Supplementary Online Content

Di Bona D, Plaia A, Leto-Barone MS, La Piana S, Di Lorenzo G. Efficacy of grass pollen allergen sublingual immunotherapy tablets for seasonal allergic rhinoconjunctivitis: a systematic review and meta-analysis. Published online June 29, 2015. *JAMA Intern Med*. doi:10.1001/jamainternmed.2015.2840.

This supplementary material has been provided by the authors to give readers additional information about their work.

eMethods. Supplemental Methods

Search strategy

The primary sources of the reviewed studies were Medline, EMBASE, Cochrane Library, ClinicalTrials.gov, and the Clinical Trial Register with the following medical subject headings: rhinitis OR hay fever OR rhinosinusitis OR rhinoconjunctivitis OR conjunctivitis OR "allergic conjunctivitis" OR pollinosis AND allergen-specific immunotherapy OR immunotherapy [tiab] OR immunotherapy [mesh] OR immunotherap* AND grass. The computer search was supplemented with manual search of reference lists for all available review articles, primary studies and abstracts from meetings. There was no language restriction.

Study quality assessment

We assessed the following six categories of potential bias: 1) lack of randomization, 2) lack of allocation concealment, 3) inadequate blinding, 4) incomplete data reporting, 5) other sources of bias and 6) participation of sponsor company in the study design and interpretation of data.[43,44] For each bias category present, a point was assigned, and depending on their point count across the six categories, studies were categorized as having a low (0-1 point), medium (2-3 points), or high (4-6 points) risk of bias. If the information was lacking or unclear one point was assigned.

To assess the methodological quality of RCTs the two domains of blinding and handling withdrawals and dropouts have now been used as suggested by Jüni and colleagues.[31] The Jadad score,[30] that includes these two domains, was used. The scale ranges from 0 to 5. A score equal to or greater than 3 indicates good quality, under 3 not good quality.

Outcome measures

Symptom score (SS)

Rhinoconjunctivitis Symptom Score (SS) comprises 6 rhinoconjunctivitis symptoms (runny nose, blocked nose, sneezing, itchy nose, gritty feeling/red/itchy eyes, and watery eyes) measured as follows: 0, no symptoms; 1, mild symptoms; 2, moderate symptoms; or 3, severe symptoms. The score ranges from 0 to 18 (0-3 points for each of the 6 symptoms).

For the Pradalier[13] and the Smith[15] studies a 0-21 point SS was used, since another symptom (likewise graded from 0 to 3) was considered (nasal itching in the Pradalier study and oropharyngeal itching in the Smith study).

Medication score (MS)

Medication Score (MS) varies greatly among the studies. It is generally composed of the sum of scores for antihistamine, ocular antihistamine, nasal corticosteroid, and oral steroid use, but the points assigned to each drug varies greatly.

For example, in the Blaiss[24] and Nelson[25] studies the following MS was used:

STEP	RESCUE MEDICATION	SCORE/DOSE UNIT	MAXIMUM DAILY SCORE
1	Loratadine tablet: 10 mg, 1 tablet QD (once daily)	6 (per tablet)	6
1B	Olopatadine hydrochloride 0.1% ophthalmic solution: 1 drop in the affected eye BID (twice daily)	1.5 (per drop)	6
2	Mometasone furoate monohydrate nasal spray: 50 mg, 2 sprays in each nostril QD	2 (per spray)	8
3	Prednisone tablet: 5 mg (day 1, 1 mg/kg/d, maximum of 50 mg/d)	1.6 (per tablet)	16
3	Prednisone tablet: 5 mg (day 21, 0.5 mg/kg/d, maximum of 25 mg/d)	1.6 x 2 (per tablet)	16
Maximum daily rhinoconjunctivitis medication score			36

In contrast the Medication Score (MS) in the Cox study[26] was derived as follows: 0, no rescue medication taken; 1, use of antihistamine (oral drops, eye drops, or both); 2, use of nasal corticosteroid; and 3, use of oral corticosteroid. If a study subject took 2 or more rescue medications on the same day, the highest score was used for the MS.

Subgroup analyses

A descriptive subgroup analysis was carried out on both symptom and medication scores. A graphical procedure was preferred to a meta-regression because of the likelihood of false-positive results positively correlated with the number of characteristics investigated. The selection of characteristics defining subgroups was motivated by clinical and methodological hypotheses. We selected two categorical variables, geographic area (Europe vs. North America) and number of allergens (1 vs. 5 allergens); and five dichotomized variables (below or above median): age, gender, asthma, size (sample size of the study), and dropout and withdrawal rate (%).

