This guidance was written prior to the February 27, 1997 implementation of FDA's Good Guidance Practices, GGP's. It does not create or confer rights for or on any person and does not operate to bind FDA or the public. An alternative approach may be used if such approach satisfies the requirements of the applicable statute, regulations, or both. This guidance will be updated in the next revision to include the standard elements of GGP's.

Statistical Aspects of Submissions to FDA: A Medical Device Perspective

Richard P. Chiacchierini, Ph.D.

and

Henry T. Lee, Jr.

Food and Drug Administration
Center for Devices and Radiological Health
Office of Science and Technology
Division of Life Sciences - HFZ-116
5600 Fishers Lane
Rockville, MD 20857

June 1984

Statistical Aspects of Submissions to FDA: A Medical Device Perspective

by R. P. Chiacchierini and H. T. Lee

Introduction

The 1976 Medical Device Amendments to the Food, Drug and Cosmetic Act established a classification and review process for medical devices which is intended to give reasonable assurance of safety and effectiveness. Devices are in Class I if general controls are sufficient to provide assurance of safety and effectiveness or, in the absence of adequate information to establish such controls, the device does not present unreasonable risk. Class II devices are those for which performance standards, in addition to general controls, are needed and can be established to assure safety and effectiveness. A device is in Class III if there is insufficient information to establish performance standards or general controls, and if the device is life-supporting or entails a significant risk. Class III devices are required to undergo a review of safety and effectiveness prior to marketing similar to that for new drugs.

This paper will discuss only submissions for Class III devices which require the most stringent review and approval process, Premarket Approval (PMA). For the sake of completeness, however, two other types of submissions for Class III devices will be mentioned. First, the "substantially equivalent" or 510(K) submission provides the manufacturer a means of demonstrating that the device intended for marketing for a specific intended use is practically the same as a device marketed prior to enactment of the Medical Device Amendments. The second type of submission is the Investigational Device Exemption (IDE) which allows a manufacturer to market a limited number of devices of one kind for study purposes if it can be shown that the device is reasonably safe for use in humans, that a reasonable expectation of benefit to the patient exists, and that scientifically valid data can be collected.

The 510(k) and IDE submissions are related to the PMA in the following manner: the Agency is legally bound to require Premarket Approval Applications for significant risk (Class III) pre-amendment and post-amendment devices. So eventually, some substantially equivalent devices [510(k)] may need to go through the PMA process. The Investigational Device Exemption is intended to provide a data gathering mechanism so that a manufacturer can test the safety and effectiveness of a device usually, in anticipation of submitting a PMA at a later time.

There are guidelines for the submission of Premarket Approval Applications which briefly outline the required elements of a submission. Generally, the Agency statistician is involved in the review of the three data sections (laboratory, animal and clinical) and the comprehensive summary of safety and effectiveness. In the latter the PMA applicant must provide a cogent demonstration of safety and effectiveness for all diagnostic or therapeutic medical claims for the device based on the data and analyses in the data sections. Therefore, it is highly desirable that the sponsor involve a statistician early in the data gathering process, preferrably in the design stages. It is essential that the study protocol, results, analyses and interpretation support and be consistent with the medical claims for the device. Though this should be intuitively obvious, a number of applicants fail to comprehend the need for this consistency.

The Medical Device Amendments provide one major difference between drug evaluation and device evaluation in the submission of data. In an attempt to allow more flexibility for the applicant in providing valid scientific evidence for demonstrating the safety and effectiveness of a device, the Medical Device Amendments allow the Commissioner of the Food and Drug Administration (FDA) to accept well-controlled investigations or "other valid scientific evidence" which has been found to be sufficient to determine the effectiveness of the device. Drug effectiveness is based primarily on well-controlled studies. While the allowance of "other valid scientific evidence" is intended to simplify the device submission process for the applicant, it often leads to confusion and delays in the review process because of different interpretations between manufacturer and the Agency in what constitutes other valid scientific evidence. Specific examples of this problem will be given in the section on Evidence of Effectiveness.

