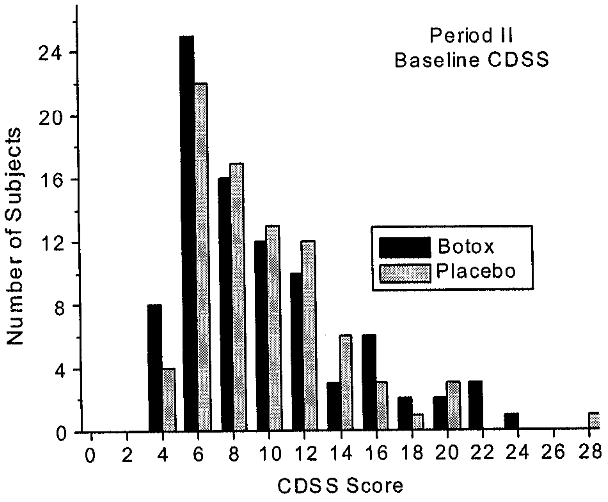
Table 5 : Demographics and Baseline Characteristics Study 140

Characteristic	Number or % of Patients		
	Period I	Period II	
		Placebo	BOTOX
	n=214	n=82	n=88
Age (yrs)	55.3	54.5	54.8
Race: White	211	82	87
Race: Other	3	0	1
Sex: Male	25%	20%	30%
Sex: Female	75%	80%	70%
Weight (kg)	72.2	69.5	74.2
Height (cm)	167	166	169
Duration of CD (mdn mo)	86	84	96
Baseline CDSS (mdn)	9	9	9
Baseline Pain Frequency (mdn)	2	2	2
Baseline Pain Intensity (mdn)	2	2	2

A histogram of the baseline severity of the two Period II groups indicates that they were generally similarly distributed. However, there is a suggestion of a small shift of the Botox group to lower baseline CDSS than the Placebo group. This is also suggested by the mean baseline CDSS scores, 9.3 in the Botox group, while 9.8 in the placebo group.

Figure 1



EFFICACY RESULTS: PRIMARY EFFICACY ENDPOINTS

CDSS Change from Baseline

One of the two primary endpoints was the change from baseline to Week 6 in CDSS during Period II. Allergan submitted the following results of their analyses:

Table 6: Period II CDSS Change from Baseline Results - Study 140

Week		Placebo	BOTOX	Tx Effect	
		n=82	n=88	pt est & CI	p-value
2	n	74	84	-1.0	0.035
	mean	-1.0	-2.0	(-2.0, -0.1)	
4	n	70	80	-1.6	0.003
	mean	-1.2	-2.7	(-2.6, -0.5)	
6	n	82	88	-1.2	0.046
	mean	-0.3	-1.5	(-2.4, 0)	
8	n	61	72	-2.1	0.002
	mean	0	-2.2	(-3.5, -0.8)	
10	n	58	76	-1.4	0.033
	mean	-0.3	-1.7	(-2.7, -0.1)	

These analyses employed ANCOVA methods, with treatment and investigator effects, and baseline CDSS as a covariate. Adjusted means are listed in the table. As per the analytic plan, only Week 6 is a true ITT analysis with value imputation for missing data. Per the analytic plan, Lack of Efficacy (LoE) missing were to be replaced with the worst observed score, other missing values were replaced by LOCF method.

Comment:

There were multiple analytic plan deviations in the analyses submitted by Allergan.

Contrary to the prospective analytic plan, interaction terms were not included in the model, and time of examination was not confirmed as within the prospectively stated permitted window.

An Analytic Plan exception occurred with Subject 653 for whom no Period II baseline score (Day 0- Visit 7) was recorded. For this subject, the baseline score of Period I (visit 0) of 10 points was assigned to Visit 7 . It is not documented if this subject was actually eligible for enrollment into period II. However, it is notable that while this was the Visit 1 CDSS score, the subject never had a score of greater than 7 any other time in the study, and had a CDSS of 4 at visit 6, approximately 4 weeks before Visit 7.

