External validation of the GREAT score to predict relapse risk in Graves' disease: results from a multicenter, retrospective study with 741 patients

Tristan Struja^{1,*}, Marina Kaeslin^{1,*}, Fabienne Boesiger^{1,*}, Rebecca Jutzi¹, Noemi Imahorn¹, Alexander Kutz¹, Luca Bernasconi², Esther Mundwiler², Beat Mueller^{1,3}, Mirjam Christ-Crain^{3,4}, Fabian Meienberg⁴, Fahim Ebrahimi⁴, Christoph Henzen^{3,5}, Stefan Fischli⁵, Marius Kraenzlin^{3,4}, Christian Meier^{3,4} and Philipp Schuetz^{1,3}

¹Medical University Department, Clinic for Endocrinology, Diabetes & Metabolism, ²Department of Laboratory Medicine, Kantonsspital Aarau, Aarau, Switzerland, ³Medical Faculty of the University of Basel, Basel, Switzerland, ⁴Clinic for Endocrinology, Diabetes & Metabolism, University Hospital of Basel, Basel, Switzerland, and ⁵Clinic for Endocrinology, Diabetes & Metabolism, Kantonsspital Luzern, Luzern, Switzerland

* (T Struja, M Kaeslin and F Boesiger contributed equally to this work)

Correspondence
should be addressed
to T Struja
Email
tristan.struja@gmail.com

Abstract

Context: First-line treatment in Graves' disease is often done with antithyroid agents (ATD), but relapse rates remain high making definite treatment necessary. Predictors for relapse risk help guiding initial treatment decisions.

Objective: We aimed to externally validate the prognostic accuracy of the recently proposed Graves' Recurrent Events After Therapy (GREAT) score to predict relapse risk in Graves' disease.

Design, setting and participants: We retrospectively analyzed data (2004–2014) of patients with a first episode of Graves' hyperthyroidism from four Swiss endocrine outpatient clinics.

Main outcome measures: Relapse of hyperthyroidism analyzed by multivariate Cox regression.

Results: Of the 741 included patients, 371 experienced a relapse (50.1%) after a mean follow-up of 25.6 months after ATD start. In univariate regression analysis, higher serum free T₄, higher thyrotropin-binding inhibitor immunoglobulin (TBII), younger age and larger goiter were associated with higher relapse risk. We found a strong increase in relapse risk with more points in the GREAT score from 33.8% in patients with GREAT class I (0–1 points), 59.4% in class II (2–3 points) with a hazard ratio of 1.79 (95% CI: 1.42–2.27, $P < 0.001$) and 73.6% in class III (4–6 points) with a hazard ratio of 2.24 (95% CI: 1.64–3.06, $P < 0.001$).

Conclusions: Based on this retrospective analysis within a large patient population from a multicenter study, the GREAT score shows good external validity and can be used for assessing the risk for relapse in Graves' disease, which influence the initial treatment decisions.

European Journal of
Endocrinology
(2017) **176**, 413–419

Introduction

Graves' disease (GD) is the most frequent cause of primary hyperthyroidism with an approximate prevalence of 0.5% (1, 2). In Europe and Asia, standard treatment for

Graves' disease includes use of antithyroid drugs (ATD) for a recommended duration of 12–18 months (3). This approach offers the possibility of disease resolution

Clinical Study	T Struja, M Kaeslin, F Boesiger and others	External validation of GREAT score	176:4	414
----------------	---	---------------------------------------	-------	-----

without residuals (i.e., without post-interventional hypothyroidism or scars) (4). Yet, risk for relapse is high (30–60%), and patients may suffer from medication side effects such as exanthema, pruritus, hepatitis and agranulocytosis. Alternative treatments such as ablation with radioactive iodine (RAI) or surgery are considered definitive treatment modalities, but often lead to persisting hypothyroidism (5, 6). Importantly, when advising patients about possible treatment options, risk for relapse is important as high-risk patients may rather choose a definite treatment as first line, whereas for low-risk patients, ATD may be preferred. Yet, there has not been a validated score for predicting relapse risk, supporting physicians and patients in treatment decisions.