Quantification of heterogeneity

The I^2 statistic, which describes the percentage of variability due to heterogeneity rather than sampling errors, was used to quantify heterogeneity. Higgins and Thompson I^2 statistics is defined as 100% (Q-df)/Q, where Q is Cochrane's heterogeneity statistic and df its degrees of freedom. I^2 ranges in 0-100, with 0 indicating no observed heterogeneity. I^2 has an appealing interpretation, being I^2=t^2/(t^2+s^2), where (t^2+s^2) is the marginal variance obtained as the sum of the within study variation s^2 and the between study variation t^2.

Influential analysis

Influential analysis—that is, the exclusion of outlying studies until homogeneity has been achieved—was also used to explore heterogeneity. This method was used to examine the effect of studies identified as being aberrant in either results or conduct.

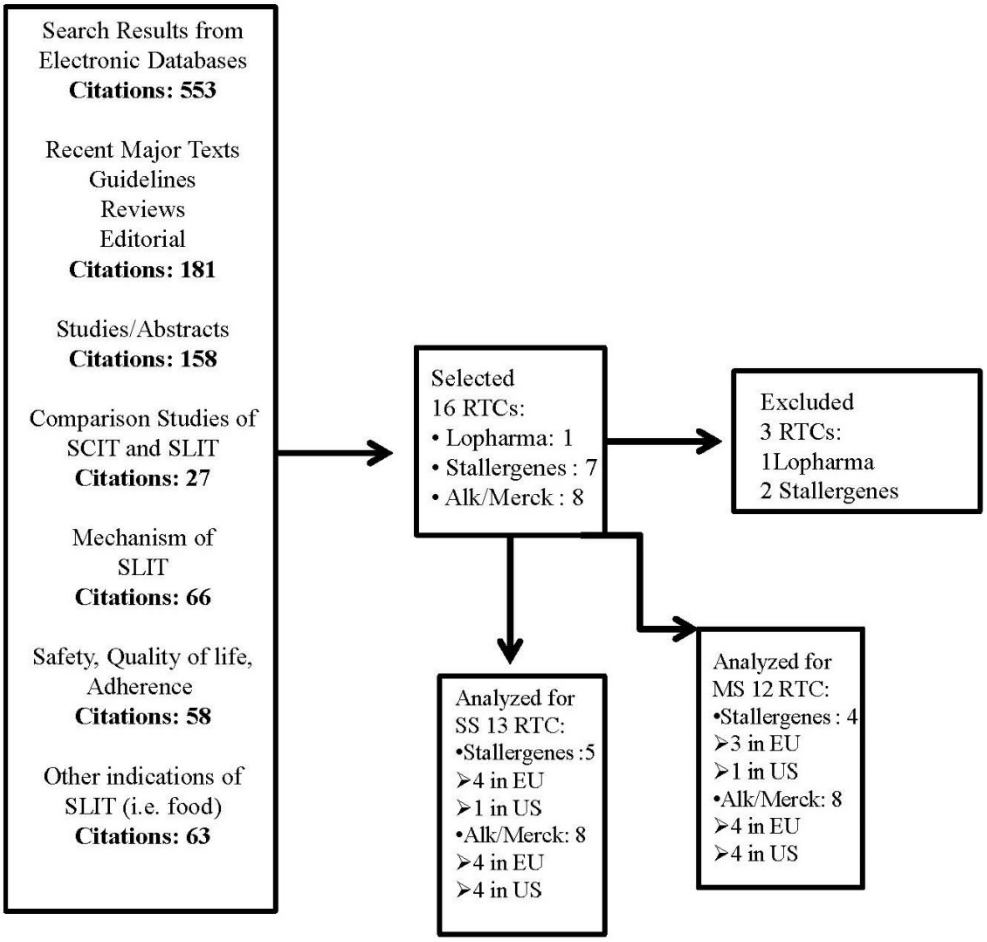
Fail-safe calculation

The fail-safe calculation is a simple procedure by which one can estimate whether publication bias (if it exists) may be safely ignored. A fail-safe number indicates the number of non-significant, unpublished (or missing) studies that would need to be added to a meta-analysis to reduce an overall statistically significant result to non-significance. If this number is large relative to the number of observed studies, one can feel fairly confident in the summary conclusions. Here the fail-safe numbers were computed following the Rosenberg approach by means of the R Metafor package.[38]

