Can we predict relapse in Graves’ disease? Results from a systematic review and meta-analysis

Tristan Struja1,*, Hannah Fehlberg1,*, Alexander Kutz1, Larissa Guebelin1, Christian Degen1, Beat Mueller1,2 and Philipp Schuetz1,2

1 Medical University Department, Clinic for Endocrinology, Diabetes & Metabolism, Kantonsspital Aarau, Aarau, Switzerland and 2 Medical Faculty of the University of Basel, Basel, Switzerland
*(T Struja and H Fehlberg contributed equally to this work)

Abstract

Background: Identification of pretreatment risk factors predicting relapse in patients with hyperthyroidism of Graves’ disease after stopping anti-thyroid drugs (ATD) is decisive to guide therapeutic options.

Purpose: We performed a systematic search and meta-analysis to study predictors for relapse after stopping ATD in patients with Graves’ disease.

Methods: Based on a pre-specified protocol, we searched PubMed, EMBASE and Cochrane in July 2015 for case-control, controlled and randomized-controlled trials reporting risk factors for relapse after stopping ATD. The primary endpoint was relapse of disease until follow-up. PRISMA and SIGN statements were used for reviewing the data and assessing the quality of included trials.

Results: We included 54 trials with a total of 7595 participants. Most trials were small with moderate-to-high risk for bias. Ten trials were assessed only qualitatively (2227 patients), genomic data were reported in 13 trials (2178 patients) and 31 trials (4346 patients) were assessed quantitatively. In total, there were 3696 relapses in 7595 patients (48.7%). By using random-effects meta-analysis, orbitopathy, smoking, thyroid volume measured by sonography, goiter size, fT4, tT3, TRAb and TBII were significantly associated with relapse, whereas male vs female sex, age and initial tT4 level did not show significant associations.

Conclusions: This analysis found several risk factors to predict relapse in Graves’ disease, which can be combined in a risk score. Prospective studies should evaluate the prognostic accuracy of such a score to guide treatment decisions.

Introduction

Graves’ disease is the most common cause of primary hyperthyroidism in iodine-sufficient areas. The overall prevalence of Graves’ disease varies but is reported to be around 0.5% (1). Graves’ disease is characterized by the autoimmune production of thyrotropin-related antibodies (TRAb) stimulating the thyroidal cells and causing an overproduction of thyroid hormones (2). Furthermore, up to 50% of patients develop Graves’ orbitopathy by the co-expression of TSH receptors by pre-adipocytes within the orbital cavities inducing proliferation and thus proptosis (3).

One standard treatment for Graves’ disease includes the use of anti-thyroid drugs (ATD) for a total duration of 12–18 months. ATD may also have immunomodulating effects and positive impact on the course of Graves’ disease (4). This approach is generally preferred in Europe and Asia, and it offers the possibility of disease resolution without residuals (i.e., post-interventional hypothyroidism, worsening of orbitopathy) (5). Yet, risk for relapse is high (around 50%) and patients may suffer from side effects such as exanthema, pruritus, urticaria, hepatitis and agranulocytosis, especially those treated
with higher doses. Due to the high risk of relapse, an alternative first-line treatment used (mostly in the United States) is radioactive iodine (RAI) ablation, a usually ablative treatment modality (3, 6).

Several risk factors predicting relapse after stopping ATD have been reported, e.g. younger age, large goiters, smoking, male sex, severe biochemical disease and higher levels of antibodies (4). Still, there is a lack of a relapse risk prediction rule, which could be helpful when counseling patients about the best first treatment option (i.e. ATD, surgery or RAI), subjecting them to a potentially unnecessary and time-consuming ATD treatment. Herein, we performed a systematic review and meta-analysis to study the possible pretreatment risk factors of relapse of patients with hyperthyroidism due to Graves’ disease.

**Methods**

**Objective**

The objective was to study pretreatment risk factor of disease recurrence after stopping ATD in patients with a first episode of hyperthyroidism due to Graves’ disease.