Several risk factors predicting relapse after stopping ATD have been reported (2). A recent systematic review and meta-analysis identified younger age, large goiter, smoking, female sex, severe biochemical disease and higher levels of antibodies before ATD treatment as risk factors for relapse (7). Recently, the Graves' Recurrent Events After Therapy (GREAT) score, a predictive model consisting of clinical and biochemical variables, has been proposed as a tool to predict relapse by a Dutch study group (8). The investigators derived clinical and biochemical variables from a prospective study in 178 patients with their first episode of Graves' hyperthyroidism treated with ATD and found higher serum free T₄ (fT₄), younger age, higher serum thyrotropin-binding inhibitor immunoglobulin (TBII) and larger goiters as independent risk factors for relapse. Other factors such as smoking, sex, orbitopathy and thyroperoxidase antibodies were omitted as they were non-significant in the univariate model. In their study, the GREAT score discriminated patients with relapse and patients without relapse with a fair prognostic accuracy (area under the curve (AUC) 0.67, 95% confidence interval (CI): 0.54–0.77) with a 68% risk of relapse in patients with class III (4–6 points in the GREAT score) compared to 16% in patients with class I (0–1 points in the GREAT score) and 44% in patients with class II (2–3 points in the GREAT score). Additionally, PTPN22 C/T polymorphism and HLA subtypes DQB1*02, DQA1*05 and DRB1*03 were also found to be predictors for relapse. These genetic markers were incorporated into an enhanced model (called GREAT+). Genotyping in patients with relapse risk classes I and III of the GREAT score did not provide substantial benefit as most patients remained in their class. Most benefit of the addition of genotyping was seen in patients with risk class II of the GREAT score. In this group, genotyping did change risk

prediction in 38% of patients: 33% are now categorized as having a low relapse risk and medical treatment seemed advisable, whereas, on the other hand, 5% would be categorized as having a high relapse risk and thus medical management would be discouraged.

Importantly, statistical models usually fit well in the initial derivation cohort. Before using such risk factors in clinical practice, models need to be externally validated to prove generalizability. Herein, we externally validated the GREAT score in a retrospective analysis including a large patient population from four Swiss endocrine outpatient clinics.

Methods

Study population

This is a 10-year retrospective, observational cohort study. We analyzed patients from four Swiss endocrine outpatient clinics over a ten-year time span (2004–2014). The four Swiss endocrine clinics are all part of a endocrine network (three tertiary hospital-based endocrine referral centers and one large endocrine private practice) in the central and northwestern region of Switzerland using the same treatment protocols for endocrine diseases including GD. Inclusion criteria were a first episode of GD defined as suppressed serum thyrotropin (TSH) (<0.01 mU/L), elevated peripheral hormones, positive TBII, and if available diffuse, increased uptake in thyroid scintigraphy (99m-Tc-pertechnetate). Exclusion criteria were follow-up after start of ATD treatment <24 months, ATD treatment <12 months and initial ablative therapy (i.e. surgery or RAI). Relapse was defined as overt disease with suppressed TSH with elevated peripheral hormones (free T₄ (fT₄)).

Study protocol was approved by the local ethics committee for all four study sites (EKNZ Nr. 2015/227).

Clinical data

We gathered clinical data by medical charts and electronic records review, complemented by phone calls to patients and general practitioners if follow-up data were missing. We collected the following clinical parameters on admission: goiter size (WHO classification, 0–III); thyroid volume assessed by sonography; date of first ATD and the type of drug used; smoking status (yes or no); presence of Graves' orbitopathy (yes or no); anti-thyroperoxidase antibodies (anti-TPO-Ab); TBII levels and whether any other autoimmune diseases were present.

Following clinical parameters were assessed at admission and at every follow-up: date of ATD withdrawal; changes in drug regimen and reason for change (i.e. adverse effects); date of relapse and if no relapse occurred, date of last consultation; in case of relapse, type of ablative treatment (RAI or surgery); fT_4 and TSH.

There was no availability of genetic data in our cohort. All data were handled in an anonymous fashion to protect the identity of patients. All patients were usually treated for 18 months with carbimazole or propylthiouracil using a titration regime. In case of no relapse and follow-up after start of ATD treatment <24 months, patients and their primary care physicians were interviewed by telephone on disease status. We did not assess follow-up after surgery or RAI.