It is essential that studies of safety and effectiveness be conducted at more than one institution by more than one investigator. Frequently, an investigator at a study site will have assisted in the design of the device, and therefore, data derived from that site may not be objectively collected and evaluated. Thus, to guard against investigator bias, data should be provided data from multiple investigators and study sites.

Finally, evidence of safety and evidence of effectiveness are discussed separately because often the data management methods differ. This distinction is made only for convenience because it is clear that safety and effectiveness cannot be separated in the device approval process.

Evidence of Safety

Valid scientific evidence of safety is given in a number of ways including physical, chemical, or biological laboratory testing; animal toxicity, carcinogenicity or teratology experiments; and human clinical investigations. Examples of laboratory testing include cell biology studies of mutagenicity or survival, chemical assay procedures for diagnostic detectability, engineering accelerated life testing, physical stress studies, and others. Animal experiments usually concern toxicity, carcinogenicity, or teratology but can also include studies on the feasibility and safety of implanted devices. The safety concerns found in the human clinical studies involve whether or not a device harms or injures the patient without demonstrating a compensating benefit. Included in this latter concern is the issue of the use of an ineffective device which prevents or postpones treatment with other therapies known to be effective. In cancer therapy, for example, this is a major concern because delayed treatment usually leads to poorer prognosis for survival.

The statistical issues involved in safety studies should be fairly routine. The laboratory and animal studies can usually be done under tight experimental control so that extraneous variability will be minimized. The usual experimental designs looking at animal, cell or product survival can be analyzed with a number of survival analysis models such as Kaplan-Meier product limit estimate, acturial life table, Cox model or the equivalent. Stratification and application of log-linear models (Bishop, 1975) or Mantel-Haenszel (Mantel, 1963) analyses are all well documented in the literature. Applications of these analyses should not lead to problems when properly applied.

The largest areas of concern with studies of safety involve three main issues. First, it is essential to have accurate and appropriate measures of dose and response. A risk assessment of the device depends on a dose-response analysis which must be characterized properly. The dose must be in the proper units derived from experimentation and must be consistent with the biological activity of the device or its by-products. For example, a certain biochemical component of a bone adhesive may leach out over time and concentrate in the kidney. One would expect to see a dose in terms of unit of weight of the material per unit of weight of the kidney. We frequently see studies where the dose is given as an average over the whole body when only a single focal point of concentration is involved. Such dose estimates severely underestimate the dose to the target organ. They are misleading because it appears that only trace amounts of the substance are in the body when in fact potentially toxic levels of the substance may be concentrated in the target organ.

The appropriateness of the response is equally important. If a substance concentrates in the kidneys and it is known that this substance may damage kidney tissue resulting in renal failure, then renal failure should be considered as an endpoint. If it is also known that the material is mutagenic, one should also look for cancer of the kidney and other associated organs such as the liver and bladder.

The second major concern in safety studies is the follow-up procedure. Whether we are concerned about morbidity or mortality in animal studies or human safety evaluations, the duration and completeness of follow-up are critical. The nature of the endpoint will dictate the type of follow-up that is needed. Acute toxicity usually requires moderate to large doses administered to a sample of animals which are then followed for days or weeks. Carcinogenicity studies among animals, however, require lifetime studies usually lasting about two years to substantiate a negative result. Serial sacrifice studies which attempt to detect gradual tissue damage are more difficult for the determination of follow-up because it is difficult to be satisfied with a negative result. One does not know if the study is truly negative or if the observation time is not long enough to detect a delayed event.

The most difficult follow-up determination concerns safety in the human clinical studies. Frequently, the Advisory Panels to the Agency and Agency staff attempt to define adequate periods of follow-up but usually these are based on clinical observation without benefit of rigorous analyses. For example, the follow-up for an implanted pacemaker is four to six months after implantation or for a laser posterior capsulotomy is a minimum of six months. On occasion such follow-up has been found to be inadequate for the analysis of specific problems, but in the absence of reliable clinical evidence it is difficult to determine what is an appropriate follow-up period. As we gain experience with different groups of devices it will become easier to determine appropriate follow-up periods.