**Protocol and eligibility criteria**

This review adheres to PRISMA guidelines (7). First, we generated a review protocol (Supplementary file, see section on Supplementary data given at the end of this article), outlining the main hypothesis, outcomes and search strategy. We included case–control, observational, randomized-controlled and controlled trials that assigned patients with a first episode of Graves’ disease to a standard treatment with any ATD. The diagnosis of Graves’ disease had to be established by a low thyrotropin (TSH), a high free or total thyroxine (f/tT4) or triiodothyronine (f/tT3), and if available, diffuse goiter and positive TRAb. Participants had to be adults of at least 16 years of age. There were no restrictions on the type and dosage of ATD used (e.g. methimazole, carbimazole, propylthiouracil, etc.). Also, we had no restrictions regarding language of publication, publication type or date and publication status. If studies were not available in a public library, we contacted the corresponding author. We also did include trials assessing the addition of T4 or T3 to standard ATD treatment as a former systematic review has shown that this does not affect the recurrence rate (6).

We excluded trials with a follow-up of <12 months, as well as case reports, reviews and trials not reporting relapse or risk factors, our primary endpoints of interest.

**Outcomes**

The primary outcome was relapse of hyperthyroidism according to pretreatment risk factors within a follow-up of at least 12 months after stopping ATD. We did not define any secondary outcomes.

**Search strategy and study selection**

We searched the electronic databases PubMed/Medline, EMBASE and the Cochrane CENTRAL Library (all until end of July 2015). We also hand-searched the references of leading articles to identify further articles.

Our search strategy included following terms: Graves’ disease, relapse, recurrence, goiter, hyperthyroidism, Basedow (Supplementary file for further details). The articles found were transferred into EndNote Version X7 (Thomson Reuters, Philadelphia, PA, USA) and then screened by title and abstract. Full articles were retrieved for further assessment if the information given indicated that the study included patients with their first episode of Graves’ disease, reported pretreatment risk factors and had well-defined criteria for the diagnosis of Graves’ disease and relapse. If there was any doubt about the relevance of the article from the information given in the title and abstract, the full article was consulted for clarification. All articles were reviewed by two independent authors (HF and TS). In case of disagreements, a solution was obtained by consensus and if not possible by a third reviewer (PS) (Fig. 1).

**Data collection**

For all included studies, a data extraction table was generated as suggested in the Cochrane Handbook for Systematic Reviews of Interventions (8). It was then tested on two pilot studies and refined afterward. Data were extracted by two authors (HF and CD) and cross-checked by another researcher (TS). Two review authors independently assessed trial quality (HF and TS). In case of disagreements, a solution was again obtained by consensus and if not possible by a third reviewer (PS). We suspected duplicate publications if the same authors published multiple studies on the same subject within two years. In November 2015, we tried to contact 12 trial authors (9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20) for suspected double publication of data and/or for missing data respectively. Four authors could not be contacted as there were no recent email addresses (15, 17, 19, 20). Two authors refused to share their data with us (10, 16). Until the end of February 2016, we received no answers from the other authors, although we contacted them twice.
Summary measures, planned methods of analysis and risk of bias across studies

Only risk factors reported in at least three studies were considered. Due to low number of studies, we omitted data on the following risk factors: urinary iodine, fT4I, fT3, thyroglobulin antibodies, scintigraphy, past history and family history. As none of the authors that we contacted about original data responded, we had to impute means and standard deviations for several studies as recommended, as only median and range had been reported (22). To impute a standard error for the within-study variation, Poison distribution was assumed. To calculate thyroid volume when only a weight was given, we assumed a density of 1 g per mL (same density as water).