Allergan chose to use worst-value imputation by within-treatment group selection of worst change from baseline value, contrary to the formal written analytic plan. This resulted in attribution of worsening Change in CDSS by 13 points for placebo LoE subjects, but only by 8 points for Botox LoE subjects.

For patient 403 in the LoE category, Allergan chose to substitute a LOCF value rather than the worst observed change value. This subject had an unplanned visit after week 6 at which time the CDSS was the same as for Week 4. Allergan then used this value for Week 6, resulting in an improvement of 5 points, in spite of being in the LoE category.

When a properly imputed per analytic plan analysis is conducted, the results are less supportive of efficacy. The observed mean change in CDSS is -0.5 points in placebo, -1.5 points in Botox groups, and the p-value using a t-Test is 0.187 (see Table XX below).

Sensitivity Analyses of the Primary Endpoint of CDSS Change

Sensitivity analyses were performed by Allergan for the missing value imputation, including no value imputation, LOCF for all values, worst or best value imputation within treatment group for all missing, worst or best value imputation over all subjects (not within treatment group) for all missing values.

Comment:

Additional sensitivity analyses were performed by CBER medical review, including one of a Proper Per Analytic Plan analysis, consisting of LOCF for missing due to other than LoE, and imputation of the worst observed score within the study as a whole for all 7 of the Week 6 LoE subjects. Also calculated were results for analysis of change in CDSS as a percentage of the baseline CDSS. These analyses were conducted using both t-Test and ranksum methods to calculate a p-value, and are shown in Table XX.

Table 7 : Change in CDSS at Week 6 Results - Comparison of Analytic Methods

	Missing Value Method	N		Means		Tx Effect	Mean	Statistic Test Method		
							Cov.			
		Placebo	Botox	Placebo	Botox	Size	Adj?	Ancova	t-Test	Ranksum
Allergan	(Improper) Analytic plan	82	88	-0.3	-1.5	-1.2	Y	0.046		
	Ignore Missing	72	79	-0.7	-2.2	-1.5	Y	0.007		
Change	LOCF all missing	82	88	-0.7	-1.9	-1.2	Y	0.013		
In	Worst change within same Tx group	82	88	-0.4	-1.7	-1.3	Y	0.019		
CDSS	Best change within same Tx group	82	88	-1.7	-2.8	-1.1	Y	0.037		
	Worst change within entire study	82	88	-0.6	-1.7	-1.1	Y	0.038		
	(Improper) Analytic plan	82	88	-0.5	-1.8	-1.3	N		0.046	
CBER Medical Review										
Change	**Proper Analytic Plan**	**82**	**88**	**-0.5**	**-1.5**	**-1.0**	**N**	**0.23**	**0.19**	**0.052**
in CDSS	Ignore Missing	72	79	-1.0	-2.2	-1.3	N		0.046	0.021
	LOCF all missing	82	88	-0.9	-2.2	-1.3	N		0.024	0.012
Percentage	LOCF all missing	82	88	-3.0	-21.9	-18.9	N		0.008	0.005
Change	Proper analytic plan	82	88	0.8	-16.6	-17.5	N		0.033	0.022

Comment:

The Proper Per Analytic Plan method yields the smallest estimate of true treatment size, and a p-value by t-test that is substantially distant from statistical significance. Other analyses yield results that are supportive of the Allergan proposed analysis of the results. The Ranksum statistic yields p-values that are generally supportive of the Allergan finding.

Allergan's selection of a covariate adjusted analysis also has an impact upon this non-robust result. When an analysis using Allergan's faulty imputation method, but without use of the covariate is performed, the p-value of the comparison is 0.075, compared to the 0.046 obtained by Allergan with their method.

