided [6, 7]. Orbital fibroblasts are activated through crosstalk between the thyroid-stimulating hormone receptor (TSHR) and the insulin-like growth factor-1 receptor (IGF-1R), which promotes cytokine release, hyaluronan accumulation, adipogenesis, and fibrosis [8]. Blockade of IGF-1R with teprotumumab was effective in a randomized trial [9].

For active moderate-to-severe TED, the European Group on Graves' Orbitopathy (EUGOGO) recommends intravenous glucocorticoids (IVGC), most commonly a cumulative dose of 4.5 g of methylprednisolone given over 12 weeks, as first-line therapy [10]. Disease activity is graded with the 7-item clinical activity score (CAS), and a CAS of 3 or more indicates active disease [10, 11]. Reported response rates to IVGC range from about 40% to 80%, and lower rates have been reported in Asian cohorts [6, 10, 12-14]. Patients who do not respond are still exposed to the adverse effects of glucocorticoids, such as hepatotoxicity, bone loss, and metabolic disturbance [15], while active inflammation may continue and lead to fibrotic sequelae. Second-line options now include teprotumumab, tocilizumab, rituximab, and IVGC combined with mycophenolate [9, 10, 16, 17]. Identifying likely non-responders at the start of treatment could help clinicians plan closer monitoring or earlier escalation. Several baseline factors

associated with IVGC response have been reported [18-21], and prediction models and nomograms have been developed [22-25]. A systematic review of these models reported a pooled area under the curve of 0.81 [24]. However, many published models depend on orbital magnetic resonance imaging or optical coherence tomography parameters [23, 25], which are not available in all clinics. A score based only on routinely available clinical and serologic data would be easier to apply at the bedside.

We hypothesized that baseline thyrotropin receptor antibody (TRAb), thyroid-stimulating immunoglobulin (TSI), and routine ophthalmic measures could predict non-response to IVGC and could be combined into a point-based score. The aims of this study were (1) to develop an integer risk score for IVGC non-response, the Predictive Risk Estimation of Steroid non-response in TED (PREST) score, and to assess it in a temporally separate pilot validation cohort; (2) to compare the score with machine-learning (ML) classifiers; and (3) to make the score available as a web-based calculator.

Methods

Study design and participants

This retrospective cohort study included patients with newly diagnosed TED who received IVGC between 2016 and 2024 at the Department of Ophthalmology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea. The study was approved by the Institutional Review Board of The Catholic University of Korea, Seoul St. Mary's Hospital (approval number KC26RISI0001), which waived the requirement for informed consent because of the retrospective design and the use of de-identified data. The study followed the Declaration of Helsinki and is reported according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement [26].

Patients were divided by period of treatment into a development cohort (2016 to 2020; n = 89) and a temporal validation cohort (2021 to 2024; n = 24) (Fig. 1). A temporal split was chosen instead of a random split so that all validation patients were treated after the development patients, which more closely reflects how a score would be applied to future patients [26].

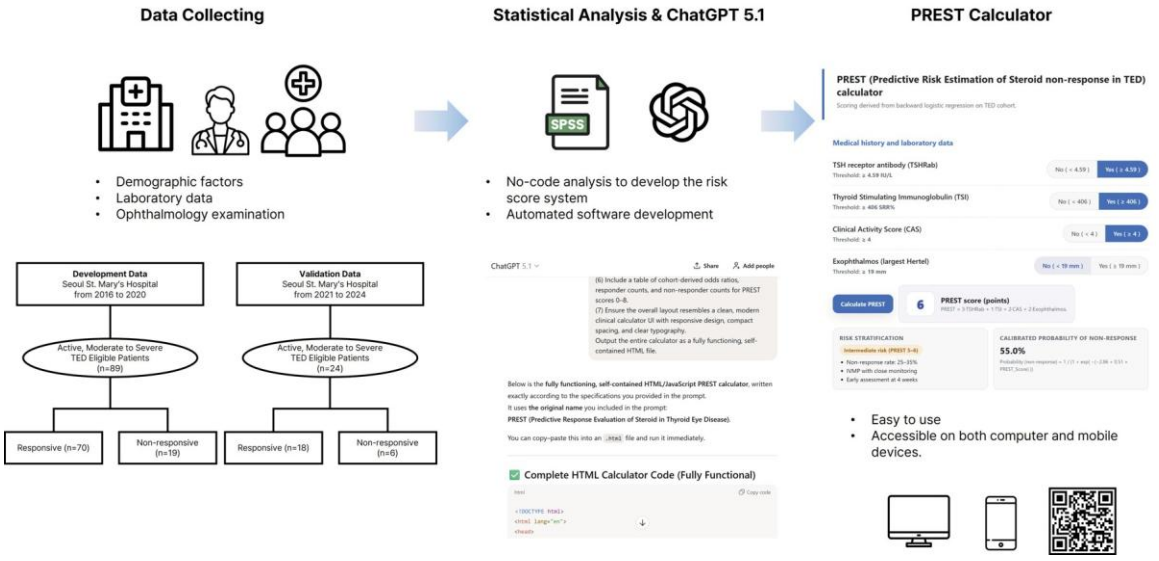


Fig. 1 Study workflow

Clinical data were collected from patients with active moderate-to-severe thyroid eye disease (TED) treated with intravenous glucocorticoids (IVGC). Patients treated in 2016 to 2020 formed the development cohort (n = 89) and those treated in 2021 to 2024 formed the temporal validation cohort (n = 24). The score was derived and validated with conventional statistical methods. A large language model (LLM; ChatGPT-5.1, OpenAI) was used to draft analysis code and the calculator interface, and all output was checked by the investigators. PREST, Predictive Risk Estimation of Steroid non-response in TED

Inclusion criteria were (1) TED diagnosed according to EUGOGO criteria [10]; (2) active (CAS of 3 or more) and moderate-to-severe disease according to EUGOGO definitions [10]; and (3) age of 19 years or older. Exclusion criteria were systemic glucocorticoid or other immunosuppressive treatment for another condition, or orbital decompression, before IVGC;

incomplete IVGC treatment; major ocular comorbidity or pre-existing strabismus; and missing values for key ophthalmic or laboratory variables (free thyroxine [free T4], triiodothyronine [T3], thyroid-stimulating hormone [TSH], TRAb, and TSI). Analyses were performed on complete cases.

