

February 5, 2003

Final Statistical Review for PMA P020012, Artecoll PMMA/Collagen Implant for Correction of Contour Deficiencies of Soft Tissue, Artes Medical USA, Inc.

Background

Artecoll is a suspension of polymethyl Methacrylate (PMMA) microspheres in a 3.5% collagen solution. It is intended as an injectable filler for the long lasting correction of dermal defects. The original IDE application (G970026) was submitted by Rofil Medical USA. One hundred fifty-seven subjects were treated with Artecoll. There was no control group. In 1999, the responsibility of the study was transferred to Artes Medical, and a new clinical study was initiated. This was a prospective, multi-center, controlled, randomized, double masked trial. Two hundred fifty-one subjects were randomized (1:1) to either Artecoll or a commercially available collagen implant (Zyderm2 for glabellar folds, Zyplast for the other 3 areas (nasolabial folds, upper lip lines, mouth corners). Follow-ups for safety and efficacy were at 1,3, and 6 months, with a final safety evaluation at 12 months. The primary objectives were to determine if the cosmetic correction provided by Artecoll was superior to the control at the end of 6 months, and to determine the safety of Artecoll. The Facial Fold Assessment (FFA) Scale was created and validated for this study. Secondary objectives were to compare the initial quality of the correction between the two treatment groups and to compare (unmasked) investigator assessment and patient satisfaction.

Reviewer's Comments

Because there were large differences in favor of Artecoll between the masked observer and unmasked investigator/patient satisfaction, I find that the unmasked analyses could be biased, and my comments will focus on the masked FFA assessments unless otherwise noted. The sponsor's argument that the differences between results in the masked and unmasked analyses could be due to looking at a 2-dimensional photo verses a 3-dimensional live person should be evaluated from a clinical perspective. The data from the early Rofil study, may be used for safety, but is of little use for efficacy since there were different outcome measures and no control. The Rofil study will not be directly addressed in this review. All references to tables and figures will apply either to Volume 12 of the original PMA, Amendment 9 (October 16, 2002), or Amendment 13 (January 23, 2003). Because all treatment areas were bilateral, scores were averaged across right and left sides of the face and also across multiple masked observers.

Changes from Baseline

Although not specifically a study endpoint, one must look first at the individual efficacy of each treatment. The sponsor has shown a statistically significant

improvement from baseline to 6 months for 3 out of the 4 Artecoll areas and 2 out of 4 control areas. It should be noted that nasolabial folds is not one of the control areas that remained improved at 6 months, which brings up the possibility of an inactive control for this area. This is particularly important should the sponsor ever decide to switch to a non-inferiority claim in the future. The issue here, however, is whether Artecoll is more effective than the control, and that is what the sponsor's analyses focused on.

Study Results

The sponsor's claim was that the magnitude of the improvement with Artecoll would be statistically superior to the magnitude of the improvement with the control. For this primary endpoint, the data did not support the claim for any treatment area except nasolabial folds. Early on, at months 1 and 3, the control was actually numerically superior to Artecoll for efficacy of glabellar folds and upper lip lines. By 6 months, Artecoll scores were statistically significantly superior for nasolabial folds ($p < 0.001$), although the difference in the mean improvements between the two treatment groups was 0.77 points on the FFA scale, which was less than the 1.0 point identified as clinically significant at the study design phase. To control type 1 error from the fact that four areas of the faces were tested on the same individual (multiplicity), the sponsor divided alpha by 4 (i.e., $0.05/4 = 0.0125$), which is akin to doing a "Bonferonni Adjustment", and their $P < 0.001$ was still statistically significant at the stricter level. The sponsor used the Mann-Whitney U Test. This is the nonparametric equivalent of the two sample t-test, and its use is appropriate for non-normally distributed data. However, because the Artecoll group had a statistically worse pretreatment wrinkle severity rating for nasolabial folds, it can be argued that the Artecoll group had more room for improvement. Thus, further analyses were necessary to adjust for these pretreatment differences.

Pretreatment Differences

The sponsor performed several supplemental analyses to adjust for these pretreatment wrinkle severity differences. The first one, a covariate analysis (ANCOVA), showed that treatment effect remained statistically significant after adjusting for differences in pretreatment wrinkle severity (Table 43, Vol. 12). The actual adjusted means are shown in Table 44, and the adjusted treatment effects for Artecoll dropped slightly from 0.77 to 0.71 points on the FFA scale. The difference between Artecoll and control also dropped from 0.77 to about 0.65. Although a covariate analysis is generally an appropriate way to adjust for baseline differences, it is a parametric procedure that is based on the assumption of normally distributed data. Since these data were not normally distributed, the validity of this adjustment remains in question. The following table shows the unadjusted and adjusted means for the masked assessment of nasolabial folds.