For dichotomous data, risk ratios with 95% CI were calculated by a random-effects model (23). For continuous data, we computed the standardized mean difference according to Hedges (g) (24) as effect size as measurements were on different scales. To facilitate clinical interpretation, we also calculated weighted raw mean differences (RMD). As test of heterogeneity, we computed the variation in risk ratio (RR) across studies attributable to heterogeneity (I²) (25). Additionally, we imputed mean values for all missing data on the relapse groups in all 31 studies using a multivariate truncated linear regression model (26). To check for internal validity, we performed univariate meta-regression and compared the results to the ones obtained by classical meta-analysis. Significant results from the univariate meta-regression were fitted into a multivariate meta-regression (27) while omitting co-linear variables. Categorical data were transformed into percentages before integration into regression models. Analyses for subgroup and sensitivity analyses were planned post hoc due to the heterogeneity of studies. Gene data were assessed qualitatively, and if sufficient data were reported, we calculated the odds ratios (OR).

As most data were available for the risk factor of age, we used these to check for small study effects. For each of these trials, we plotted the effect by the inverse of its standard error (28). The symmetry of these ‘funnel plots’ was assessed both visually and formally with Egger’s test to see if the effect decreased with increasing sample size.

The statistical analysis was conducted using Stata software version 12.1 (Stata Corp.). All significance tests were two-sided, and \( P<0.05 \) was considered to be statistically significant.
Results

Study selection and study characteristics

A total of 1859 articles were evaluated by the accordingly skilled members of our study group (Fig. 1 according to suggestions of the PRISMA group (7)). Of these, 375 full-text articles were assessed for eligibility and 54 trials were included in the final data analysis (9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70). There were 31 trials (of which 5 randomized-controlled and 18 controlled trials) with 2322 relapses in a total of 4346 participants (overall relapse rate 53.4%) for quantitative meta-analysis (Supplementary Tables 1 and 2). Of note, the number of patients does not add up to the grand total as data reported were both used for the genomic data table and the quantitative analysis. Trials reporting genomic data were only evaluated qualitatively. Trials were published between 1977 and 2015. Mainly propylthiouracil and methimazole were used as ATDs.

Risk of bias within studies

Trials were mostly small with diverse risk of bias (10 high, 15 medium and 29 low risk of bias; Supplementary Table 3). Of the 31 trials included in the quantitative analysis, 12 were of low risk, 14 were of medium risk and 5 were of high risk of bias.

Qualitative reporting of gene data

Gene data were available from 13 of 54 studies (24.1%) with a total of 2178 patients being genotyped. Number of subjects ranged from 30 to 451 per trial, and trials had a moderate-to-low risk of bias (one high, five medium and seven low risk of bias; Supplementary Table 3). Results are summarized in Table 1 (9, 14, 20, 32, 35, 40, 41, 42, 51, 54, 57, 60, 70). The effect size for most markers were rather modest with OR around 1–2 in general. Many were below a size, which could be deemed clinically significant (i.e. >2), but with a wide range from 0.1 to 16.1.

Out of 35 studies testing HLA types, only 4 studies reported their result in detail. Mostly, negative or non-significant associations were omitted, apparently with high risk of publication bias. De Bruin (14) showed, contradictorily to others, no increased relapse rate for DR3-positive patients.

There were also investigations on the composition of peripheral leukocytes. One study did not find a difference in peripheral T-, B- and NK-lymphocyte subsets, to sufficiently predict relapse, even not after combination of factors (66). Van Ouwerkerk (34) reported the amount of subsets of T-lymphocytes (helper and suppressor type) without any predictive value.