Treatment protocol

All patients received intravenous methylprednisolone according to the EUGOGO regimen: 0.5 g weekly for 6 weeks followed by 0.25 g weekly for 6 weeks (cumulative dose 4.5 g) [10]. Liver function and blood glucose were monitored during treatment, and thyroid function was managed to maintain euthyroidism. Orbital radiotherapy was given at the discretion of the treating physician and was not part of a fixed protocol.

Clinical and laboratory assessments

Ophthalmic examinations were performed at baseline and at 2, 6, and 12 months after the start of IVGC. They included Snellen visual acuity, proptosis measured by exophthalmometry (mm), diplopia graded with the Bahn-Gorman scale (0 to 3), extraocular motility, and the 7-item CAS described by Mourits et al. [11], which scores spontaneous retrobulbar pain, pain on attempted upward or downward gaze, eyelid redness, conjunctival redness, eyelid swelling, chemosis, and inflammation of the caruncle or plica. Laboratory tests included TRAb, TSI, thyroid function tests, and a lipid profile. TSI was reported as the specimen-to-reference ratio (SRR%).

Candidate predictors

Candidate predictors were baseline variables selected before analysis on the basis of clinical plausibility and previous studies [18, 22-24]: age, sex, smoking status (never or ever), family history, TED duration, Graves' disease (GD) duration, prior orbital radiotherapy, CAS, increased intraocular pressure (IOP), proptosis, radiologic muscle-predominant or fat-predominant type,

extraocular muscle (EOM) limitation, diplopia, visual field index, free T4, T3, TSH, TRAb, TSI, and lipid parameters. For bilateral disease, values from the more severely affected eye were used. Prior orbital radiotherapy was defined as external-beam orbital radiotherapy given before the start of IVGC. Previous radioactive iodine therapy for hyperthyroidism was not recorded separately in the study database and could not be analyzed.

Outcome definition

The primary outcome was non-response to IVGC, defined as either of the following. (1) Early treatment escalation during the IVGC course because of new or progressive dysthyroid optic neuropathy (DON; decreased visual acuity, relative afferent pupillary defect, abnormal color vision, or a visual field defect consistent with optic nerve compression on imaging), sight-threatening corneal exposure, worsening activity (an increase in CAS of 1 point or more from baseline), or intolerable adverse effects requiring discontinuation. (2) After completion of the 12-week course, improvement in fewer than two of the following five components: reduction in proptosis of 2 mm or more in the more affected eye; reduction in CAS of 2 points or more or a final CAS below 3; improvement in visual acuity of at least one Snellen line; improvement in diplopia on the Bahn-Gorman scale; and no recurrence and no need for additional orbital radiotherapy within 6 months after IVGC. Post-treatment CAS was recorded 3 to 6 months after treatment. This definition follows the multicomponent assessment of response recommended by EUGOGO and by the consensus statement of the American Thyroid Association and the European Thyroid Association, and a similar two-of-five definition has been used in Korean patients [6, 10, 14]. The components that defined non-response in each non-responder are listed in Additional file 1: Table S8.

Statistical analysis

Continuous variables are presented as mean (standard deviation [SD]) or median (interquartile range [IQR]) according to normality assessed with the Shapiro-Wilk test, and were compared with Welch's t-test or the Mann-Whitney U test. Categorical variables were compared with the chi-square test or Fisher's exact test. Trends in response rate across score categories were tested with the Cochran-Armitage test. Ninety-five percent confidence intervals (CI) for performance measures were obtained by bootstrap resampling with 1,000 replicates. All tests were two-sided, and $P < 0.05$ was considered significant. Analyses were performed with Python version 3.9.

Score development

For score development, continuous variables were dichotomized at clinical thresholds or at the cohort median. Increased IOP was defined as 21 mmHg or higher and proptosis as 19 mm or more. TRAb was dichotomized at 4.59 IU/L (assay cutoff) and TSI at 406 SRR% (manufacturer decision limit). Thyroid function and lipid variables were dichotomized according to reference ranges (Additional file 1: Table S1). Collinearity between TRAb and TSI was assessed with the variance inflation factor (VIF) and the Pearson correlation coefficient.

Variables were screened with univariable logistic regression, and a multivariable model was built with backward stepwise elimination. Adjusted odds ratios (OR) from the final model were converted to integer points: 3 points for an OR of 5 or more, 2 points for an OR of 2 to less than 5, and 1 point for an OR of 1 to less than 2. Variables whose direction of association differed between the univariable and multivariable analyses were not assigned points.

Model performance and validation

Discrimination was assessed with the area under the receiver operating characteristic curve (AUC), and sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were calculated. The cutoff that maximized the Youden index in the

development cohort was applied unchanged to the validation cohort. Calibration was assessed with the Brier score and a calibration plot, and clinical usefulness with decision curve analysis (DCA). The performance of score versions with and without TSI was compared (Additional file 1: Table S4).

Comparison with machine-learning models

Five ML classifiers (logistic regression, support vector machine [SVM], random forest, XGBoost, and LightGBM) were trained on all candidate predictors in their continuous form. Class weights were balanced, and hyperparameters were tuned by grid search with repeated stratified 5-fold cross-validation (10 repeats) to maximize the AUC. To reduce overfitting in the tree-based models, the maximum tree depth was limited to 2 or 3 and the minimum child weight was increased. Decision thresholds were adjusted where needed to favor sensitivity.

Use of a large language model

ChatGPT-5.1 (OpenAI) was used to draft templates of the analysis code and the HTML and JavaScript code of the web calculator. All analyses were run by the investigators, and all output of the large language model (LLM) was checked against the investigators' own code and results before use. No patient-level data were entered into the LLM. Representative prompts are listed in Additional file 1: Table S2, the calculator code is provided in Additional file 1: Table S7, and the calculator is publicly available [27].