Another question concerning follow-up is the comparability between study and control subjects. To compare 500 person-years among the controls by following 500 people for one year may not be comparable to 500 person-years in a device treated group in which 100 people were followed for 5 years, unless some very stringent assumptions can be made about the endpoints. When such an assertion of equivalence is made, the investigator must provide supporting evidence.

A third area of concern with safety studies involves sample size. The investigator must consider the combination of the number of subjects followed for an adequate time period to have a reasonable chance of detection of the endpoint of concern. Though the Agency with the advice from some advisory panels has provided guidelines on the number of follow-up subjects, we are working to allow the investigator flexibility to determine the appropriate sample size and to give adequate justification for the choice. Methodology for sample size determination will differ between categorical and quantitative studies but sufficient documentation exists for both types so that it should not be a burden to incorporate a sample size determination as part of the study protocol.

Evidence of Effectiveness

The demonstration of effectiveness differs for different types of devices. For a diagnostic or monitoring device it may be sufficient to show that the device records a measure of bodily function reasonably accurately or that it produces diagnostically useful information. Medical knowledge may lag behind technological advance in such devices and it usually requires time and experience for the diagnostic utility of a device to be fully explored. Therapeutic devices, on the other hand, usually require more than a demonstration that the device physically, chemically, or biologically functions as it was designed to. In addition to proper function, therapeutic devices are required to show that said function results in a benefit to the patient. For example, in a recently approved PMA for a hyperthermia device for palliative cancer management, it was not enough to show that the device delivered a known amount of heat to a tumor site. The sponsor also had to show that the heat delivered resulted in a reduction in tumor size.

Devices which are designed to provide diagnostic or monitoring information must provide diagnostically useful information or data that are potentially diagnostically useful. For example, if a diagnostic imaging device shows specific tissue structure more clearly than other imaging devices, such images may be useful for the study of degenerative diseases of the tissue even though no specific disease entity has been detected. It is important to understand that the instructions for use are just as important as the device characteristics. If a device is intended to detect disease and if its effectiveness depends on a specific regimen of follow-up procedures, the device is not considered effective unless it is used in accordance with that regimen.

Clinical studies of diagnostic devices should be conducted with blind evaluations but on rare occasions unblinded studies may be submitted if it can be shown that the bias brought about by unblindedness is inconsequential. In any case, for diagnostic devices designed to detect disease or abnormality, the sponsor must submit the whole detection picture by estimating sensitivity, specificity, false positive rate and false negative rate with appropriate confidence intervals. Indices of agreement or concordance shoud be accompanied by a suitable test via the Kappa Statistic (Cohen, 1960) or an equivalent procedure to determine that the agreement is not due to chance (Fleiss, 1981). Simple correlation coefficients usually do not provide much information on agreement so these should only provide supporting evidence and not the main element of the demonstration.

Let us consider a hypothetical example typical of our collective experience of real Premarket Approval Applications for diagnostic devices. A manufacturer submits a PMA for FIND-A, a fictitious device designed to detect a disease condition or abnormality A. The device is much cheaper to purchase and use than the expensive accurate standard method for detecting disease condition or abnormality A. The applicant enlists four investigators, each at a different facility, two of which are to use the standard technique and two are to use FIND-A. Typical data from such a PMA are given in Table 1. The manufacturer probably has not consulted a statistician and does not realize that each facility/investigator should employ both methods under blind conditions of observation, that results from two or more facility/investigators should not be pooled without proper justification, and that the target population for the device should be both specified and represented by the sample. There is no mention of sensitivity, specificity, false positive rates or false negative rates. The statistics provided were probably done by the clinicians or other scientists. There is no description of the statistical methodology nor is there an explanation of the chi-square statistic other than to say that the non-significant difference indicates a similarity in effectiveness between the two methods.