Mean Nasolabial Fold Masked FFA Scale Improvement at Six Months

Treatment	N	Unadj. Means	Tx Diff (A-C)	Adj. Means	Adj.Tx Diff (A-C)
Artecoll	92	0.77	0.77 (p<0.001)	0.712	0.654 (p<0.001)
Control	91	0.00		0.058	

The sponsor performed 3 other analyses to adjust for pretreatment severity, 2 of which were non-parametric and not subject to assumptions of normality (an analysis restricting the cases to nasolabial folds with pretreatment severity ratings of at least 1, and an analysis of 55 pairs randomly matched on pretreatment severity). The statistical superiority of Artecoll was maintained for nasolabial folds for all adjustment techniques. As shown in the table below, when the analysis was restricted to only those cases who had the ability to improve 1 point or more, the mean improvement was 0.99 points for Artecoll and 0.28 points for the control, the difference being 0.71, virtually identical to the adjustment using the covariate analysis above.

Mean Nasolabial Fold Masked FFA Scale Improvement at Six Months For Subjects with Initial Pretreatment Severity of least 1.0

Treatment	N	Mean	Tx Diff (A-C)	Mann-Whitney Test	
Artecoll	67	0.99	0.71	U	p-value
Control	53	0.28		935.5	<0.001

Follow-up and Attrition

Six-month follow-up was available on 229 (91.2%) of the subjects. The sponsor performed an ANOVA showing subjects without 6-month follow-up had similar results at 1 and 3 months to patients having 6-month follow-up (Vol.12, Table 39). There were no significant differences between those with 6-months data and those without, and no "treatment by attrition interaction". This means that, for a given area of the face, you didn't have one device working better for those with 6-month follow-up and the other working better for those without. There was a consistent pattern. Although the same problem of using parametric statistics exists as with the discussion above, I don't think attrition is an issue here given that it was generally less than 9% of the study population.

Timing of Follow-up Visits

Another concern is all the protocol deviations in the timing of follow-up visits. One-third to one-half of the subjects, depending on treatment area, were lost when the analysis is limited to only cases meeting protocol timing restrictions for every follow-up visit (compare Vol. 12, Tables 55 and 26). However, the results for this subgroup analysis are still statistically significant for nasolabial folds. Further, the primary efficacy endpoint was an analysis of change from baseline to

6 months, so deviations in timing of earlier follow-up visits are of little consequence.

Poolability across Centers

There were some significant differences in masked pretreatment wrinkle severity scores across centers, but these occurred for the treatment areas of upper lip lines and mouth corners. Since these areas did not meet the endpoint for the masked analysis, it is really a moot point. There were no significant study center differences in outcome. Therefore, I consider the data to be poolable across centers.

Averaging across Masked Observers

As mentioned earlier, scores from masked assessments for a given treatment area on a given face were averaged across masked observers and between sides of the face. Because all this averaging could result in a loss of information, I requested an analysis stratified by masked observer for 8 of the tables I felt to be particularly pertinent (Vol. 12, Tables 26, 28, 34, 36, 46, 48, 55, 63). The results were presented in Amendment 13. The statistical superiority of Artecoll was maintained separately for each observer, in every place where it occurred when they were pooled. Thus, the sponsor's conclusions are the same as would have been inferred by use of any single masked observer.

Relationships among Efficacy Measures

The sponsor also performed a correlation analysis between the primary efficacy endpoint (masked observer rating) and the two secondary study objectives (investigator ratings of success (unmasked) and patient ratings of satisfaction (masked)). The Spearman rank-order correlations (Vol. 12, Table 59) are not impressive. Table 60 shows that there were statistically significant differences between masked and unmasked pretreatment FFA scores for all 4 treatment areas, with unmasked ratings being about 1 point higher. This is more reason to consider only the masked analysis, because the unmasked analysis could be biased.

A more relevant measure would be the agreement, or consistency among the three masked observers. The sponsor's use of the intraclass correlation coefficient as a measure of inter-rater reliability is appropriate given that these are interval data. However, the coefficient actually measures inter-rater "consistency", and not agreement, per se. If one rater rates consistently higher or lower than another, and this pattern is maintained across treatment groups, the intraclass correlation will be high, even if the scores for a given wrinkle do not seem to agree. Since interpretation of the FFA scale has an element of subjectivity in it, consistency is sufficient for evaluating inter-rater reliability. The intraclass correlation was about 90% for most treatment areas, which is good.