Results

Patient characteristics

In the development cohort, 70 of 89 patients (78.7%) responded to IVGC and 19 (21.3%) did not (Additional file 1: Table S3). Age, sex, smoking status, IOP, EOM limitation, diplopia, and disease duration did not differ significantly between the groups. Non-responders had a higher

baseline CAS (P = 0.02), higher TRAb (P = 0.004) and TSI (P = 0.04) levels, a lower TSH level (P = 0.04), more frequent prior orbital radiotherapy (P = 0.02), and a higher post-treatment CAS (P < 0.001). In the validation cohort, the 6 non-responders had higher TRAb levels (P = 0.02), a higher post-treatment CAS (P < 0.001), and more frequent prior orbital radiotherapy (P = 0.003) than the 18 responders.

Score derivation

TRAb and TSI were moderately correlated (Pearson r = 0.35; VIF 1.14), so both could be included in the same model. In univariable analysis, CAS of 4 or more (OR 4.24; 95% CI 1.13 to 15.87; P = 0.03), TRAb of 4.59 IU/L or more (OR 7.54; 95% CI 2.01 to 28.27; P = 0.003), TSI of 406 SRR% or more (OR 5.00; 95% CI 1.51 to 16.60; P = 0.01), abnormal TSH (OR 0.27; 95% CI 0.09 to 0.83; P = 0.02), and prior orbital radiotherapy (OR 4.36; 95% CI 1.41 to 13.51; P = 0.01) were associated with non-response, and proptosis of 19 mm or more showed a borderline association (OR 2.57; 95% CI 0.88 to 7.54; P = 0.09) (Table 1).

Table 1 Univariable and multivariable logistic regression analyses of non-response to intravenous glucocorticoids and derivation of the PREST score (development cohort, n = 89)

		Univariable logistic regression		Multivariable logistic regression		Points
Variables	**Category**	**Odds Ratio (95% CI)**	***p*-value**	**Odds Ratio (95% CI)**	***p*-value**	
Sex	Male	1.00 (reference)				
	Female	0.86 (0.31–2.41)	0.78			
Age, y	<53	1.00 (reference)				
	≥53	1.11 (0.40–3.07)	0.84			
Smoking habit	Never	1.00 (reference)				
	Ever	0.48 (0.16–1.47)	0.20			
Family history	No	1.00 (reference)				
	Yes	0.98 (0.28–3.39)	0.97			
TED duration, mo	<5	1.00 (reference)				
	≥5	0.44 (0.15–1.24)	0.12			
GD duration, mo	<9	1.00 (reference)				
	≥9	0.44 (0.15–1.24)	0.12			
Orbital radiotherapy	No	1.00 (reference)		1.00 (reference)		
	Yes	**4.36 (1.41–13.51)**	**0.01**	0.40 (0.10–1.57)	0.19	
CAS, score	<4	1.00 (reference)		1.00 (reference)		
	≥4	**4.24 (1.13–15.87)**	**0.03**	**3.82 (0.94–15.53)**	**0.05**	**2**
IIOP	No	1.00 (reference)				
	Yes	0.71 (0.14–3.53)	0.67			
Proptosis, mm	<19	1.00 (reference)		1.00 (reference)		
	≥19	2.57 (0.88–7.54)	0.09	**3.41 (1.00–11.58)**	**0.05**	**2**
Muscle predominant	No	1.00 (reference)				
	Yes	1.50 (0.44–5.07)	0.51			
EOM limitation	No	1.00 (reference)				

	Yes	0.81 (0.29–2.28)	0.69			
Diplopia	No	1.00 (reference)				
	Yes	0.67 (0.20–2.26)	0.51			
Visual Field Index	<97	1.00 (reference)				
	≥97	1.11 (0.40–3.07)	0.84			
Thyroid function test						
free T4, ng/dL	Normal	1.00 (reference)				
	Abnormal	3.36 (0.41–27.79)	0.26			
T3, ng/dL	Normal	1.00 (reference)				
	Abnormal	0.64 (0.20–2.09)	0.46			
TSH, mIU/L	Normal	1.00 (reference)		1.00 (reference)		
	Abnormal	**0.27** (0.09–0.83)	**0.02**	4.06 (0.62–26.75)	0.15	
TRAb, IU/L	<4.59	1.00 (reference)		1.00 (reference)		
	≥4.59	**7.54** (2.01–28.27)	**0.003**	**7.40 (1.87–29.26)**	**0.004**	**3**
TSI, SRR%	<406	1.00 (reference)		1.00 (reference)		
	≥406	**5.00** (1.51–16.60)	**0.01**	**1.53 (0.34–6.97)**	**0.58**	**1**
Lipid profile, mg/dL						
Total cholesterol	<200	1.00 (reference)				
	≥200	0.57 (0.18–1.76)	0.33			
LDL cholesterol	<100	1.00 (reference)				
	≥100	1.20 (0.38–3.76)	0.75			
HDL cholesterol	<60	1.00 (reference)				
	≥60	0.78 (0.25–2.43)	0.67			
Triglycerides	<150	1.00 (reference)				
	≥150	0.42 (0.14–1.31)	0.14			

CI, confidence interval; CAS, clinical activity score; EOM, extraocular muscle; GD, Graves' disease; HDL, high-density lipoprotein; IIOP, increased intraocular pressure; LDL, low-density lipoprotein; OR, odds ratio; PREST, Predictive Risk Estimation of Steroid non-response in Thyroid Eye Disease; SRR, specimen-to-reference ratio; TED, thyroid eye disease; TRAb, thyrotropin receptor antibody; TSH, thyroid-stimulating hormone; TSI, thyroid-stimulating immunoglobulin. Bold values indicate $P < 0.05$ or $P = 0.05$. Points were assigned from the adjusted odds ratios: OR ≥5, 3 points; OR 2 to <5, 2 points; OR 1 to <2, 1 point. Orbital radiotherapy and TSH were not assigned points because the direction of association differed between the univariable and multivariable analyses.