The obvious statistical deficiencies in this submission require an inordinate amount of staff time to improve the quality. After many meetings, the manufacturer may be persuaded that new studies need to be done and that a statistician should be involved. In a subsequent resubmission the data might look like that given in Table 2 with estimation given in Table 3. Generally, these data can be compared to other existing detection methods or additional statistical analyses can be provided depending on the severity of errors in diagnosis. If the standard method is not error-free, then one needs to find the false positive and false negative rates for a sequence of true incidence rates of the disease or abnormality, as discussed in Fleiss (1981). If it is a very serious matter to miss a disease case or to falsely diagnose a negative, then an analysis by McNemar's test (McNemar, 1947) or equivalent procedures would be in order. Finally, if the populations at the four facilities were substantially similar and it is found that pooling can be justified, a single table similar to the form of Table 2 can be obtained with overall rates based on 200 people.

Clinical studies of the effectiveness of therapeutic devices can be either controlled clinical trials or clinical investigations which have controls of different types. The statistics of controlled clinical trials has a rich literature to which I cannot do justice in this discussion but I must touch on some important issues. Once clear definitions of the study endpoints are established, sample size determinations consistent with the endpoints should be done. This is especially crucial in studies where a non-significant statistical comparison will be used to imply equivalence with existing therapies. Patient follow-up should be given careful consideration with adequate duration and completeness and with comparability in important characteristics between the treated and control patients. The data from several study centers should be reported separately unless an adequate justification for pooling exists.

Clinical investigations of other types may be employed if the endpoints are suitable. For a device to be used for the palliative treatment for a progressive degenerative disease, proper historical controls can be used. Cross-over designs can be done when residual effects of the treatment are short-lived and an adequate design can be implemented. Other types of studies may be attempted if the result is so dramatic that statistical procedures would only demonstrate the obvious.

Let us consider a couple of typical hypothetical examples of therapeutic device PMA's again based on real PMA data with fictitious names, numbers, and device characteristics. Let us assume that the REMEDY-B device is claimed to be effective in improving mobility for four basic conditions identified as Ache 1, 2, 3, and 4. Typical data from such an experiment may be as given in Tables 4 and 5. It is highly probable that the observations were not done under blind conditions which matters little here anyway because no control group is provided or even discussed. No statistical analysis is provided because, it is obvious that marked or moderate improvement occurs in nearly 80% of the patients. It is not infrequent to have a "clinical trial" such as this where 90% or more of the patients come from a single investigator with some limited case histories from elsewhere rounding out the patient population. One can say very little about the statistical treatment of the data set because so many rules have been compromised that a meaningful analysis is impossible.

What would this study look like if it were done properly in a controlled clinical trial? It might look like the data given in Table 6. These follow-up data should be consistent with the sponsor's claim. For example, if the claim is for long-term relief of symptoms, say six months, these data need to be gathered a minimum of six months after treatment. If the effect is temporal and the claim is for temporary relief, then perhaps a Kaplan-Meier product-limit approach or actuarial life table analysis of the data is needed to find changes in distribution over time. The improvement index should be determined by an objective measure under blind conditions. In any case, a properly controlled design gives the sponsor and the reviewer the best overall data presentation and analysis and generally provides the sponsor with the best prospects for approval.

Conclusion

There is no substitute for both good science and good statistics. Submissions to FDA should be based on good science and this should be evidenced by the following four elements.

First, the objectives of the study must be clearly and concisely stated and must be determined in a manner which will support the sponsor's claims for the device. The more specific the objectives the better because such specificity allows for a more direct evaluation.

Second, the statistical procedures should be completely described. Included in this description are verification of assumptions, population selection criteria, pooling justification, statistical model selection and justification, description or reference of special statistical procedures, sample size justification, follow-up criteria, randomization criteria, how blind evaluation was accomplished, etc. Again, the more specific the better because the reviewer needs to determine precisely what was done to whom and how.

Third, the report of the results should be detailed. This should include hypothesis tests, confidence intervals, estimates of variability, method of variability estimation, patients lost to follow-up, reasons for loss of follow-up, specification of statistical tests, stratification criteria, etc. For categorical data analyses, the actual tabulations should be submitted and for quantitative data, enough data needs to be submitted to allow verification of test statistics and confidence intervals.

Finally, the sponsor should provide an interpretation of the data and analyses to show how these studies support the medical claim for the device. In this area, the sponsor must link together the results of hypothesis tests to the study objectives to the medical claims for the device.

Providing adequate information regarding these four elements removes much guesswork on the part of the reviewers and will expedite the review process through the agency. The sponsor should provide all relevant information, data, and analyses in a concise discussion, explaining precisely how the data demonstrate the safety and effectiveness of the device for stated medical claims. Such a presentation, while it cannot guarantee approval, will provide an optimal climate for discussion between the applicant and FDA throughout the evaluation process.

References:

Bishop, Y., S. Feinberg, and P. Holland. (1975). Discrete multivariate analysis: Theory and practice. M.I.T. Press, Cambridge, Mass.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20:37-46.

Fleiss, J. (1981). Statistical methods for rates and proportions, (Second ed.). J. Wiley and Sons, New York.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12:153-157.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. J. Am. Stat. Assoc. 58:690-700.

Table 1. Summary of Effectiveness of FIND-A

	Disease Status-A		
	Present	Absent	Total
Standard Method -	93	7	100
FIND-A	91	9	100
	184	16	200

Chi-square = 0.27 (P = 0.7, not significant)

Table 2. Summary of Effectiveness of FIND-A

a. Facility/Investigator 1

		FIND-A		
		A	Not A	Total
Standard Method	A	44	1	45
	Not A	2	3	5
	Total	46	4	50

b. Facility/Investigator 2

		FIND-A		
		A	Not A	Total
Standard Method	A	45	0	45
	Not A	1	4	5
	Total	46	4	50

c. Facility/Investigator 3

		FIND-A		
		A	Not A	Total
Standard Method	A	44	2	46
	Not A	2	2	4
	Total	46	4	50

d. Facility/Investigator 4

		FIND-A		
		A	Not A	Total
Standard Method	A	45	0	45
	Not A	0	5	5
	Total	45	5	50

Table 3. Analysis of FIND-A Effectiveness Data

Facility/ Investigator	Sensitivity	Specificity	False Pos. Rate	False Neg. Rate
1	0.98(0.89,1.00)*	0.60(0.15,1.00)	0.04	0.25
2	1.00(0.92,1.00)	0.80(0.28,1.00)	0.02	0.00
3	0.96(0.85,1.00)	0.50(0.07,1.00)	0.04	0.50
4	1.00(0.92,1.00)	1.00(0.48,1.00)	0.00	0.00

* Values inside the parentheses are upper and lower 95% confidence limits.

Table 4. Summary Data for REMEDY-B

Facility	Ache 1	Ache 2	Ache 3	Ache 4	Total
1	1	2	1	2	6
2	2	1	4		7
3		1		2	3
4	3		1		4
5	78	62	57	83	280
Total	84	66	63	87	300

Table 5. Improvement Data for REMEDY-B

Improvement	Ache 1	Ache 2	Ache 3	Ache 4	Total	%
Marked	21	23	31	39	114	38
Moderate	34	23	25	38	120	40
Mild	21	17	4	8	50	17
None	8	3	3	2	16	5
TOTAL	84	66	63	87	300	100

Table 6. Effectiveness Data for REMEDY-B

a. Facility 1

Improvement (Mobility Index)

	< 25%	25-50%	50-75%	75-100%	Total
Treated	6	17	37	40	100
Controls	35	35	20	10	100
Total	41	52	57	50	200

b. Facility 2

Improvement (Mobility Index)

	< 25%	25-50%	50-75%	75-100%	Total
Treated	6	18	39	37	100
Controls	40	34	18	8	100
Total	46	52	57	45	200

Appendix
Observed Uses and Abuses of Statistical Procedures in Medical Device Submissions

by

Statistics Branch, DLS, OST
Center for Devices and Radiological Health, FDA

The following is a listing of commonly used acceptable and unacceptable practices each in alphabetical order gathered from the experience of the staff of the Statistics Branch. The list is not complete but is intended to indicate the most common statistical features present in good and bad PMA submissions.