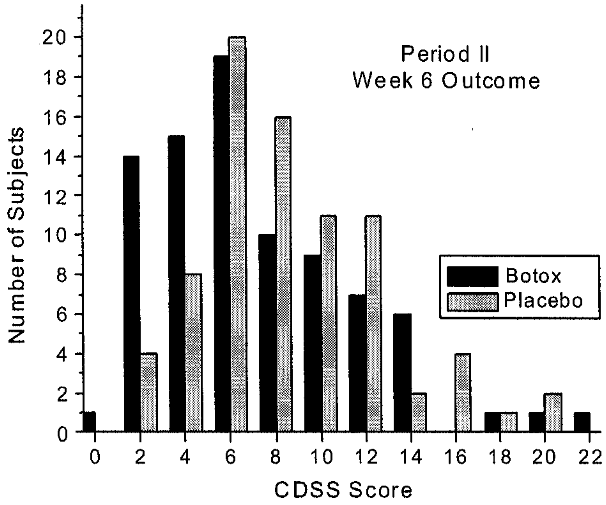
Although the analyses differ substantially in the p-value assessing the statistical significance, all analyses give estimates of treatment effect size that are similar. There is little real difference between a treatment effect of 1.0 points and 1.3 points on the CDSS. As 1.0 point implies a mean effect of at most 5 degrees of head deviation different, and this is 1 point out of a mean baseline severity of 9.3 points in the Botox group, 9.8 points in the placebo group. This modest degree of benefit is confirmed by the analysis of change as a percent of baseline, which suggests that an average of only approximately 18% of baseline deviation was alleviated by toxin treatment.

Comment:

The treatment effect was reasonably consistently present across centers. No single center was individually overly responsible for the study-wide observed treatment effect. Of 21 centers total, 16 had at least 2 subjects in both treatment groups. Of these 16, 10 centers showed a favorable treatment effect associated with Botox, while 6 showed unfavorable treatment associated effects. Of 7 centers with at least 5 subjects in each treatment group (and comprising 101 of the 170 total subjects), all 7 showed beneficial treatment associated effects.

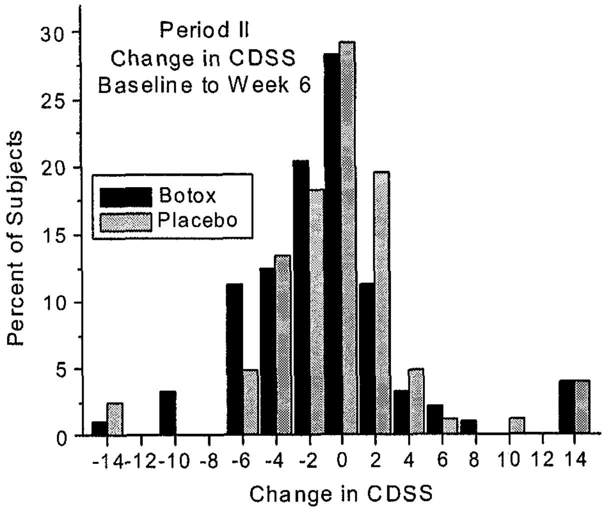
As shown in Figure XX the treatment effect appears as a general shift of the entire Botox treatment group in CDSS score. This figure, however, does not adjust for the previously seen mild difference in baseline CDSS.

Figure 2



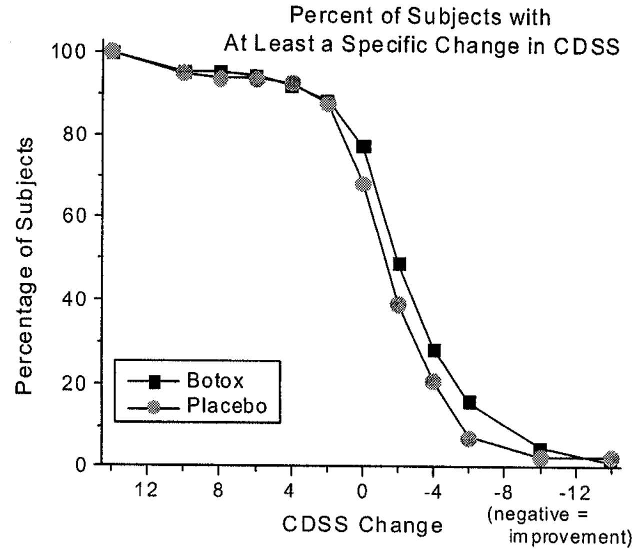
The histogram of change in CDSS at week 6 more clearly shows that the difference between treatments is modest, and involves a broad tendency of the group to shift to larger change in CDSS scores in the beneficial direction with toxin treatment. There is no indication of a bimodal distribution. The group of subjects clustered at a CDSS change of 13-14 is due to the Lack of Efficacy imputation for missing values.

