

Retrospective assessment of seasonal allergic symptoms: over-rating but useful

U. Bodtger, L. K. Poulsen and H.-J. Malling

Allergy Clinic, National University Hospital, Copenhagen, Denmark

Summary

Background The history of the severity of seasonal allergic symptoms is often obtained post-seasonally as a retrospective assessment. Correct rating is essential when determining the efficacy of pharmaceutical treatment, indications for allergen-specific immunotherapy (SIT), or inclusion into controlled clinical studies.

Objectives To investigate the agreement between in- and post-seasonal ratings of seasonal symptoms, and to investigate whether the effect of SIT could be detected retrospectively.

Material and methods Thirty-five birch pollen-allergic patients were allocated to SIT or placebo in a double-blind study. Assessment of severity of symptoms from the nose, eyes and lungs were performed daily during the season 2000, and post-seasonally 6 months after the season in 1999 and 2000. A four-point verbal descriptor scale (VDS-4) was used at all occasions. A mean in-seasonal symptom rating was calculated for four periods: the day, the week and the 2 weeks with the highest symptoms score, and the arithmetic season (the period covering the mid-90% of the accumulated pollen count). In- and post-seasonal ratings were compared with Cohen's weighted kappa (κ_w).

Results Agreement between in-seasonal and retrospective ratings was fair to moderate (κ_w : 0.30–0.60). Post-seasonal ratings were most related to symptoms experienced in the week with the highest symptom scores, and least related to the arithmetic season. The post-seasonal ratings were significantly skewed towards higher symptom scores than the mean of in-seasonal ratings in periods ≥ 2 weeks. Despite being comparable before intervention, only in the SIT-treated group was a significant decrease in post-season ratings of severity of rhinoconjunctivitis apparent ($P < 0.05$). Asthma scores were not reduced but fewer patients in the SIT group reported lung symptoms ($P < 0.001$).

Conclusion Post-seasonal assessment of seasonal allergic symptoms generally describes a shorter period than the arithmetic season. Post-season assessment tends to over-rate average symptom severity, but appears sufficiently sensitive to detect treatment efficacy.

Keywords asthma, conjunctivitis, immunotherapy, rating scales, recollection, retrospective, rhinitis, symptom score, VDS-4

Submitted 2 August 2002; revised 31 October 2002; accepted 20 November 2002

Introduction

A proper history on the nature and severity of seasonal allergic symptoms is a keystone in the diagnosis of hayfever, and a major outcome parameter in the assessment of treatment efficacy in both everyday clinical practice and in clinical trials [1]. The severity of hayfever symptoms assessed post-seasonally has been used as a randomization criterion in clinical trials [2, 3], and annual post-seasonal rating of in-seasonal symptoms to monitor efficacy of allergen-specific immunotherapy (SIT) is recommended by the EAACI guidelines [4]. However, at present it is not known if post-seasonal assessment agrees with the symptoms experienced during the season. The perception of allergic symptoms tends to decrease over the year, as fewer

subjects confirm suffering from birch pollen allergy when interviewed during the winter than during the summer or autumn [5]. Pain research has shown that retrospective ratings are skewed after a few days, mainly towards the recollection of higher pain levels than originally perceived [6–9].

Symptoms can be assessed on a variety of scales that predominantly are ordered categorical/ordinal (e.g. verbal descriptor scales or numerical rating scales), or continuous (e.g. visual analogue scale, VAS) [10]. A widely used scale that has a clear logical terminology for each category, is the four-point verbal descriptor scale (VDS-4) where symptoms can be rated as absent, mild, moderate, or severe [8, 11–13]. At present, it is not known if this scale is sufficiently sensitive to detect the impact of therapeutic intervention. Cohen's kappa (κ) is the recommended test to assess agreement between two observations on ordinal scales and provides a change-adjusted measure of agreement [11].

In the present study, we followed a cohort of birch pollen-allergic subjects for 1 year. We used the VDS-4 for both in- and

Correspondence: Uffe Bodtger MD, Allergy Clinic TA 7551, National University Hospital, Blegdamsvej 9, DK-2100 Copenhagen, Denmark.
E-mail: ub-all@rh.dk

post-seasonal assessment of severity of symptoms from the nose, eyes and lungs. The patients were treated with birch pollen SIT or placebo injections, and assessed symptom severity of the relevant seasons in 1999 and 2000.