Laboratory measurements

After phlebotomy, samples were directly centrifuged and analyzed, as every study center had its own laboratory at hand. Serum TSH, fT_4 , anti-TPO-Ab and TBII levels were assessed by standard commercially available laboratory kits as part of the clinical routine in the different participating sites. Assays used are listed in [Supplementary Table 1](#) (see section on [supplementary data](#) given at the end of this article).

Statistical analysis

Primary objective of our study was to externally validate the GREAT score. For this purpose, we performed a similar statistical analyses as described before (8). In brief, we used univariate and multivariate Cox proportional hazard regression models to study the association of previously suggested risk factors separately and combined in the GREAT score with the risk for time to relapse. For dichotomization of variables, we used the same cut-offs as in the original report. We also calculated the GREAT score risk classes as suggested. Kaplan–Meier method was used to graphically display data with use of the log-rank test. Area under the receiver operator curve (AUC) over the whole follow-up time and restricted to 24 months of follow-up after ATD stop was used to assess the discriminative power of the GREAT score.

All significance tests were two-sided, and $P < 0.05$ was considered to be statistically significant. Categorical variables are expressed as percentages (counts) or vice versa, and continuous variables as mean and standard deviation. If applicable, a 95% confidence interval

(95% CI) is provided. Continuous variables are represented as means with standard deviation and were analyzed by two-sided tests. Categorical variables were analyzed by Pearson's chi-squared test. The statistical analysis was conducted using Stata software, version 12.1 (Stata Corp.).

Results

We included 741 patients in this cohort (79.9% females and mean age 49 ± 16 years) of which 371 (50.1%) had a relapse within a mean follow-up time of 25.6 ± 33.5 after ATD start or 14 ± 22 months after ATD withdrawal respectively ([Table 1](#)). [Table 1](#) shows details of the patient population stratified by relapse risk. Patients with relapse had higher fT_4 and TBII levels, larger goiters in clinical examination and larger thyroid volume in sonography. After relapse, 66 (33.3%) patients chose RAI of which three (1.5%) were not successfully treated. A total of 129 (65.2%) patients were referred for thyroidectomy. Graves' orbitopathy was present in roughly 2/3 of all patients (37.0% in the relapse group and 34.2% in the non-relapse group respectively). There was a tendency for thyroidectomy in patients with orbitopathy (72%, $P = 0.01$ by Pearson's chi-squared test, data not shown).

94.3% of the patients were initially treated with carbimazole, whereas the remainder received propylthiouracil. 92 patients were switched from carbimazole to propylthiouracil or vice versa. Most changes occurred because of pregnancies or skin rashes ([Supplementary Table 2](#)). Besides one case of thrombocytopenia and one of dyspnea, no serious adverse effects occurred. Especially, there was no case of liver failure or agranulocytosis.

Regression analysis

[Table 2](#) shows the results of univariate analysis and multivariate cox regression analysis. We used the same predictors and same cut-offs as used in the previous study deriving the GREAT score (8). In univariate analysis, significant associations were found for $fT_4 > 40$ pM, TBII > 6 U/L and younger age < 40 years. Also, there was a non-significant trend for more relapses with larger goiter sizes (WHO classes II and III). In multivariate analysis, there were no statistically significant associations between age, fT_4 and goiter sizes and relapse. Higher TBII (> 19.9 U/L) levels showed a statistically significant association with relapse (hazard ratio (HR) 1.76, 95% CI 1.21; 2.56), whereas those between 6 and 19.9 U/L did not.

Clinical Study	T Struja, M Kaeslin, F Boesiger and others	External validation of GREAT score	176:4	416
-----------------------	---	---------------------------------------	--------------	------------

Table 1 Baseline characteristics of patients with first episode of Graves' hyperthyroidism before the start of titration ATD therapy per outcome.