In the multivariable model, TRAb remained associated with non-response (OR 7.40; 95% CI 1.87 to 29.26; P = 0.004), and CAS of 4 or more (OR 3.82; 95% CI 0.94 to 15.53; P = 0.05) and proptosis of 19 mm or more (OR 3.41; 95% CI 1.00 to 11.58; P = 0.05) showed borderline associations (Table 1). TSI was not significant after adjustment (OR 1.53; 95% CI 0.34 to 6.97; P = 0.58) but was retained because it increased specificity and has a plausible biological role (Additional file 1: Table S4 and Figure S1). Orbital radiotherapy and TSH changed direction between the univariable and multivariable analyses and were not assigned points.

The PREST score assigned 3 points for TRAb of 4.59 IU/L or more, 2 points for CAS of 4 or more, 2 points for proptosis of 19 mm or more, and 1 point for TSI of 406 SRR% or more (Fig. 2). Scores ranged from 0 to 8 in the development cohort and from 0 to 7 in the validation cohort (Additional file 1: Figure S2).

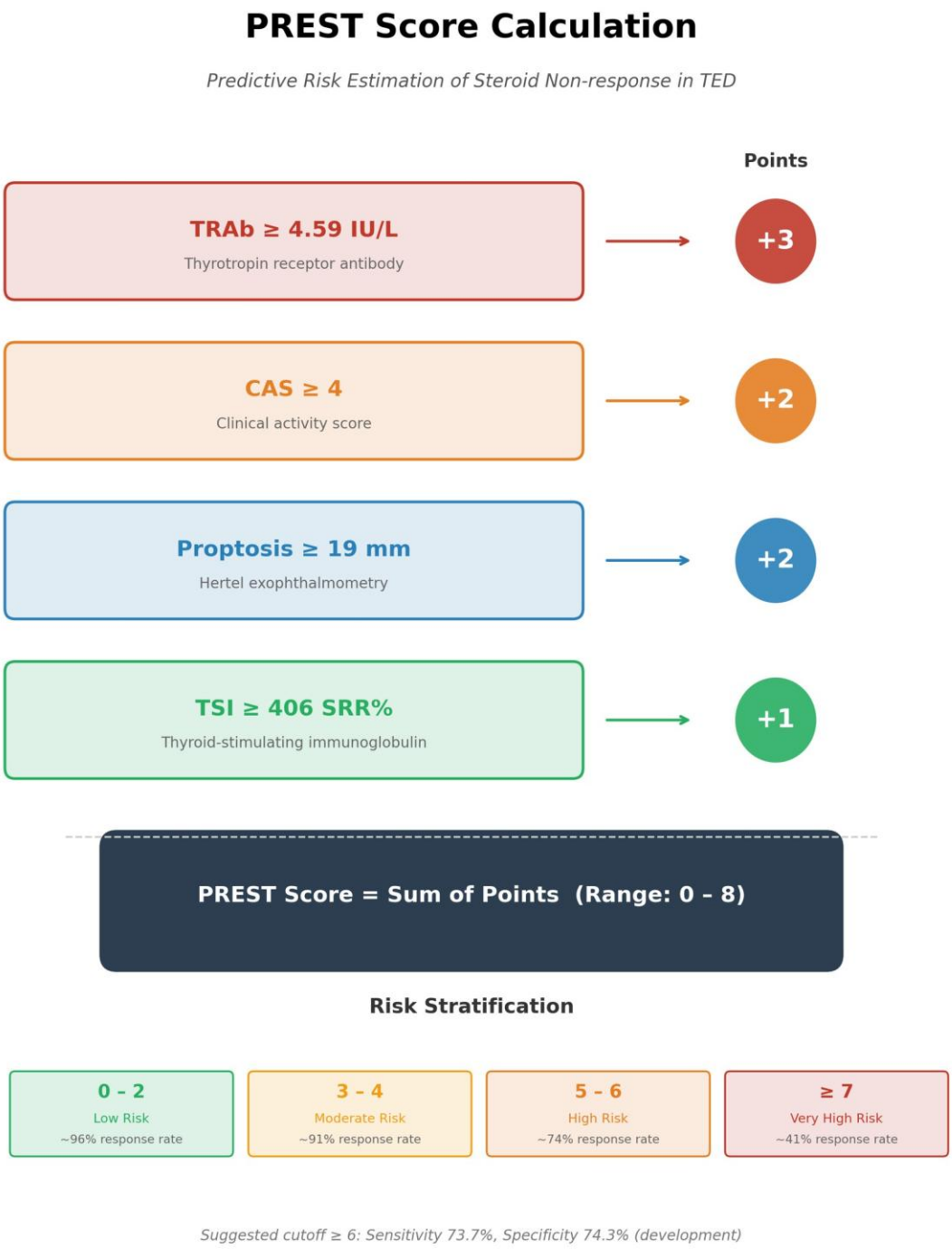


Fig. 2 Calculation of the PREST score

The score includes four baseline binary items: thyrotropin receptor antibody (TRAb) of 4.59 IU/L or more, 3 points; clinical activity score (CAS) of 4 or more, 2 points; proptosis of 19 mm or more, 2 points; and thyroid-stimulating immunoglobulin (TSI) of 406 SRR% or more, 1 point.

Total scores range from 0 to 8 and are grouped into four risk categories: low (0 to 2), moderate (3 to 4), high (5 to 6), and very high (7 or more). PREST, Predictive Risk Estimation of Steroid non-response in Thyroid Eye Disease; SRR, specimen-to-reference ratio

Performance, calibration, and clinical usefulness

In the development cohort, the AUC of the PREST score was 0.809 (95% CI 0.696 to 0.909). The Youden-optimal cutoff was 6 points, which gave a sensitivity of 73.7% and a specificity of 74.3%. In the validation cohort, the AUC was 0.870 (95% CI 0.721 to 0.974), with a sensitivity of 66.7% and a specificity of 77.8% at the same cutoff (Additional file 1: Table S5). The ML models had AUCs of 0.883 to 1.000 in the development cohort but 0.648 to 0.815 in the validation cohort (Fig. 3). The Brier score was 0.159 (Fig. 4a). In DCA, the PREST score provided a higher net benefit than treating all or no patients at threshold probabilities of about 10% to 20% (Fig. 4b).