Estimation of treatment benefit

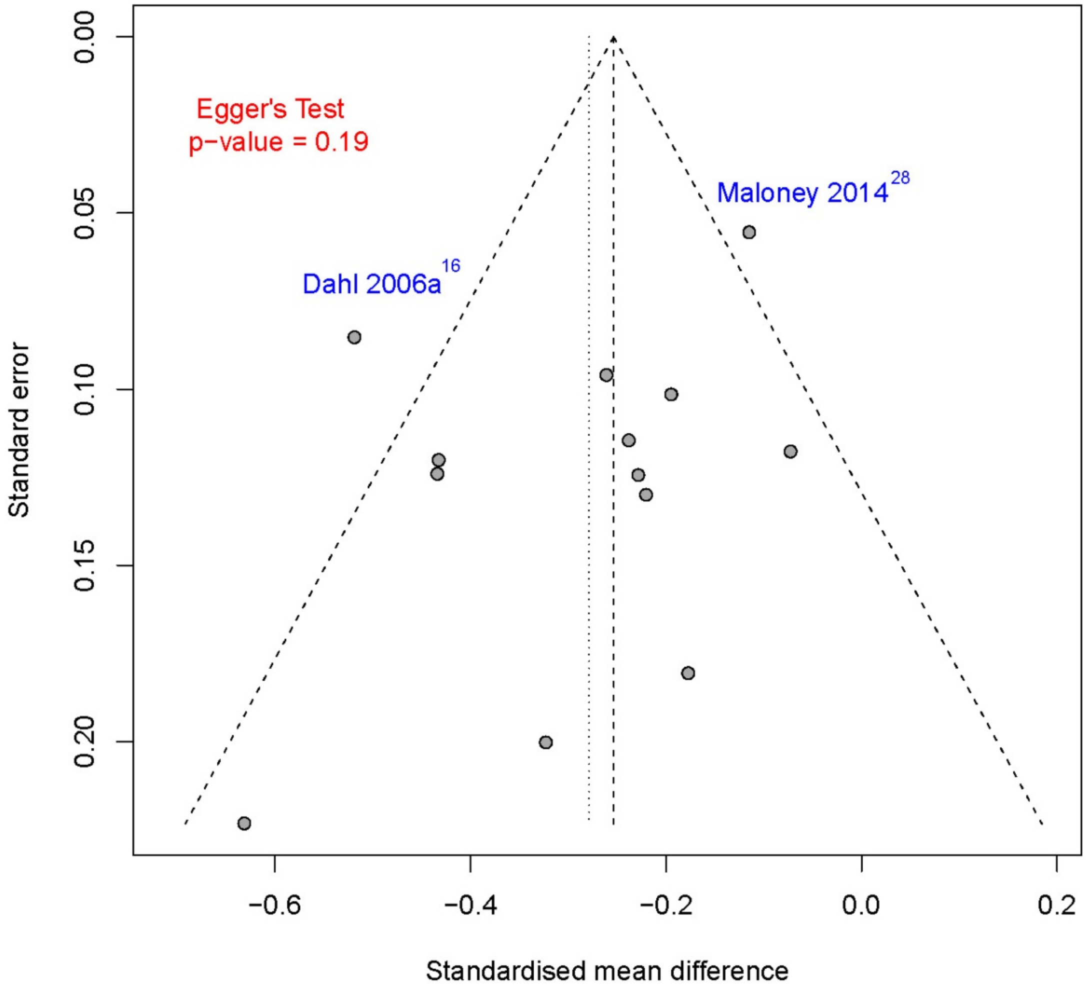
The percentage of improvement of SLIT and placebo is calculated as follows: (SSu-SSt)/SSu; where SSu and SSt represent respectively SS without any treatment (SS baseline values that are retrospectively reported in absence of any treatment for each group, SSu) and during the pollen season under treatment (SSt). SSu should be the same between active and placebo groups due to randomization. Thus, difference between the improvement in each group is the treatment benefit. Alternatively, this difference can be calculated as follows: SSt (SLIT) - SSt (Placebo) /SSu. This calculation allows us to incorporate the scale range in the evaluation of the clinical improvement, in contrast to the usual method (SLIT SS-Placebo SS /Placebo SS) that does not take into account the scale range used, which overstates the treatment effect.