Materials and methods

Patients

In a specialized allergy clinic, 35 patients (14 male) complaining of allergic rhinoconjunctivitis in April–May (corresponding to the local birch pollen season) were included in a double-blind, placebo-controlled clinical study on birch pollen SIT conducted from 1999 to 2000 [2]. Descriptions of demographics, randomization, treatment protocol and safety and efficacy estimates of SIT have previously been published [2]. Briefly, all patients had a median duration of birch pollinosis of 10 years (range 2–31). All had a positive skin prick test to birch (mean weal diameter ≥ 3 mm, Soluprick, ALK-Abelló, Hørsholm, Denmark), a CAP-class ≥ 3 to birch (CAP system, Pharmacia, Uppsala, Sweden) and a positive conjunctival challenge to birch (Aquagen birch, ALK-Abelló). Thirteen subjects reported seasonal asthmatic symptoms. Twenty patients reported being allergic to grass pollen. None suffered from perennial rhinitis, asthma, or other diseases requiring anti-inflammatory treatment. There were no demographic differences between the treatment groups. The study was approved by the local ethics committee of Copenhagen, Denmark, and written informed consent was obtained from all patients prior to admission to the study.

Birch pollen season and total observation period

Pollen grains were collected by a standard Burkard Volumetric 7-Days Spore Trap in Copenhagen, Denmark (55°43'N, 12°34'E) 15 meters above ground level. Pollen grains were counted daily (expressed as number/m³ air) by the Danish Meteorological Institute, Copenhagen [14].

The arithmetic birch pollen season was defined as the period including the mid-90% of the total birch pollen count, and started 18 April and ended 11 May [2]. Symptoms were ascribed exclusively to birch until 16 May where grass pollen counts exceeded 10 grains/m³, which is sufficient to elicit symptoms in grass pollen-allergic subjects [12].

In-season symptom registration

From 13 March to 25 May 2000, all patients completed daily diary cards at home. The severity of symptoms from each target organ (nose, eyes and lungs) were assessed using a the VDS-4: 0 = none, 1 = mild, 2 = moderate and 3 = severe symptoms [10, 11]. A mean of the area under the curve defined the individual symptom score for each target organ for four periods of different duration: the single day, the single week and the 2 weeks with the highest symptom score, and the arithmetic season. The mean value was adjusted to the nearest discrete value. Data on medication and efficacy of SIT are presented elsewhere, but briefly, the SIT-group used 55% less medication and experienced 36% fewer symptoms than in the placebo group [2].

Post-season symptom registration

At inclusion (Autumn 1999) and at follow-up (Autumn 2000), patients assessed the severity of the symptoms of each target

organ of the preceding season using the VDS-4. The patients were not informed of the retrospective rating of the season 2000 prior to the follow-up visit. The post-seasonal VDS-4 was completed in the presence of an investigator (UB), in absence of symptom diaries and before unblinding of the injection therapy.

Statistics

All data were analysed using statistical software (SPSS for Windows, version 10.0.7, SPSS, Chicago, IL, USA). Symptom score medians for groups were compared with Mann–Whitney *U*-test, and intra-rater agreement was investigated by Cohens's weighted kappa (κ_w) [11]. The calculation of weighted kappa is described by Altman [15]. The weights used are shown in Table 1a. Kappa equals 1.0 when agreement is perfect, and this is obtained along the diagonal of the matrix. The magnitude of disagreement increases with distance from the diagonal (an example of a matrix is shown in Table 1b). It is generally accepted to quantify the strength of agreement as follows: $\kappa_w < 0$ (poor agreement), 0–0.20 (slight agreement), 0.21–0.40 (fair agreement), 0.40–0.60 (moderate agreement), 0.61–0.80 (substantial agreement) and 0.81–1.00 (almost perfect agreement) [11]. Standard error of κ_w could be calculated since n was $\geq 2g^2$, where g equals the number of categories. The κ_w was tested against the null hypothesis, but differences in κ_w between groups were not tested as the criterion of $n \geq 3g^2$ for comparison was not fulfilled [11].