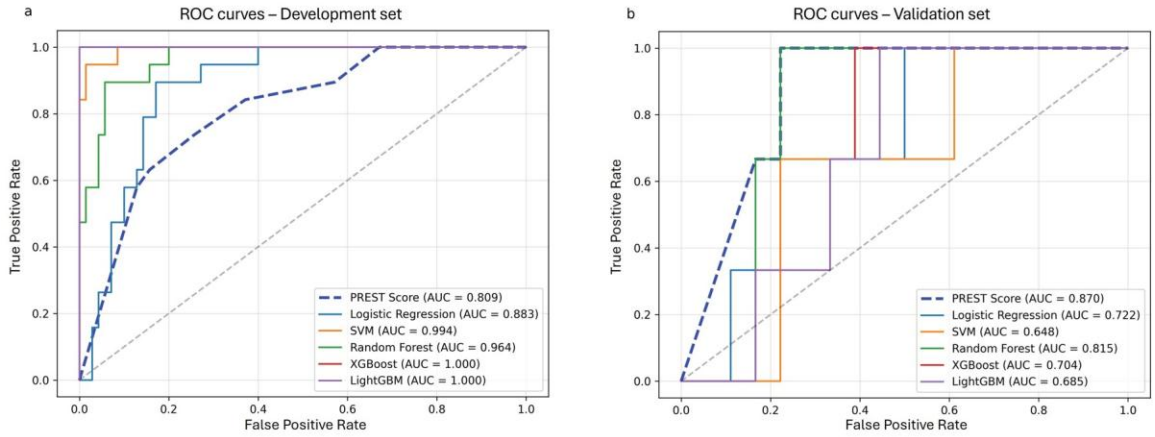


Fig. 3 Receiver operating characteristic curves of the PREST score and machine-learning models

a Development cohort. b Temporal validation cohort. The machine-learning models were logistic regression, support vector machine (SVM), random forest, XGBoost, and LightGBM. The

dashed diagonal line indicates no discrimination. AUC, area under the curve; PREST, Predictive Risk Estimation of Steroid non-response in Thyroid Eye Disease

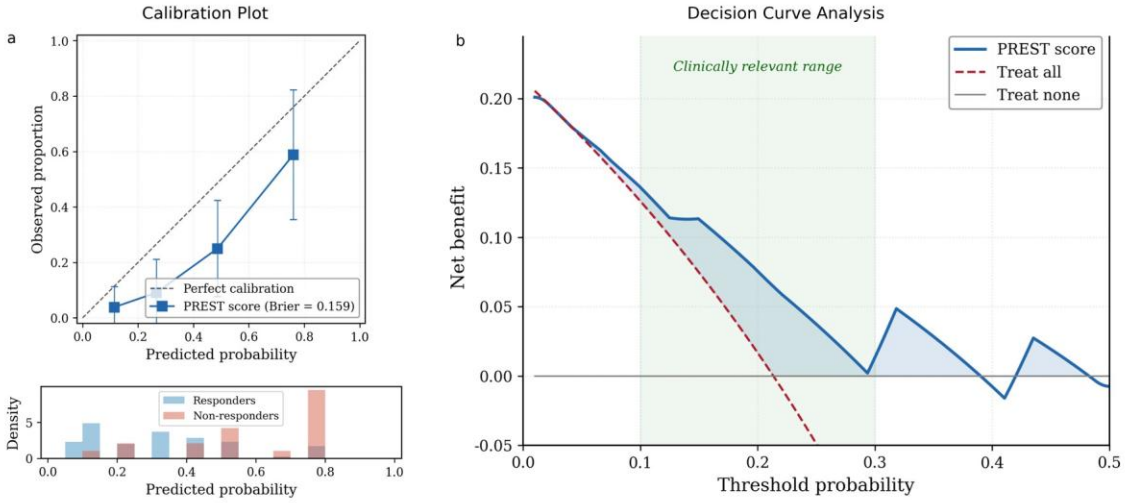


Fig. 4 Calibration and decision curve analysis of the PREST score

a Calibration plot of observed against predicted probabilities of non-response in the development cohort. The dashed diagonal line indicates perfect calibration (Brier score 0.159). b Decision curve analysis showing the net benefit of the PREST score compared with treating all or no patients across threshold probabilities. PREST, Predictive Risk Estimation of Steroid non-response in Thyroid Eye Disease

Risk categories

In the development cohort, response rates were 96.3%, 90.9%, 73.9%, and 41.2% for scores of 0 to 2, 3 to 4, 5 to 6, and 7 or more, respectively (P for trend < 0.001). In the validation cohort, the corresponding response rates were 100.0%, 100.0%, 50.0%, and 33.3% (P for trend = 0.001) (Fig. 5; Additional file 1: Table S6).