Quantitative synthesis of results

Random-effects meta-analysis did show a significant predictive value for orbitopathy RR 1.16 (1.08; 1.25), smoking RR 1.13 (1.02; 1.25), thyroid volume assessed by sonography g 0.54 (0.18; 0.90)/RMD 5.46mL (1.92; 9.00), goiter size according to WHO grades g 0.18 (0.07; 0.30), fT4 g 0.16 (0.05; 0.28)/RMD 3.98pM (0.53; 7.44), tT3 g 0.25 (0.04; 0.46)/RMD 0.59nM (0.10; 1.09), TRAb g 0.33 (0.04; 0.62)/RMD 16.93U/L (1.94; 31.93), TBI g 0.36 (0.09; 0.63)/RMD 8.27% (2.61; 13.92), TSAb g 0.26 (0.06; 0.45)/RMD 127.39% (18.01; 236.77), whereas male vs female sex RR 0.91 (0.82; 1.00), age g −0.14 (−0.35; 0.07)/RMD −0.80year (−2.54; 0.94), and initial tT4 level g −0.04 (−0.31; 0.23)/RMD 2.11nM (−13.84; 18.06) did not show significant associations (Table 2).

As we computed standardized mean difference according to Hedges, values are comparable across different units of measurement. This gives a crude idea of effect size between different factors. A large drawback of this method is its difficult interpretation. Generally, effect sizes with g values around 0.1 are viewed as small, around 0.3 as medium and above 0.5 as strong (24).

For better clinical comparison, we calculated raw mean differences. Thus, an increase of the units depicted above results in a 1% increase in relapse risk (e.g. thyroid volume 1% relapse risk for every 5.46mL increase).

Additionally, we depicted the two risk factors age and goiter stratified by the subgroups relapse vs remission for age and WHO grades for goiter (Fig. 2, additional figures in Supplementary file).

In univariable meta-regression values for TBI, tT3, tT4, sonographic thyroid volume and smoking status became non-significant, whereas tT4 turned significant. As this might be due to imputation, we refitted the models with the original data, but observed no difference.

To further check for internal validity, we refitted models with values for remission. These models showed similar results except for the gender of participants were there was a collinear increase of remission and relapse with both sexes. This might be explained by the low statistical power of meta-regression and the small effect of sex on relapse risk. Also, age and tT3 showed contradictory effect sizes, which we interpreted were due to small effect sizes.
## Table 1  Associated genes for relapse.

<table>
<thead>
<tr>
<th>Gene locus</th>
<th>SNP polymorphism of gene and location</th>
<th>Receptor or transcript</th>
<th>Allele</th>
<th>Effect size (95% confidence interval) for risk of relapse</th>
<th>Reported by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>HLA-DR3</td>
<td>N/A</td>
<td>RR 0.29 (P 0.016)/OR 0.1 (0.0; 0.7) RR 2.05 (P 0.009)/OR 5.0 (1.4; 17.7)</td>
<td>de Bruin (1988), The Netherlands</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>HLA-A1</td>
<td>N/A</td>
<td>OR 0.5 (0.2; 1.3)</td>
<td>Allanic (1983), France</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>HLA-A1</td>
<td>N/A</td>
<td>OR 1.1 (0.4; 3.1)</td>
<td>Irvine (1977), UK</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>HLA-B8</td>
<td>N/A</td>
<td>OR 1.8 (0.6; 4.9)</td>
<td>Schifferdecker (1991), Germany</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>HLA-Dw12</td>
<td>N/A</td>
<td>RR 0.23 (P 0.02)/OR 0.2 (0.0; 1.1) RR 9.85 (P 0.03)/OR 9.9 (1.2; 438.0)</td>
<td>Schleusener (1987), Germany</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>HLA-DR3</td>
<td>N/A</td>
<td>RR 0.27 (P 0.05)/OR 0.3 (0.1; 1.3)</td>
<td>Schleusener (1989), Germany</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>HLA-Dw3</td>
<td>N/A</td>
<td>OR 1.2 (0.7; 2.4)</td>
<td>Wilson (1986), Scotland</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>CC</td>
<td>OR 1.2 (0.5; 2.9)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.6 (0.3; 1.4)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 3.7 (0.5; 166.2)</td>
<td>Wang (2013), Taiwan</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>CT</td>
<td>OR 2.2 (0.4; 12.3)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>CC</td>
<td>OR 1.2 (0.5; 2.9)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.6 (0.3; 1.4)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>CC</td>
<td>OR 0.9 (0.3; 2.5)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.4 (0.0; 31.9)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>CC</td>
<td>OR 1.1 (0.4; 3.5)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.9 (0.4; 2.2)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.4 (0.0; 31.9)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>CC</td>
<td>OR 1.1 (0.4; 3.5)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.9 (0.4; 2.2)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.4 (0.0; 31.9)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>CC</td>
<td>OR 1.1 (0.4; 3.5)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.9 (0.4; 2.2)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.4 (0.0; 31.9)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>CC</td>
<td>OR 1.1 (0.4; 3.5)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.9 (0.4; 2.2)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.4 (0.0; 31.9)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>CC</td>
<td>OR 1.1 (0.4; 3.5)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.9 (0.4; 2.2)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.4 (0.0; 31.9)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>CC</td>
<td>OR 1.1 (0.4; 3.5)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.9 (0.4; 2.2)</td>
<td>Kim (2007), South Korea</td>
</tr>
<tr>
<td>Ch 6p21.31</td>
<td>N/A</td>
<td>CT40</td>
<td>TT</td>
<td>OR 0.4 (0.0; 31.9)</td>
<td>Kim (2007), South Korea</td>
</tr>
</tbody>
</table>
Table 2  Summary of results.