Results

Patients

One female patient in the placebo-treated group dropped out for unknown reasons, and one female patient in the SIT-treated group was excluded because of non-compliance in completion of diaries. The missing data in the diaries constituted less than 4%. Rhinitis and conjunctivitis were reported by virtually all patients (94–100%) in both treatment groups. SIT reduced the prevalence of out-seasonally reported asthma from 38 to 19%,

Table 1. Agreement matrix for weighted kappa calculations

In-season assessment	Post-season assessment			
	None	Mild	Moderate	Severe
(a)				
None	1	0.66	0.33	0
Mild	0.66	1	0.66	0.33
Moderate	0.33	0.66	1	0.66
Severe	0	0.33	0.66	1
(b)				
None	2	3	2	0
Mild	0	14	7	2
Moderate	0	0	1	2
Severe	0	0	0	0

Perfect agreement is seen along the diagonal. The farther from the diagonal, the greater the disagreement, and the lower the weights (Table 1a). An example is shown in Table 1b: the assessments of eye symptoms during the arithmetic season. Agreement is seen in 2 + 14 + 1 + 0 cases. On each side of the diagonal we find the retrospective over-rating (3 + 2 + 0 + 7 + 2 + 2 cases), and retrospective underrating (0 cases): weighted kappa = 0.25.

whereas asthma prevalence increased in the placebo group from 33 to 47%. In-seasonally, asthma was reported by 13% ($n = 2$) of SIT-treated and 59% ($n = 10$) of placebo-treated subjects.

Pollen count

Figure 1 depicts the pollen count during the seasons 1999 and 2000. The pollen counts were representative for a low and an average season, respectively [16].

Agreement between in- and post-season rating

The κ_w -values for post-seasonal assessment and different in-season measures are shown in Fig. 2. For the arithmetic season, all kappa-values were insignificantly above 0, thus the null hypothesis could not be excluded. The relatively high values for lung symptoms possibly reflect the low prevalence of seasonal asthma (< 35%). Table 2 shows the distribution of disagreeing ratings. Furthermore, it shows that the trend for increased retrospective over-rating with length of in-seasonal period was significant for all target organs, and that retrospective over-rating was significantly more frequent than under-rating for in-seasonal periods ≥ 2 weeks. Concerning gender, injection therapy and the presence of grass pollinosis, agreement differed less than one standard error (data not shown), but correct statistical testing of the difference between groups was not possible due to insufficient number of patients [11].

Ratings before and after immunotherapy

In Table 3, the median values of post-seasonal ratings before (season 1999) and after (season 2000) 9 months of immunotherapy are presented. A significant reduction in the post-seasonal assessment of nasal and eye symptoms was observed in the SIT-treated group, thus post-seasonal rating on a four-step verbal scale was sufficiently sensitive to detect the efficacy of treatment. Concerning lung symptoms, the reduction in post-seasonal score was insignificant, possibly due to the low number of patients experiencing lower airway symptoms in the SIT-group. Actually, patients receiving SIT were less likely to report lung symptoms during the season (two vs. 10; Fisher's exact test: $P < 0.001$). Figure 1 suggests that the decrease in symptom severity was not due to decreased pollen exposure.

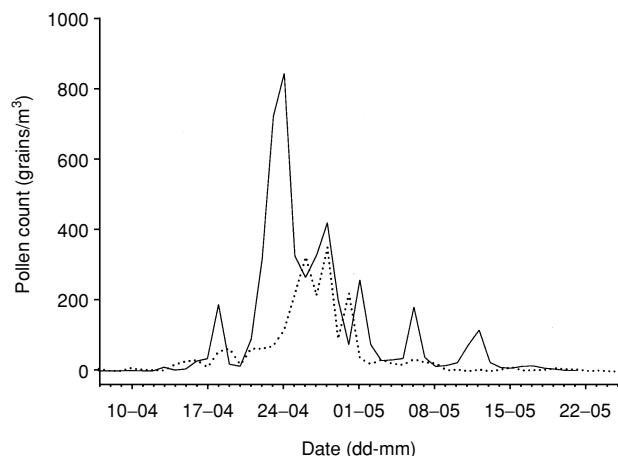


Fig. 1. Daily birch pollen counts in the seasons 1999 (dotted line), and 2000 (full line).

The medication usage was lower in the SIT than in the placebo group as described in our previous report [2].