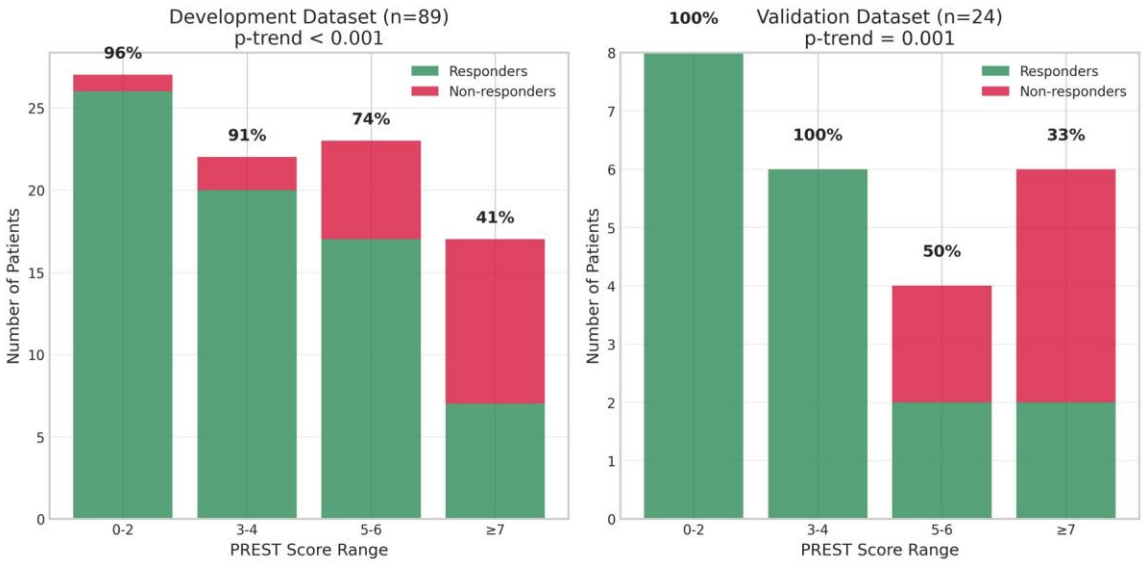


Fig. 5 Response rates by PREST score category in the development and validation cohorts

Bars show the numbers of responders and non-responders in each score category (0 to 2, 3 to 4, 5 to 6, and 7 or more) in the development cohort (n = 89) and the validation cohort (n = 24). Percentages above the bars are response rates. P values were obtained with the Cochran-Armitage test for trend (development cohort, P < 0.001; validation cohort, P = 0.001). PREST, Predictive Risk Estimation of Steroid non-response in Thyroid Eye Disease

Components of non-response

Of the 25 non-responders in both cohorts, 22 (88%) did not meet the CAS component: CAS increased in 4, and in 18 it decreased by less than 2 points and remained 3 or more. The remaining 3 patients had a CAS reduction of 2 points or more. One received triamcinolone injection for persistent activity, one showed no improvement in proptosis, ocular motility, or diplopia, and one developed DON during treatment with worsening motility and diplopia. Treatment escalation consisted of orbital decompression in 7 patients, tocilizumab in 3, adjunctive orbital radiotherapy in 7, an additional course of methylprednisolone in 2, and triamcinolone injection in 1 (Additional file 1: Table S8).

Web-based calculator

The PREST score was implemented as a web-based calculator that shows the score, the estimated probability of non-response, the risk category, and the observed response rates in the development cohort [27] (Additional file 1: Figure S3).

Discussion

In this single-center retrospective study, we developed a four-item score for predicting non-response to IVGC in active moderate-to-severe TED. The score uses TRAb, TSI, CAS, and proptosis, which are measured routinely, and can be calculated without software. Its discrimination was similar in the development and temporal validation cohorts, whereas the ML models fit the development data closely but performed less well in validation. Because the validation cohort was small and came from the same center, these results should be considered preliminary.

A higher baseline CAS was associated with non-response. This may appear to contradict the use of CAS to select patients for anti-inflammatory treatment. However, all patients in this study had active disease (CAS of 3 or more), so the comparison was between patients with active and highly active inflammation. Previous studies have reported conflicting results. Baczewska et al. found that a higher CAS before IVGC was associated with a lower likelihood of response [21], whereas Wang et al. reported that a higher pretreatment CAS was associated with a better response [19]. Differences in response definitions, timing of assessment, and disease duration may explain these discrepancies. Our finding does not suggest that patients with low CAS or inactive disease should receive glucocorticoids, and the score is intended only for patients in whom IVGC is already indicated.

TRAb was the strongest predictor of non-response. TRAb reflects the autoimmune response against TSHR, which drives orbital fibroblast activation [8], and TSI measures the stimulating

fraction of these antibodies [28, 29]. Park et al. found that TSI contributed strongly to an XGBoost model of steroid response [22], and Zloto et al. reported higher TSI levels in patients who needed second-line treatment after intravenous methylprednisolone [30]. In our cohort, TSI was not significant after adjustment, and adding it raised the AUC only slightly (0.809 vs 0.796) while increasing specificity from 68.6% to 74.3% (Additional file 1: Table S4). The cohort was too small to show whether this difference is clinically meaningful, and the version without TSI may be preferable when sensitivity is the priority. Because TRAb and TSI values depend on the assay, the thresholds used here may need to be adjusted in laboratories that use different assays.

Proptosis of 19 mm or more showed a borderline association with non-response. Greater proptosis may reflect established tissue expansion that responds less to glucocorticoids alone [12]. Prior orbital radiotherapy was associated with non-response in univariable analysis but not after adjustment. This pattern is consistent with confounding by indication, because radiotherapy is usually given to patients with more severe or persistent disease. In the CIRTED trial, orbital radiotherapy added to oral prednisolone did not improve outcomes [31].

Composite outcomes can be criticized as subjective, so we reviewed the components that defined non-response in each patient. Most non-responders did not achieve the CAS component, and the remaining patients either required additional treatment for persistent activity or DON or showed no improvement in proptosis, motility, or diplopia (Additional file 1: Table S8). In this cohort, escalation followed recorded clinical findings such as DON or persistent activity.

The discrimination of PREST was similar to the pooled AUC of 0.81 reported for published IVGC response models [24]. Unlike models that require orbital imaging or optical coherence tomography [23, 25], PREST uses four binary items that can be added at the bedside, and the

same cutoff of 6 points gave similar sensitivity and specificity in both cohorts without recalibration.

In the development cohort, non-response rates were 3.7%, 9.1%, 26.1%, and 58.8% for scores of 0 to 2, 3 to 4, 5 to 6, and 7 or more, and the validation cohort showed a similar gradient. Patients with scores of 0 to 4 may be treated with standard IVGC and routine follow-up. Patients with higher scores may benefit from closer monitoring and earlier consideration of other options, such as teprotumumab, tocilizumab, rituximab, mycophenolate, or orbital radiotherapy, according to current guidelines and local availability [9, 10, 16, 17]. The score is meant to support, not replace, the decision to start IVGC, which remains the standard first-line treatment for most patients with low scores.