Figure 3



The modest size of the shift seen in Figure 3 is clearly indicated in Figure 4, showing the percentage of subjects in each treatment group who achieve a certain amount of change in CDSS. These curves show a very modest degree of separation. For any specific amount of beneficial change, only a few percent more subjects achieve this score with Botox than with placebo.

Figure 4



Physician Global Assessment

The other co-primary endpoint was the physician global assessment at week 6, analyzed as percentage of subjects who show improvement of any amount.

Table 8: Period II Percent of Subjects with Improvement on Physician Global Assessment Results - Study 140

Week		Placebo n=82	BOTOX n=88	Tx Effect pt est & CI	 p-value
2	n	75	83	20.5	0.012
	%	33.2	53.7	(4.6, 36.4)	
4	n	71	79	23.6	0.004
	%	34.6	58.3	(7.9, 39.4)	
6	n	82	88	19.5	0.009
	%	31.1	50.6	(4.9, 34.1)	
8	n	61	72	17	0.05
	%	35.6	52.6	(-0.2, 34.2)	
10	n	58	76	6.4	0.44
	%	34.2	40.6	(-10.0, 22.7)	

These analyses were conducted by Allergan, employing the same ANCOVA method, and including baseline CDSS as a covariate. Adjusted values are shown.

Comment:

Again per the analytic plan, only Week 6 was to have missing data imputed. Allergan performed an analysis of this endpoint committing the same Analytic Plan deviations as was done with CDSS. LoE values were attributed differently for subjects in the two treatment groups. One LoE subject in the Botox group had LOCF for the missing value rather than worst value imputation.

Additionally, one subject in the missing for non-LoE category in the placebo group (#552) did not have any post baseline Global Assessment evaluations to use for LOCF. Allergan chose to impose the attribution of worst observed score in the treatment group for this subject.

These analytic plan deviations are less important for this endpoint's analysis, as the primary analysis of this endpoint for the primary EP relied on then dichotomizing subjects into improved (by any amount; i.e. a Global score of 1 or higher) vs those who did not (score 0 to –4). As the degree of worsening was not important in this analysis, attribution of either –4 or 0 would not be any different for purposes of this endpoint. However, the analyses of mean Global Assessment which were also performed by Allergan would be different.

Allergan conducted sensitivity analyses for this endpoint as well. For week 6 outcomes alternative methods of addressing missing values (ignoring; LOCF for all, replacement by non improved for all, replacement by improved for all) yielded analyses with point estimates of treatment effect of 19% to 23% of subjects with benefit, and p-values (ANCOVA) of 0.005 to 0.013. Thus, this endpoint was quite robust to alternative methods of missing value incorporation.

Exploratory analyses

Not stated as a primary endpoint, but supplied by Allergan was an analysis of CDSS that dichotomized subjects into success or failure based on the change in CDSS. For this analysis Allergan employed their prospectively stated, but unsubstantiated, claim that a 2 point change on CDSS was a meaningful change on a per-patient basis. Allergan's improper application of the analytic plan was again employed in the Week 6 analysis (ANCOVA analysis, adjusted means shown).

Table 9: Period II Percent of Subjects with 2 Point Improvement on CDSS - Study 140

Week		Placebo	BOTOX	Tx Effect	
		n=82	n=88	pt est & CI	p-value
2	n	74	84	10.9	0.11
	%	48.6	59.5	(-4.6, 26.4)	
4	n	70	80	22.3	0.005
	%	41.4	63.8	(6.7, 37.9)	
6	n	82	88	12.2	0.07
	%	37.8	50	(-2.6, 27.0)	
8	n	61	72	24.2	0.004
	%	32.8	56.9	(7.7, 40.6)	
10	n	58	76	17.6	0.03
	%	31	48.7	(1.3, 34.0)	

Comment:

The week 6 analysis is modestly supportive in this analysis. Estimated treatment effect size (12% of subjects) is less than in the Physician Global Assessment analysis, and the p-value is not statistically significant here. A proper application of the analytic plan would not substantially change the Week 6 results in this analysis, since a biased imputation of the degree of worsening does not influence this analysis.