Table 2. The distribution of disagreeing ratings and length of in-seasonal period (n retrospective over-rating/ n retrospective under-rating)*

	Worst day	Worst week	Worst 2 weeks	Season	<i>P</i> -value*
Nose	1/20	8/10	9/6	17/1	< 0.0001
Eyes	1/18	9/5	12/4	16/0	< 0.0001
Lungs	5/5	9/1	9/0	10/0	< 0.05
<i>P</i> -value†	0.08	> 0.5	< 0.05	< 0.05	

*Total number of patients = 33. Retrospective over-rating: post-seasonal VDS-4 score > in-seasonal VDS-4 score; retrospective under-rating: in-seasonal VDS-4 score > post-seasonal VDS-4 score. χ^2 for trend.

†Difference between n retrospective over-rating and n retrospective under-rating (Mann-Whitney *U*-test).

Table 3. Post-seasonal rating of symptom severity during the seasons 1999 and 2000 [median (interquartile range)]

Target organ	1999	2000	<i>P</i> -value*
Nose			
SIT	2 (1–3)	1 (1–2)	< 0.05
Placebo	2 (1–2)	2 (1–2)	> 0.4
<i>P</i> -value†	> 0.6	0.053	
Eyes			
SIT	2 (1–3)	1 (1–2)	< 0.005
Placebo	2 (1–2)	2 (1–2.5)	> 0.4
<i>P</i> -value†	> 0.15	0.063	
Lungs			
SIT	0 (0–1)	0 (0–0)	> 0.15
Placebo	0 (0–1)	0 (0–1)	> 0.15
<i>P</i> -value†	> 0.8	> 0.2	

SIT: $n = 16$; Placebo: $n = 17$. *Mann-Whitney test for paired comparison.

†Mann-Whitney *U*-test for comparison between groups.

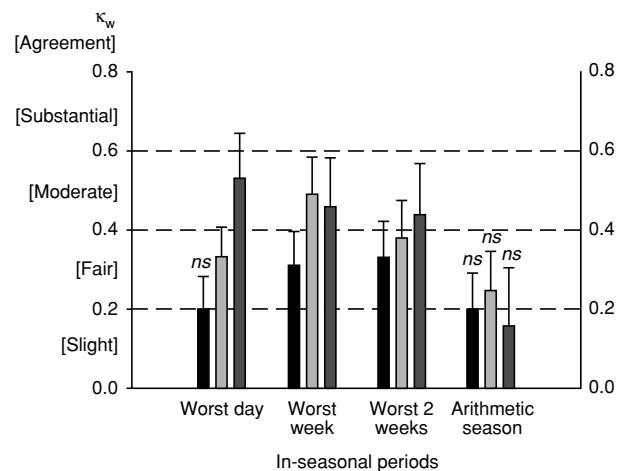


Fig. 2. Agreement between retrospective and in-seasonal assessment of severity of symptoms from the nose (black), eyes (light grey) and lungs (dark grey), and the impact of increasing length of the in-seasonal period (see text for definitions). κ_w = weighted kappa; error bars = standard errors; ns = insignificantly different from $\kappa_w = 0$.

Discussion

A proper history is essential for characterization and quantification of symptoms of seasonal rhinoconjunctivitis (hayfever) which is characterized by annual recurrence of symptoms separated by long symptom-free intervals [1, 17]. Data on severity of preceding seasons are important when considering indications for SIT but, rarely, has consecutive in-season registration been performed, thus ratings are conducted retrospectively. However, to our knowledge, no studies on the agreement between in- and post-seasonal ratings of allergic symptom have been published.

All symptoms, or subjective perceptions at large (i.e. pain, discomfort, functional impairment, well-being, feelings), can be rated on discrete or continuous scales [10]. In our study, we used a verbal descriptor scale consisting of four levels of increasing symptom severity: none, mild, moderate and severe symptoms (VDS-4) from nose, eyes and lungs. The VDS-4 is widely used in allergological research and each point has a clear verbal correlate as opposed to seven- or 10-point scales [10, 12]. To assess agreement between in- and post-seasonal ratings, we used Cohen's κ_w analysis that is a chance-corrected measure of agreement of categorical data [11]. Erroneous analyses of agreement are reviewed by Altman, and includes correlation tests and χ^2 (both are association tests), and comparisons of group means (a high P -value does not imply agreement) [15]. The κ_w has a few restrictions: the magnitude depends on the number of categories (the fewer the categories the higher the κ_w), and on the prevalence of subjects in each category (which influences the expected frequencies). These shortcomings are mainly of importance when comparing agreement between different studies [11, 15].