We used an LLM to draft analysis code and the calculator interface. All analyses were run and checked by the investigators, and the prompts are reported so that the process can be reproduced. LLM-assisted work in clinical research requires human verification and does not replace external validation [32].

This study has several limitations. First, there was no untreated control group, so spontaneous improvement cannot be separated from the effect of glucocorticoids. Proptosis, motility, CAS, and antibody levels can improve spontaneously, and patients with more severe disease and higher antibody levels may have a longer active phase that could appear as non-response. Second, both cohorts were small and came from a single center in Korea. Temporal validation is not external validation, and the score needs to be tested in other centers, ethnic groups, and laboratories. Third, the composite outcome combines early escalation and lack of improvement after treatment; time-to-event analysis and separate analysis of each component in larger cohorts would clarify which type of failure the score predicts. Fourth, CAS is observer-dependent, and

inter-observer variability could not be assessed in this retrospective study. Fifth, complete-case analysis may have introduced selection bias. Sixth, previous radioactive iodine therapy, a known risk factor for the progression of TED, was not recorded separately. Seventh, five of the six non-responders in the validation cohort received adjunctive orbital radiotherapy, and the sequence between the decision for radiotherapy and the classification of non-response could not always be determined from the records. Eighth, the number of non-response events was small relative to the number of candidate predictors, so overfitting cannot be excluded, particularly for the ML models. Finally, dichotomizing continuous variables makes the score easy to use but loses information, and continuous or spline-based models should be explored in larger datasets.

Conclusions

PREST is a four-item score based on routinely available baseline TRAb, TSI, CAS, and proptosis that stratified the risk of non-response to IVGC in patients with active moderate-to-severe TED in this single-center pilot study, with similar performance in a temporally separate validation cohort. Validation in larger multicenter cohorts is required before the score is used to guide treatment.

List of abbreviations

AUC: Area under the receiver operating characteristic curve; CAS: Clinical activity score; CI: Confidence interval; DCA: Decision curve analysis; DON: Dysthyroid optic neuropathy; EOM: Extraocular muscle; EUGOGO: European Group on Graves' Orbitopathy; GD: Graves' disease; IGF-1R: Insulin-like growth factor-1 receptor; IOP: Intraocular pressure; IQR: Interquartile range; IVGC: Intravenous glucocorticoids; LLM: Large language model; ML: Machine learning; NPV: Negative predictive value; OR: Odds ratio; PPV: Positive predictive value; PREST: Predictive Risk Estimation of Steroid non-response in TED; SD: Standard deviation; SRR:

Specimen-to-reference ratio; SVM: Support vector machine; T3: Triiodothyronine; T4: Thyroxine; TED: Thyroid eye disease; TRAb: Thyrotropin receptor antibody; TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; TSH: Thyroid-stimulating hormone; TSHR: Thyroid-stimulating hormone receptor; TSI: Thyroid-stimulating immunoglobulin; VIF: Variance inflation factor

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Review Board of The Catholic University of Korea, Seoul St. Mary's Hospital (approval number KC26RISI0001). The Institutional Review Board waived the requirement for informed consent because of the retrospective design and the use of de-identified data. The study was conducted in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. The source code of the PREST calculator is provided in Additional file 1 and is publicly available [27].

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

SK contributed to conceptualization, data curation, formal analysis, and software development and wrote the first draft of the manuscript. JP conceived and designed the study, curated the data, developed the methodology, performed the formal analysis and software development, supervised the study, and revised the manuscript. SWY contributed to data curation and supervised the study. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Bahn RS. Graves' ophthalmopathy. N Engl J Med. 2010;362:726-38.

2. Smith TJ, Hegedüs L. Graves' disease. N Engl J Med. 2016;375:1552-65.

3. Chin YH, Ng CH, Lee MH, Koh JWH, Kiew J, Yang SP, et al. Prevalence of thyroid eye disease in Graves' disease: a meta-analysis and systematic review. Clin Endocrinol (Oxf). 2020;93:363-74.

4. Bartley GB, Fatourechi V, Kadrmas EF, Jacobsen SJ, Ilstrup DM, Garrity JA, et al. The incidence of Graves' ophthalmopathy in Olmsted County, Minnesota. Am J Ophthalmol. 1995;120:511-7.

5. Rachmasari KN, Hamadi D, Thapa P, Bradley EA, Stan MN. The epidemiology of thyroid eye disease in Olmsted County, Minnesota, 2005-2020. Thyroid. 2024;34:1522-30.

6. Burch HB, Perros P, Bednarczuk T, Cooper DS, Dolman PJ, Leung AM, et al. Management of thyroid eye disease: a consensus statement by the American Thyroid Association and the European Thyroid Association. Thyroid. 2022;32:1439-70.

7. Schovanek J, Radvansky M, Karhanova M, Schovanek M, Cibickova L, Karasek D. Long-term impact of thyroid eye disease on quality of life: insights from a retrospective cohort study. Endocr Pract. 2025;31:607-13.

8. Wiersinga WM, Eckstein AK, Žarković M. Thyroid eye disease (Graves' orbitopathy): clinical presentation, epidemiology, pathogenesis, and management. Lancet Diabetes Endocrinol. 2025;13:600-14.

9. Douglas RS, Kahaly GJ, Patel A, Sile S, Thompson EHZ, Perdok R, et al. Teprotumumab for the treatment of active thyroid eye disease. N Engl J Med. 2020;382:341-52.

10. Bartalena L, Kahaly GJ, Baldeschi L, Dayan CM, Eckstein A, Marcocci C, et al. The 2021 European Group on Graves' orbitopathy (EUGOGO) clinical practice guidelines for the medical management of Graves' orbitopathy. Eur J Endocrinol. 2021;185:G43-67.

11. Mourits MP, Koornneef L, Wiersinga WM, Prummel MF, Berghout A, van der Gaag R. Clinical criteria for the assessment of disease activity in Graves' ophthalmopathy: a novel approach. Br J Ophthalmol. 1989;73:639-44.