Allergan also conducted an analysis of the Physician Global Assessment for the amount of change in the global score. This analysis was again based on ANCOVA, with missing value imputation only at week 6. Allergan's analysis again used an improper imputation method for some of the week 6 missing values, as described above for the primary analysis of this assessment.

Table 10: Period II Physician Global Assessment Mean Score Results Study 140

Week		Placebo	BOTOX	Tx Effect	
		n=82	n=88	pt est & CI	p-value
2	n	75	84	0.7	0.001
	mean	0.07	0.82	(0.32, 1.18)	
4	n	71	79	0.84	0.001
	mean	-0.07	0.78	(0.34, 1.35)	
6	n	82	88	0.68	0.014
	mean	-0.37	0.31	(0.14, 1.22)	
8	n	61	72	0.72	0.022
	mean	-0.33	0.39	(0.11, 1.34)	
10	n	58	76	0.49	0.077
	mean	-0.33	0.16	(-0.05, 1.05)	

Estimated mean treatment associated benefit was less than 1 point on the Global Assessment Scale at all timepoints of evaluation. A treatment effect size of less than the difference of 0 to 1 suggests that on average, the effect is less than even a mild benefit. While statistically significant, the effect size is quite modest.

Comment:

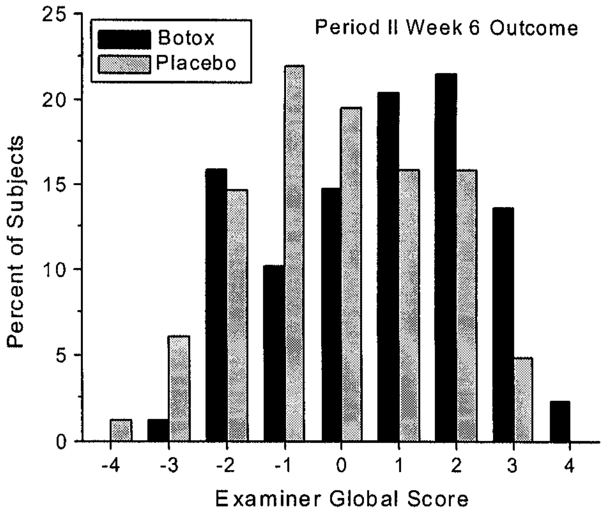
CBER analyses were also performed as sensitivity analyses. These are shown in Table XX below. Note that both Examiner (Physician) and Patient Global Assessment analyses are shown (see below for Allergan analyses of Patient Global Score). These analyses are largely consistent with that presented by Allergan. The analyses indicate statistically significant treatment effects, of a size similar to that suggested by Allergan. There are approximately 20% more subjects with successful treatment associated with Botox treatment than with placebo. This is true whether a liberal criterion of success is used (score > 0) or one that restricts success to somewhat more substantial amounts of improvement (score > 1). For the dichotomized success/failure analysis, Chi-Square and Fisher's Exact tests are more appropriate. For analysis of the full range of scores on the Global Assessments, p-values from t-test are shown to assist comparison with Allergan's ANCOVA analysis. However a ranksum analysis is more appropriate for this ordered category scale. The Ranksum analysis remains fully supportive of a beneficial treatment effect associated with Botox treatment. The actual amount of improvement suggested by this analysis remains modest.