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Classical meta-analysis</th>
<th>Univariable meta-regression</th>
<th>Multivariable meta-regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR or SMD/Hedges' g</td>
<td>Weighted raw mean difference (95% CI)</td>
<td>Heterogeneity (I²) (%)</td>
</tr>
<tr>
<td>Male vs female sex</td>
<td>RR 0.91 (0.82; 1.00)</td>
<td>N/A</td>
<td>40.30</td>
</tr>
<tr>
<td>Age (years)</td>
<td>g −0.14 (−0.35; 0.07)</td>
<td>−0.80 (−2.54; 0.94)</td>
<td>85.30</td>
</tr>
<tr>
<td>Graves' orbitopathy (y/n)</td>
<td>RR 1.16 (1.08; 1.25)*</td>
<td>N/A</td>
<td>0.00</td>
</tr>
<tr>
<td>Goiter (WHO grades)</td>
<td>g 0.18 (0.07; 0.30)*</td>
<td>N/A</td>
<td>0.00</td>
</tr>
<tr>
<td>Smoking (y/n)</td>
<td>RR 1.13 (1.02; 1.25)*</td>
<td>N/A</td>
<td>7.80</td>
</tr>
<tr>
<td>Thyroid vol. (mL)</td>
<td>g 0.54 (0.18; 0.90)*</td>
<td>5.46 (1.92; 9.00)</td>
<td>75.60</td>
</tr>
<tr>
<td>fT4 (pmol/L)</td>
<td>g 0.16 (0.05; 0.28)*</td>
<td>3.98 (0.53; 7.44)</td>
<td>2.00</td>
</tr>
<tr>
<td>tT4 (nmol/L)</td>
<td>g −0.04 (−0.31; 0.23)</td>
<td>2.11 (−13.84; 18.06)</td>
<td>83.50</td>
</tr>
<tr>
<td>tT3 (nmol/L)</td>
<td>g 0.25 (0.04; 0.46)*</td>
<td>0.59 (0.10; 1.09)</td>
<td>66.70</td>
</tr>
<tr>
<td>TRAb (U/L)</td>
<td>g 0.33 (0.04; 0.62)*</td>
<td>16.93 (1.94; 31.93)</td>
<td>75.20</td>
</tr>
<tr>
<td>TBI (% activity)</td>
<td>g 0.36 (0.09; 0.63)*</td>
<td>8.27 (2.61; 13.92)</td>
<td>75.80</td>
</tr>
<tr>
<td>TSAb (% activity)</td>
<td>g 0.26 (0.06; 0.45)*</td>
<td>127.39 (18.01; 236.77)</td>
<td>46.40</td>
</tr>
</tbody>
</table>

*P < 0.05; TRAb (thyrotropin receptor antibodies), TBII (thyrotropin-binding inhibiting immunoglobulins), TSAb (thyroid-stimulating antibody), f/tT3/4 (free/total triiodothyronine/thyroxine), thyroid volume (assessed by sonography).
size and scarcity of data respectively. These results and measurement of heterogeneity by $I^2$ fitted well with the classical meta-analysis, except the ones noted previously.