Factor	No relapse	Relapse	P-value*
Numbers	370 (49.9%)	371 (50.1%)	
Female	305 (82.4%)	287 (77.4%)	0.10
Male	65 (17.6%)	84 (22.6%)	
Age (years)	49 ± 16	49 ± 16	0.93
Treatment duration (months)	20 ± 8.7	22 ± 20	0.45
Follow-up after ATD start	36 ± 38	14 ± 23	<0.01
Follow-up after ATD withdrawal (months)	30 ± 44	14 ± 23	<0.01
BMI (kg/m ²)	25 ± 5.5	25 ± 4.8	0.86
Free T ₄ (pmol/L)	37 ± 22	45 ± 26	<0.01
TBII (U/L)	13 ± 30	19 ± 31	0.02
TPO-AK (U/L)	678 ± 2557	1101 ± 2739	0.30
Goiter size (WHO grades)			0.015
0	131 (41.9%)	109 (36.1%)	
I	115 (36.7%)	95 (31.5%)	
II	56 (17.9%)	76 (25.2%)	
III	11 (3.5%)	22 (7.3%)	
Thyroid volume (mL)	19 ± 13	22 ± 16	0.03
Orbitopathy	125 (34.2%)	136 (37.0%)	0.44
Smoking	127 (41.1%)	130 (43.5%)	0.57
Mode of ablative treatment			
RAI		66 (33.3%)	
Both		3 (1.5%)	
Surgery		129 (65.2%)	
Other autoimmune disorder			0.84
APS II	6 (20%)	6 (21%)	
GIT (IBD, CD, PA)	6 (20%)	9 (31%)	
Musculoskeletal disorders (SLE, PR)	3 (10%)	3 (10%)	
Type I diabetes mellitus			
Skin (vitiligo, psoriasis)	4 (13%)	5 (17%)	
Other	7 (23%)	4 (14%)	
	4 (13%)	2 (7%)	

*Categorical outcomes assessed by Fisher's exact test, continuous outcomes by two sample t-test.

APS, autoimmune polyglandular syndrome; CD, celiac disease; GIT, gastrointestinal tract; IBD, inflammatory bowel disease; PA, pernicious anemia; PR, polymyalgia rheumatic; RAI, radioactive iodine ablation; SLE, systemic lupus erythematosus.

Table 2 Univariate and multivariate analyzed hazard ratios for relapse after ATD withdrawal.

Risk factor	Relapse % (n/N)	Univariate		Multivariate	
		HR (95% CI)	P value	HR (95% CI)	P value
Age (years)					
≥40	33.4 (246/490)	Ref.		Ref.	
<40	33.2 (125/251)	1.35 (1.09–1.68)	<0.01	1.18 (0.85–1.63)	0.33
fT ₄ (pM)					
<40	39.1 (148/378)	Ref.		Ref.	
≥40	53.7 (131/244)	1.38 (1.09–1.75)	<0.01	1.16 (0.86–1.57)	0.34
TBII (U/L)					
<6	32.6 (79/242)	Ref.		Ref.	
6–19.9	44.4 (88/198)	1.42 (1.05–1.92)	0.03	1.36 (0.95–1.94)	0.09
>19.9	63.3 (69/109)	2.05 (1.48–2.83)	<0.01	1.76 (1.21–2.56)	<0.01
Goiter (WHO size)					
0–I	31.2 (204/450)	Ref.		Ref.	
II–III	37.3 (98/165)	1.22 (0.96–1.55)	0.11	1.05 (0.76–1.46)	0.77
GREAT class					
Class I (0–1 points)				Ref.	
Class II (2–3 points)				1.79 (1.42–2.27)	<0.001
Class III (4–6 points)				2.24 (1.64–3.06)	<0.001

Table 3 GREAT score.

Risk factor	GREAT score	Distribution of relapse
Age (y)		
≥40	0	
<40	+1	
fT ₄ (pM)		
<40	0	
≥40	+1	
TBII (U/L)		
<6	0	
6–19.9	+1	
>19.9	+2	
Goiter WHO size		
0–I	0	
II–III	+2	
Total	6 points	
AUC (95% CI)	0.68 (0.63–0.70)	
GREAT class		
Class I (0–1 points)		33.8% (107/317)
Class II (2–3 points)		59.4% (200/337)
Class III (4–6 points)		73.6% (64/87)

In addition, we investigated the prognostic performance of the GREAT score to predict relapse. We found a strong increase in relapse risk with more points in the GREAT score from 33.8% in patients with GREAT class I (0–1 points), 59.4% in class II (2–3 points) with a HR of 1.79 (95% CI: 1.42–2.27, $P < 0.001$) and 73.6% in class III (4–6 points) with a HR of 2.24 (95% CI: 1.64–3.06, $P < 0.001$) (Table 3). These findings are illustrated in a Kaplan–Meier survival curve (Fig. 1).