In our study, we found a fair to moderate agreement between the in- and post-seasonal assessment of rhinoconjunctivitis assessed 6 months apart. Agreement with the daily mean of the arithmetic season did not differ significantly from no agreement ($\kappa_w = 0$). For nose and eye symptoms, retrospective ratings tended to agree best with symptoms experienced during the worst week (Fig. 2), but the number of cases were too low to allow statistical testing of differences between the in-seasonal periods [11]. Retrospective over-rating of symptoms increased significantly with the length of the in-seasonal period investigated, and was significantly more common than under-rating for periods ≥ 2 weeks (Table 2). This suggests that patients do not recollect the symptom severity of the entire season but merely the period shorter than 14 days with the severest symptoms. An explanation might be acceptance of symptoms that do not infer with daily activities. Such symptoms are probably not recollected as days of disease and tend to 'dilute' the efficacy of intervention. Acceptance of a certain level of symptoms is apparent in intervention studies where the relative reduction in medication usage exceeds that of symptom severity [2, 3]. A consequence of accepted symptoms is an increased focus on peak-seasonal symptoms, and rescue medication-free days in future studies.

Despite the apparently disappointing findings, our results resemble findings from pain research, where ratings of pain performed before (pre-treatment) and after relief (post-treatment) have been compared. Post-treatment perception of symptoms was altered within days after pain relief, resulting in

poor to slight agreement with pre-treatment ratings ($\kappa < 0.2$) with a tendency of post-treatment over-rating [6–8]. When symptoms were still present, agreement was substantial ($\kappa > 0.6$) [18, 19]. These findings have been supported by studies of disordered mood [9, 20].

Theoretically, several factors may disturb the recollection of symptoms. The presence of few symptoms could be neglected because of an extremely high pollen count in the preceding season and *vice versa*. Figure 1 depicts that this was not true in our study. Allergy of other causes (e.g. grass pollinosis) might hamper the perception of severity of the birch season. In our study, agreement of the birch + grass pollen-allergic patients was within the standard error of the exclusively birch pollen-allergics. Likewise, no differences in agreement was observed concerning gender or treatment.

The items included in the rating scale are of importance. We included only ratings of symptoms from the target organs, and it has clearly been demonstrated by the Rhinitis Quality of Life Questionnaire (RQLQ) that hayfever includes impairment of several non-target organ symptoms as sleep, daily activities, social/emotional skills and overall quality of life [21]. We have previously shown that, retrospectively, perception of in-seasonal general well-being – and not the reduction of symptoms, or medicine intake – was significantly associated with relief of in-seasonal nose and eye symptoms [2]. This indicates that non-target organ symptoms are integrated in the retrospective perception of target organ symptoms, thereby over-rating these retrospectively. Retrospective use of the RQLQ needs validation but seems promising as a tool for post-seasonal assessment of hayfever [22].

Despite the only fair to moderate agreement with in-seasonal ratings, post-seasonal VDS-4 was sufficiently sensitive to detect the impact of SIT on rhinoconjunctivitis (Table 3). The VDS-4 could not detect that fewer patients in the SIT-group had lung symptoms, but this might be a type II error due to the low number of patients. Our findings suggest that the reproducibility of post-seasonal ratings using the VDS-4 is acceptable, and support that post-seasonal ratings can be used both as a measure of the efficacy of an intervention and as the basis of randomization to a clinical study on seasonal rhinoconjunctivitis [2, 3]. Thus, careful matching of patients based on post-seasonally obtained data might abrogate the need for run-in seasons in clinical studies.

In conclusion, retrospective assessment of symptom severity agrees moderately with in-seasonal assessment, and post-seasonal ratings describes a shorter period than the arithmetic season. In an intervention study, post-seasonal ratings could detect the impact of SIT on nasal and eye symptoms. Future studies are needed to optimize a rating scale for post-seasonal assessment of seasonal symptoms.

Acknowledgements

We would like to thank all patients for their exquisite patience, and the study nurse Anne-Sofie Lassen for her practical assistance. Supported by grants from Mogens Ambt Balslevs Mindefond, and the Danish Allergy Research Centre, Denmark.