eFigure 1. Flow Diagram of Sublingual Immunotherapy Studies

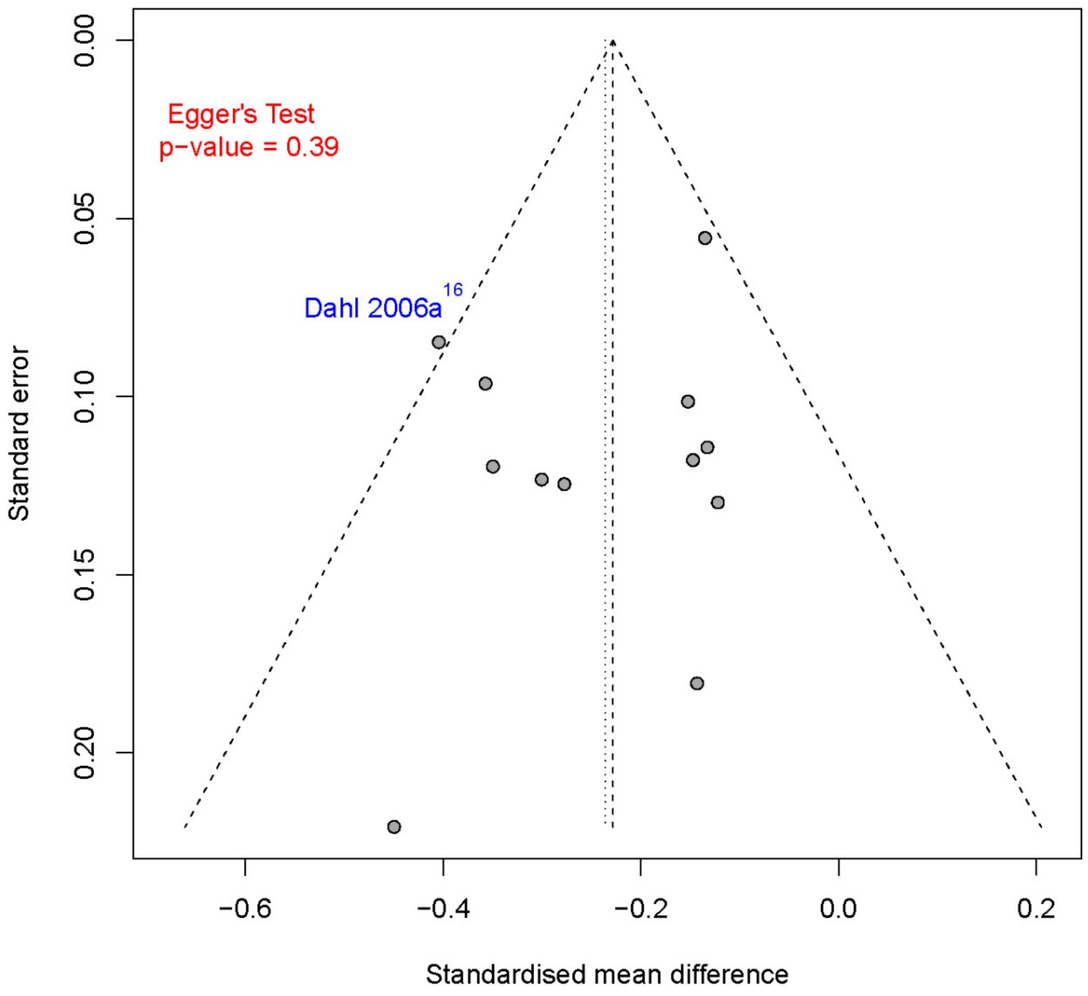


eFigure 2. Funnel Plots of (A) Symptom Score and (B) Medication Score Data, Testing for Publication Bias