The GREAT score also showed a fair overall discrimination with an AUC of 0.66 (95% CI: 0.62–0.70) and slightly better (AUC 0.68; 95% CI: 0.63–0.70) if follow-up was restricted to the first 24 months (Table 3).

Figure 1

Kaplan Meier–curve of relapses.

Discussion

Based on this retrospective analysis within a large patient population from a multicenter study, the GREAT score shows good external validity. The score separated patients into three distinctive categories with a relapse risk of 33.8%, 59.4% and 73.6% each. Because objective risk assessment may help physicians when counseling patients about best initial option for GD (medical vs definite treatment), the GREAT score may be used for individually assessing the risk for relapse.

We observed an overall recurrence rate of 50.1% compared to 37% in the study by Vos *et al.* (8). Although slightly higher, our observations are in line with frequencies reported from other cohorts in the past (30–60%) (9–11). One might argue that many patients relapsed shortly after our observation. But median time to relapse was 4 months, and 43.3% of the relapses happened within 2 months. This is very similar to the replicated report (5 months and 36%) and the published literature (11). We had a high follow-up rate and performed follow-up interviews with patients and/or their primary care physicians in case there had not been a contact within the last 6 months with a study center. In Switzerland, patients typically stay with their general practitioner for many years. We thus believe to have a good data quality, and results are valid.

We were able to replicate all risk factors analyzed and reported by Vos *et al.* (8) and found univariate associations with relapse, although effect sizes were slightly smaller than those in the original cohort. Initial derivation studies are often more optimistic in regard to effect size, which makes external validation of utmost importance before risk scores are used in clinical practice. Importantly, we still found good prognostic accuracy despite differences in our data set compared to the one from Vos *et al.* (8). Most importantly, we did use different laboratory assays across study centers, which might explain some discrepancies. Nevertheless, results were still in line with the original study, which underscores the consistency of the GREAT score.

Previous attempts to predict relapse risk with scores have failed (4, 7, 11–17). Attempts were made using genetic (e.g. HLA, CD40, CTLA-4, thyroglobulin, programmed cell death-1 ligand and G alpha subunit of G protein genes), clinical (e.g. gender, age, ophthalmopathy, smoking status, treatment duration, goiter size), technical (e.g. thyroid appearance and blood perfusion assessed by sonography) and biochemical markers (e.g. free hormones, TSH, various assays of TRAb or TBII). One reason may have been that the used tools did not provide

Clinical Study	T Struja, M Kaeslin, F Boesiger and others	External validation of GREAT score	176:4	418
----------------	---	---------------------------------------	-------	-----

enough discriminative power. Another option might have been flaws in study design such as retrospective trials or small sample size leading to limited statistical power (7). It could still be that these two recent findings represent only a finding by chance. Hence, optimization of the GREAT and GREAT+ scores by additional factors with more discriminative power would be ideal.

We acknowledge several limitations in our study. First, this study is retrospective in design. However, we were able to gather most data from medical records and we did have a very high rate of follow-up. Second, we did not have access to genomic data as needed for the proposed GREAT+score. But, it should be kept in mind that genomic testing in the original cohort was most beneficial for GREAT class II patients with an intermediate risk of relapse. Genotyping was less useful in patients assigned to class I and III as it did not substantially change relapse risk prediction. Thus, our results for the other two classes should be unaffected. Further, genetic backgrounds in Switzerland and the Netherlands are expected to be rather similar.

Third, we present a mixed population from three tertiary hospital-based endocrine referral centers and one large endocrine private practice. This might rather represent a strength as it resembles everyday clinical practice.