In multivariate meta-regression, we first fitted a model with all significant covariates from univariate regression. As these results could be influenced by imputation, we refitted the models with the original data, but observed no difference again. Therefore, we concluded that influential, biased studies attributed to these observed effects.

As data were handled as percentages, Table 2 shows a percentage increase or decrease for every unit of risk factor respectively. As an example, it can be depicted that a male with Graves’ orbitopathy and goiter grade II has a 47% increased risk for relapse (cumulative

Figure 2
(A) Relapse risk according to age at diagnosis. (B) Relapse risk according to WHO goiter grade.
risk = 14% + 7% + (2 × 13%) = 47%) compared to a female patient without orbitopathy and no goiter.

**Risk of bias across studies**

As there was strong evidence for heterogeneity for most risk factors, we computed a funnel plot (Supplementary Fig. 2) for age as a risk factor showing slight asymmetry. We chose age as most data were available on this risk factor. Egger’s test for no small study effects did show a significant result, thus rejecting publication bias across studies (P = 0.006). Of course, these results may also be influenced by small study effects, poor design of the included studies and true heterogeneity considering the genetic differences in the studied populations (28).

**Summary of evidence**

This is the first systematic review and meta-analysis focusing on pretreatment risk factors predicting relapse before the first treatment course with ATD in patients with hyperthyroidism due to Graves’ disease. The 54 included trials had moderate-to-low trial quality and risk of bias. The pooled analysis showed a good prediction of relapse with several risk factors, but in most trials, these factors by themselves had no enough power. Risk factors were similar in randomized trials compared to case-controlled and controlled trials.

Graves’ disease is an autoimmune disease caused by the stimulation of antibodies directed against the thyrotropin receptor in most patients (4). The risk for disease occurrence depends by part on environmental factors such as smoking and psychological factors such as stressful life events, and by part on genetic factors as previously demonstrated in twin studies (4, 71). Although environmental factors for relapse could be altered, genetic factors could not be directly influenced. But it would be possible to modify the immune system’s reaction by immunosuppressive drugs (72) until the environmental co-factors subside.

Our analysis has two main clinical benefits. First, it offers clinicians an up-to-date qualitative overview of genetic factors and other risk factor associated with the recurrence of Graves’ disease. Second, the quantitative analysis also allows to calculate the risk for relapse in an individual patient. Such numbers may be used when discussing the therapeutic options with patients, i.e. whether medical therapy or more definite treatment options (radio-iodine and surgery) should be used. Still, the effect sizes for most risk factors assessed in our analysis were small and statistically non-significant. Thus, calculation of relapse risk based on this analysis should be used with caution and only in conjunction with good clinical judgment. Also, a prospective study should validate our findings before more widespread use in clinical practice.

**Strengths and limitations**

The strengths of our study are the predefined study protocol adhering to PRISMA guidelines for systematic reviews, the inclusion from studies all around the globe without language restrictions and thus reflecting different ethnicities and genetic backgrounds (2) and the relevance of this topic in every day practice. As one might argue that the response to ATD depends on genetic background, our robust findings of the included trials from different countries suggest otherwise. This increases the clinical applicability and external validity. Although we do not provide insights into new risk factor and there have already been various systematic reviews assessing the different treatment options for Graves’ disease (1, 5, 6, 73), we provide the very first systematic review on pretreatment risk factors so far.