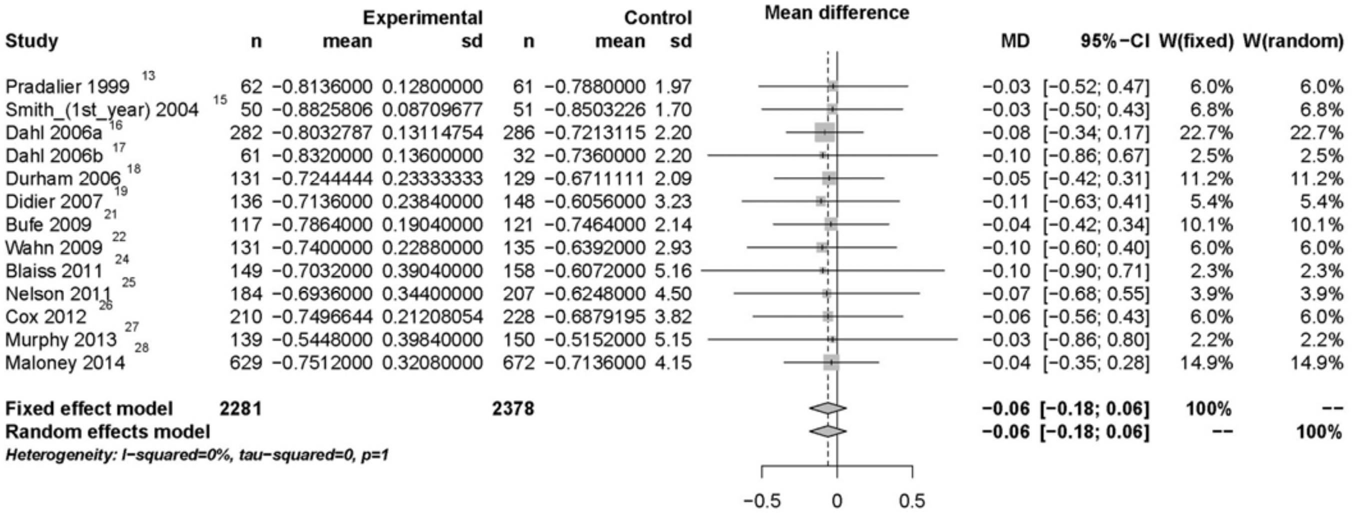
eFigure 2A



eFigure 2B


Egger's Test
p−value = 0.39
Dahl 2006a[16]
0.00
0.05
0.10
0.15
0.20
Standard error
−0.6
−0.4
−0.2
0.0
0.2
Standardised mean difference

eFigure 3. Meta-analysis of the Efficacy of SLIT vs. Placebo for ARC (SS, Mean Difference)



		Experimental			Control					
Study	**n**	**mean**	**sd**	**n**	**mean**	**sd**	**MD**	**95%−CI**	**W(fixed)**	**W(random)**
Pradalier 1999 [13]	62	−0.8136000	0.12800000	61	−0.7880000	1.97	−0.03	[−0.52; 0.47]	6.0%	6.0%
Smith_(1st_year) 2004 [15]	50	−0.8825806	0.08709677	51	−0.8503226	1.70	−0.03	[−0.50; 0.43]	6.8%	6.8%
Dahl 2006a [16]	282	−0.8032787	0.13114754	286	−0.7213115	2.20	−0.08	[−0.34; 0.17]	22.7%	22.7%
Dahl 2006b [17]	61	−0.8320000	0.13600000	32	−0.7360000	2.20	−0.10	[−0.86; 0.67]	2.5%	2.5%
Durham 2006 [18]	131	−0.7244444	0.23333333	129	−0.6711111	2.09	−0.05	[−0.42; 0.31]	11.2%	11.2%
Didier 2007 [19]	136	−0.7136000	0.23840000	148	−0.6056000	3.23	−0.11	[−0.63; 0.41]	5.4%	5.4%
Bufe 2009 [21]	117	−0.7864000	0.19040000	121	−0.7464000	2.14	−0.04	[−0.42; 0.34]	10.1%	10.1%
Wahn 2009 [22]	131	−0.7400000	0.22880000	135	−0.6392000	2.93	−0.10	[−0.60; 0.40]	6.0%	6.0%
Blaiss 2011 [24]	149	−0.7032000	0.39040000	158	−0.6072000	5.16	−0.10	[−0.90; 0.71]	2.3%	2.3%
Nelson 2011 [25]	184	−0.6936000	0.34400000	207	−0.6248000	4.50	−0.07	[−0.68; 0.55]	3.9%	3.9%
Cox 2012 [26]	210	−0.7496644	0.21208054	228	−0.6879195	3.82	−0.06	[−0.56; 0.43]	6.0%	6.0%
Murphy 2013 [27]	139	−0.5448000	0.39840000	150	−0.5152000	5.15	−0.03	[−0.86; 0.80]	2.2%	2.2%
Maloney 2014 [28]	629	−0.7512000	0.32080000	672	−0.7136000	4.15	−0.04	[−0.35; 0.28]	14.9%	14.9%
Fixed effect model	**2281**			**2378**			**−0.06**	**[−0.18; 0.06]**	**100%**	**--**
Random effects model							**−0.06**	**[−0.18; 0.06]**	**--**	**100%**

Heterogeneity: I−squared=0%, tau−squared=0, p=1

The percent improvement from baseline to treatment period was calculated for each group and the comparison (the pooled estimate of -0.06 represents a 6% difference) is reported in the forest plot.

eTable 1. Study Quality Assessment by the Jadad Score. A score ≥3 indicated good quality, under 3 not-good quality.