Treatment regime between our study and the one from Vos *et al.* (8) was very similar. Mean treatment time was longer (12 vs 20 and 22 months). This is mostly due to our retrospective design, as most physicians and patients in Switzerland try an extended medical therapy before embarking to irreversible thyroid ablative procedures. We do not believe that this might have influenced the results, as a Cochrane meta-analysis found an optimal treatment duration of 12–18 months with no benefit in regard to longer durations (5). In contrast to the work of Vos *et al.* (8), we used a titration regimen instead of a block and replace approach. Aforementioned meta-analysis also compared these two treatment strategies. The authors found a slightly lower relapse rate in the block and replace studies (Peto fixed odds ratio of 0.86 (0.68, 1.08) if relapses were assumed in all lost to follow-up and Peto fixed odds ratio of 0.79 (95% CI: 0.64–0.98) in study completers only). This was achieved at the expense of a higher rate of adverse effects (5). Considering this evidence, we do not believe that the two different treatment regimens did introduce any bias. Also, our works can so far only be applied to patients from Caucasian descent. Finally, due to our inclusion criteria, seronegative patients with Graves'

hyperthyroidism are underrepresented in our study, and it remains unclear how well the GREAT score performs in this patient population.

Although much larger (741 vs 178 patients), our patient population is similar to that of Vos *et al.* (8) in regard to gender, age, smoking status and fT₄ distribution. On the other hand, we had a higher occurrence of Graves' orbitopathy, more and larger goiters and more relapses. Of note, anti-TPO-Ab and TBII levels were also higher, but we used different assays. We explain these differences by the prospective design of the study by Vos *et al.* (8). As the authors noted, a substantial number of patients were excluded ($n=67/263$). Most of them ($n=51$) could not be included as they were of non-Caucasian descent. Twelve patients preferred RAI or surgery, four were switched to propylthiouracil because of pregnancy wish and two were excluded due to adverse effects. This might also have been the main reason for their rather low relapse rate of only 36%. With the retrospective design of our work, we had the possibility to categorize these patients as non-responders to ATD therapy. Thus, we provide a relapse rate that resembles more closely every day practice. Also, this might have led to the slightly lower HR (i.e. non-significant findings for fT₄, goiter size and age) in our predictive model as we had a higher ratio of patients resistant to medical therapy. One could account differences in laboratory methods for this finding. However, risk stratification by GREAT class and findings for clinical parameters (i.e. age and goiter size) remained the same.

Conclusions

Based on this retrospective analysis within a large patient population from a multicenter study, the GREAT score shows good external validity and can be used for assessing the risk for relapse in Graves' disease, which influence initial treatment decisions. To further refine this score, future studies should focus on new biomarkers from the immune system, which may help to improve the scores ability to better identify patients at high risk for relapse.

Supplementary data

This is linked to the online version of the paper at <http://dx.doi.org/10.1530/EJE-16-0986>.

Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding

This study was supported in part by the Swiss National Science Foundation (SNSF Professorship, PP00P3_150531/1) and the Research Council of the Kantonsspital Aarau (1410.000.044). Funders had no role in the design, analysis or writing of this article.

Author contribution statement

M K, F B and T S analyzed data and wrote the first draft of the manuscript with primary responsibility for the final content. All authors read and approved the final manuscript.