Table 11: Sensitivity Analyses for Week 6 Global Assessment Scores

Type of Assessment	Missing Value Method	N		% Improvement or Mean			Statistical test	
		Placebo	Botox	Placebo	Botox	Obsv Tx Effect	p-value	
	Global score as % with success vs no improvement						Chi-Sq	Fisher Ex.
Examiner	LOCF all	82	88	37%	58%	21%	0.008	0.006
Patient	LOCF all	82	88	26%	55%	29%	0.0002	0.0002
	Global score as % with success of score of 2 or greater						Chi-Sq	Fisher Ex.
Examiner	LOCF all	82	88	21%	38%	17%	0.026	0.019
Patient	LOCF all	82	88	18%	38%	20%	0.009	0.006
	Global Assessment Analyzed as Mean Score						t-test	Ranksum
Examiner	Ignore	72	79	-0.07	0.76	0.83	0.004	
	LOCF all	82	88	-0.12	0.68	0.80	0.003	0.004
Patient	Ignore	71	79	-0.42	0.71	1.13	0.0006	
	LOCF all	82	88	-0.52	0.65	1.17	0.0001	0.0002

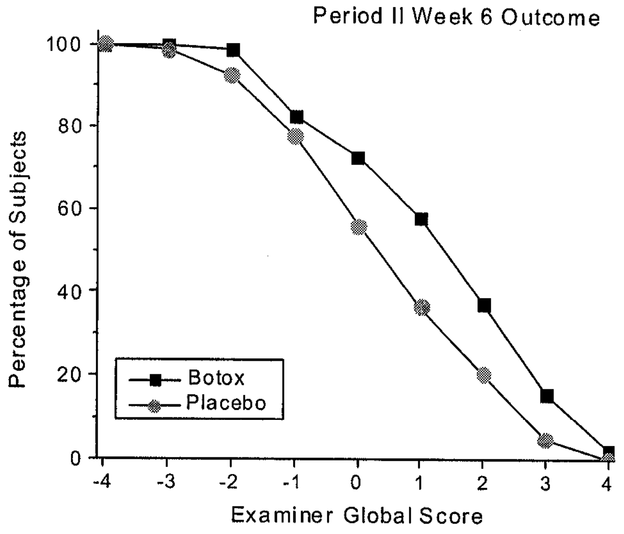
The histogram of distribution of Global Assessment scores at week 6 does not suggest any bimodal pattern to the outcome, just a modest difference in subjects achieving the more positive scores.

Figure 5



The curves of numbers of subjects who achieve at least a specific outcome on the Physician Global Assessment is as would be expected from this histogram. There is separation of the curves in a manner indicating that Botox treated subjects achieved higher global Assessment scores. While this figure indicates a greater degree of Botox efficacy in terms of separation of the curves, overall this is still showing a minority of subjects achieve identifiable benefit with Botox.

Figure 6



Comparison of Period II Outcomes with Observations During Period I

Allergan has also submitted summary analyses of the efficacy outcome measures as obtained during Period I. For purposes of comparison with the primary efficacy endpoint analyses, the responses observed in Period II subjects during their Period I run-in observations are shown in the following table. Also shown are the Period I results for these subjects for secondary endpoints of the pain assessments.

Table 12: Period I Week 6 Outcomes in Period II Subjects by Period II Treatment Group

Outcome	value descrip.	Placebo	BOTOX
	Period II n # of Period I evaluations	n=82 80 - 81	n=88 86
CDSS Change from baseline	mean	-4.4	-4.2
Physician Global Assessment	% with any imprv	100	100
CDSS Change from baseline	% with 2 points imprv	78	87
Physician Global Assessment	mean score	2.1	2.2
Pain Frequency change from baseline	mean	-0.65	-0.48
Pain Intensity change from baseline	mean	-0.57	-0.45

Comment:

The open label responses observed during Period I are considerably greater in size and extent than the responses observed during the double blind Period II. Open label responses to Botox were approximately three times as great in size on the CDSS change from baseline, and twice as widespread in extent on the Physician Global Assessment percentage with improvement. The size of the improvement on the mean of Physician Global Assessment score was more than four times as large in Period I as in Period II. In comparison with the results on the Pain assessments in Period II (see Secondary Endpoints discussion , below), the open label results were also somewhat greater than the blinded Period II effects were observed to be. These comparisons suggest that there is a considerable component of placebo effect within the apparent results obtained in usual medical practice.