Study	Randomization used	Double-blinding	Dropouts and withdrawals	Generation of random numbers	Allocation concealment	Score
Pradalier 1999[13]	1	1	1	n.r.	?	3
Smith 2004[15]	1	1	1	n.r.	?	3
Dahl (a) 2006[16]	1	1	1	n.r.	?	3
Dahl (b) 2006[17]	1	1	1	n.r.	?	3
Durham 2006[18]	1	1	1	1	?	4
Didier 2007[19]	1	1	1	1	?	4
Bufe 2009[21]	1	1	1	n.r.	?	3
Wahn 2009[22]	1	1	1	1	?	4
Blaiss 2011[24]	1	1	1	1	?	4
Nelson 2011[25]	1	1	1	1	?	4
Cox 2012[26]	1	1	1	1	?	4
Murphy 2013[27]	1	1	1	1	?	4
Maloney 2014[28]	1	1	1	1	?	4

n.r., not-reported; ?, unclear.

eTable 2. Potential Bias

	1	2	3	4	5	6	Total
Authors							
Pradalier 1999[13]	?	?	No	No	No	Yes	3
Smith 2004[15]	?	?	No	Yes	No	Yes	4
Dahl (a) 2006[16]	?	?	No	No	No	Yes	3
Dahl (b) 2006[17]	Yes	?	No	No	No	Yes	2
Durham 2006[18]	Yes	?	No	No	No	Yes	2
Didier 2007[19]	Yes	?	No	No	No	Yes	2
Bufe 2009[21]	?	?	No	No	No	Yes	3
Wahn 2009[22]	Yes	?	No	No	No	Yes	2
Blaiss 2011[24]	Yes	?	No	No	No	Yes	2
Nelson 2011[25]	Yes	?	No	No	No	Yes	2
Cox 2012[26]	Yes	?	No	No	No	Yes	2
Murphy 2013[27]	?	?	No	No	No	Yes	3
Maloney 2014[28]	Yes	?	No	No	No	Yes	2

Items (1) Adequate sequence generation, (2) lack of allocation concealment, (3) inadequate blinding, (4) incomplete data reporting, (5) other sources of bias, (6) participation of sponsor company in the study design and interpretation of data. For each bias category present, a point was assigned, and, depending on their point count across the six categories, studies were categorized as having a low (0-1 point), medium (2-3 points), or high (4-6 points) risk of bias. If the information was lacking or unclear (?) a point was assigned.

eTable 3. Most Common Treatment-Related Adverse Events (TRAE), Occurring in at Least 5% of Patients in the Treatment Group

TRAD	SLIT n (%)	Placebo	*p*	#Studies
Oral pruritus	689/2228 (30.9)	84/2126 (3.9)	<0.00001	11/13
Throat irritation	418/2045 (20.4)	71/2006 (3.5)	<0.00001	9/13
Mouth edema	226/2105 (10.7)	17/2033 (0.8)	<0.00001	9/13
Ear pruritus	181/1524 (11.9)	32/1444 (2.2)	<0.00001	6/13
Eye pruritus	81/852 (9.5)	20/768 (2.6)	<0.00001	4/13
Oropharingeal pain	122/1306 (9.3)	33/1309 (2.5)	<0.00001	4/13

Other side effects, such as headache, cough, tongue pruritus, sneezing, rhinorrhea, nasal discomfort, naso-pharyngitis have not been reported in the table since they were reported in less than four studies.

eTable 4. Proposed Calculation for Quantification of Symptom Score Differences Between Groups

Cox study[26]	SLIT	Placebo	Difference	Percentage improvement
RRTSS	14.9	14.9		
Mean SS during treatment	3.21	4.16	-0.95	(3.21-4.16)/4.16=**22.9%** **(not including the scale)**
Difference	-11.69	-10.74	-0.95	
Percentage improvement	78%	72%		78%-72%=**6%** **(10.74-11.69)/14.9=6%** **(including the scale)**