12. Xu Y, Liu R, Huang L, Qin Y, Liu W, Huang S, et al. Comprehensive comparisons of different treatments for active Graves' orbitopathy: a systematic review and Bayesian model-based network meta-analysis. J Clin Endocrinol Metab. 2025;110:1792-801.

13. Zang S, Ponto KA, Kahaly GJ. Clinical review: intravenous glucocorticoids for Graves' orbitopathy: efficacy and morbidity. J Clin Endocrinol Metab. 2011;96:320-32.

14. Ahn HY, Lee JK. Intravenous glucocorticoid treatment for Korean Graves' ophthalmopathy patients. J Korean Med Sci. 2020;35:e177.

15. Pofi R, Caratti G, Ray DW, Tomlinson JW. Treating the side effects of exogenous glucocorticoids; can we separate the good from the bad? Endocr Rev. 2023;44:975-1011.

16. Salvi M, Vannucchi G, Currò N, Campi I, Covelli D, Dazzi D, et al. Efficacy of B-cell targeted therapy with rituximab in patients with active moderate to severe Graves' orbitopathy: a randomized controlled study. J Clin Endocrinol Metab. 2015;100:422-31.

17. Boutzios G, Nikolopoulou A, Chatzi S, Tzioufas AG, Goules AV. Current concepts of tocilizumab efficacy in active moderate-to-severe corticosteroid-resistant cases of Graves' orbitopathy. Front Endocrinol (Lausanne). 2025;16:1632531.

18. Zhang H, Fan J, Qu J, Han Q, Zhou H, Song X. Predictive markers for anti-inflammatory treatment response in thyroid eye disease. Front Endocrinol (Lausanne). 2023;14:1292519.

19. Wang Y, Zhang S, Zhang Y, Liu X, Gu H, Zhong S, et al. A single-center retrospective study of factors related to the effects of intravenous glucocorticoid therapy in moderate-to-severe and active thyroid-associated ophthalmopathy. BMC Endocr Disord. 2018;18:13.

20. Hu S, Wang Y, He M, Zhang M, Ding X, Shi B. Factors associated with the efficacy of intravenous methylprednisolone in moderate-to-severe and active thyroid-associated ophthalmopathy: a single-centre retrospective study. Clin Endocrinol (Oxf). 2019;90:175-83.

21. Baczewska N, Alexopoulou O, Constantinescu SM, Daumerie C, Coutel M, Boschi A, et al. Factors associated with response to intravenous glucocorticoids in active moderate-to-severe thyroid eye disease. Thyroid. 2025;35:424-32.

22. Park J, Kim J, Ryu D, Choi HY. Factors related to steroid treatment responsiveness in thyroid eye disease patients and application of SHAP for feature analysis with XGBoost. Front Endocrinol (Lausanne). 2023;14:1079628.

23. Ma J, Leng Z, Zuo D, Zhang X, Zhang C, Chen H, et al. A novel nomogram to predict glucocorticoid response in thyroid-associated ophthalmopathy: findings from a pilot study. Endocrine. 2024;86:824-33.

24. Zhang H, Wu S, Hu S, Fan X, Song X, Feng T, et al. Prediction models of intravenous glucocorticoids therapy response in thyroid eye disease. Eur Thyroid J. 2024;13:e240122.

25. Zhang H, Lu T, Wang H, Hu S, Qu J, Zhou H, et al. Effectiveness of intravenous glucocorticoids in active thyroid eye disease patients with low clinical activity scores. Endocrine. 2025;90:719-29.

26. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med. 2015;13:1.

27. PREST calculator. https://justgivemeaname-oh.github.io/PREST-Calculator/. Accessed 7 Oct 2026.

28. Khamisi S, Lundqvist M, Engström BE, Larsson A, Karlsson FA, Ljunggren Ö. Comparison between thyroid stimulating immunoglobulin and TSH-receptor antibodies in the management of Graves' orbitopathy. Exp Clin Endocrinol Diabetes. 2023;131:236-41.

29. Moledina M, Roos J, Murthy R. Thyrotropin receptor autoantibody assessment in thyroid eye disease: does the assay type matter? Korean J Ophthalmol. 2023;37:147-56.

30. Zloto O, Rosset A, Priel A, Landau-Prat D, Cukierman-Yaffe T, Shavit R, et al. Elevated serum thyroid stimulating immunoglobulin linked to failure of first-line intravenous methylprednisolone monotherapy in moderate-to-severe thyroid eye disease. Eye (Lond). 2024;38:687-90.

31. Rajendram R, Taylor PN, Wilson VJ, Harris N, Morris OC, Tomlinson M, et al. Combined immunosuppression and radiotherapy in thyroid eye disease (CIRTED): a multicentre, 2 x 2 factorial, double-blind, randomised controlled trial. Lancet Diabetes Endocrinol. 2018;6:299-309.

32. Artsi Y, Sorin V, Glicksberg BS, Korfiatis P, Nadkarni GN, Klang E. Large language models in real-world clinical workflows: a systematic review of applications and implementation. Front Digit Health. 2025;7:1659134.

Additional files

Additional file 1 (.docx): Supplementary tables and figures. Figure S1. Comparison of ROC curves for PREST models developed with and without TSI. Figure S2. Distribution of PREST scores and steroid response rates. Figure S3. PREST calculator interface. Table S1. Cutoffs used for dichotomization of thyroid function and lipid parameters. Table S2. Prompts provided to ChatGPT-5.1 to develop the PREST calculator. Table S3. Baseline characteristics of patients with active moderate-to-severe thyroid eye disease by response to intravenous glucocorticoids. Table S4. Comparison of predictive performance between TSI-inclusive and TSI-exclusive PREST score models. Table S5. Comparative predictive performance of the PREST score and machine-learning models. Table S6. PREST score bands and steroid response rates. Table S7. HTML code for the PREST calculator. Table S8. Components that defined non-response in each non-responder.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx