July 9, 2001

Mathematical Statistician (P. Silverman) HFZ-542
Division of Biostatistics, OSB

Statistical Summary for PMA P000057, Ascension MCP Finger Joint Prosthesis, Ascension, Inc.

John Goode HFZ-410
Division of General, Restorative, and Neurological Devices, ODE

Background

The study was retrospectively constructed from data from 53 patients, implanted with 147 joints, at the Mayo Clinic between 1979 and 1987. There was no pre-set follow-up schedule, although follow-up is as long as ten years for some patients. Thirty-one literature references spanning 1975-1999 were selected as historical controls. Twenty-two of these articles used the Swanson Silastic spacer, and this was selected as the main control. The primary endpoint was implant fracture or implant removal. Subjective endpoints, such as cosmesis, activity level, and patient satisfaction, were constructed from physician notes and patient comments in the clinical records. Sample size justification was based on arbitrarily set precision levels for adverse event reporting and limits on mean changes from baseline for measurable events. The sponsor initially claimed non-inferiority to the historical control.

A Statistical Checklist Review performed by myself (February 6, 2001) identified three major deficiencies: (1) concern over the appropriateness of the literature controls, (2) failure to define the window of non-inferiority (i.e., delta), and (3) lack of a statistical comparison to the control to support the non-inferiority claim. The sponsor was asked to address those three issues, and thus submitted the March 15, 2001 Amendment.

Reviewer's Comments

Primary Endpoint

The sponsor's primary endpoint is implant survival (as measured by incidence of fracture or revision). An offset value, or delta, of 10% was chosen. This means that the 10-yr survival for the Ascension group could be up to 10 percentage points less that the control before it would be considered statistically inferior. This comparison is based, however, not on the observed rates, but the lower limit

of the 95% confidence interval. Unfortunately, only one of the 22 literature studies using the silastic spacer had a survival curve (Hansraj, 1997). This was used as the sole control for the primary endpoint. The 10-yr survival was 84.3% for the Ascension and 90.3% for the control. Problems with this control are discussed below.

Follow-up for the Hansraj study started at 2 yrs post-op, and 51% of the data was lost either by patient death or dropout in the first two years. Thus, data on only 170/348 implants (49%) were available for analysis. One of the assumptions for comparing two Kaplan-Meier survival curves is that the censoring is comparable in each group i.e., the rate at which subjects (or in this case, implants) exit the lifetable for reasons other than the event in question (implant failure) is similar between the treatment and control. The dropout rate for the two groups is shown in the table below. These data are extracted by from pages 849 and 3438 of the PMA.

Implants At Risk (%)

Years	Silastic Spacer	Ascension
0	348 (100%)	147 (100%)
2	170 (49%)	109 (74%)
5	103 (29%)	98 (66%)
8	28 (8%)	86 (58%)
10	6 (1.7%)	78 (53%)

As you can see, the censoring distributions are not at all similar, with the control dropping off much more quickly. Therefore, the sponsor could not perform a log-rank analysis of the survival curves, but instead performed a statistical comparison of the 10-year survival estimates. As I mentioned above, the 10-yr survival was 84.3% for Ascension and 90.3% for the control. The sponsor performed the statistical comparison in two ways (3/15 amendment), treating the Hansraj data as a historical control with and without variance. Without variance (Table 1.1), the p-value was 0.12, and with variance (Table 1.2) it was 0.20. Since there were only 170 implants in the Hansraj analysis, I would hardly consider the success rate estimate to be without variance. However, either way, the sponsor has not met their primary endpoint as both p-values are quite higher than the 0.05 convention for statistical significance. (Please note that in equivalence testing, the hypothesis are set up so that the null hypothesis states things are different or inferior, and the alternative hypothesis is the hypothesis of equivalence or non-inferiority. The goal is still to reject the null. The sponsor states (p.000014 of the 3/15 amendment) that a "low probability suggests non-

inferiority of the MCP implant". Traditional statistics requires it to be at least as low as 0.05. The "with variance" result of p=0.20 implies that 20% of the time we would get these results by chance alone. Therefore, the sponsor has not statistically met their primary endpoint.

Secondary Endpoints

For the secondary endpoints, the sponsor also did a "with and without variance " analysis. Again, I don't think one can assume a "no variance" situation because the sample sizes were not that large. For variables that could be pooled across control studies, the sponsor ran the comparisons to the Ascension two ways, assuming homogenous and heterogeneous variances. Separate tables were presented for the binary measures (subjective, delta=25%, and continuous) measures (objective, delta=10°), and the results are broken down by 5-6 time intervals. The p-values did not differ much whether one assumed homogeneous or heterogeneous controls. There were many p-values that were less than 0.05, and there were many that were greater, with no consistent pattern across time intervals. What really stood out was that the data were very sparse for these secondary endpoints for the Ascension.