EFFICACY RESULTS: SECONDARY ENDPOINTS

The formal stated secondary endpoints were the Range of Motion endpoints (3 assessments), the Patient Pain Frequency and Pain Intensity evaluations, and the Functional Disability Assessments (Physician and Patient). These 7 endpoints were not given an explicit order of importance. Note that since these were secondary endpoints, Allergan's analytic plan called for no data imputation for missing values at any time for these analyses. Pain Assessments employed p-values from exact Smirnov tests.

Pain Assessments

Pain has long been recognized as an important aspect of this disease. Allergan chose to assess this with two separate scales, assessing frequency of pain and intensity. There has not been any validation supplied to demonstrate that these two evaluations are truly applied in an independent manner. These scales were all analyzed without a true intent to treat analysis.

Table 13: Period II Patient Frequency of Pain Assessment Change from Baseline - Study 140

Week		Placebo n=82	BOTOX n=88	p-value
2	n	74	83	0.55
	mean	-0.2	-0.33	
4	n	71	79	0.51
	mean	-0.18	-0.37	
6	n	72	78	0.018
	mean	-0.01	-0.31	
8	n	61	71	0.929
	mean	-0.2	-0.27	
10	n	58	75	0.488
	mean	-0.15	-0.19	

Table 14: Period II Patient Intensity of Pain Assessment Change from Baseline - Study 140

Week		Placebo n=82	BOTOX n=88	p-value
2	n	74	83	0.026
	mean	-0.07	-0.39	
4	n	71	79	0.107
	mean	-0.18	-0.47	
6	n	72	78	<0.001
	mean	0.06	-0.36	
8	n	61	71	0.178
	mean	-0.06	-0.34	
10	n	58	75	0.334
	mean	0	-0.2	

Comment:

For both the Pain Frequency and Intensity evaluations, there were not statistically significant effects in Allergan's analyses except for the Week 6 evaluations and week 2 for intensity. This is different than the CDSS and Global Assessment scales, where the assessment had good consistency between one evaluation timepoint and the succeeding timepoint. These stand out as being anomalous from the other analyses, and warrant further exploration. The estimated treatment effect is quite modest at all timepoints, but is approximately 25% larger at week 6 than at other timepoints. The cause of this inconsistency is unidentified. A true change in the efficacy of Botox this large at week 4 vs 6 or 6 vs 8 seems unlikely. Sensitivity analyses performed by CBER, shown following in Table XX include a proper ITT analysis. For these ordered category scale endpoints, a ranksum method of analysis appears more appropriate. These sensitivity analyses suggest a consistency of the treatment effect size irrespective of the analytic method of addressing missing values. For both endpoints, the week 6 (and likely all other weeks) outcome is statistically significant, but quite small in actual magnitude.

Table 15 : Sensitivity analyses on Week 6 Pain Assessments

		N		Means		Tx Effect	t-test	Ranksum
		Placebo	Botox	Placebo	Botox	Estim.	p-value	
Frequency	Ignore missing	72	78	-0.014	-0.314	-0.3	0.032	0.026
	LOCF all	82	88	-0.012	-0.282	-0.27	0.04	0.027
	Proper Analytic Plan	82	88	0.122	-0.109	-0.231	0.18	0.043
Intensity	Ignore missing	72	78	0.062	-0.359	-0.421	0.0006	0.001
	LOCF all	82	88	0.079	-0.362	-0.441	0.0001	0.0002
	Proper Analytic Plan	82	88	0.116	-0.264	-0.38	0.002	0.002

These analyses do not explain why the Week 6 outcomes were substantially better than the outcomes only 2 weeks prior or later. Particularly for the Pain Frequency outcome, the difference between week 6 and adjacent evaluations is a change in the placebo group scores. Exploratory analyses of the dataset show that there were a substantial (11) number of subjects in the placebo group who changed from scores of –1 to –1.5 at week 4 to scores of –0.5 or 0 at week 6. These changes within the placebo group are the cause of the change in apparent significance of the treatment effect. The cause of such changes in scores are not readily apparent.

Functional Disability and Range of Motion

The other Secondary Endpoints selected by Allergan were the Functional Disability Assessment and the head Range of Motion assessments.