References

- 1 Druce HM. Allergic and non-allergic rhinitis. In: Middleton EJR, Reed, CE, Ellis, EF, Adkinson, NF Jr, Yunginger, JW, Busse, WW, eds. *Allergy. Principles & Practice*, pp. 1005–16. St Louis: Mosby-Year Book 1998.
- 2 Bodtger U, Poulsen LK, Jacobi HH, Malling HJ. The safety and efficacy of subcutaneous birch pollen immunotherapy: a 1-year, randomised, double-blind, placebo-controlled study. *Allergy* 2002; 57:297–305.
- 3 Varney VA, Gaga M, Frew AJ, Aber VR, Kay AB, Durham SR. Usefulness of immunotherapy in patients with severe summer hay fever uncontrolled by antiallergic drugs. *Br Med J* 1991; 302:265–9.
- 4 Malling HJ, Weeke B. Immunotherapy. *Allergy* 1993; 48 (Suppl. 14):9–35.
- 5 Remes ST, Korppi M, Kajosaari M, Koivikko A, Soininen L, Pekkanen J. Prevalence of allergic rhinitis and atopic dermatitis among children in four regions of Finland. *Allergy* 1998; 53:682–9.
- 6 Macarthur C, Macarthur A, Weeks S. Accuracy of recall of back pain after delivery. *Br Med J* 1996; 313:467.
- 7 Feine JS, Lavigne GJ, Dao TT, Morin C, Lund JP. Memories of chronic pain and perceptions of relief. *Pain* 1998; 77:137–41.
- 8 Porzelius J. Memory for pain after nerve-block injections. *Clin J Pain* 1995; 11:112–20.
- 9 Beese A, Morley S. Memory for acute pain experience is specifically inaccurate but generally reliable. *Pain* 1993; 53:183–9.
- 10 Svensson E. Concordance between ratings using different scales for the same variable. *Stat Med* 2000; 19:3483–96.
- 11 Kramer MS, Feinstein AR. Clinical biostatistics. LIV. The biostatistics of concordance. *Clin Pharmacol Ther* 1981; 29:111–23.
- 12 Winther L, Malling HJ, Moseholm L, Mosbech H. Allergen-specific immunotherapy in birch- and grass pollen-allergic rhinitis. I. Efficacy estimated by a model reducing the bias of annual differences in pollen counts. *Allergy* 2000; 55:818–26.
- 13 Breivik EK, Bjornsson GA, Skovlund E. A comparison of pain rating scales by sampling from clinical trial data. *Clin J Pain* 2000; 16:22–8.
- 14 Hodal L, Rasmussen A. Measurements of Pollen and Spores in Denmark, Season 1999. Technical Report 00–12. Copenhagen: AAF and DMI 2000.
- 15 Altman DG. *Practical Statistics for Medical Research*. New York: Chapman & Hall 1992.
- 16 Taudorf E, Moseholm L. Pollen count, symptom and medicine score in birch pollinosis. A mathematical approach. *Int Arch Allergy Appl Immunol* 1988; 86:225–33.
- 17 Bousquet J, Lockey R, Malling HJ. Allergen immunotherapy: therapeutic vaccines for allergic diseases. A WHO position paper. *Allergy* 1998; 53 (Suppl. 44):1–42.
- 18 Staes F, Stappaerts K, Vertommen H, Nuyens G. Visual analogue scale for the perceived influence of exertion and movements/positions on low back problems in surveys of adolescents. *Acta Paediatr* 2000; 89:713–6.
- 19 Roach KE, Brown MD, Dunigan KM, Kusek CL, Walas M. Test–retest reliability of patient reports of low back pain. *J Orthop Sports Phys Ther* 1997; 26:253–9.
- 20 Hinton J. How reliable are relatives' retrospective reports of terminal illness? Patients and relatives' accounts compared. *Soc Sci Med* 1996; 43:1229–36.
- 21 Juniper EF, Guyatt GH. Development and testing of a new measure of health status for clinical trials in rhinoconjunctivitis. *Clin Exp Allergy* 1991; 21:77–83.
- 22 Wolf H, Schnitker J, Petermann F, Bergmann KC. The LQC study: quality of life and compliance with specific immunotherapy against grass and rye allergens after 3 years of treatment. *Allergy* 2002; 57 (Suppl. 73):77–8.