Bias from Smoking and Sun Exposure

The question arose as to whether there could be biases in patient enrollment from baseline differences in smoking and sun exposure. Looking at the distribution of smoking in the table on page 1-069 of Amendment 9, and the correlation coefficients given on page 1-070, I am satisfied that smoking was not a source of potential bias. It appears the sponsor used a 3 or 4 point ordinal scale for smoking and sun exposure (e.g., 0-2, or 0-3) and then used the Mann-Whitney U-test. This is the counterpart of the t-test for independent samples, and its use here is acceptable. There was less sun exposure among the controls, and the negative Spearman correlation with treatment outcome shows that lower sun exposure is correlated with greater improvement from treatment. Therefore, I would have to agree that, if anything, the bias was in favor of the control group. To help control for differences in sun exposure, I requested a subgroup analysis of just the low-exposure group (49 Artecoll, 58 controls). The statistical superiority of Artecoll remained, although the actual treatment difference was only about ½ point on the FFA scale.

Mean Nasolabial Fold Masked FFA Scale Improvement at Six Months - Low Sun Exposure Group

Treatment	N	Mean	Mann-Whitney Test	
Artecoll	49	0.64	U	p
Control	58	0.15	913.0	0.001

Categorical Analyses

Because there were problems with pretreatment differences in wrinkle severity, non-normal data and achieving a prespecified value for clinical improvement, the sponsor also performed some categorical analyses not subject to these conditions. The sponsor compared the proportions of patients who had a pretreatment wrinkle severity rating of at least “1”, and improved at least 1 point on the masked observer FFA scale at 6 months. As shown in the table below, a statistically significantly higher percentage of Artecoll patients improved 1 point or more in nasolabial fold severity as compared to the control (47.8% (32/67) vs 13.2% (7/53)). The other treatment areas were not statistically significant.

Percentage of subjects with pretreatment Masked Observer FFA Scale ratings of at least 1.0 showing improvement of at least 1.0 points

Treatment Area	Artecoll		Control		Chi-Square	p-value
	N	% Significantly Improved	N	% Significantly Improved		
Glabellar Folds	46	28.3	48	33.3	0.095	.757
Nasolabial Folds	67	47.8	53	13.2	14.569	<.001
Upper Lip Lines	31	12.9	27	7.4	0.064	.800
Mouth Corners	55	18.2	58	13.8	0.144	.704

Another subgroup analysis focused on those subjects who had a pretreatment severity of at least “2” and improved at least 1 point (Vol. 12, Table 66). For nasolabial folds, 71% (22/31) of the Artecoll as opposed to 24% (6/25) of the controls met this criteria (p=0.001). Thus, Artecoll appeared to work better than the control in a reasonably high percentage of patients with moderate initial severity, although these results are based on rather small numbers. This study population was such that over 70% had pretreatment wrinkle severity scores less than 2.0 for all treatment areas. Because the primary endpoint spelled out in the protocol was statistically significant for nasolabial folds, these additional subgroup analyses can be performed without being considered exploratory data analysis.

Safety

As far as safety is concerned, I don't see a statistical difference between the 2 treatment groups. There were more subjects who had adverse effects with Artecoll (21 vs 16), but the controls tended to have multiple AE's per person (36 AE's for 16 controls versus 26 AE's for 21 Artecoll). The safety profile has been extensively discussed in the clinical review.

Summary

In summary, the main weaknesses of this study, along with a brief description of how each was addressed are summarized below.

- ?? Large differences between masked and unmasked FFA Scale ratings that suggest potential for bias in wrinkle assessment
 - o FDA statistical review focused on masked analyses only
- ?? For each wrinkle, scores averaged across multiple observers and two sides of the face
 - o Perform analyses stratified by masked observer

- ?? Use of parametric statistics when assumptions of normality are not met
 - Perform additional nonparametric analyses

- ?? Enrolling patients that had, for the most part, mild defects that afforded little room for improvement
 - Subgroup analysis of pretreatment wrinkle severity =1
 - Small subgroup analysis of pretreatment wrinkle severity =2

- ?? Significant difference between Artecoll and control in masked pretreatment severity for best performing area (nasolabial folds)
 - Adjust for baseline differences in several ways

- ?? Abundance of timing violations for follow-up visits (? to ½ of patients)
 - Subgroup analysis of those meeting timing restrictions

- ?? A mean improvement (masked FFA Scale) for the best performing area (nasolabial folds) of 0.77, when the clinically meaningful improvement was predetermined to be 1.0
 - Must be clinically assessed

Conclusion

In conclusion, I feel that, although the study has several weaknesses, the highly significant result at 6 months for nasolabial folds holds up even after adjustment of the significance level for multiple treatment areas, and the effect was corroborated by additional categorical analyses. Baseline differences in pretreatment severity, and other potential sources of bias (e.g., attrition, follow-up timing, center differences, smoking, sun exposure) have all been addressed by statistical analyses and found not to have a significant impact. Therefore, I find that the data to support a claim of statistical superiority for nasolabial folds at 6-months post-treatment, particularly for defects of moderate pretreatment severity. Whether this difference is clinically significant will have to be addressed by subject matter experts.