Overall Assessment of Statistical Analysis

My overall assessment was that I could not give this device my statistical blessing. The sponsor had not statistically substantiated their claim of non-inferiority for the primary endpoint, or any of their secondary endpoints except possibly active flexion. Lack of statistical power is likely the reason, since sample sizes were small. A retrospective power analysis showed that the sponsor would have needed 122 observations per treatment group in any analysis in order to have 80% power to reject the null hypothesis, using a 10% delta. Safety was addressed by descriptive statistics (i.e., reporting proportions), without any statistical analysis. Small sample sizes would make a statistical demonstration of safety nearly impossible. In light of all this, I had to conclude that the sponsor had not statistically supported their claims for this device.

FDA then suggested a different approach, based on case studies. The sponsor has stratified the population based on two baseline medical conditions: (1) osteoarthritis/post traumatic (OA/PT) and (2) rheumatoid arthritis/ systemic lupus erythematosus (RA/SLE). They have defined success criteria for each subgroup, and have presented specific safety and effectiveness data on both a per patient and per implant basis. They discussed at length the risks associated with the control device. Limited data on implant fracture, pain, and the incidence of reactive synovitis is given for the control device.

Assessment of Data Base as Case Series

I would consider the 53 patients and 147 implants to represent a *case series*. This is a particularly large case series, with follow-up more extensive than is typically required in prospective studies. 29/53 patients were still being followed after 10 years. All patients were treated by one of two physicians at a single clinic (Mayo) over an approximate 7year period. Because this case series was retrospectively constructed, there are many "holes" in the data and information on pain and function was not available at all timepoints. However, there was enough information for the sponsor to classify each patients as a success or failure based on the criteria given on pages 11 and 15 of the amendment and I did not see anything in the amendment to make me think the process was slanted or biased. However, this process needs to be evaluated from a clinical perspective. With case series, the investigator does not control treatment assignment, endpoint ascertainment, selection biases, or confounding factors. Case reports and case series are typically used to generate hypotheses, not to test them.

The success/failure classification that the sponsor used was based on the last follow-up, but if an RA/SLE patient got worse after 5 years, this did not alter their classification because of the natural progression of the disease. With these classification criteria, approximately 60% of the implants were classified successful (i.e., excellent or good rating), and 62% of the patients had all of their implants considered successful. Seventy-two percent of successful implants in the RA/SLE group were followed more than 2 years, but not necessarily for all endpoints.

The only problem I have with the sponsor reanalysis of the data using descriptive statistics is that the OA/PT group is too small to support the sponsor's claim. There were only 9 implants in this group (as opposed to 138 RA/SLE). A number such as this would only be sufficient to suggest when a device is <u>not</u> working. Even though 7/9 implants were rated excellent or good, I would not consider this conclusive. Also, 2/9 implants fractured during the insertion process, which raises another concern.

The sponsor gives an extensive discussion of the control literature. There's a lot of variability in the pain and fracture rates reported and it is difficult to make all but the most cursory comparisons to the study device. Timepoints don't correspond. One point that the sponsor drives home, however, with good literature support, is that the silicone spacers have a problem with inflammatory synovial, erosive and cystic changes in the bone, and also have a high post-operative fracture rate. They also do little to increase joint function. These facts can be considered important in the risk benefit argument.

Longterm Analysis (Amendment 5)

The sponsor was asked to provide an additional longitudinal analysis so that reductions in treatment improvements at follow-up times greater than five years were considered. This concerned the RA/SLE cohort only since the 5-year restriction was not applied to the OA/PT cohort. With the modified criteria, 36 implants moved from being rated excellent or good to the unsatisfactory category. This left 37% of the RA/SLE cohort rated Successful (i.e., excellent or good) and 40% of the total study population (RA and OA combined) rated successful. On a per patient basis, 57% of subjects (N=53) had 1 or more successful implants and 43% had all implants successful. If you restrict the denominator to implants with known outcome, the percentages are about 4 percentage points higher.

The sponsor states that the longterm results represent a potential worst case analysis, since the rate of disease progression and soft tissue degradation is not known. While this may be true, one must also keep in mind that this whole case series analysis is based on the assumption that "no news is good news". In any case series where information is not systematically sought out, one cannot be sure that the absence of recording of a condition (e.g., pain) means that it didn't exist. A patient may have had a pain so long that he/she doesn't bother mentioning it to the doctor anymore, or the doctor may not have written it in the file. Therefore, there are a lot of unknowns here and this longterm analysis must be considered from a clinical perspective and risk/benefit perspective.

Conclusion

In summary, the information presented in support of this PMA has come full circle – from a nonstatistical argument to a statistical argument and back again. Amendments #3 and #5 represent a case series for a device whose benefits seem to outweigh the risks, given the limitations of the currently available treatment for these conditions, at least for the RA/SLE population.
Since the sponsor's claims cannot be supported on a statistical basis, I feel my role in this process is limited. The panel will have to look at this data, drawing on their clinical and scientific expertise, and make their recommendation to FDA.

Phyllis M. Silverman, M.S.