RRTSS, Retrospective Rhino-conjunctivitis Total Symptom Score.[26] Horizontal arrow: calculation of improvement in RCTs: Vertical Arrow: calculation of improvement including the scale. With the calculation shown in RCTs (horizontal arrow) only the mean SS during the treatment is considered, ignoring the scale range. In our proposed calculation (vertical arrow) the scale range is included. The inclusion of the scale in the calculation changes the percentage of the improvement, even if the difference between the two groups remains the same.

eResults. Supplemental Results

Subgroup analyses

Figure 2 (in the text) summarizes the effect of SLIT on symptom and medication score in subgroups of trials categorized according to study characteristics. These data suggest that the benefit of SLIT is more pronounced in European studies (SS median SMD, -0.38, I^2=18%; MS median SMD, -0.30, I^2=0%)[13,15-19, 21,22] than in studies conducted in North America (SS median SMD, -0.19, I^2=0%; MS median SMD, -0.15, I^2=7%),[24-28] with no evidence of heterogeneity, and that the 5 allergen grass extract tablets are more effective (SS median SMD, -0.32, I^2=0; MS median SMD, -0.32, I^2=0)[13,15,19,22,26] than tablets with 1 pollen extract (SS median SMD, -0.22, I^2=66%; MS median SMD, -0.15, I^2=30%).[16-18,21,24,25,27,28] However, this latter difference may reflect the inclusion in this group of both European and North American studies, leading to the high heterogeneity reported in this subgroup for symptom score.

One study reported the difference in the symptom score between SLIT and placebo according to a per-protocol analysis.[17] Five other studies are at high[15] or moderate risk of bias.[13,16,21,27] A sensitivity analysis excluding these six studies produced similar results (SMD -0.25, 95% CI, -0.34, -0.15; p<0.0001), suggesting that trial quality affects outcomes only marginally.

The completion rates of RCTs were very different, raising the possibility of withdrawal bias. The effect of withdrawal or dropouts was analyzed in two categories: studies with a dropout/withdrawal rate below the median value of 14%,[13,18,19,21,22,25,26] and studies with a dropout/withdrawal rate of 14% or greater.[15,16,17,24,27,28] We found no significant difference between studies with high dropout/withdrawal rate (SS median SMD, -0.22, I^2=72%; MS median SMD, -0.15, I^2=46%) and studies with a low dropout/withdrawal rate (SS median SMD, -0.24, I^2=0%; MS median SMD, -0.29, I^2=0%) in particular for SS, suggesting no evidence of bias due to a different number of dropouts or withdrawal in the individual studies.

The individual studies featured very different sample sizes. Analysis by sample size (below/above the median sample size of 311 participants) showed a greater benefit of the treatment in small studies (SS median SMD, -0.32, I^2=0%; MS median SMD, -0.29, I^2=0%) than in bigger studies (SS median SMD, -0.21, I^2=72%; MS median SMD, -0.15, I^2=52%) (FIGURE 2a, 2b).

Analysis of efficacy according to the new proposed metric

It was possible to calculate the improvement, as change between baseline and during treatment period, for both groups only for the Cox study,[26] since this is the only one precisely reporting the symptom score at baseline (Retrospective Rhinoconjunctivitis Total Symptom Score, RRTSS, as defined in the Cox study). In any case, the other studies qualitatively reported the basal symptom score. In fact, patients included in the studies are affected by a moderate-severe or by moderate, or by mild-to-severe rhinoconjunctivitis, as shown in Table 1. Therefore, we repeated the analysis, although with a certain approximation (due to the assumption of a homogeneous proportion of subjects with moderate or severe disease). We assumed, for example, that patients with a moderate-severe rhinoconjunctivitis have a baseline SS of 12.5, that is the center of the moderate (6-12 SS points) and severe (13-18 SS points) classes. A specific baseline score was then assigned to each study depending on the severity of the disease at baseline (shown in Table 1). Only two studies[13,15] did not report the baseline severity of the disease. For these studies we assumed a moderate-severe disease, that should be the common inclusion criteria.

The percent of improvement from baseline to treatment period was calculated for each group and the comparison, expressed as percent difference, is reported in the forest plot (eFIGURE 3; MD, mean difference). No study reported a statistically significant difference using this metric (the CIs crosses the reference line), and no study reported a difference equal to or greater than 15%, as required by the FDA.