References

- 1 Genovese BM, Noureldine SI, Gleeson EM, Tufano RP & Kandil E. What is the best definitive treatment for Graves' disease? A systematic review of the existing literature. *Annals of Surgical Oncology* 2013 **20** 660–667. (doi:10.1245/s10434-012-2606-x)
- 2 Longo DL, Smith TJ & Hegedüs L. Graves' disease. *New England Journal of Medicine* 2016 **375** 1552–1565. (doi:10.1056/NEJMr1510030)
- 3 Franklyn JA & Boelaert K. Thyrotoxicosis. *Lancet* 2012 **379** 1155–1166. (doi:10.1016/S0140-6736(11)60782-4)
- 4 Sundaresh V, Brito JP, Wang Z, Prokop LJ, Stan MN, Murad MH & Bahn RS. Comparative effectiveness of therapies for Graves' hyperthyroidism: a systematic review and network meta-analysis. *Journal of Clinical Endocrinology and Metabolism* 2013 **98** 3671–3677. (doi:10.1210/jc.2013-1954)
- 5 Abraham P, Avenell A, McGeoch SC, Clark LF & Bevan JS. Antithyroid drug regimen for treating Graves' hyperthyroidism. *Cochrane Database of Systematic Reviews* 2010 CD003420. (doi:10.1002/14651858.CD003420.pub4)
- 6 Orgiazzi J. Thyroid autoimmunity. *Presse Médicale* 2012 **41** 25. (doi:10.1016/j.lpm.2012.10.002)
- 7 Struja T, Fehlberg H, Kutz A, Guebelin L, Degen C, Muller B & Schuetz P. Can we predict relapse in Graves' disease? Results from a systematic review and meta-analysis. *European Journal of Endocrinology* 2017 **176** 87–97. (doi:10.1530/EJE-16-0725)
- 8 Vos XG, Endert E, Zwinderman AH, Tijssen JGP & Wiersinga WM. Predicting the risk of recurrence before the start of antithyroid drug therapy in patients with Graves' hyperthyroidism. *Journal of Clinical Endocrinology and Metabolism* 2016 **101** 1381–1389. (doi:10.1210/jc.2015-3644)
- 9 Allahabadia A, Daykin J, Holder RL, Sheppard MC, Gough SC & Franklyn JA. Age and gender predict the outcome of treatment for Graves' hyperthyroidism. *Journal of Clinical Endocrinology and Metabolism* 2000 **85** 1038–1042. (doi:10.1210/jcem.85.3.6430)
- 10 Schleusener H, Schwander J, Fischer C, Holle R, Holl G, Badenhop K, Hensen J, Finke R, Bogner U & Mayr WR *et al.* Prospective multicentre study on the prediction of relapse after antithyroid drug treatment in patients with Graves' disease. *Acta Endocrinologica* 1989 **120** 689–701. (doi:10.1530/acta.0.1200689)
- 11 Vitti P, Rago T, Chiovato L, Pallini S, Santini F, Fiore E, Rocchi R, Martino E & Pinchera A. Clinical features of patients with Graves' disease undergoing remission after antithyroid drug treatment. *Thyroid* 1997 **7** 369–375. (doi:10.1089/thy.1997.7.369)
- 12 Preus M, Frecker MF, Stensky V, Balazs C & Farid NR. A prognostic score for Graves' disease. *Clinical Endocrinology* 1985 **23** 653–661. (doi:10.1111/j.1365-2265.1985.tb01126.x)
- 13 Wei S, Pan X, Liang Z, Xu S, Cheng Y & Li J. [Evaluation of a logistic regression model in predicting the prognosis of Graves' disease treated by antithyroid drugs]. *Journal of West China University of Medical Sciences* 1991 **22** 314–317.
- 14 Rave S de, Goldschmidt HM, Bravenboer B, Loosveld OJ & Lockfeer JH. Logistic multiple regression analysis of factors predicting recurrence of hyperthyroidism after medical treatment of toxic diffuse goitre. *Netherlands Journal of Medicine* 1991 **39** 131–135.
- 15 Saleh A, Cohnen M, Furst G, Modder U & Feldkamp J. Prediction of relapse after antithyroid drug therapy of Graves' disease: value of color Doppler sonography. *Experimental and Clinical Endocrinology & Diabetes* 2004 **112** 510–513. (doi:10.1055/s-2004-821308)
- 16 Wallaschofski H, Miehle K, Mayer A, Tuschy U, Hentschel B & Paschke R. Prediction of remission or relapse for Graves' hyperthyroidism by the combined determination of stimulating, blocking and binding TSH-receptor antibodies after the withdrawal of antithyroid drug treatment. *Hormone and Metabolic Research* 2002 **34** 383–388. (doi:10.1055/s-2002-33470)
- 17 Anagnostis P, Adamidou F, Polyzos SA, Katargari S, Karathanasi E, Zouli C, Panagiotou A & Kita M. Predictors of long-term remission in patients with Graves' disease: a single center experience. *Endocrine* 2013 **44** 448–453. (doi:10.1007/s12020-013-9895-0)

Received 1 December 2016

Revised version received 11 January 2017

Accepted 17 January 2017