Comment:

While the Functional Disability assessments appeared to provide statistically significant differences, it is most unclear what the difference between these evaluations and the Global evaluations are. There was too little direction and explanation provided to the physician or patient to be able to distinguish between the two evaluations in meaning. The ROM evaluations were just the reverse. While they may be readily understood as individual scales, there were no significant treatment effects observed with these evaluations.

Table 16: Period II Other Secondary Endpoint Outcomes at Week 6 - Study 140

Evaluation		Placebo n=82	BOTOX n=88	p-value
Physician Assm. of	n	69	77	
Functional Disability	mean	-0.01	-0.38	0.008
Patient Assm of	n	72	78	
Functional Disability	mean	0.11	-0.31	0.005
ROM: Lateral	n	72	79	
	median	7	6	0.55
ROM: Rotational	n	72	79	
	mdn	9	9	0.65
ROM: Ant/Posterior	n	72	79	
	mdn	8	8	0.74

Comment:

> These 5 additional analyses provide little support for the utility of Botox treatments. While the "functional disability" assessments produce statistically significant results, the assessment tool is not interpretable. The physician assessment was most likely highly dependent upon the subject's report to the physician of their impression, thus duplicative of the patient evaluation. The meaning of this evaluation is unclear, since there was apparently no guidance of what the term "functional disability" was to include.

The ROM assessments did not provide any evidence of benefit. However, these were coarse assessments of range of motion, and unlikely to have been sensitive to small effects.

EFFICACY RESULTS: TERTIARY ENDPOINTS

The numerous tertiary endpoints included the Patient Global Assessment. Also included is an Activities of Daily Living evaluation (analyzed as multiple separate activities) as well as several other evaluations even less validated or interpretable. These were not analyzed in an ITT manner, and provide little additional information. They were not considered further.

Patient Global Assessment

The Patient Global Assessment was designated as a tertiary endpoint. However, this is likely to be an informative endpoint. The co-primary endpoint of the Physician Global Assessment can be expected to actually be dependent upon the patient's assessment, as physician discussion on the amount of benefit and satisfaction with the amount of benefit are likely to occur in order to enable the physician to make their global assessment. Allergan reports the results as shown in Table XX below.

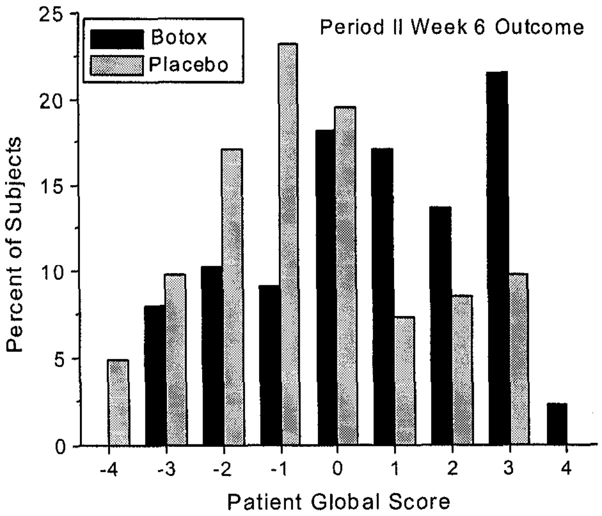
Table 17: Period II Patient Global Assessment Results - Study 140

Week		Placebo	BOTOX	
		n=82	n=88	p-value
2	n	75	84	
	mean	-0.19	0.74	0.003
4	n	71	80	
	mean	-0.24	0.83	0.001
6	n	71	79	
	mean	-0.42	0.71	0.001
8	n	61	72	
	mean	-0.52	0.43	0.006
10	n	58	76	
	mean	-0.6	0.05	0.014

Comment:

The Patient Global Assessment provides an impression largely similar to that derived from the Physician Global Assessment. Although Allergan provided no ITT analysis at any timepoint, the magnitude of benefit is modest at all times, but there would appear to be statistical significance to the effect. Neither of these conclusions is likely to be altered were a true ITT analysis to be performed.

Figure 7


■ Botox
▨ Placebo
Period II Week 6 Outcome
Percent of Subjects
Patient Global Score