Acceptable Procedures

Statistical Procedure	Use
Analysis of Variance	Hypothesis testing, variance estimation, model development
Calculation of Survival Probabilities	Life Testing, Morbidity Analyses, Mortality Analyses, testing in situ device longevity
Distribution Tests (Kolmogorov-Smirnov,etc.)	Verification of underlying distribution
Early Study Termination	Significance test indicates device treated patients worse than comparison group.
Equality of Variance Tests (Bartlett, F-test method, etc)	Verify equal variance assumption for least squares, Anova, etc.
Exact Binomial Distribution Tests	Tests and confidence limits for rates and proportions
Fisher's Exact Test (Hypergeometric Distri.)	Test two independent categorical samples when one response cell is small.
Linear Regression	Significance tests, correlation, and confidence intervals for trends over time or space and "before and after" observations.

Appendix - 2

Statistical Procedure	Use
Log-rank, Peto, or other tests	Test differences in survival functions among several study groups.
Mantel-Haenszel, Log-Linear models, etc.	Adjusted rates for stratified samples, homogeneity among stratified samples, cumulative survival comparisons.
Multiple Comparison Procedures	Significance tests for many means either paired or unpaired
Non-parametric Tests (Rank, Rank Sum, Sign ,etc.)	Compare properties of small samples of unknown underlying structure.
Paired t-Test	Comparison of "before and after" treatment responses
Patient Comparability Testing	Assure similarity in important parameters such as age, sex, disease condition, etc., via appropriate statistical test for comparison groups.
Stepwise Discriminant Analysis or Equivalent	Prediction for categorical response variable data
Stepwise Multiple Regression Analysis	Prediction for quantitative data.
Transformation (Log, Square Root, etc.)	Stabilize variance for least squares, Anova, etc.
Two Sample t-Test	Compare two means from samples with equal variances.
Welch t-Test	Compare two means from samples with unequal variances.

Appendix - 3

Unacceptable Procedures

Statistical Procedure	Abuse
Analysis of Variance	Test for categorical or other data for which equal variance assumption is violated.
Clinical Trial	Lack of controls or choice of improper historical control, use of prototype device which evolves during study period, use of only one investigator, use of several investigators each with a few patients, and/or under different protocols.
Comparability	Patients in treated and comparison groups not similar in important characteristics.
Confidence Intervals, Variances and Standard Errors	Failing to provide estimates and adequate data for verification, failing to label what is provided, using standard error and standard deviation interchangeably, confusing confidence intervals with tolerance intervals.
Crossover Analyses	Crossovers done without regard for acceptable design practices, carryover effects ignored.
Data Selection	Presenting selected data from treated and comparison subjects (especially for historical controls).
Enumeration of Complications or Study Endpoints	Pooling variables of mostly different types (e.g., one cancer and one infection equals two abnormalities).

Appendix - 4

Statistical Procedure	Abuse
Exclusion of Patients	Poorly designed and unequally applied exclusion criteria lead to exclusion of complications and early deaths or device failure from survival analyses, justification for exclusion by saying event was "not device related", exclusion of patients whose response is difficult to analyze, etc.
Follow-up	Unequal between comparison groups, patient lost to follow-up unaccounted for in analysis or simply excluded without explanation.
Interpretation	Imprecise and ill-defined terminology such as use of "reflect" or "trend" when model does not adequately predict, implication of causality when association is presented, etc.
Measurement of Variables	Undefined measurement protocol leading to guesstimates, points, intervals, or all of above in the same data set.
Multiple Comparisons	No adjustment for numerous pairwise tests in overall significance level.
Outliers	Rejecting boundary data such as background readings or extremes over time as outliers in trend tests, rejecting data points without proper testing or valid reason.
Pooling	Combining data from different strata or diverse study centers without proper justification.
Regression and Correlation Estimates	Combining estimates from several studies without a test of homogeneity, using incorrect method of combination when valid.

Appendix - 5

Statistical Procedure	Abuse
Stratification	Lack of stratification in samples which are obviously unequal in critical characteristics.
Tests of Hypotheses	Using inappropriate test and test statistic because of incorrect distribution, unequal variance, correlated samples, design induced bias, incorrect model, etc.