We are aware of several limitations in regard to this analysis. First, retrieved data were very heterogeneous and authors of eligible studies could either not be contacted or refused to share their date with us. Thus, we had to rely on several imputation methods, which could have influenced our results. For our data set, we assumed that data were missing at random. A complete-case analysis would introduce bias and lower the power of the analysis. By using multiple imputation, we assume that bias is minimal as the approach is generally considered to be conservative. Still, imputation has limitations particularly if data are not normally distributed.

Secondly, trials were mostly monocentric, heterogeneous and a lot had a rather high risk for bias. Many of the included studies were not randomized or case-controlled trials. Thirdly, there were many different laboratory assays over a wide time span of 38 years, making comparability difficult, especially for the immunological antibody assays. In addition, only two studies had blinding of participants or assessors (placebo controlled). No trial had an intention-to-treat analysis performed and adverse effects were rarely, if at all, systematically reported. Still, studies reported eligibility criteria, and mean follow-up was at least one year after the withdrawal of ATD, mainly 24 months. Notably, most relapses occur...
within the first two years. Still, studies reported eligibility criteria and mean follow-up was at least one year after the withdrawal of ATD, mainly 24 months. Heterogeneity of data did not allow for formal quantitative analysis (e.g. mean time to relapse). Hence, we qualitatively present these data without formally analyzing it (Supplementary Table 2). Roughly 2/3 of the relapses occurred between months 6 and 18 after stopping ATD, being in line with current literature (1).

Meta-regression is prone to over-interpretation of findings. However, in our case, results were similar to classical meta-analysis. Furthermore, by adjusting P values for multiple testing, we are certain that the results reflect true findings.

Overall, the markers identified by our systematic review, either alone or combined, were not strong enough to predict the clinical outcome of a single patient. Ideally, fresh markers would fill this gap. In this regard, a very recent publication from the Netherlands might shed new light. The authors generated a risk score based on a multivariate analysis of 173 prospectively enrolled patients subjected to their first ATD course. They identified lower age, higher fT4, higher TBII, large goiters, PTPN22 C/T SNP and HLA types DQB1*02, DQA1*05 and DRB1*03 as independent risk factors (74). Still, these results need to be externally validated and replicated in a larger cohort first. However, this approach would lead to a new perspective and could be the first step to a form of individualized treatment in Graves’ disease.

Conclusions

In conclusion, this systematic review and meta-analysis found several pretreatment risk factors of relapse in patients with standard treatment of Graves’ disease. The occurrence of orbitopathy, smoking, larger thyroid volume and biochemically more severe disease (fT4, tT3, TRAb, TBII and TSAb levels) did show a higher risk of relapse.

Yet, the small predictive power of a single risk factor is insufficient to predict the outcome of a single patient. This calls for novel markers to be identified and tested in future trials.

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
T S, B M and P S designed the study and H F and C D analyzed the data. H F and T S wrote the first draft of the manuscript and had primary responsibility for the final content. All authors read and approved the final manuscript.

Acknowledgments
The authors want to express their gratitude to Alexander Litke, MD, and Merih Guglielmetti for their translation of Russian and Turkish articles respectively.

References

46 Park SM, Cho YY, Joung JY, Sohn SY, Kim SW & Chung HJ. Excessive iodine intake does not increase the recurrence rate of Graves’ disease after withdrawal of the antithyroid drug in an iodine-replete area. European Thyroid Journal 2015 4 36–42. (doi:10.1159/000375261)


51 Kim KW, Park YJ, Kim TY, Park DJ, Park KS & Cho BY. Susceptible alleles of the CD40 and CTLA-4 genes are not associated with the relapse after antithyroid withdrawal in Graves’ disease. Thyroid 2007 17 1229–1234. (doi:10.1089/thy.2007.0011)


61 Choo YK, Yoo WS, Kim DW & Chung HK. Hypothyroidism during antithyroid drug treatment with methimazole is a favorable prognostic indicator in patients with Graves’ disease. Thyroid 2010 20 949–954. (doi:10.1089/thy.2009.0126)


